Pathogen-host omics analyses of human papillomavirus type 16 sub-lineages in a human epithelial organoid model

by
ROBERT JACKSON

A dissertation submitted in partial fulfillment
of the requirements of the degree of
Doctor of Philosophy in Biotechnology

Biotechnology PhD Program, Lakehead University
Thunder Bay, Ontario, Canada

8 May 2019

ABSTRACT

Pathogens such as human papillomaviruses (HPVs) have co-evolved with their hosts and form a molecular basis for common diseases. Persistent infection with the "high-risk" HPV type 16 (HPV16) is a potent cause of anogenital and oropharyngeal cancers. Taxonomic HPV16 sub-lineages, based on geographic origin of discovery, are noteworthy due to their variable tumourigenicity. In this dissertation, I present basic research and the resulting biotechnologies we developed, improved, and utilized to study their fascinating pathogen-host relationship with human stratified epithelia. A small number of variations in the E6 gene of HPV16, found in the D2 and D3 sub-lineages, lead to increased tumourigenic risk compared to the prototype A1 sub-lineage. Using an organotypic human epithelial model (or *in vitro* organoid) we recapitulated the viral life cycle and used "-omics" analyses to assess viral and host molecular differences due to sub-lineage variation. Sub-lineage variants of E6 were associated with host genome instability and viral integration into host DNA. Following these initial findings, I provide perspective on epithelial organoids, namely that the trade-off between model complexity and feasibility should be sensibly considered based on its utility for answering the biological research question at hand. Model applications and improvements are presented, including time-series epithelial stratification measurements, strategies for introducing full-length sub-lineage HPV16 genomes into host keratinocytes, and experiments to study innate immune evasion. These wet-lab works are accompanied by software to aid biologists in analyzing sequencing data. As well, we present current work using The Cancer Genome Atlas to test the association between HPV16 sub-lineage and integration. Overall, this interdisciplinary and interconnected collection has significance for basic researchers, providing insight on how a small number of natural viral variations can lead to increased tumourigenic risk, as well as for experimentalists to gain insight on organoid modelling and novel bioinformatics tools. More broadly, characterizing these molecular interactions between pathogen and host enables us to form a basis for diagnosis, treatment, and ultimately prevention of disease. Future research should aim to closely integrate biological and computational sciences for improving experimental approaches and our ability to make meaningful biological interpretations given the complexity and variability of biological systems.

ACKNOWLEDGEMENTS

Foremost I thank my mentor and supervisor, Dr. Ingeborg Zehbe, for the incredible opportunity to pursue my scientific passions within her lab, supporting both me and my research (financially and philosophically) during this adventure, and for cultivating me into the researcher I am today. I am fortunate to have strengthened my skillset and refined my craft through this rigorous training, locally and abroad, coupled with exposure to teaching, mentorship, management, grant-writing, conferences, and networking events.

My dissertation committee members, Dr. Wely Floriano and Dr. Marina Ulanova, provided wise guidance and vital feedback. I am also thankful to Dr. Neelam Khaper, who served as the external member for my comprehensive examination, as well as Dr. Heidi Schraft, who chaired the examination. Thank you to my external examiner, Dr. James Pipas, for agreeing to review my dissertation and provide critique with their expertise.

From Lakehead University I thank Dr. Justin Jiang for his informative seminar class, Dr. Brenda Magajna, Eleanor Maunula, and the Faculty of Graduate Studies for administrative support, Darryl Willick for access to the Lakehead University High Performance Computing Cluster (LUHPCC), computer scientists Dr. Jinan Fiaidhi and Dr. Sabah Mohammed, as well as all undergraduate and graduate students I have had the pleasure to work with and instruct.

Past and present Thunder Bay Regional Health Research Institute (TBRHRI) and Thunder Bay Regional Health Sciences Centre (TBRHSC) students, scientists, staff, and volunteers were instrumental in my dissertation work. I am especially grateful to my long-time friend and lab companion Dr. Melissa Togtema for battling in the trenches, side-by-side, for the better part of the past decade. Thank you for showing me the way! As well, thanks to friends and colleagues Sarah Niccoli, Sean Cuninghame, Dr. Chris Abraham, Jessica Grochowski, Vanessa Masters, Peter L. Villa, Dr. Guillem Dayer, Christopher Gibb and the PHAT Boyz, Josee Bernard, Kathlyn Alexander, Mehran Masoom, Dallas Nygard, Alejandro Ortigas Vásquez, Anirudh Shahi, Statton Eade, Jordan Lukacs, Carmen Dore, Shannon Maki, and many more that supported me.

I am grateful to collaborators worldwide that have contributed ideas and provided inspiration: Dr. Bruce Rosa, Dr. Alain Nicolas, Dr. Allyson Holmes, Sonia Lameiras, Dr. Paul Lambert, Dr. Carmen Lía Murall, and Dr. Samuel Alizon. I am also thankful for discussions, materials, and methodologies from the following scientists: Dr. Louise Chow, Dr. Tom Broker,

Dr. Andras Nagy, Dr. John Lee, Dr. Michael Dean, Dr. Robert Burk, Dr. Zigui Chen, Dr. Martin Müller, and Dr. Koenraad Van Doorslaer.

Finally, I thank the following for their generous funding support: Natural Sciences and Engineering Research Council of Canada (NSERC) for awarding me an Alexander Graham Bell Canada Graduate Scholarship-Doctoral (CGS-D, #454402-2014) as well as grants awarded to Dr. Zehbe in support of her research program, TBRHRI's Elekta travel awards, CUPE and Lakehead University professional development bursaries, CIHR for their travel award to attend a Canadian Bioinformatics Workshop, Lakehead University's Graduate Student Association (GSA) for supporting an award and travel for the provincial Three Minute Thesis (3MT) Ontario competition, and the Faculty of Science and Environmental Studies for supporting an award at the Biotechnology and Allied Sciences Symposium.

DEDICATION

I am eternally grateful to my life partner, Jessie Jones, for being a stalwart supporter. I dedicate this dissertation to her and to all those who undertake the methodical pursuit of truth. Jessie's backing, combined with reinforcement from my family, friends, comrades, colleagues, and critters, has made this journey worthwhile and joyful.

TABLE OF CONTENTS

ABSTRACT ..... **i**
ACKNOWLEDGEMENTS ..... **ii**
DEDICATION ..... **iv**
TABLE OF CONTENTS ..... **v**
LIST OF TABLES ..... **ix**
LIST OF FIGURES ..... **x**
CHAPTER 1 – INTRODUCTION **1**
1.1 – Literature Review **1**
1.2 – Research Rationale, Question, Hypothesis, and Objectives **4**
1.3 – Original Scientific Contributions **6**
CHAPTER 2 – VIRAL-HOST INTEGRATION DUE TO SUB-LINEAGE **8**
2.1 – Abstract **9**
2.1.1 – Background *9*
2.1.2 – Results *9*
2.1.3 – Conclusions *9*
2.2 – Background **10**
2.3 – Results and Discussion **13**
2.3.1 – Viral integration in the HPV16 AAE6 but not EPE6 epithelium *13*
2.3.2 – The HPV16 AAE6 epithelium has a unique transcriptional profile *16*
2.3.3 – Nature of viral-human fusion transcripts detected in HPV16 AAE6 epithelium *20*
2.3.4 – The HPV16 AAE6 epithelium reveals a signature of chromosomal instability conducive to host genome integration *23*
2.3.5 – HPV16 AAE6 epithelium exhibits a proliferating phenotype as a consequence of viral integration into the host genome *25*
2.4 – Conclusions **31**
2.5 – Methods **32**
2.5.1 – Cell lines *32*
2.5.2 – Detection of integrated papillomavirus sequences by DNA-Seq: Capt-HPV *32*
2.5.3 – RNA-Seq library preparation and sequencing *32*
2.5.4 – Viral variant read alignment, mapping, and coverage plotting *33*

2.5.5 – Identification of viral-human fusion transcripts *34*
2.5.6 – Human read alignment, mapping, and count generation *34*
2.5.7 – Differential expression analysis of human transcriptome *35*
2.5.8 – CIN70 scoring and micronuclei detection *35*
2.5.9 – Gene set enrichment analysis and networks *36*
2.6 – Declarations **36**
2.6.1 – Acknowledgements *36*
2.6.2 – Funding *36*
2.6.3 – Availability of data and materials *37*
2.6.4 – Authors' contributions *37*
2.6.5 – Competing interests *37*
2.6.6 – Open access *37*
2.7 – Additional Files **38**
2.7.1 – Additional file 1: Viral and human read tables *38*
2.7.2 – Additional file 2: DESeq plots *40*
2.7.3 – Additional file 3: DESeq output *45*
2.7.4 – Additional file 4: Follow-up discussion of host expression analysis *45*
2.7.5 – Additional file 5: GO output *51*
2.7.6 – Additional file 6: Pearson correlations *51*
CHAPTER 3A – EPITHELIAL ORGANOID MODEL **52**
3A.1 – Abstract **52**
3A.2 – Introduction **53**
3A.3 – The "Silent Killer": How HPV Evades Host Immune Recognition **56**
3A.3.1 – The HPV clearance hypothesis: the role of LCs *57*
3A.3.2 – Implication of HPV16 variants in modulating the host immune system *59*
3A.4 – An Eclectic Methodological Approach for an Epithelial Organoid **60**
3A.4.1 – Preparation of the revised model *63*
3A.4.2 – Organoid cultivation and characterization *68*
3A.4.3 – Immune-competent component *69*
3A.4.4 – NGS and the Pathogen-Host Analysis Tool (PHAT) *70*
3A.5 – Limitations of Existing Research Models **71**

3A.6 – Conclusion ... **72**
3A.7 – Declarations ... **73**
3A.7.1 – Acknowledgements ... *73*
3A.7.2 – Data accessibility ... *73*
3A.7.3 – Authors' contributions ... *73*
3A.7.4 – Competing interests ... *73*
3A.7.5 – Funding ... *73*
CHAPTER 3B – THEORETICAL APPLICATION OF THE MODEL ... **74**
3B.1 – Abstract ... **74**
3B.1.1 – Author summary ... *75*
3B.2 - Introduction ... **75**
3B.3 – Results ... **76**
3B.3.1 – Uninfected epithelial dynamics ... *76*
3B.4 – Discussion ... **77**
3B.4.1 – Dynamical implications of ecological features ... *77*
3B.4.2 – Perspectives ... *77*
3B.5 – Materials and Methods ... **78**
3B.5.1 – Ethics statement ... *78*
3B.5.2 – Cell culture data ... *78*
3B.6 – Declarations ... **78**
3B.6.1 – Acknowledgements ... *78*
3B.6.2 – Funding ... *79*
3B.6.3 – Data availability ... *79*
CHAPTER 3C – ENHANCING THE MODEL'S BIOLOGICAL RELEVANCE ... **80**
3C.1 – Full-Length HPV16 Sub-Lineage Genomes ... **80**
3C.2 – Strategies for Introducing HPV16 Genomes to Host Keratinocytes ... **85**
3C.3 – Adapting the Organoid Model to Study Innate Immunity ... **87**
CHAPTER 4A – PATHOGEN-HOST ANALYSIS TOOL ... **90**
4A.1 – Abstract ... **91**
4A.1.1 – Summary ... *91*
4A.1.2 – Availability and implementation ... *91*

4A.2 – Introduction **91**
4A.3 – Features **92**
4A.4 – Future Work **95**
4A.5 – Declarations **95**
4A.5.1 – Acknowledgements *95*
4A.5.2 – Funding *95*
CHAPTER 4B – HPV16 SUB-LINEAGE AND INTEGRATION IN BIG DATA **96**
4B.1 – Abstract **96**
4B.2 – Introduction **97**
4B.3 – Methodology **100**
4B.3.1 – Data acquisition and storage *100*
4B.3.2 – HPV16 sub-lineage genotyping analysis *103*
4B.3.3 – Viral-human integration detection and characterization *105*
4B.3.4 – Statistical analyses *105*
4B.4 – Preliminary Results **106**
4B.4.1 – HPV16 sub-lineages in cervical and head and neck cancer datasets *106*
4B.4.2 – Integration histograms reveal different patterns of integration *109*
4B.4.3 – Circular visualizations reveal sequence differences between sub-lineages *110*
4B.4.4 – TCGA data used for functional analysis other than integration *111*
4B.5 – Discussion **112**
4B.6 – Conclusion **114**
4B.7 – Declarations **114**
4B.7.1 – Acknowledgements *114*
4B.7.2 – Authors' contributions *115*
4B.8 – Supplemental Information **115**
4B.8.1 – Supplementary Script 1: Custom R script for sub-lineage pipeline *115*
4B.8.2 – Supplementary Script 2: Custom R script for histogram creation *115*
CHAPTER 5 – CONCLUSIONS **116**
LITERATURE CITED **119**
APPENDIX A **145**

LIST OF TABLES

Table 2.1 – Integration loci detected by ViralFusionSeq……...……22
Table 2.S1 – Viral reads summary……..38
Table 2.S2 – Human RefSeq alignment statistics for all samples……..39
Table 2.S3 – Human library size factor for all samples……..39
Table 2.S4 – Top-ten most significant down-regulated genes in AAE6 compared to NIKS....49
Table 2.S5 – Top-ten most significant up-regulated genes in AAE6 compared to NIKS…….49
Table 2.S6 – Top-ten most significant down-regulated genes in AAE6 compared to EPE6…50
Table 2.S7 – Top-ten most significant up-regulated genes in AAE6 compared to EPE6…....50
Table 4B.1 – HPV16 sub-lineage reference genomes used for genotyping…….104
Table 4B.2 – Sample summary breakdown for each TCGA case set……..107
Table A.1 – Plasmid information…….....145

LIST OF FIGURES

Figure 1.1 – Hallmarks of cancer…….....3

Figure 1.2 – Experimental scheme…….....5

Figure 1.3 – Dissertation layout…….....7

Figure 2.1 – The HPV16 genome and our experimental epithelial model……12

Figure 2.2 – Characterization of viral integration and viral-human fusion transcripts in AAE6 epithelia……..15

Figure 2.3 – The HPV16 transcriptome in EPE6 and AAE6 organotypic rafts……...18

Figure 2.4 – Chromosomal instability signature and micronuclei in AAE6 epithelia…….24

Figure 2.5 – Gene Ontology (GO) terms enriched in highly significant differentially expressed genes in AAE6 vs. NIKS……27

Figure 2.6 – Gene Ontology (GO) terms enriched in highly significant differentially expressed genes in AAE6 vs. EPE6……28

Figure 2.7 – Co-expression networks of highly significant [a] down-regulated and [b] up-regulated genes in AAE6 vs. EPE6……29

Figure 2.8 – Venn diagram of differentially expressed genes common and unique to each pairwise comparison……30

Figure 2.S1 – Plot of normalized mean counts versus $\log_2$ fold change for the contrast NIKS versus EPE6…….....40

Figure 2.S2 – Plot of normalized mean counts versus $\log_2$ fold change for the contrast NIKS versus AAE6…….....41

Figure 2.S3 – Plot of normalized mean counts versus $\log_2$ fold change for the contrast EPE6 versus AAE6…….....42

Figure 2.S4 – Empirical and fitted dispersion values plotted against the mean of the normalized human gene-level counts……..43

Figure 2.S5 – Heatmap of Euclidean distances between human gene-level counts of sample.44

Figure 3A.1 – Historical overview of epithelial models……..55

Figure 3A.2 – Epithelial microenvironment and immune landscape…….....58

Figure 3A.3 – Flow diagram of the epithelial model…….....61

Figure 3A.4 – Characterization of the epithelial model…….66

Figure 3B.1 – Epithelial cell growth in 3D raft cultures…….76

Figure 3C.1 – HPV16 A1 vs D3 single-nucleotide polymorphisms (SNPs)…….....81
Figure 3C.2 – Predicted 3D protein structures for EP (A1, red) and AA (D2/D3, blue)……..82
Figure 3C.3 – Location of preferred LoxP sites in HPV16's URR…….....84
Figure 3C.4 – Immunocompetent epithelia trials…….....89
Figure 4A.1 – Pathogen-Host Analysis Tool (PHAT) and visualization…….....94
Figure 4B.1 – Authorized access process for TCGA data …….....101
Figure 4B.2 – Summary of the analytical workflow…….....102
Figure 4B.3 – HPV16 sub-lineage reference tree for phylogenetic analyses…….....104
Figure 4B.4 – Preliminary analysis of CESC samples……108
Figure 4B.5 – Integration histograms…….....110
Figure 4B.6 – Circular genome visualizations of HPV16…….....111

CHAPTER 1 – INTRODUCTION

1.1 – Literature Review

Viruses and their hosts have complex, albeit, fascinating relationships: "two-stepping" together through evolutionary time [Meyerson and Sawyer, 2011]. These molecular interactions are influenced by a mixture of viral, host, as well as environmental factors, and can manifest in host disease, including cancers [Chang *et al*., 2017]. While seven groups of human oncoviruses are established (*i.e*., human papillomavirus, hepatitis B and C viruses, human T-lymphotropic virus-1, Kaposi sarcoma herpesvirus, Epstein-Barr virus, and Merkel cell polyomavirus) [Chang *et al*., 2017], human papillomaviruses (HPVs) are the most common pathogens associated with human malignancies [zur Hausen, 1996; 2002]. HPVs are double-stranded DNA viruses, a subset of which, the so-called "high-risk" HPVs, cause 5.2% of human cancers [Parkin and Bray, 2006]. Although 500+ types of human and non-human PVs exist [Van Doorslaer *et al*., 2013; 2017a; 2017b], HPV type 16 (a member of the viral species *Alphapapillomavirus 9*) is the most prevalent of the high-risk human types, causing anogenital cancers (*e.g*., of the cervix) as well as oropharyngeal cancers as a result of persistent infection. Its potent tumourigenicity is reflective of the activities of virally encoded oncoproteins, primarily E6 [reviewed by Vande Pol *et al*., 2013] and E7 [reviewed by Roman and Münger, 2013]. However, the role of the entire viral genome and the viral life cycle needs to be considered when studying viral tumourigenesis due to cis-acting viral proteins (such as E2, which is involved in regulating E6/E7 expression) as well as the complexity of viral-host interactions (such as the emerging understanding of an oncogenic role for E5) [Doorbar *et al*., 2012]. A useful unifying framework for studying the molecular and cellular basis of how cancers form, named the "hallmarks of cancer" [**Figure 1.1**], is described by Hanahan and Weinberg [2000; 2011] and were further expanded by Mesri *et al*. [2014] to oncoviruses.

Intratypic variants (or sub-lineages) of HPV16 have co-evolved with human populations and differ in their persistence and frequency of detection in pre-cancers and cancers [Burk *et al*., 2013]. The tumourigenic differences of these variants have been ascribed, at least partially, to genomic variation within the E6 oncogene (but also include interaction effects with host genome variability [Togtema *et al*., 2015]). The Asian-American (AA, also known as the D2 and D3 sub-lineage) and European Prototype (EP, also known as the A1 sub-lineage) are common genomic variants of HPV16 with their E6 genes differing by only six single-nucleotide polymorphisms (SNPs), three of which are non-synonymous leading to the 151-residue AAE6 protein differing by

three amino-acids: Q14H, H78Y, and L83V [Cornet *et al*., 2012]. Epidemiological studies revealed that coding changes in E6 have differential tumourigenic risk [Zehbe *et al*., 1998a; 1998b; 2001; Grodzki *et al*., 2006] and that AA is a higher risk factor for dysplasia as well as earlier onset invasive tumours than EP [Xi *et al*., 1997; 2007; Villa *et al*., 2000; Berumen *et al*., 2001; Schiffman *et al*., 2010; Freitas *et al*., 2014]. As well, AAE6 alone has a greater transforming, migratory, and invasive potential than EPE6 when retrovirally transduced into primary human keratinocytes during recent long-term *in vitro* immortalization studies [Zehbe *et al*., 2009; Richard *et al*., 2010; Niccoli *et al*., 2012; Togtema *et al*., 2015], as well as altered cellular energetics [Cuninghame *et al*., 2017], consistent with the hypothesis that coding changes in E6 contribute to differences in cancer risk. Protein-based binding experiments are being conducted to determine unique properties and cellular partners of HPV16 E6 variant proteins [Mehran Masoom, unpublished observations]. Our lab is also investigating anti-E6 therapeutics, including monoclonal antibodies introduced to target cells via sonoporation [Togtema *et al*., 2012], RNA-interference using small-interfering RNA (siRNA) [Togtema *et al*., 2018b], and camelid-derived single-domain antibodies [Togtema *et al*., 2018a].

In our previous investigation of these two common HPV16 E6 variants we demonstrated that AAE6 drives tumourigenesis in an early infection scenario (using 3D epithelial cultures) by increasing cellular proliferation, disrupting differentiation and apoptosis, decreasing innate immune gene expression, and promoting immortalization [Jackson *et al*., 2014]. The differences in host epithelia were reflective of increased oncogene expression in AAE6 cultures (E6 and E7) and loss of productive life cycle (decreased E2, E1^E4, and L2), both suspected to be a result of integration of the AAE6 viral DNA into the host genome [Jackson *et al*., 2014]. To further address this hypothesis, and characterize the molecular pathways involved, it became essential to use high-throughput "-omics" techniques coupled with bioinformatics. Additional chapter-specific literature review addressing these topics is provided throughout the dissertation [**2.2 – Background**, **3A.2 – Introduction**, **4A.2 – Introduction**, and **4B.2 – Introduction**] to supplement this section's concise introduction to the literature.

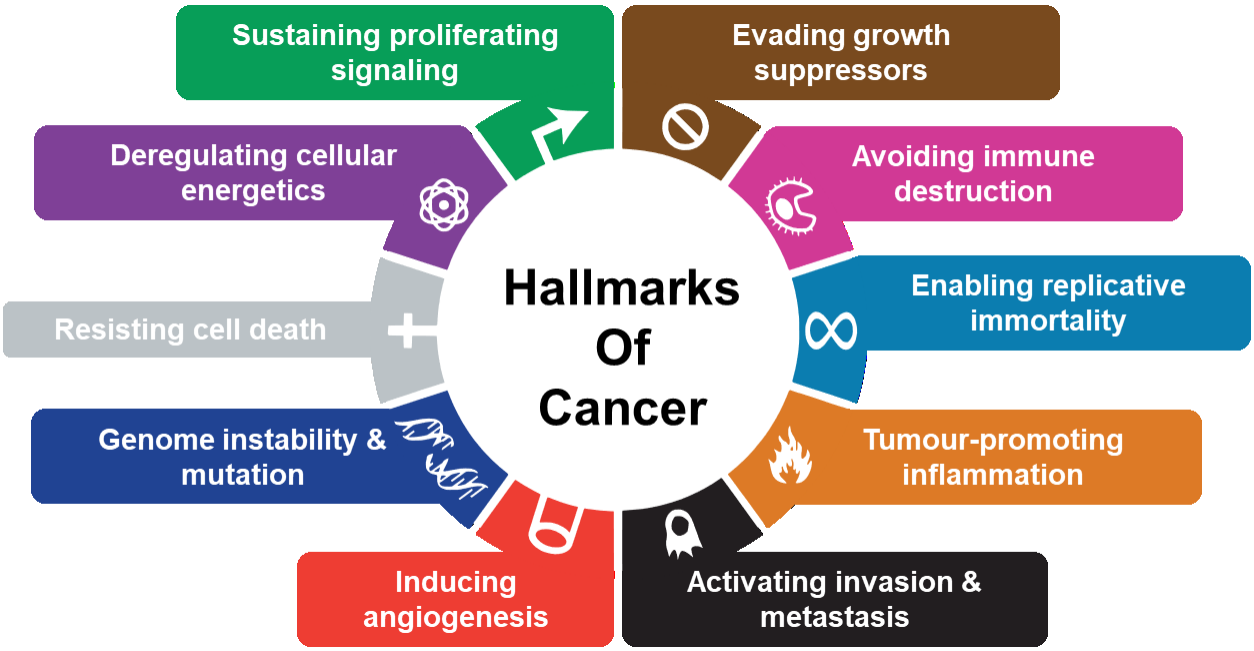


Figure 1.1 – Hallmarks of cancer. Hanahan and Weinberg [2000; 2011] describe ten useful traits for characterizing the abilities that cancer cells gain during tumourigenesis. These hallmarks and enabling characteristics can be used as a unifying framework for studying cancers, including those caused by pathogens [Mesri *et al.*, 2014]. This figure is based on Figure 6 from Hanahan and Weinberg [2011] and is re-used with permission from Elsevier (License Number 4517840162596).

1.2 – Research Rationale, Question, Hypothesis, and Objectives

While a considerable amount of information has been gathered in the past few decades about the tumourigenic relationship between HPV16 and human keratinocytes, much of the fundamental work has been performed in classical monolayer experiments (preventing a full viral life cycle from occurring) and using single genes (preventing a full viral life cycle, but also missing the interactions due to the full viral genome being present). As well, given that a low proportion of HPV16 infections persist to invasive cancer [Stanley, 2012], factors affecting persistence and progression to carcinogenesis (especially those associated with viral variation), deserve scrutiny in biologically relevant experimental models. To understand these factors, the concept of a "pathogenic lifestyle" from Dr. Stanley Falkow comes to mind: "Falkow came to think of the 'pathogenic lifestyle' as being not about causing disease, but rather about subtly manipulating a host. He became best friends with his prey" [Amieva, 2018]. Given the subtly of interactions, where pathogens have adapted mechanisms to subvert cellular machinery and processes, the comprehensive study of these pathogen-host relationships requires experimental approaches that provide a realistic host environment (epithelial organoids), full-length "competent" viral genomes, and sensitive analytics ("-omics" techniques and tools such as next-generation sequencing and bioinformatics). Overall, my goal was to fill these gaps and share with the scientific community how small genomic variations in pathogens, such as that between HPV16 sub-lineages, cause significant changes in host mechanisms manifesting in disease. In other words, to answering the question: why is the Asian-American (AA, D2/D3 sub-lineage) variant of HPV16 more tumourigenic than the European Prototype (EP, A1 sub-lineage) variant? We hypothesize that virus-host interactions differ between HPV16 sub-lineages (A1 vs D2/D3, otherwise known as variants, EP vs AA) and that specifically D2/D3 promotes tumourigenic molecular pathways (*e.g.*, genomic instability) while suppressing anti-tumourigenic molecular pathways (*e.g.*, innate immune response) compared to A1. To address these hypotheses, basic (fundamental) research was conducted on the pathogen-host relationship of HPV16 sub-lineages. Whilst doing so, we developed, enhanced, and utilized novel biotechnologies, including: three-dimensional organotypic human epithelia (*in vitro* organoids) capable of hosting a productive and reproducible viral life cycle, and bioinformatics tools and "-omics" analyses for pathogen-host relationships. These objectives were accomplished via an experimental scheme using HPV16 variant genomes, epithelial organoids, and molecular analyses of both viral and human nucleic acids [**Figure 1.2**].

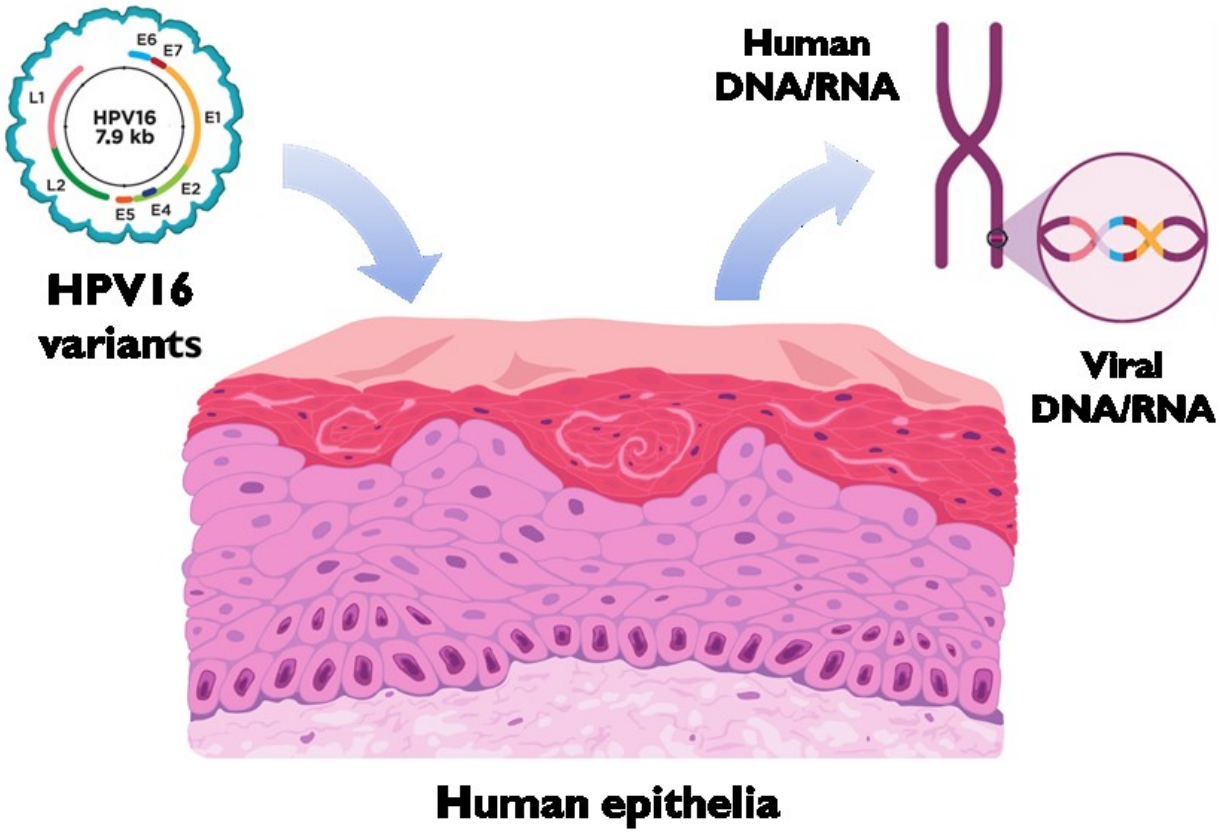


Figure 1.2 – Experimental scheme. HPV16 variant genomes were introduced to human epithelial organoids permitting an active viral life cycle alongside host keratinocyte differentiation. Molecular analyses were performed on DNA and RNA, characterizing both viral and human sequences, including viral-human integrations and fusion transcripts. This figure has been adapted with permission for re-use within this dissertation [Jackson *et al*., 2016].

1.3 – Original Scientific Contributions

This dissertation is structured as a hybrid thesis [**Figure 1.3**], with chapters including published scientific contributions as well as unpublished complementary material. This original work included interdisciplinary collaboration throughout the duration of my doctoral studies. Given my prior work on papillomaviruses it is essential that I first demarcate the continuum of my contributions. Prior to my current degree studies, four co-authored articles were published [DeCarlo *et al*., 2012; Togtema *et al*., 2012; Jackson *et al*., 2013; Pichardo *et al*., 2013]. Since starting doctoral work, an additional eleven co-authored articles were published (or accepted for publication), four of which (**underlined and bolded**) are included within this dissertation [Cuninghame *et al*., 2014; 2017; **Jackson *et al*.**, 2014; **2016**; **2019**; Togtema *et al*., 2015; 2018b; Zehbe *et al*., 2016a; Villa *et al*., 2018; **Gibb *et al*., 2019**; **Murall *et al*., 2019**]. These contributions span basic science, therapeutics, as well as community-based screening and prevention studies.

Initial findings on viral-host integration due to HPV16 sub-lineage are presented in **CHAPTER 2** in the form of a first-author research article [Jackson *et al*., 2016]. This chapter lays the foundation for all other chapters, as it includes both biological and computational aspects which are each further expanded. **CHAPTER 3A**, **CHAPTER 3B**, and **CHAPTER 3C** are based on the epithelial organoid model. Perspectives on using epithelial organoids for studying the viral life cycle of human papillomaviruses are discussed in **CHAPTER 3A**, as a first-author commentary article [Jackson *et al*., 2019]. Following these, epithelial organoids are applied in **CHAPTER 3B** to study stratification dynamics, informing parameters of ecologically-based *in silico* mathematical models of infection which includes excerpts from a co-authored research article [Murall *et al*., 2019]. Additional enhancements to the organoid model are presented in **CHAPTER 3C**, including recent progress on full-length HPV16 variant genomes, transfection strategies into host keratinocytes of different anatomical origin, and studying innate immunity. **CHAPTER 4A** and **CHAPTER 4B** focus on bioinformatics and analysis of next-generation sequencing data for papillomaviral sequences. In **CHAPTER 4A**, the development of the Pathogen-Host Analysis Tool (PHAT) software is presented as a co-first-author article [Gibb *et al*., 2019]. Current progress on analyses of The Cancer Genome Atlas (TCGA) datasets, with tumours containing HPV16, are described in **CHAPTER 4B**. **CHAPTER 5** concludes the dissertation, highlighting its primary significance for basic researchers and experimentalists, as well as the broad impact of studying variability within pathogen-host relationships.

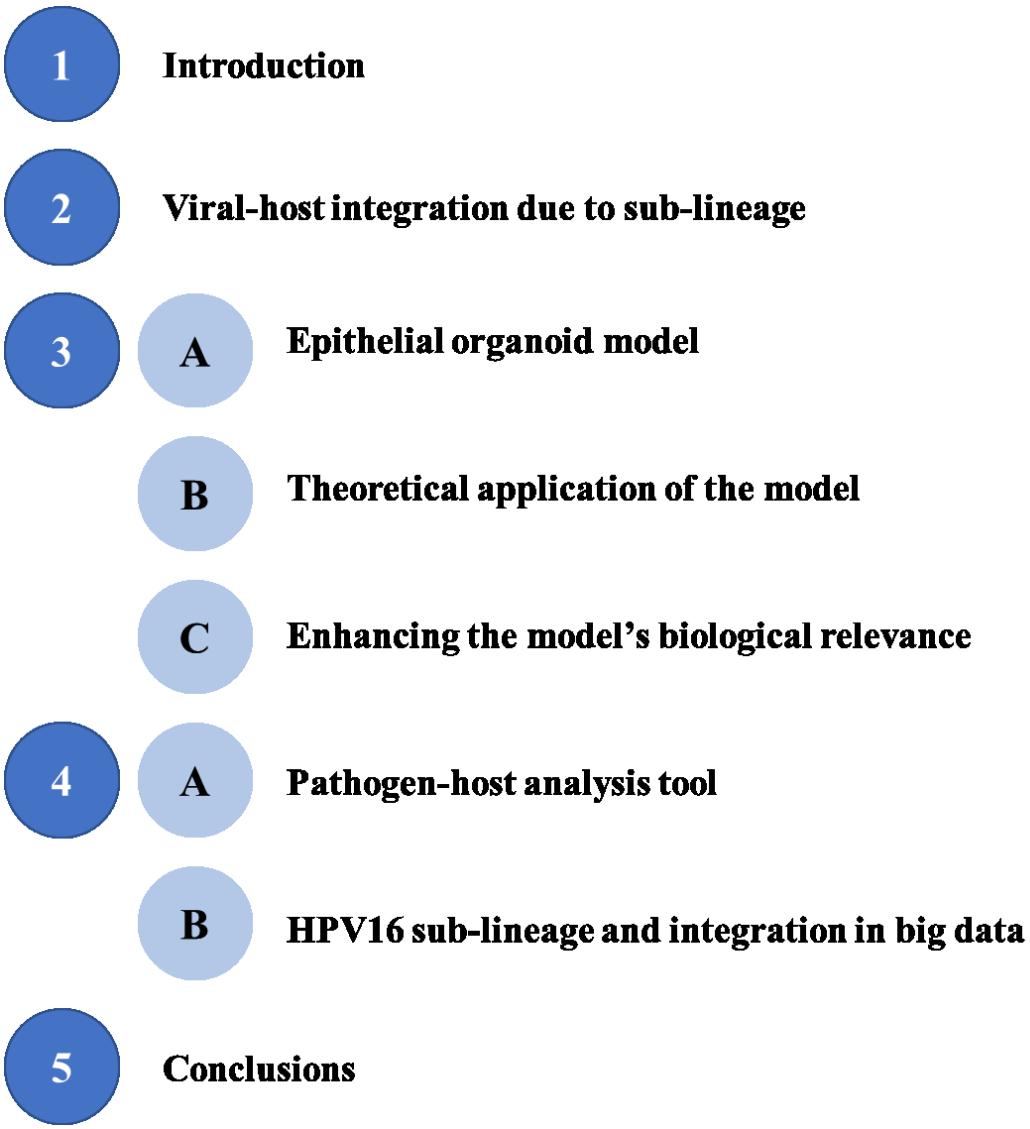


Figure 1.3 – Dissertation layout. The dissertation contains five main-body chapters: an introduction (Chapter 1: literature review, research rationale and objectives), initial findings (Chapter 2: a first authored research article [Jackson *et al*., 2016]), three interconnected epithelial organoid model chapters (Chapter 3A: a first authored perspective article [Jackson *et al*., 2019]; Chapter 3B: an excerpt from a collaborative research article [Murall *et al*., 2019]; and Chapter 3C: unpublished progress on enhancing the model's biological relevance), two interconnected bioinformatics chapters (Chapter 4A: a co-first authored software article [Gibb *et al*., 2019]; and Chapter 4B: unpublished progress on analyzing The Cancer Genome Atlas data), as well as conclusions of this interdisciplinary work (Chapter 5: summary and impact of findings, future directions).

CHAPTER 2 – VIRAL-HOST INTEGRATION DUE TO SUB-LINEAGE

This chapter represents initial findings from my doctoral work and was published as a research article in *BMC Genomics* on 2 Nov 2016 in volume 17, as article 851 (DOI: 10.1186/s12864-016-3203-3) [Jackson *et al*., 2016]. It has been adapted with permission for re-use within this dissertation, as it was published under a CC-BY license and the authors retain the original copyright.

Functional variants of human papillomavirus type 16 demonstrate host genome integration and transcriptional alterations corresponding to their unique cancer epidemiology

Robert Jackson[1,2], Bruce A. Rosa[3], Sonia Lameiras[4], Sean Cuninghame[1,5], Josee Bernard[1,6], Wely B. Floriano[7], Paul F. Lambert[8], Alain Nicolas[9], Ingeborg Zehbe[1,5,6]

[1]Probe Development and Biomarker Exploration, Thunder Bay Regional Research Institute, Thunder Bay, Ontario, Canada
[2]Biotechnology Program, Lakehead University, Thunder Bay, Ontario, Canada
[3]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA
[4]NGS platform, Institut Curie, PSL Research University, Paris, France
[5]Northern Ontario School of Medicine, Lakehead University, Thunder Bay, Ontario, Canada
[6]Department of Biology, Lakehead University, Thunder Bay, Ontario, Canada
[7]Department of Chemistry, Lakehead University, Thunder Bay, Ontario, Canada
[8]McArdle Laboratory for Cancer Research, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA
[9]Institut Curie, PSL Research University, Centre National de la Recherche Scientifique UMR3244, Sorbonne Universités, Paris, France

Keywords: Human papillomavirus, HPV16, E6 oncogene variants, Organotypic rafts, Viral integration, Transcriptomics, Pathogen-host relationship

2.1 – Abstract

2.1.1 – Background

Human papillomaviruses (HPVs) are a worldwide burden as they are a widespread group of tumour viruses in humans. Having a tropism for mucosal tissues, high-risk HPVs are detected in nearly all cervical cancers. HPV16 is the most common high-risk type but not all women infected with high-risk HPV develop a malignant tumour. Likely relevant, HPV genomes are polymorphic and some HPV16 single nucleotide polymorphisms (SNPs) are under evolutionary constraint instigating variable oncogenicity and immunogenicity in the infected host.

2.1.2 – Results

To investigate the tumourigenicity of two common HPV16 variants, we used our recently developed, three-dimensional organotypic model reminiscent of the natural HPV infectious cycle and conducted various "omics" and bioinformatics approaches. Based on epidemiological studies we chose to examine the HPV16 Asian-American (AA) and HPV16 European Prototype (EP) variants. They differ by three non-synonymous SNPs in the transforming and virus-encoded E6 oncogene where AAE6 is classified as a high- and EPE6 as a low-risk variant. Remarkably, the high-risk AAE6 variant genome integrated into the host DNA, while the low-risk EPE6 variant genome remained episomal as evidenced by highly sensitive Capt-HPV sequencing. RNA-seq experiments showed that the truncated form of AAE6, integrated in chromosome 5q32, produced a local gene over-expression and a large variety of viral-human fusion transcripts, including long distance spliced transcripts. In addition, differential enrichment of host cell pathways was observed between both HPV16 E6 variant-containing epithelia. Finally, in the high-risk variant, we detected a molecular signature of host chromosomal instability, a common property of cancer cells.

2.1.3 – Conclusions

We show how naturally occurring SNPs in the HPV16 E6 oncogene cause significant changes in the outcome of HPV infections and subsequent viral and host transcriptome alterations prone to drive carcinogenesis. Host genome instability is closely linked to viral integration into the host genome of HPV-infected cells, which is a key phenomenon for malignant cellular transformation and the reason for uncontrolled E6 oncogene expression. In particular, the finding of variant-specific integration potential represents a new paradigm in HPV variant biology.

2.2 – Background

Approximately 20% of human cancers are caused by infectious agents [Bouvard *et al*., 2009], including >500,000 patients diagnosed annually with human papillomavirus (HPV) associated cancers. Oncogenic HPV, denoted as "high-risk", is the primary risk factor for cervical cancer due to its exclusive tropism for mucosal tissues [zur Hausen, 1996; 2002]. Upon persistent infections of the cervical mucosa, oncogenic HPVs can cause progression from low- to high-grade cervical intraepithelial neoplasias that, without ablative treatment, may develop into invasive carcinomas. At the molecular level HPV is a double-stranded DNA virus and, to date, the sequences of over 200 types have been described [Kocjan *et al*., 2015]. The ~8 kbp genome of HPV contains 8 functional open reading frames (ORFs) that encode 5 early gene products (E1, E2, E5, E6 and E7) and 3 late gene products (E4, L1 and L2). While E1 and E2 are involved in DNA replication and transcriptional regulation of the viral genome [Doorbar *et al*., 2012], HPV's potent tumourigenicity is primarily due to E6 [Vande Pol and Klingelhutz, 2013], E7 [Roman and Münger, 2013], and E5 [Maufort *et al*., 2007]. L1 and L2 are structural proteins that self-assemble to form icosahedral capsids [Conway and Meyers, 2009], while the fused product of ORFs E1 and E4 (E1^E4) is most abundant in the productive viral life cycle, coinciding with the onset of viral DNA amplification [Middleton *et al*., 2003].

Among the HPV types, HPV16 (a member of species *Alphapapillomavirus 9*) is the most prevalent in cervical cancers. Intriguingly, and perhaps related to its prevalence, the HPV16 genome is polymorphic. Evolutionary analyses have revealed that the worldwide diversity of HPV16 genomes evolved for over 200,000 years [Bernard, 2005], leading to five phylogenetic branches representing isolates from Africa, Europe, Asia and the Americas [Yamada *et al*., 1997]. Furthermore, each branch can be further dissected into intratypic single nucleotide polymorphisms (SNPs) or variants differing in their host persistence and frequency of detection in human pre-cancers and cancers (reviewed in [Burk *et al*., 2013]). The tumourigenic differences of these SNPs have been ascribed largely to those within the E6 oncogene [Zehbe *et al*., 1998a; 1998b; 2001; Grodzki *et al*., 2006]. The Asian-American (AAE6) and European Prototype (EPE6) are common HPV16 genome variants which differ by six SNPs in their E6 genes, three of which are non-synonymous, leading to the 151-residue AAE6 protein differing by three amino-acids: Q14H, H78Y, and L83V [Cornet *et al*., 2012] (with residue 14 and 83 being under Darwinian constraint [Chen *et al*., 2005]).

Epidemiological studies showed that the AAE6 genome variant is a higher risk factor for dysplasia as well as an earlier onset of invasive tumours than EPE6 [Xi *et al*., 1997; 2007; Villa *et al*., 2000; Berumen *et al*., 2001; Zuna *et al*., 2009; Schiffman *et al*., 2007; Freitas *et al*., 2014]. As well, AAE6 has a greater transforming, migratory, and invasive potential than EPE6 when retrovirally transduced into primary human keratinocytes during recent long-term *in vitro* immortalization studies [Zehbe *et al*., 2009; Richard *et al*., 2010; Niccoli *et al*., 2012; Togtema *et al*., 2015]. These results suggested that coding changes in E6 have strong mechanistic and functional consequences for infection and thus contribute to marked differences in cancer risk of HPV16 variants.

To decipher the fundamental biology of HPVs and their tumourigenic features in a model system, the organotypic 3D infection model (raft culture) has the advantage of allowing reproducible and simultaneous epithelial differentiation and hence the occurrence of an active viral life cycle [Jackson *et al*., 2014; **Figure 2.1**]. Thus, using engineered human epithelium resembling *in vivo* conditions based on near-diploid immortalized keratinocytes (NIKS) [Allen-Hoffmann *et al*., 2000] we recently elucidated the phenotypic characteristics of both E6 gene variants in the context of the full HPV16 genome [Jackson *et al*., 2014], building upon previous work on the effects of transduction with the E6 or E6/E7 genes only [Zehbe *et al*., 2009; Richard *et al*., 2010; Schütze et al., 2014]. Using the organotypic model we observed that the AAE6 genome drives tumourigenesis by increasing epithelial proliferation, disrupting routine differentiation and apoptosis, evading the innate immune system and promoting immortalization [Jackson *et al*., 2014]. Interestingly, we also observed that the differences in host epithelia histologically classified as mild keratinizing (EPE6) or moderate (AAE6) dysplasia were reflective of increased oncogene (E6 and E7) expression in AAE6 cultures and loss of productive life cycle (decreased E2, E1^E4, and L2). Together these observations lead us to suspect integration of the AAE6 viral DNA into the host genome [Jackson *et al*., 2014], a common phenomenon during HPV-induced tumourigenesis (reviewed in [Poreba *et al*., 2011]).

Here, to further advance our mechanistic understanding of the impact of these common but epidemiologically and clinically important E6 SNPs, we conducted an "-omics" analysis on the NIKS-based organotypic epithelia containing the HPV16 variants AAE6 and EPE6 [**Figure 2.1**]. Modern deep sequencing techniques have been used to study HPV [Mine *et al*., 2013; Khoury *et al*., 2013; Bryant *et al*., 2014; Chandrani *et al*., 2015; Cullen *et al*., 2015], but only recently in the

context of intratypic variants [Lavezzo *et al*., 2016], and not using an organotypic epithelial model with full viral variant genomes. Instead, our complete approach allowed a comparison of these variants with regards to their integration capacity and subsequent transcriptional consequences in close to *in vivo* conditions, resulting in viral integration and a molecular signature of host chromosomal instability for AAE6 only.

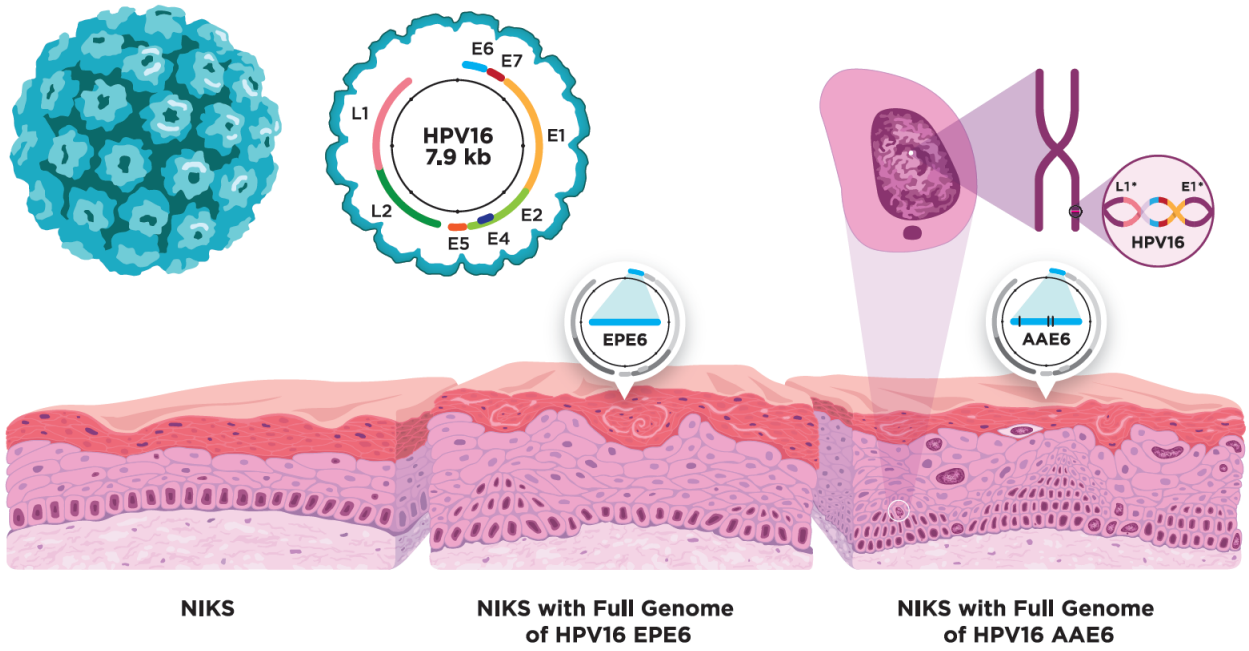


Figure 2.1 – The HPV16 genome and our experimental epithelial model. Contained within the viral protein capsid (*top left*, not to scale relative to skin) is the 7.9 kb HPV16 genome, comprised of eight viral genes. Over a 14-day differentiation process we grew three-dimensional organotypic epithelia, or raft cultures, using near-diploid immortalized keratinocytes (NIKS) and primary human fibroblasts embedded in collagen-based dermal matrix. To permit the viral life cycle in this culture system we transfected the keratinocytes, prior to rafting, with complete viral genomes containing either the European Prototype or Asian-American variant of HPV16 E6 (EPE6 or AAE6, respectively). NIKS represented normal epithelia, NIKS with HPV16 EPE6 was a mild dysplasia (indicated by thickening and some suprabasal proliferation), whereas NIKS with HPV16 AAE6 was a moderate dysplasia (indicated by a greater number of suprabasal proliferating cells and abnormal cellular phenotypes, including micronuclei). Additionally, HPV16 viral integration was detected in AAE6 epithelia.

2.3 – Results and Discussion

2.3.1 – Viral integration in the HPV16 AAE6 but not EPE6 epithelium

To permit the viral life cycle in a raft culture system, we transfected the keratinocytes, prior to rafting, with complete viral genomes containing either the HPV16 EPE6 or AAE6 variant. A similar technique was used in a recent study to successfully study varicella zoster virus [Jones *et al*., 2014], providing a keratinocyte model and a "global" perspective of all changes in host transcription in response to a pathogen. As illustrated in **Figure 2.1**, over a 14-day differentiation process, we observed that the NIKS were normal epithelia whereas NIKS with HPV16 EPE6 exhibited a mild dysplasia and NIKS with HPV16 AAE6 exhibited a moderate dysplasia.

To examine the HPV status of these cells we used the highly sensitive and high-throughput DNA capture and sequencing technique named Capt-HPV [Holmes *et al*., 2016]. We prepared genomic DNA from epithelia of both EPE6 and AAE6. Then, after double capture on the HPV probes, we performed 2 x 151 nt paired-end sequencing (see **2.5 – Methods**). As expected, we readily identified numerous HPV reads in both epithelial cultures. The sequencing reads of the E6 coding region confirmed the positive infection of the epithelia by the AAE6 and EPE6 variants. However, as we hypothesized [Jackson *et al*., 2014], the physical genomic status of HPV was clearly different. In the EPE6 epithelia, the reads covered the entire HPV genome indicative of its episomal state [**Figure 2.2a**] whereas only a fraction of the virus genome was detectable in the AAE6 epithelia, indicative of its integration into the host genome. Furthermore, in the case of EPE6, no human-viral junction reads were detected while the integrated AAE6 viral genome was truncated and several human-viral junction reads were identified in AAE6 epithelia. The integrated viral sequence was from nt 2453 (within HPV16 E1 gene) and nt 5780 (within HPV16 L1 gene) and thus includes the E6 and E7 oncogenes. Precisely, the insertion of the HPV16 AAE6 variant occurred between the nt position 149,347,294 and 149,347,305 of chromosome 5. Mechanistically, this is a simple "end-out" integration event with a typical two junction, co-linear (2J-COL) signature [Holmes *et al*., 2016], associated with a very short 11 bp deletion of the host genome, and two overlapping nucleotides between viral and human sequence at each junction [**Figure 2.2a**]. Functionally, the insertion occurred within the 5q32 sub-band region, and more precisely, within the first intron of the SLC26A2 gene, approximately 13 kb upstream of its third exon.

Based on the Dr.VIS (Viral Integration Site) v2.0 database of HPV16 integration sites [Yang *et al*., 2015], this exact region (5q32) of integration is not frequent, but potentially recurrent

as it was found in 2 out of 878 previously documented sites. The nearest fragile site was 13 Mb upstream of this integration site: FRA5C, 5q31.1. Since repeated regions might be prone to genome rearrangements and therefore prone to HPV integration, we scanned the adjacent regions using the UCSC hg19 genome browser RepeatMasker track for human repeat elements and found a nearby 158 bp long interspersed nuclear element (LINE): L1MB5 located from Chr5 nt position 149,347,143 to 149,347,300. Indeed, L1MB5-derived sequences have been documented as breakpoints, such as in the human genes HPRT [Williams *et al*., 2002], CYP2C [Zhou, 2016], and in proximity of genes containing the ubiquitin ligase Mib-herc2 domain, which mediates Notch signalling [del Rosario *et al*., 2014]. Strikingly, this domain contains the Hect region, homologous to the E6-associated protein carboxyl terminus, raising the question of whether or not the underlying homology could play a role in this target site selection. Another, non-exclusive hypothesis is that the frequent hypo-methylation of LINE elements plays a role to facilitate access to the chromosomal DNA and associated genomic instability [Richards *et al*., 2009; Baba *et al*., 2014]. Altogether, our three-dimensional organotypic cultures demonstrated that the HPV16 AAE6 variant had integrated into the host genome while the EPE6 variant remained episomal, suggesting an increased propensity towards integration due to AAE6. A previous study of HPV16 integration propensity with respect to the variants did not demonstrate a statistically significant difference ($P = 0.28$, two-tailed Fisher's exact test) between EPE6 (3 episomal and 20 integrated cases) and the E-T350G variant (6 episomal and 16 integrated cases, responsible for one of the residue changes also found in AAE6: L83V) [Xu *et al*., 2013]. Only one tumour sample in their set contained the AA variant, therefore precluding a formal analysis of its propensity to integrate, but notably it was in integrated form.

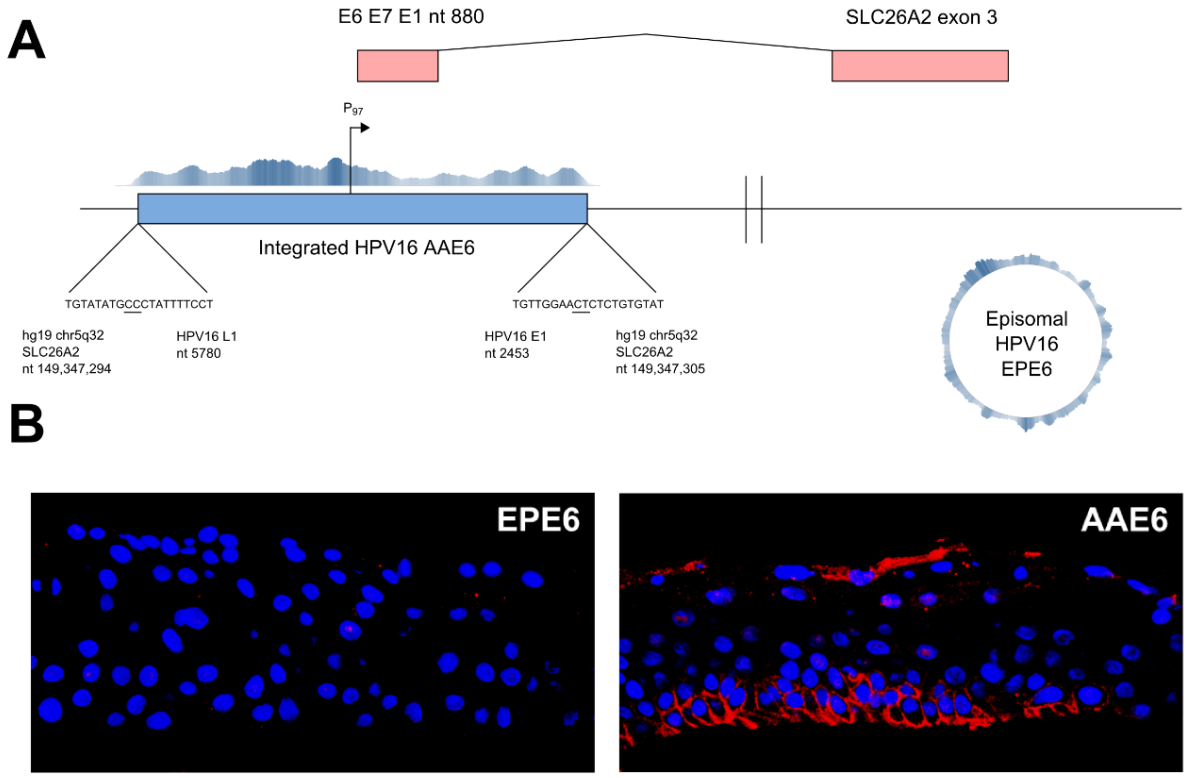


Figure 2.2 – Characterization of viral integration and viral-human fusion transcripts in AAE6 epithelia. [a] Integration site schematic showing viral and human junctions, including nucleotide positions, the early promotor, as well as viral-human fusion transcript between HPV16 early region and SLC26A2 exon 3. A coverage plot above the integrated HPV16 genome demonstrates coverage across the junction sites within SLC26A2 (5482 reads across a 4978 nt assembly containing the AAE6 integrated form flanked by 200 nt of SLC26A2), while a circular coverage plot on the right shows the full episomal assembly of the EPE6 episomal form. **[b]** Immunofluorescence overlays of EPE6 and AAE6 raft cultures (400x magnification). Nuclei are indicated by blue DAPI staining while SLC26A2 is indicated by red fluorescence.

2.3.2 – The HPV16 AAE6 epithelium has a unique transcriptional profile

Another essential feature that may differentiate the behaviour of the HPV16 EPE6 and AAE6 variants is expression of the viral genome, viral-human fusion transcripts when integrated, as well as downstream host effects due to expression of the E6/E7 oncogenes. To assess these, we performed a genome-wide RNA-Seq analysis of the EPE6 and AAE6 epithelia using Illumina sequencing of total RNAs (see Methods), mapping first against our reference HPV16 W12E genome [GenBank AF125673]. Viral transcriptomes were visualized with the Integrative Genomics Viewer (IGV) [Robinson *et al*., 2011], while viral gene counts and variant calls were performed using SAMtools [Li *et al*., 2009]. The average sequencing depth of 40.4 million total reads per sample (~20 to 25 million fragments producing paired-end reads) was appropriate to detect the small proportion of total reads of both HPV variant genomes (~0.0001 to 0.01%, **Additional file 1: Table 2.S1**), while none were detected in the HPV-negative control epithelium. The variant-specific non-synonymous SNPs (relative to the reference HPV16 W12E genome) present in EPE6 (G350T) and AAE6 (G145T+C335T) were confirmed with depth of reads of 6x for EPE6 and with 14x to ~300x depth of reads for AAE6. Among the EPE6 epithelial samples, we detected few E6, E6*I (spliced transcript), E7, E1, E2, E1^E4, and E5 transcripts, with even fewer L2 and L1 reads, as confirmed by L2 RT-qPCR and L2 protein immunohistochemistry results from the same independent set of rafts reported previously [Jackson *et al*., 2014]. Among the three individual epithelial raft cultures for EPE6 samples the viral transcriptional landscape appeared similar but the read coverage was higher in raft #2 due to an overall higher abundance of viral transcripts in this sample [**Figure 2.3a**]. In contrast, the transcriptional landscape for the three AAE6 samples was more homogenous [**Figure 2.3a**], further emphasized in a clustered heatmap [**Figure 2.3b**]. Abundant full-length E6, E6*I, E7, and only truncated E1 and L1 transcripts were detected. Full-length E1, E2, E1^E4, and L2 reads were absent in AAE6 epithelia, consistent with the Capt-HPV data reported above and our previous RT-qPCR results and DNA copy number analyses on these molecules [Jackson *et al*., 2014].

To quantitatively account for sample variance, we also performed differential expression analysis of the viral gene counts using DESeq [Anders and Huber, 2010]. DESeq software tests for differential expression in library size-corrected count data using a negative binomial distribution model. In agreement with our previous RT-qPCR results [Jackson *et al*., 2014], we found significantly more E6 (24.05 fold higher, $P < 10^{-10}$) and E7 (17.30 fold higher, $P < 10^{-10}$)

counts in triplicate AAE6 rafts in comparison to triplicate EPE6 rafts [**Figure 2.3c**]. Taken together, analyses of viral transcriptome data revealed that the AAE6 viral transcriptome significantly differs from that of EPE6 in a manner that is indicative of integration, with increased E6 and E7 levels [Dürst *et al*., 1991; Jeon *et al*., 1995; Daniel *et al*., 1997]. Evidently, AAE6 transcriptome profiles are lacking E2 and have increased E6/E7 oncogene expression, perhaps due to loss of transcriptional repression by E2. We therefore reasoned that the increased levels of E6/E7 expression between the variants were ultimately due to their viral integration status, as we hypothesized in our phenotypic study, and confirmed by Capt-HPV, leading to a significant effect on the host transcriptome [Jackson *et al*., 2014].

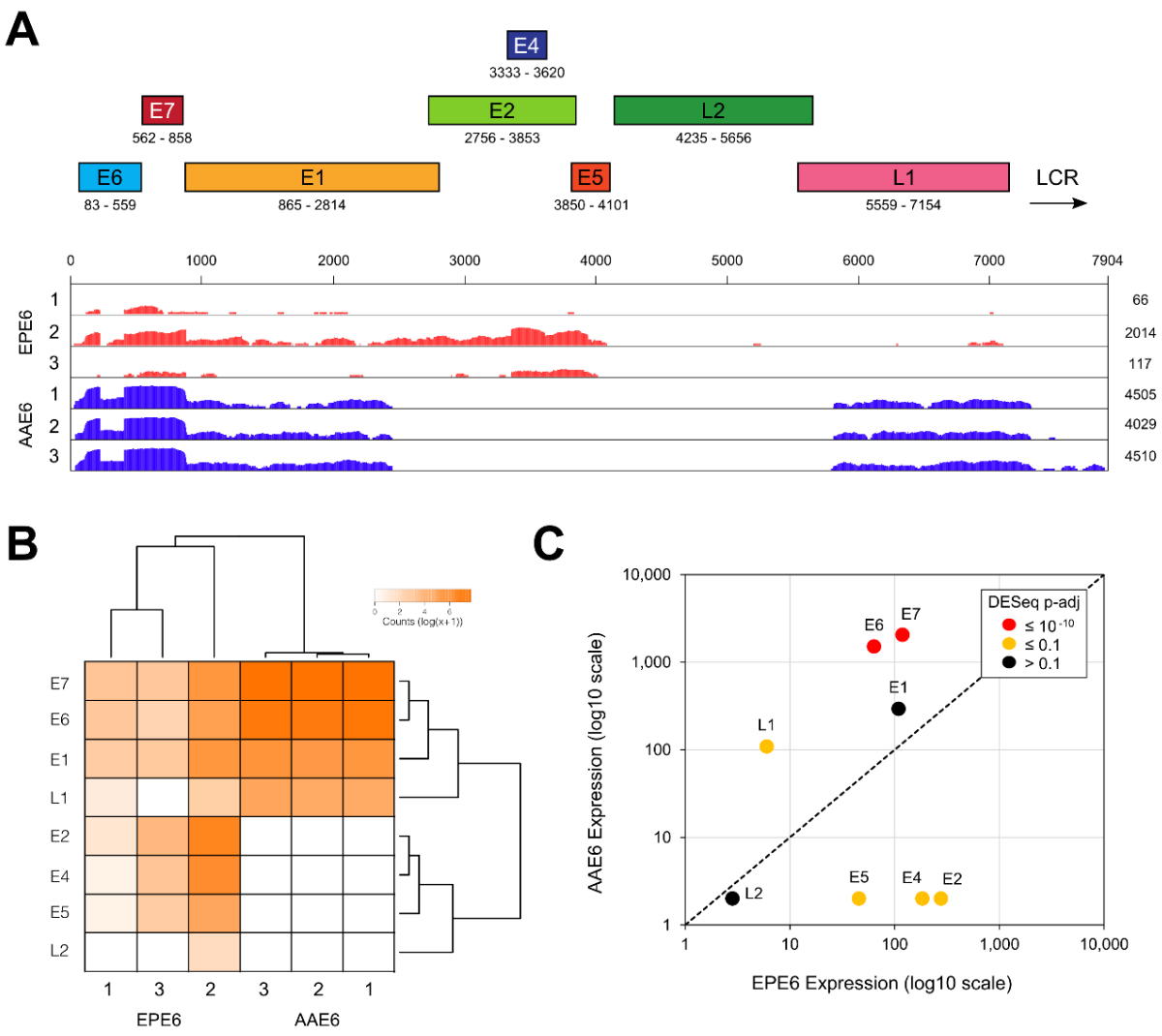


Figure 2.3 – The HPV16 transcriptome in EPE6 and AAE6 organotypic rafts. [a] Linear viral gene map. Viral RefSeq ([GenBank: AF125673], HPV16 W12E genome) alignment from each individual raft culture was visualized using IGV [Robinson *et al*., 2011]. The y-axis (coverage) is $\log_2$ scaled. Total number of viral reads are given on the right-hand side of each track. **[b]** Heatmap & clustering analysis of viral transcriptome on DESeq normalized counts: viral genes vs sample replicates. Two distinct sample clusters matched EPE6 and AAE6 replicates respectively, clustering independently of each other. Within the high-variability EPE6 cluster, replicate 1 and 3 were clustered together. Within the low-variability AAE6 cluster, replicate 1 and 2 were clustered together. As well, AAE6 epithelia converged on consistently high viral transcription (specifically E6/E7). From the viral gene perspective, two distinct clusters were identified: E6, E7, E1, and L1

in one, and E2, E4, E5, and L2 in another. Within the first primary cluster, E6 and E7 cluster close together, as expected given they are expressed together as a multi-cistronic transcript. E1 and L1 also cluster together, constituting the truncated transcripts on the periphery of the non-transcribed region within AAE6 samples. In the second primary cluster, E5 and L2 cluster together, independent of E2 and E4 which is transcribed only in EPE6 samples. E2 and E4 expression unsurprisingly clusters together given that E4 is contained within the E2 ORF. **[c]** Scatterplot of average viral gene expression for EPE6 samples (x-axis) and AAE6 samples (y-axis). The axes (DESeq normalized gene counts) are $\log_{10}$ scaled. Significant differential gene expression is denoted by marker colour. *Dashed line* represents equal expression.

2.3.3 – Nature of viral-human fusion transcripts detected in HPV16 AAE6 epithelium

The integration of HPV16 genomes into host chromosomes is a frequent phenomenon associated with carcinogenesis, and not only modifies the expression of HPV-encoded E6 and E7 oncogenes [**Figure 2.3a**] but can also trigger the expression of fusion viral-human mRNAs [Poreba *et al*., 2011; Lace *et al*., 2011]. Since the virus can integrate into a variety of positions in the human genome, these fusion transcripts are specific to each integration site. In recent years, following the introduction of high-throughput sequencing techniques, multiple software for detecting pathogen sequences in host sequence data have become available [Hawkins *et al*., 2011; Westermann *et al*., 2012; Bonfert *et al*., 2013; Chen *et al*., 2013; Li *et al*., 2013; Wang *et al*., 2013; Katz and Pipas, 2014; Chandrani *et al*., 2015]. Here, to identify the viral-human fusion transcripts expressed in our epithelia, we used the ViralFusionSeq (VFS) software [Li *et al*., 2013; Lau *et al*., 2014]. VFS was chosen over alternatives due to its optimization for RNA-Seq data from the Illumina platform, the ability to define our own reference virus genome, as well as the full suite of fusion transcript discovery techniques it uses. Using this technique, only the AAE6 rafts yielded viral-human fusion transcripts [**Table 2.1**], providing further evidence of viral integration as well as its transcriptional impact.

In accordance with the structure of the HPV integration, the transcript breakpoints mapped to either the E1 or L1 HPV16 ORF. Alternative splicing was detected with the viral nucleotide position at the fusion site of one class of the viral-human fusion transcripts [**Figure 2.3a**]: nt 880 (splice donor, SD) in the E1 gene [Johansson and Schwartz, 2013]. This is the same SD site for the E1^E4 splice transcript typically expressed in the late stage of the viral life cycle [Doorbar, 2005], and previously shown to be expressed in our EPE6 epithelia [Jackson *et al*., 2014]. HPV16 viral-human fusion transcripts are often detected with a breakpoint at this natural splice donor site [Wentzensen *et al*., 2002; Kraus *et al*., 2008; Lace *et al*., 2011], and the coverage plot for AAE6 shows decreased coverage for transcripts downstream of this E1 SD site, supporting the hypothesis of alternative splicing. With respect to the L1 breakpoints, the typical L1 splice acceptor (SA) site is at nt 5639 [Johansson and Schwartz, 2013], but notably in our study, the viral-human fusion transcripts here had a putative downstream SA site at nt 5778. Interestingly, the coverage plot of the viral transcriptome shows nt 5778 as the site where L1 coverage begins to be detected in AAE6 rafts [**Figure 2.3a**], so we reasoned that this discrepancy in SA site could be due to either a cryptic

SA site in the HPV16 W12E genome (although not found previously in the literature) or simply due to integration truncating the upstream region of L1.

Next, we mapped the human portion of the fusion transcripts using VFS's clipped-seq (CS) and read-pair (RP) methods. Confirmed by both these methods, two fusions mapped to the human chromosome location 5q32, occurring within the solute carrier family 26 (anion exchanger), member 2 (SLC26A2) and phosphodiesterase 6A, cGMP-specific, rod, alpha (PDE6A) human ORFs [**Table 2.1**]. Strikingly, along with detection of fusion transcripts with these genes, we detected a significant increase in the expression of human genes from this region in AAE6 epithelia compared to normal epithelia, namely SLC26A2 (114.19 fold increase, $P = 2.14 \times 10^{-173}$) and colony-stimulating factor 1 receptor (CSF1R, 407.82 fold increase, $P = 4.70 \times 10^{-112}$, which was only detected as RP fusion reads by VFS, and not confirmed by CS). This observation is in agreement with others who have found that, in numerous cervical carcinomas across multiple high-risk HPV types, HPV integration leads to an increase in the expression of genes adjacent to integration loci [Ojesina *et al*., 2014]. To explain the molecular basis of this cis-effect, it has been proposed to be the result of viral promotor-driven expression or somatic genome amplification at the integration site [Peter *et al*., 2010; Akagi *et al*., 2014]. In the present case, this last hypothesis is unlikely because the AAE6 integration produced a clean 11 bp deletion of the target region that led to two co-linear viral-human junctions (2J-COL), which is not associated with gene amplification [Holmes *et al*., 2016].

Functional human fusion proteins can be formed due to chromosomal translocations in cancer cells [Rabbitts, 1994]. The elucidation of novel protein-coding viral-human fusion transcripts is particularly intriguing due to their potentially functional roles within host cells. Using immunofluorescence for the expressed portion of the SLC26A2 protein in formalin-fixed and paraffin embedded (FFPE) rafts, we determined that SLC26A2 protein expression was aberrantly high in AAE6 compared to EPE6, supposedly as a result of its viral-human fusion and increased transcription [**Figure 2.2b**]. This translated fusion protein contains exon 3 of the transmembrane protein SLC26A2, previously known as diastrophic dysplasia sulfate transporter (DTDST) [Hästbacka *et al*., 1999], which encodes the carboxy-terminal cytoplasmic sulfate transporter and anti-sigma factor (STAS) domain [Sharma *et al*., 2011]. We cannot find any evidence in the literature of this unique viral-human fusion protein in other HPV-integrated samples. Overall, these chimeric molecules are unique for each sample and to the specific integration site, with

presently unknown effect on host cell functions, an aspect to be further researched due to its importance for understanding mechanisms of tumourigenesis as well as in the emerging field of personalized medicine.

Table 2.1 – Integration loci detected by ViralFusionSeq. Viral-human fusion transcripts were discovered using ViralFusionSeq's [Li *et al*., 2013]: clipped-sequence (CS) and read-pair (RP) modules. Detected by at least 1 RP and CS event (†). As detected by CS method (‡). VFS uses two methods to detect viral-human fusion transcripts. The Clipped-Seq (CS) method detects viral fusion transcript breakpoints with a read that maps to both viral and human sequences, while the Read-Pair (RP) analysis detects transcripts with read ends mapped separately to the viral and human genome [Li *et al*., 2013]. We required candidate viral fusion transcripts to be supported by at least 1 CS and 1 RP event in order to improve its stringency [Lau *et al*., 2014]. Although RP events were more abundant in our samples, CS analysis provided single-base resolution of viral-human fusion transcript breakpoints. In particular, we identified an average of 1.33 +/− 1.53 CS transcripts in EPE6 and 7.66 +/− 6.66 in AAE6. We detected no RP transcripts in EPE6, while 118.66 +/− 7.23 were found in AAE6 rafts. While one RP transcript was detected in a NIKS control culture, this read was not confirmed by the CS method of VFS and therefore not considered as a valid event.

Sample	Mapped human transcript†	Gene description	Chromosome location	HPV transcript breakpoint(s)‡
AAE6	SLC26A2	Solute carrier family 26, member 2	5q32	E1, L1
	PDE6A	Cyclic GMP- Phosphodiesterase 6A alpha subunit	5q32	E1, L1
EPE6	None	–	–	–
NIKS	None	–	–	–

2.3.4 – The HPV16 AAE6 epithelium reveals a signature of chromosomal instability conducive to host genome integration

Integration of HPV DNA into the host genome is considered to be a key factor for cervical cancer development [Wentzensen *et al*., 2002; Pett and Coleman, 2007; Bodelon *et al*., 2016], but the cellular events that initiate the integration process (and selection of insertion sites) remain to be better understood. A reasonable hypothesis is that the integration is triggered by a rare and stochastic target site event, such as a replicative fork stalling or an accidental chromosome double-strand break, leading to an ultimate use of the viral DNA for repair via recombination, template switching (FoSTeS) and/or microhomology-mediated break-induced replication (MMBIR) ([Akagi *et al*., 2014; Hu *et al*., 2015; Holmes *et al*., 2016], and references within each). Indeed, infections with pathogens can cause chromosomal instability by inactivating the host DNA damage response [Weitzman and Weitzman, 2014]. For HPV, this has been linked to the expression of both HPV16 E6 and E7 oncoproteins, affecting the infected cell's genome integrity [White *et al*., 1994; Kessis *et al*., 1996; Duensing *et al*., 2000; Duensing and Münger, 2002]. A model of early carcinogenesis due to HPV16 E6 and E7 suggests that this chromosomal instability is caused by uncontrolled proliferation, leading to an insufficient nucleotide pool that cannot support normal replication [Bester *et al*., 2011]. Alternatively, E6 alone, through the inactivation of p53, can promote chromosomal instability, at least during early onset of carcinogenesis [Havre *et al*., 1995]. Presently, HPV16 AAE6 demonstrated enhanced integration propensity over EPE6 and exhibited increased E6 and E7 oncogene expression, which is in accordance with elevated E6 and E7 levels reported in other studies [Dürst *et al*., 1991; Jeon *et al*., 1995; Daniel *et al*., 1997]. This enhanced integration ability is based on AAE6's greater proliferation ability, leading to chromosomal instability. The underlying mechanism of its increased cell growth is the result of a deregulated sugar metabolism (Warburg effect), as we reported previously [Richard *et al*., 2010] and currently under study [Cuninghame *et al*., 2017].

To assess the host chromosomal instability in our HPV16 variant epithelia, we examined our RNA-Seq data to detect the CIN70 gene expression signature [Carter *et al*., 2006], which has been applied as a prognostic marker in cervical cancer [How *et al*., 2015] and more generally as a significant indicator to predict clinical outcome across multiple cancer types [Carter *et al*., 2006]. This signature is derived from 18 gene expression datasets (with genes ranked based on their correlation to functional aneuploidy). The CIN70 score relative to HPV-negative NIKS was

significantly higher in AAE6 compared to EPE6 epithelia (2.32 fold higher, $P = 0.02$ by Welch's *T*-test), indicating a signature of host chromosomal instability in AAE6 epithelia [**Figure 2.4a**]. Furthermore, as a morphological sign of chromosomal instability, we detected micronuclei (MN) in AAE6 but not EPE6 or NIKS FFPE H&E-stained epithelia [**Figure 2.4b**]. MN were reported to be present in higher grade cervical intraepithelial neoplastic lesions and invasive cervical cancer [Samanta *et al*., 2011] and mechanistically have been associated with hallmarks of genomic instability [Zhang *et al*., 2015].

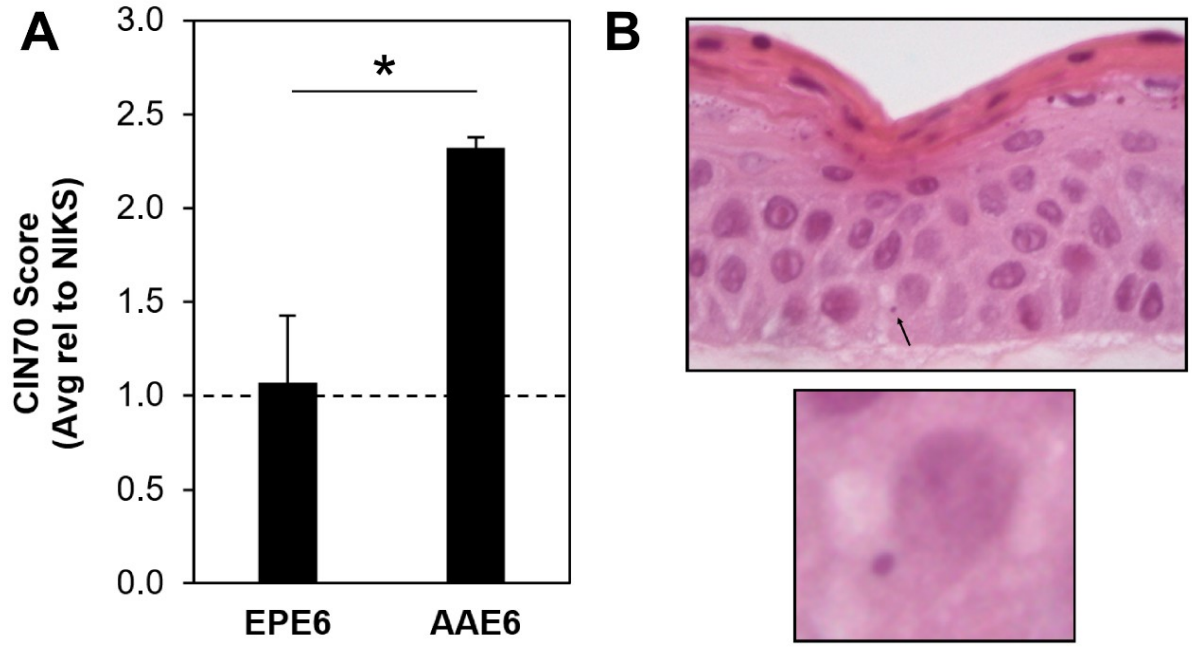


Figure 2.4 – Chromosomal instability signature and micronuclei in AAE6 epithelia. [a] The CIN70 score relative to HPV-negative NIKS was significantly higher in AAE6 compared to EPE6 epithelia (2.32 fold higher, $P = 0.02$ by Welch's *T*-test). Mean values are shown with error bars representing standard deviation ($n = 3$). Statistical significance ($P < 0.05$) denoted by "*". **[b]** Haematoxylin and eosin micrographs of FFPE AAE6 epithelia, 400x cropped, micronuclei indicated by *arrow*. Close-up shows micronucleus and normal-sized nucleus within same cell.

2.3.5 – HPV16 AAE6 epithelium exhibits a proliferating phenotype as a consequence of viral integration into the host genome

More broadly, our RNA-Seq data led us to examine global changes in host gene expression. Our previous study demonstrated enhanced tumourigenesis by the full HPV16 genome with AAE6 [Jackson *et al*., 2014], while another study presented altered gene expression by the AA variant [Sichero *et al*., 2012]. Work by other groups have studied the downstream pathways in the AA variant [Hochmann *et al*., 2016; Zacapala-Gómez *et al*., 2016], and have utilized high-throughput techniques to investigate genetic variation within HPV16 [Cullen *et al*., 2015; Muller *et al*., 2015; Lavezzo *et al*., 2016], but this is the first study investigating the downstream pathways affected by the HPV16 variants in an organotypic epithelial model using next-generation sequencing. We hypothesized two scenarios that can be associated with these findings and analyzed in our present study: *i*) the global gene expression profile within AAE6-infected epithelium would differ significantly from that of EPE6 and *ii*) significant gene expression differences in the host due not only to the actions of the viral oncogenes E6 and E7, but also as a result of integration [Lace *et al*., 2011]. A global "-omics" technique, RNA-Seq, was required to sufficiently address our hypotheses around the functional relevance of the AA variant in epithelia. We assessed host differential gene expression using DESeq [Anders and Huber, 2010] to determine how it reflected the unique viral gene expression profiles induced in human epithelium undergoing differentiation. Strikingly, NIKS, which contain no virus genome, had zero significant differentially expressed genes compared to EPE6, at a false-discovery rate (FDR) of 10% [**Additional file 2: Figure 2.S1**]. NIKS to AAE6 had 3006 significant differentially expressed genes [**Additional file 2: Figure 2.S2**, **Additional file 3** for list of differentially expressed genes between NIKS and AAE6]. Of these genes, 1312 were down-regulated while 1694 were up-regulated in AAE6 compared to NIKS. The lack of any differentially expressed genes between NIKS and EPE6 organotypic epithelial cultures was surprising, but consistent with the similarity between the NIKS and EPE6 cultures monitored with respect to basal and suprabasal keratinocyte proliferation assessed by BrdU-incorporation, p53 and $p16^{INK4A}$ by immunohistochemistry and IFN-κ by RT-qPCR [Jackson *et al*., 2014]. Phenotypically, these results suggest that the episomal expression of the EPE6 variant in our model does not have a significant tumourigenic effect. Since our 3D culture model specifically captures early tumourigenesis, with only a 2-week growth period and low initial viral copy number, very small gene expression differences in a homogenized epidermal sample are not expected to be easily

detected with global transcriptomic techniques. On the other hand, AAE6 significantly perturbed a high number of human genes, demonstrating its ability to cause a wide-range of host molecular changes consistent with tumourigenesis. Compared to EPE6, AAE6 had 1666 significant differentially expressed genes [**Additional file 2: Figure 2.S3**, **Additional file 3** for list of differentially expressed genes between EPE6 and AAE6]. Of these genes, 666 were down-regulated while 1000 were up-regulated in AAE6 compared to EPE6. Additional discussion of the top-ten most significant down- and up-regulated genes for each pair-wise comparison is provided in **Additional file 4**. To further investigate the differential gene expression data we applied two additional bioinformatics analyses: gene ontology (GO) biological process term enrichment [**Additional file 5** for GO output, **Figures 2.5** and **2.6**], as well as co-expression analysis and visualization using networks [**Figure 2.7**]. Finally, we also compared the pair-wise lists of differentially expressed genes to determine the number of common and unique genes among each set [**Figure 2.8**]: 1541 genes unique to the NIKS comparison, 201 unique to the EPE6 comparison, and 1465 common between them. Overall, these bioinformatics analyses highlight the global effects of AAE6 on host epithelia due to its integration event, increased E6/E7 expression, and perhaps in part functional differences due to the AAE6 oncoprotein itself: increased proliferation and decreased differentiation.

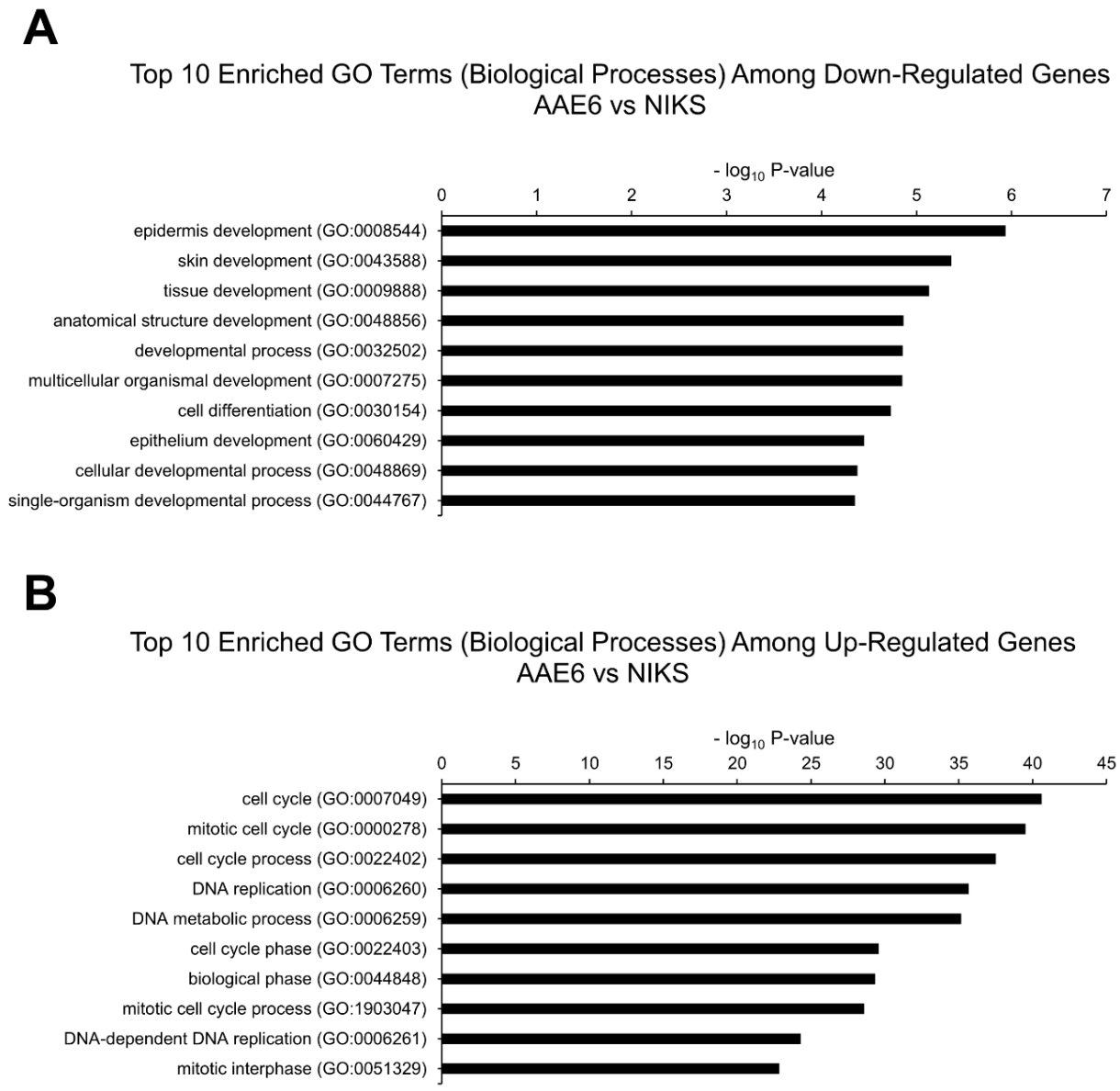


Figure 2.5 – Gene Ontology (GO) terms enriched in highly significant differentially expressed genes in AAE6 vs. NIKS. The Term Enrichment Service available on the AmiGO 2 website [Carbon *et al*., 2009] was used to determine enriched GO (biological process) terms among **[a]** down-regulated and **[b]** up-regulated genes. Only the top ten GO terms are shown for each. See **Additional file 4** for discussion.

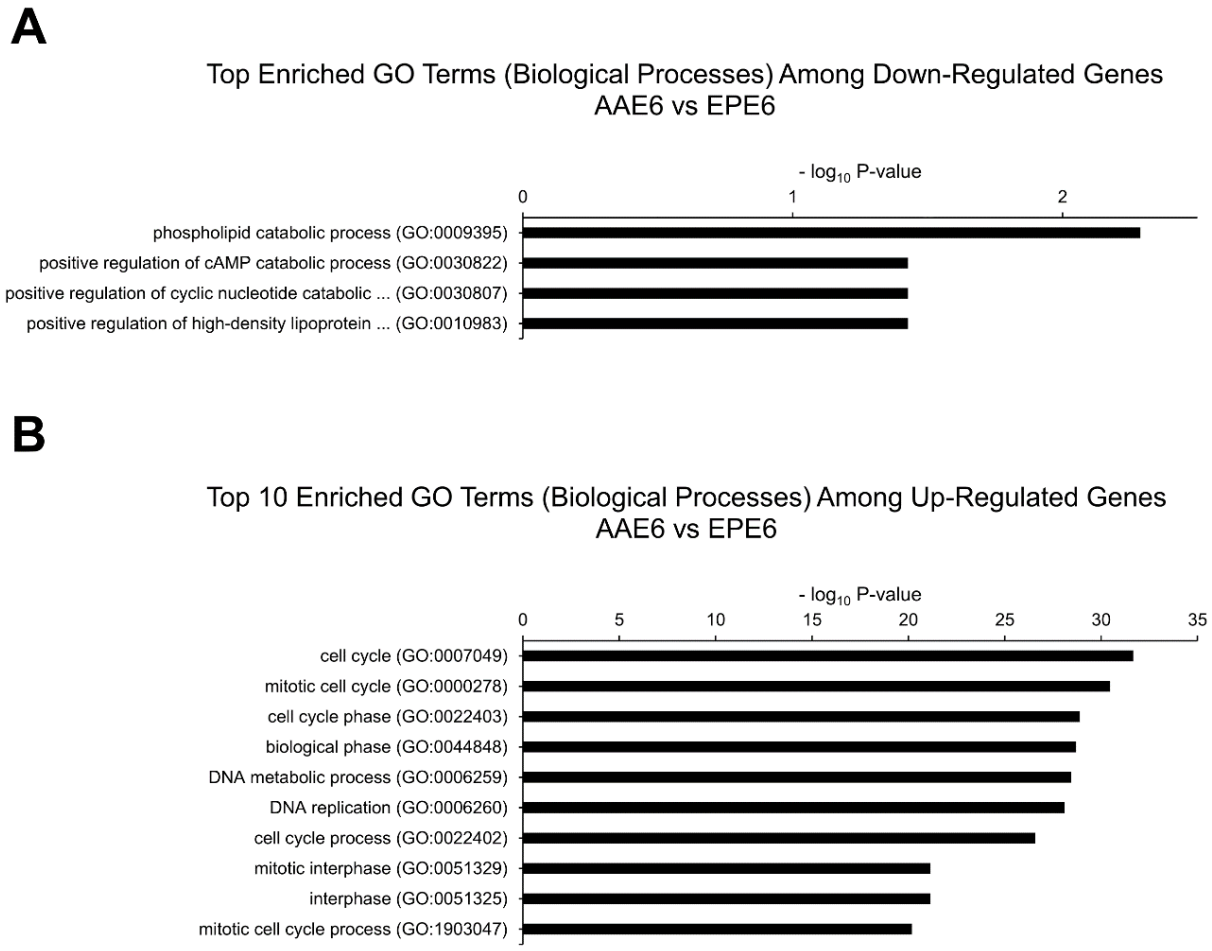


Figure 2.6 – Gene Ontology (GO) terms enriched in highly significant differentially expressed genes in AAE6 vs. EPE6. The Term Enrichment Service available on the AmiGO 2 website [Carbon *et al*., 2009] was used to determine enriched GO (biological process) terms among **[a]** down-regulated and **[b]** up-regulated genes. Only the top ten GO terms are shown for each. See **Additional file 4** for discussion.

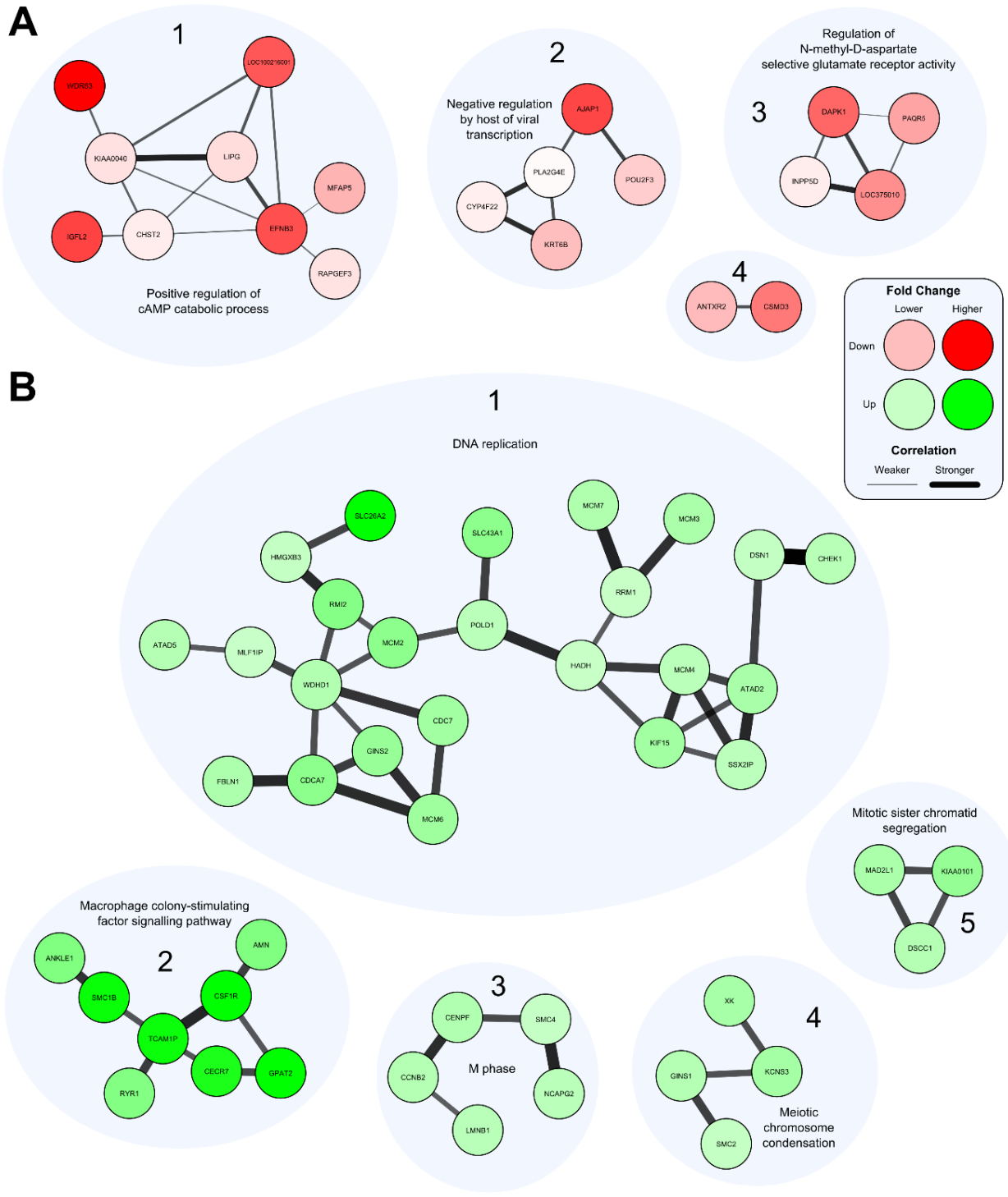


Figure 2.7 – Co-expression networks of highly significant [a] down-regulated and [b] up-regulated genes in AAE6 vs. EPE6. [a] Four discrete clusters of down-regulated and co-expressed genes were observed. Only co-expressed genes with a Pearson correlation coefficient greater than 0.95 are shown. Clusters are labelled by number and functionally annotated with their

significantly enriched biological process. Nodes = gene, denoted by gene symbol; node colour = white to red with down-regulation (fold change) in AAE6 from EPE6; edge thickness = increases with Pearson correlation coefficient. **[b]** Five discrete clusters of up-regulated and co-expressed genes were observed. Only clusters co-expressed genes with a Pearson correlation coefficient greater than 0.996 and are shown, to narrow down the number of genes displayed. Clusters are labelled by number and functionally annotated with their significantly enriched biological process. Nodes = gene, denoted by gene symbol; node colour = white to green with up-regulation (fold change) in AAE6 over EPE6; edge thickness = increases with Pearson correlation coefficient. See **Additional file 4** for discussion.

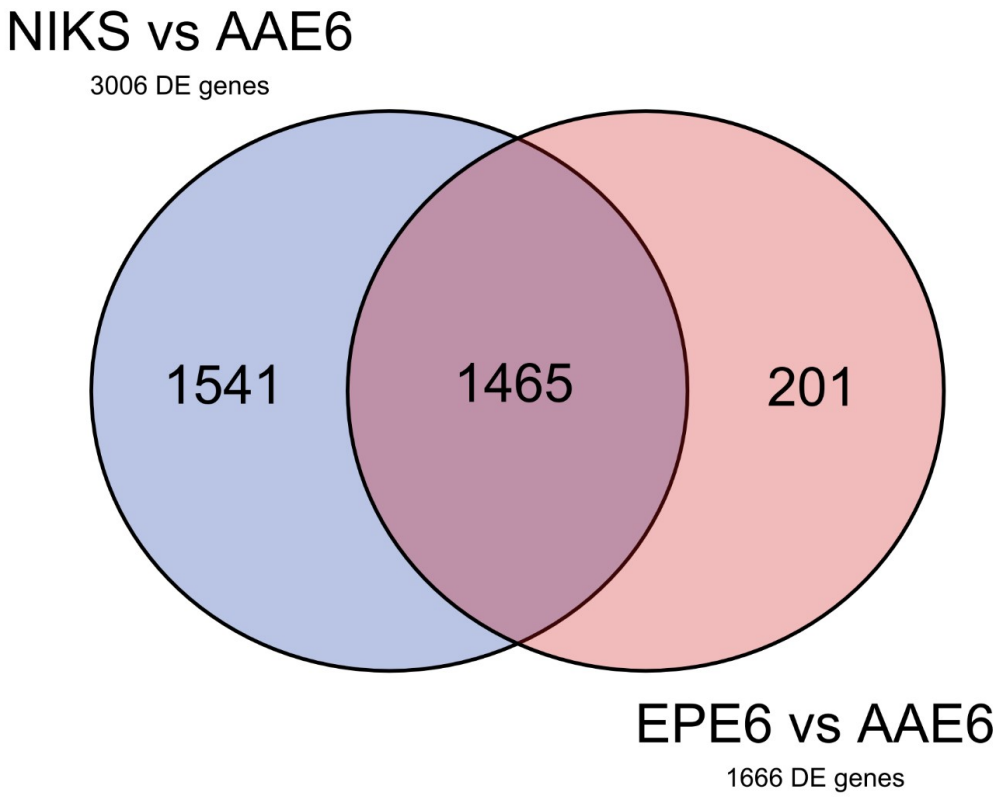


Figure 2.8 – Venn diagram of differentially expressed genes common and unique to each pairwise comparison. Of the 3006 differentially expressed (DE) genes in NIKS vs AAE6 and the 1666 differentially expressed (DE) genes in EPE6 vs AAE6 there were 1541 genes unique to the NIKS comparison, 1465 common between them, and 201 unique to the EPE6 comparison. No genes were up-regulated in one set of a pair-wise comparison (either NIKS vs EPE6 or EPE6 vs AAE6) while down-regulated in the other.

2.4 – Conclusions

We have systematically characterized the viral integration process of a common high-risk HPV16 variant and its consequences for the affected host cell. This and earlier work lend themselves to propose a model of increased tumourigenicity in human keratinocyte epithelia where AAE6's enhanced ability to proliferate leads to chromosomal instability. In such an environment, the host genome may be susceptible to viral integration subsequently increasing E6/E7 oncogene expression and ultimately driving additional tumourigenic changes. Previously, we performed phenotypic studies of the EPE6 and AAE6 variants in a 3D raft model of early carcinogenesis [Jackson *et al*., 2014] and determined the functional differences of these variants in longitudinal monolayer cell cultures [Zehbe *et al*., 2009; Richard *et al*., 2010; Niccoli *et al*., 2012; Togtema *et al*., 2015]. While necessary for studying the viral life cycle, limitations of the current organotypic model are the lack of immune components, vasculature, and the complexity of tissue heterogeneity that arises. Our current study builds on the foundation of these investigations. We have applied a wide range of molecular analyses, creating a framework which can benefit future virus-host interaction studies with various organotypic cell culture models. A variant-specific integration is worth reporting and should be further investigated, with additional samples from independent donors, as it represents a new paradigm in HPV variant biology. Here we report a viable integration mechanism in a robust viral life cycle model for AAE6. The findings of the current and other studies reported by us [Zehbe *et al*., 2009; Richard *et al*., 2010; Niccoli *et al*., 2012; Jackson *et al*., 2014; Togtema *et al*., 2015], and others [Sichero *et al*., 2012; Hochmann *et al*., 2016; Zacapala-Gómez *et al*., 2016], are consistent with cancer epidemiology studies demonstrating that the HPV16 AA variant is a higher risk factor for high-grade intraepithelial neoplasia and progression to invasive cervical cancer [Berumen *et al*., 2001; Xi *et al*., 2007; Zuna *et al*., 2009; Sichero *et al*., 2012]. In the future, HPV variant genotyping could be used as a clinical prognostic factor for patient-centered health services, while the role of individual host genomics on integration, including characterization of integration sites, will be important to consider for personalized medicine approaches.

2.5 – Methods

2.5.1 – Cell lines

As described by us previously [Jackson *et al*., 2014], we used the Normal/Near-Diploid Immortalized Keratinocytes (NIKS) cell line [Allen-Hoffmann *et al*., 2000] to establish 3D organotypic epithelia cultures. These spontaneously immortalized cells were originally derived from neonatal human foreskin and are non-tumourigenic, though contain an additional long arm piece of chromosome 8 (8q). In monolayer they are grown on mitomycin-C-treated Swiss mouse J2/3T3 fibroblast feeder layers [Allen-Hoffmann *et al*., 2000], while primary human foreskin fibroblasts (ATCC CRL-2097) are incorporated into the dermal equivalent of organotypic NIKS cultures [Jackson *et al*., 2014].

2.5.2 – Detection of integrated papillomavirus sequences by DNA-Seq: Capt-HPV

DNA-Seq was used to confirm the presence and location of the viral integration sites in the human genome using DNA extracted from formalin-fixed paraffin embedded (FFPE) samples which had been prepared previously [Jackson *et al*., 2014]. DNA was extracted using the DNeasy Blood and Tissue Kit (QIAGEN, Cat# 69504) with the recommended pre-treatment for FFPE samples and the optional RNase treatment. To overcome the limitations of traditional techniques, such as DIPS-PCR (Detection of Integrated Papillomavirus Sequences by ligation mediated PCR), we used an unbiased and state-of-the-art next-generation DNA sequencing technique for detecting HPV viral integration sequences in our samples [Holmes *et al*., 2016]. Library preparation, sequence capture, and high-throughput sequencing was carried out at the Institut Curie on an Illumina MiSeq platform with a V2 Nano chip (~1 x 10^6 total reads) with 2 x 151 base pair read length. Analysis of sequencing data was performed using the Galaxy platform [Giardine *et al*., 2005; Blankenberg *et al*., 2010b; Goecks *et al*., 2010], with the primary goal of detecting the viral-human junction site locations. Packages used were FASTQ Groomer [Blankenberg *et al*., 2010a], Bowtie2 [Langmead and Salzberg, 2012], Picard MarkDuplicates [Picard Tools], SAMtools BAM-to-SAM and Filter SAM [Li *et al*., 2009].

2.5.3 – RNA-Seq library preparation and sequencing

Isolation of high-quality total RNA from the epithelium of organotypic keratinocyte cultures containing full-length HPV16 E6 variant genomes, European Prototype (EPE6) and

Asian-American (AAE6), was described previously [Jackson *et al*., 2014]. Our keratinocyte model was grown for 14 days to allow simultaneous epithelial differentiation and occurrence of an active viral life cycle. Total RNA for EPE6, AAE6, and HPV16 negative cultures (NIKS), three organotypic raft cultures ($n = 3$) each, were sent for library preparation and sequencing at The Centre for Applied Genomics, Hospital for Sick Children, Toronto, Canada. RNA-Seq libraries were prepared by Illumina TruSeq® RNA Sample Preparation kit followed by sequencing using an Illumina HiSeq® 2500 platform with Illumina v3 chemistry. One lane of multiplexed, paired-end, 2 x 101 base pair sequencing was performed with nine samples: yielding an average of 40.4 million total reads (~20 to 25 million fragments) per sample [**Additional file 1: Table 2.S2**].

2.5.4 – Viral variant read alignment, mapping, and coverage plotting

The human papillomavirus type 16 W12E isolate genome [GenBank: AF125673] [Jeon *et al*., 1995; Flores *et al*., 1999] was used as a viral reference sequence since it was the parental sequence modified by site-directed mutagenesis to generate the EPE6 and AAE6 viral genomes used in this study [Jackson *et al*., 2014]. Only the three non-synonymous nucleotide changes differentiated EPE6 and AAE6 genomes: EPE6 was made by mutating the parental W12E genome at G350T while AAE6 was mutated at G145T and C335T. Prior to alignment and mapping, Bowtie2 [Langmead and Salzberg, 2012] was used to build a reference index for HPV16 using the AF125673 W12E isolate RefSeq. TopHat2 [Trapnell *et al*., 2012] was used for alignment to our viral RefSeq. Variant-specific non-synonymous SNPs were confirmed by variant calling with SAMtools [Li *et al*., 2009]. The Broad Institute's Integrative Genomics Viewer (IGV) [Robinson *et al*., 2011] was used to visualize alignment coverage for each sample. Gene-level counts of the HPV16 W12E ORF's were generated using SAMtools [Li *et al*., 2009], and normalized with library-size correction factors using the Bioconductor project DESeq [Anders and Huber, 2010] in the statistical environment R [R Core Team, 2018]. DESeq was also used for differential viral gene expression analysis. DESeq uses a default false discovery rate (FDR) of 10% for its binomial statistical inference tests to determine differentially expressed genes. Clustered heatmaps of normalized viral gene counts were generated using the gplots package [gplots package for R].

2.5.5 – Identification of viral-human fusion transcripts

ViralFusionSeq (VFS) [Li *et al*., 2013] was used, with default parameters, to identify any viral-human fusion transcripts in each of our sample RNA-Seq datasets. As with viral alignment by TopHat2 (described above), the W12E genome was used as a reference sequence for VFS. Briefly, VFS is a Perl script that searches in high-throughput sequencing data (RNA or DNA-Seq) for viral-human fusion transcripts, which are present as a result of viral integration events into host DNA. This software uses read pair (RP) and clipped sequences (CS) to accurately discover and identify viral-fusion sequences [Li *et al*., 2013]. Additionally, VFS is able to reconstruct fusion transcripts by a targeted *de novo* assembly process. These methods allow us to identify, with single-base resolution, viral-human fusion transcripts present within our epithelial cultures. Viral-human fusion transcripts were compared to known HPV16 integration sites and fusion transcripts with assistance from the database of disease related viral integration sites (Dr. VIS v2.0, [Yang et *al*., 2015]).

We sought to perform protein-level confirmation of highly expressed viral-human fusion transcripts containing exons from human targets SLC26A2 and CSF1R. SLC26A2 protein expression was detected in raft cultures by immunofluorescence, as described previously [Jackson *et al*., 2014]. Based on the viral-human fusion RNA-Seq data, the primary antibody (rabbit polyclonal, 1:500 dilution, Bethyl Laboratories Inc., Cat. No. A304-467A) was chosen to have specificity for translated exon 3 (epitope between amino acid residue 689 and 739). Although also highly up-regulated, no suitable commercial antibody was found for CSF1R exons 20 to 22.

2.5.6 – Human read alignment, mapping, and count generation

Read alignment, mapping, and count generation for the human reference genome (hg19, UCSC nomenclature for GRCh37) was performed by The Centre for Applied Genomics, Hospital for Sick Children, Toronto, Canada. TopHat2 [Trapnell *et al*., 2012] was used for RefSeq while gene- and exon-level counts were generated using HTSeq [Anders *et al*., 2015]. Number of reads and percentage of human RefSeq reads defined as aligned, exon, and exon-exon are reported in **Additional file 1: Table 2.S2** for each sample analyzed.

2.5.7 – Differential expression analysis of human transcriptome

Differential analysis of pair-wise human gene-level counts between NIKS and EPE6, NIKS and AAE6, and EPE6 and AAE6 were performed using the Bioconductor project DESeq [Anders and Huber, 2010] package implemented in the statistical environment R [R Core Team, 2018]. Raw gene counts from HTSeq were first normalized by estimating the sample library sizes [**Additional file 1: Table 2.S3**] and applying the size-factor correction to all counts within a given sample. A dispersion plot was made to visualize the variance estimation step prior to differential expression inference [**Additional file 2: Figure 2.S4**]. A clustered heatmap with hierarchical dendrograms was used to show overall sample and biological replicate clustering: the gene expression profile of AAE6 samples was distinct from EPE6 and NIKS (control) samples [**Additional file 2: Figure 2.S5**]. Although EPE6 replicate 3 and NIKS replicate 1 cluster outside of their specific sample group, viral RNA-Seq analysis has confirmed these sample ID's are correct, and that their grouping is likely a result of the minor host transcriptomic difference between NIKS and EPE6 cultures. DESeq uses a default false discovery rate (FDR) of 10% for its binomial statistical inference tests to determine differentially expressed genes. However, for downstream analyses of down- and up-regulated genes we used a more stringent adjusted *P*-value cut-off of 10^{-5}.

2.5.8 – CIN70 scoring and micronuclei detection

Host chromosomal instability was assessed, using normalized human gene count data from our RNA-Seq experiments, by calculating a CIN70 gene expression signature score [Carter *et al*., 2006] for EPE6 and AAE6 relative to NIKS epithelia. For each of the 70 genes, a normalized human gene count ratio was calculated for all EPE6 and AAE6 samples relative to the average of the NIKS samples. Relative ratio values were then averaged for all 70 genes in each sample and a Welch's *T*-test, for unequal variance, was used to determine whether there was a statistically significant difference in host chromosomal instability signature between EPE6 and AAE6 epithelia. We used a significance level of $P < 0.05$. As a morphological assessment of chromosomal instability we screened haematoxylin and eosin-stained sections from formalin-fixed and paraffin-embedded NIKS, EPE6, and AAE6 epithelia for micronuclei (MN). These aberrant nuclei structures [Zhang *et al*., 2015] were detected using light microscopy with high-magnification (at least 400x).

2.5.9 – Gene set enrichment analysis and networks

Enrichment of host biological processes of differentially expressed human genes was determined using the Gene Ontology (GO) Term Enrichment Service hosted on the AmiGO 2 website [Carbon *et al*., 2009]. Only biological processes were included. Terms were considered significantly enriched if the Bonferroni-corrected *P*-value was less than 0.05. To aid in the visual interpretation of down- and up-regulated gene sets, co-expression networks were constructed with Cytoscape software [Smoot *et al*., 2011]. Pearson correlation coefficients were calculated for each gene-gene pairwise comparison in highly significant down- and up-regulated genes between AAE6 and EPE6 [**Additional file 6** for down- and up-regulated gene-gene pairwise comparisons, respectively]. Pearson correlation coefficient cut-offs used for networking were selected strategically to produce small distinct clusters of genes, since setting the threshold too low results in all nodes connected, and setting the threshold too high results in a lack of clusters.

2.6 – Declarations

2.6.1 – Acknowledgements

Thank you to Dr. Allyson Holmes at the Institut Curie for her valuable feedback and collaboration on the DNA-Seq experiments. Special thanks go to Dr. Melissa Togtema for her insightful comments while preparing the manuscript as well as Darryl Willick for his help in setting up and maintaining the Galaxy platform hosted at the Lakehead University High Performance Computing Centre (LUHPCC).

2.6.2 – Funding

This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) grants to IZ (#355858-2008, #435891-2013, #RGPIN-2015-03855), NSERC Alexander Graham Bell Canada Graduate Scholarship-Doctoral (CGS-D) to RJ (#454402-2014), NSERC Alexander Graham Bell Canada Graduate Scholarship-Masters (CGS-M) to SC (#442618-2013), and an NSERC Undergraduate Student Research Award (USRA) to JB (#483630-2015). Work on the NGS platform of the Institut Curie was supported by the Agence Nationale de la Recherche (ANR Investissement d'Avenir) (ANR-10-EQPX-03) and by the France Génomique National infrastructure (ANR-10-INBS-09). The funding bodies had no role in study design, data collection, data analysis and interpretation, or preparation of the manuscript.

2.6.3 – Availability of data and materials

Raw sequence data used in this article can be accessed via the Sequence Read Archive (SRA), study accession number SRP055094 (http://www.ncbi.nlm.nih.gov/sra/SRP055094) and National Center for Biotechnology Information (NCBI) BioProject, accession number PRJNA275642 (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA275642). Remaining supporting data can be accessed as Additional files, while software and tools used have been cited throughout the Methods section.

2.6.4 – Authors' contributions

This interdisciplinary study was initially conceived by IZ and RJ refined the bioinformatics portion in collaboration with BR, WF, SL, and AN. RJ, PL, and IZ designed and carried out the 3D organotypic skin culturing experiments. IZ and PL contributed reagents, materials, and methods for culturing experiments. RJ, BR, SC and JB performed RNA-Seq and follow-up data analyses. AN contributed reagents, materials, and methods for DNA sequencing. SL and RJ performed DNA-Seq and follow-up analyses. RJ, BR, SL, SC, JB, WF, PL, AN, and IZ contributed to data interpretation. All authors contributed to writing the paper with RJ being the lead author and IZ having considerable input into the writing. All authors have read and approved the final manuscript.

2.6.5 – Competing interests

The authors declare that they have no competing interests.

2.6.6 – Open access

2.7 – Additional Files

2.7.1 – Additional file 1: Viral and human read tables

These three tables are embedded below and available for download (DOCX file, 15 kb): https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-016-3203-3/MediaObjects/12864_2016_3203_MOESM1_ESM.docx

Table 2.S1 – Viral reads summary. Overall, viral reads make up ~0.0001 to 0.01% of the total reads, while human reads make up 80 to 85% of the total reads (the remaining reads are unmapped, to either viral or human sequences).

Sample	Viral Reads	Total Reads	%
EPE6-1	66	46.029×10^6	0.00014
EPE6-2	2014	37.344×10^6	0.00539
EPE6-3	117	36.182×10^6	0.00032
AAE6-1	4505	37.518×10^6	0.01201
AAE6-2	4029	39.224×10^6	0.01027
AAE6-3	4510	42.050×10^6	0.01073

Table 2.S2 – Human RefSeq alignment statistics for all samples. NIKS were HPV16 negative organotypic keratinocyte cultures while EPE6 and AAE6 were cultures containing the full genome of HPV16 with either European Prototype E6 or Asian-American E6 variants, respectively. "Aligned" refers to reads overlapping exons, "Exon" refers to reads completely within an exon, and "Exon-Exon" refers to reads overlapping exon junctions.

Sample	Reads ($x10^6$)	Aligned ($x10^6$)	Exon ($x10^6$)	Exon-Exon ($x10^6$)	Aligned (%)	Exon (%)	Exon-Exon (%)
NIKS-1	41.438	33.133	16.198	16.935	79.958	48.888	51.112
NIKS-2	41.729	33.103	15.930	17.172	79.328	48.124	51.876
NIKS-3	42.202	33.082	15.943	17.140	78.390	48.191	51.809
EPE6-1	46.029	36.888	17.940	18.948	80.143	48.634	51.366
EPE6-2	37.344	30.845	15.137	15.708	82.596	49.074	50.926
EPE6-3	36.182	30.983	15.348	15.635	85.630	49.538	50.462
AAE6-1	37.518	31.803	15.400	16.403	84.769	48.423	51.577
AAE6-2	39.224	33.213	16.117	17.096	84.674	48.527	51.473
AAE6-3	42.050	35.122	17.114	18.008	83.525	48.728	51.272

Table 2.S3 – Human library size factor for all samples. Library size factors derived from DESeq [Anders and Huber, 2010].

Sample	Library Size Factor
NIKS-1	0.9913870
NIKS-2	0.9666649
NIKS-3	0.9699978
EPE6-1	1.0824037
EPE6-2	0.9072066
EPE6-3	0.8962464
AAE6-1	1.0532805
AAE6-2	1.1016102
AAE6-3	1.1151592

2.7.2 – Additional file 2: DESeq plots

These five figures are embedded below and available for download (DOCX file, 204 kb): https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-016-3203-3/MediaObjects/12864_2016_3203_MOESM2_ESM.docx

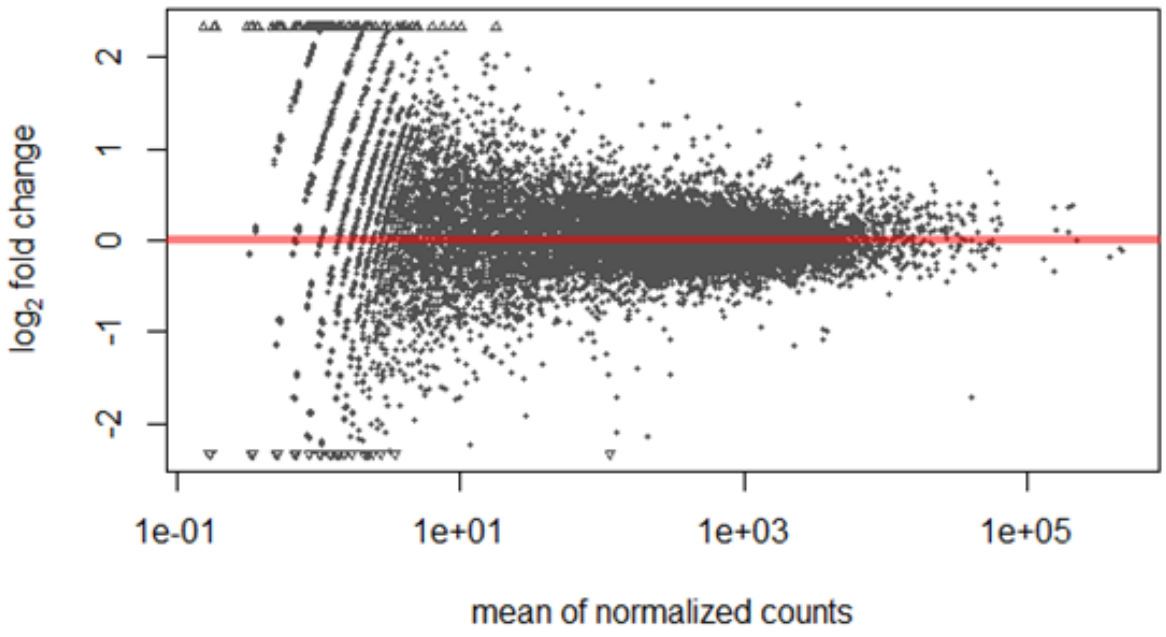


Figure 2.S1 – Plot of normalized mean counts versus $\log_2$ fold change for the contrast NIKS versus EPE6. Red points would represent genes that have significant differential expression between the two conditions (false-discovery rate of 10%, adjusted $P < 0.1$). No genes were significantly differentially expression between NIKS and EPE6.

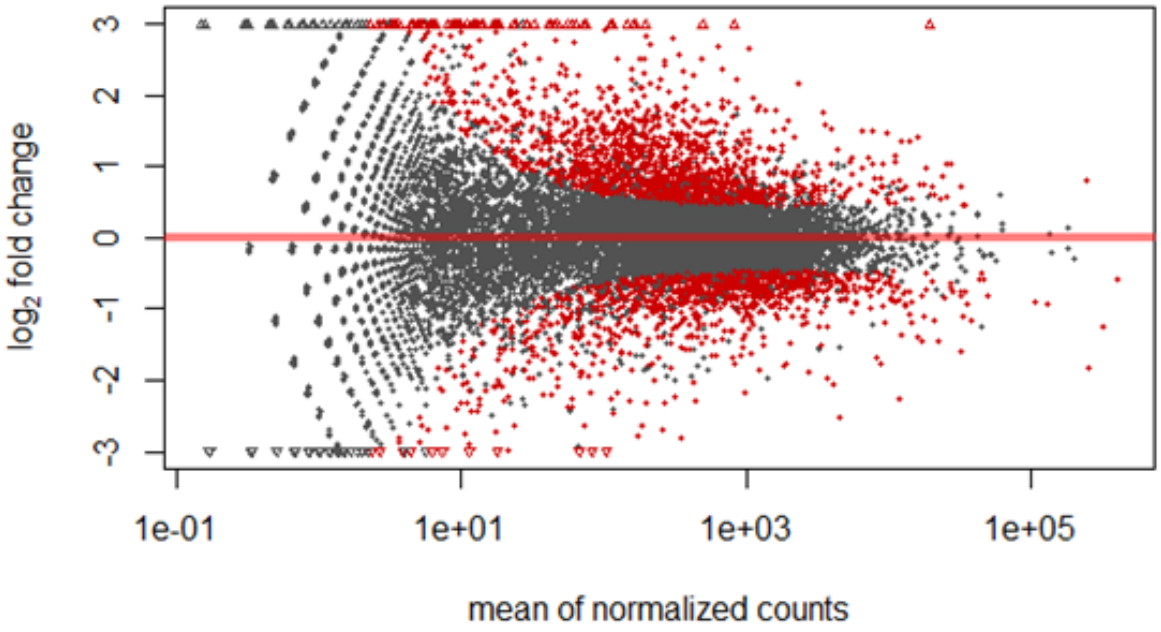


Figure 2.S2 – Plot of normalized mean counts versus $\log_2$ fold change for the contrast NIKS versus AAE6. Red points represent genes that have significant differential expression between the two conditions (false-discovery rate of 10%, adjusted $P < 0.1$). In total, 3006 genes were significantly differentially expression between NIKS and EPE6.

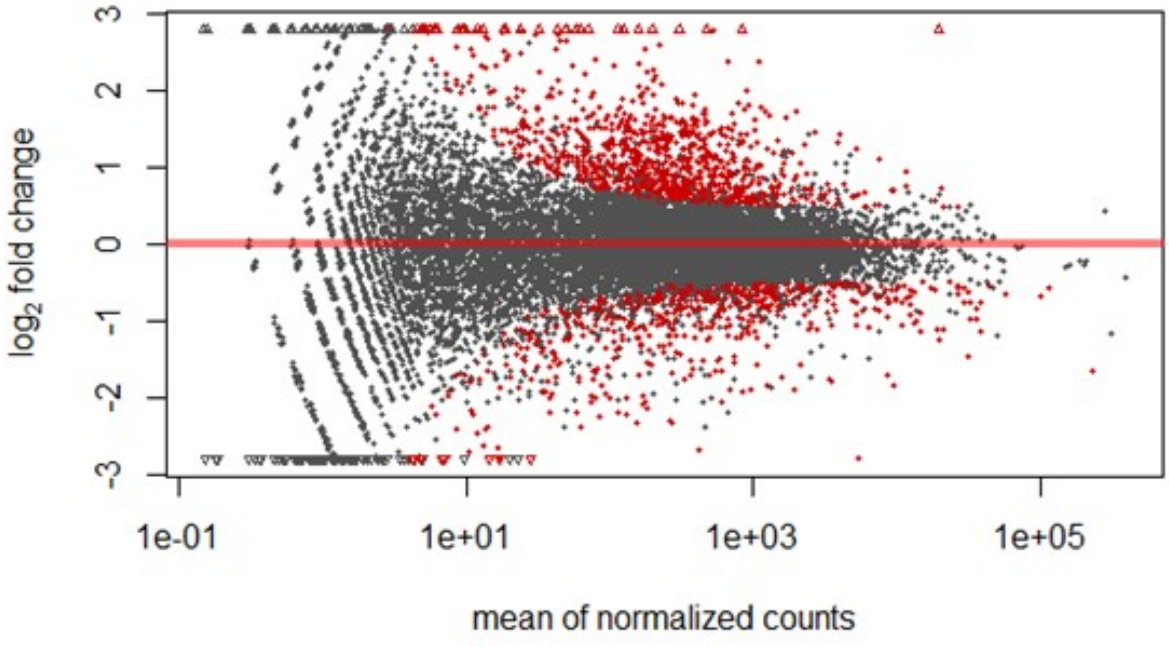


Figure 2.S3 – Plot of normalized mean counts versus $\log_2$ fold change for the contrast EPE6 versus AAE6. Red points represent genes that have significant differential expression between the two conditions (false-discovery rate of 10%, adjusted $P < 0.1$). In total, 1666 genes were significantly differentially expressed between NIKS and EPE6.

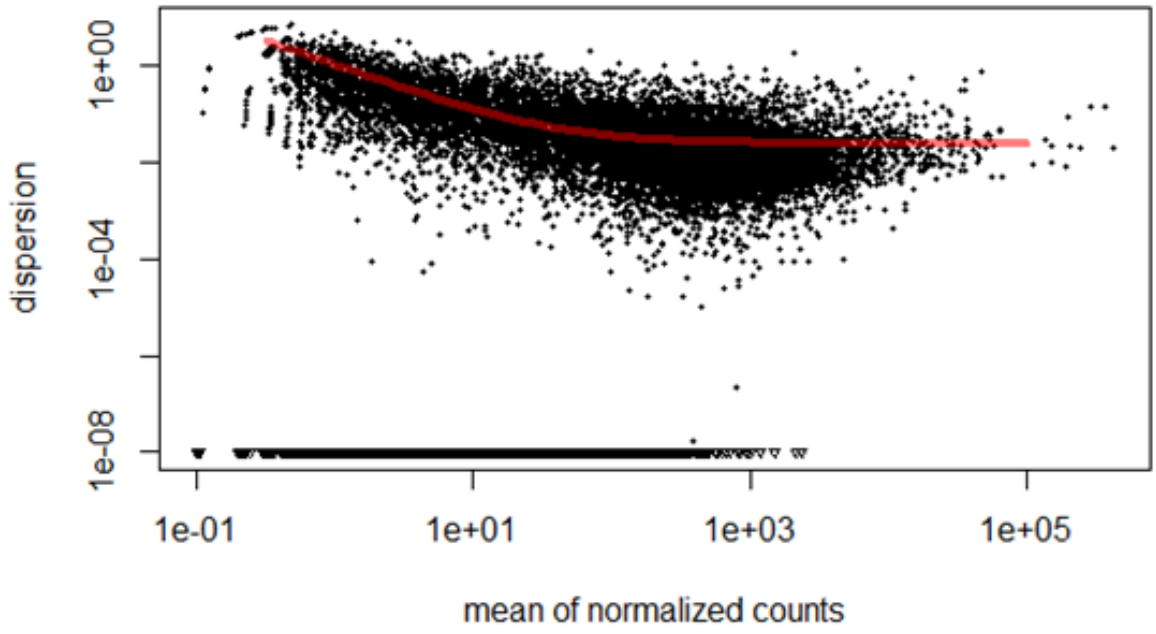


Figure 2.S4 – Empirical and fitted dispersion values plotted against the mean of the normalized human gene-level counts. Red line represents fitted dispersion over the empirical values (black dots).

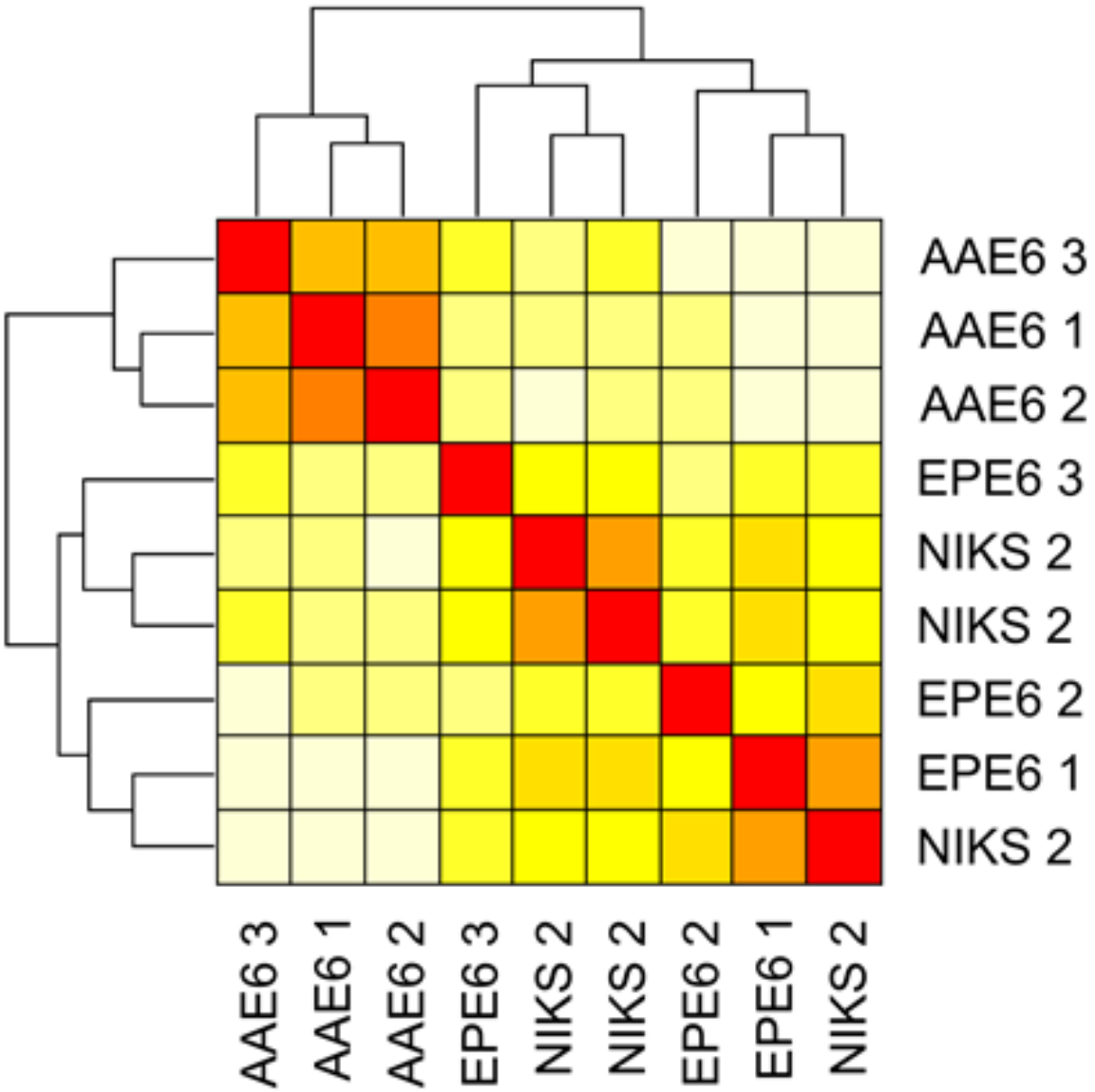


Figure 2.S5 – Heatmap of Euclidean distances between human gene-level counts of samples. Heatmap and clustering was performed after DESeq variance-stabilizing transformation of human gene-level count data.

2.7.3 – Additional file 3: DESeq output

Significant differential expression output for NIKS and AAE6 contrast as well as EPE6 and AAE6 contrast; available for download (XLSX file, 430 kb): https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-016-3203-3/MediaObjects/12864_2016_3203_MOESM3_ESM.xlsx

2.7.4 – Additional file 4: Follow-up discussion of host expression analysis

Additional discussion of differential gene expression analysis, pathway-level enrichment, and co-expression networks; available for download (DOCX file, 30 kb): https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-016-3203-3/MediaObjects/12864_2016_3203_MOESM4_ESM.docx

From the top-ten most significant down-regulated genes in AAE6 compared to NIKS [**Table 2.S4**], CYFIP2 encodes a known p53 target [Jackson II *et al*., 2007] and may be suppressed to prevent apoptosis, or perhaps even be down-regulated as a result of integration as it is located on chromosome 5q33.3. Kelley *et al*., [2005] noted CYFIP2 as significantly altered by siRNA against E6 or E6AP in HeLa and CaSki cells, likely due to p53 inactivation. A potential biomarker for cervical cancer, AJAP1, has been previously found to be silenced by methylation and implicated in beta-catenin signaling [Chen *et al*., 2014]. The transcription factor POU2F3 regulates keratinocyte differentiation and proliferation, encodes a candidate tumour suppressor protein, with silencing by promoter methylation suspected to play a role in cervical cancer [Zhang *et al*., 2006]. Of the top up-regulated genes in AAE6 compared to NIKS [**Table 2.S5**], both SLC26A2 and CSF1R are likely increased as a consequence of viral integration into chromosome 5. Interferon-induced IFITM1 is down-regulated in our AAE6 epithelia, which is consistent with a previous report that assessed gene expression in HPV-infected head and neck cancers [Schlecht *et al*., 2007]. The detection of down-regulated RMI2, encoding for a protein involved in homologous recombination [Deans and West, 2011], could be related to increased genome instability (a possible mechanism for integration). EPE6 to AAE6 had 1,666 significant differentially expressed genes [**Additional file 2: Figure 2.S3**, **Additional file 3** for list of differentially expressed genes between EPE6 and AAE6]. Of these genes, 666 were down-regulated while 1,000 were up-regulated in AAE6 compared to EPE6. The top-ten down-regulated genes in AAE6 compared to EPE6 are presented in **Table 2.S6** and include 6 of the same genes as found in the AAE6 to NIKS

comparison. The top-ten significant up-regulated genes in AAE6 compared to EPE6 are presented in **Table 2.S7** and include 8 of the same genes as found in the AAE6 to NIKS comparison.

To increase the stringency of differentially expressed genes we separated down-regulated and up-regulated genes from each pairwise comparison below the adjusted *P*-value threshold of 10^{-5}. For genes differentially expressed in AAE6 compared to NIKS, the stringent cut-off yielded 61 down-regulated genes (61/1,312 = 4.65% highly significant down-regulated genes) and 248 up-regulated genes (248/1,694 = 14.64% highly significant up-regulated genes). For genes differentially expressed in AAE6 compared to EPE6, the stringent cut-off yielded 26 down-regulated genes (26/666 = 3.90% highly significant down-regulated genes) and 152 up-regulated genes (152/1,000 = 15.20% highly significant up-regulated genes). With only highly significant differentially expressed genes identified, we then proceeded to pathway-level analysis to determine which host biological processes were enriched given down-regulated and up-regulated sets of genes.

Enrichment of host biological processes within the highly significant sub-sets of differentially expressed human genes was determined using the Gene Ontology (GO) Term Enrichment Service hosted on the AmiGO 2 website [Carbon *et al*., 2009]. Terms were considered significantly enriched if the Bonferroni-corrected *P*-value was less than 0.05. In total, 50 GO terms were significantly enriched among highly significant down-regulated genes in AAE6 compared to NIKS [**Figure 2.5a** for top-ten, **Additional file 5** for full list]. Enrichment of these biological processes among highly significant down-regulated genes reflects the poor differentiation and tumourigenic tissue phenotype caused by AAE6 [Jackson *et al*., 2014], now evidenced by transcriptome-level data. Additionally, 176 GO terms were significantly enriched among highly-significant up-regulated genes [**Figure 2.5a** for top-ten, **Additional file 5** for full list], reflecting the proliferative phenotype caused by AAE6 [Jackson *et al*., 2014] and providing further evidence for variant-specific transcriptome-wide changes.

When comparing AAE6 to EPE6, only 4 GO terms were significantly enriched among highly significant down-regulated genes [**Figure 2.6a, Additional file 5** for full list]. Enrichment of these lipid metabolism biological processes among highly significant down-regulated genes is a finding that sheds new light on HPV-centered host-pathogen interactions and HPV-driven tumourigenesis. Notable down-regulated genes were ALDH1A1/A2. These aldehyde dehydrogenases catalyze the synthesis of retinoic acid which interestingly suppresses viral

oncogene expression [Faluhelyi *et al*., 2004]. A decrease of their expression could perhaps be permitting, at least in part, to the continued over-expression of oncogenes by AAE6. For highly significant up-regulated genes in AAE6 compared to EPE6, 231 GO terms were significantly enriched [**Figure 2.6a** for top-ten, **Additional file 5** for full list]. Enrichment of cell cycle and proliferation biological processes was confirmed in AAE6 over EPE6, further demonstrating AAE6's enhanced tumourigenic potential over EPE6 [Jackson *et al*., 2014], thought to be due in part to an enhanced Warburg effect [Cuninghame *et al*, 2017]. Pathway-level analysis revealed that many of the changes due to AAE6 were related to increased cell cycle and DNA synthesis, which are commonly promoted pathways in tumourigenesis [Hanahan and Weinberg, 2000; 2011], and more specifically, viral tumourigenesis [Mesri *et al*., 2014]. In the synthesis of DNA pathway, for example, nearly every gene is up-regulated, with the exception of lower CDKN1A (coding for p21). This potent cyclin-dependent kinase inhibitor is functionally regulated by p53, with its down-regulation associated with invasive cervical cancer [Bahnassy *et al*., 2006].

In addition to pathway-level analysis, Cytoscape was used for visualization of co-expressed genes from the highly significant down- and up-regulated genes in AAE6 compared to EPE6. Visualizations and interpretations with networks, including co-expression networks, can play an important role in analyzing the large amount of data generated from high-throughput experiments [Merico *et al*., 2009] since multiple levels of information can be presented simultaneously (such as the degree of gene co-expression, fold change, and functional annotations). The co-expression network for down-regulated genes reveals four distinct clusters of genes [**Figure 2.7a**]. The strongest correlation in the first cluster was between KIAA0040 (an uncharacterized protein) and LIPG (lipase, endothelial), while WDR63 (WD repeat domain 63) was the most down-regulated due to AAE6. In the second cluster, the strongest correlation was between CYP4F22 (cytochrome P450, family 4, subfamily F, polypeptide 22) and KRT6B (keratin 6B). CYP4F22, a cytochrome P450, is likely involved in keratinocyte differentiation [Sasaki *et al*., 2012], so it is not surprising that it is strongly correlated to a cytokeratin. Mutations of this cytochrome gene lead to the skin disorder ichthyosis (scaly skin) [Fischer, 2009], which further supports it has a role in keratinocyte differentiation. The most down-regulated gene in this second cluster was AJAP1, discussed above as a methylated gene in cervical cancer. In the third down-regulated cluster the strongest correlation was between INPP5D (inositol polyphosphate-5-phosphatase) and LOC375010 (ankyrin repeat domain 20 family, member A pseudogene). INPP5D is a negative regulator of the

PI3K (phosphoinositide 3-kinase) pathway [Huang *et al*., 2012], a common pathway involved in proliferation and tumourigenesis, so INPP5D's down-regulation coincides with the tumourigenic potential of AAE6. The most down-regulated gene in this cluster was DAPK1 (death-associated protein kinase 1), which is another gene that is known to have its promoter methylated in cervical cancer [Banzai *et al*., 2014]. The fourth cluster of down-regulated genes contained only two genes: ANTXR2 (anthrax toxin receptor 2) and CSMD3 (CUB and sushi multiple domains). ANTXR2 binds to extracellular matrix (ECM) proteins collagen IV and laminin, which suggests it may have a role in ECM adhesion in mouse studies [Reeves *et al*., 2013]. Interestingly, CSMD3 has been associated with the fragile site FRA8C at chromosome 8q24 in cervical carcinoma with HPV integration [Ferber *et al*., 2004].

Looking at the highly significant and co-expressed up-regulated genes, five distinct clusters were observed [**Figure 2.7a**]. The first cluster was significantly enriched for "DNA replication" (GO:0006260) due to the high number of up-regulated and strongly co-expressed gene such as the MCM's. The most up-regulated gene was SLC26A2, which we previously discussed as up-regulated as a result of viral integration. The strongest correlations were between MCM7 and MCM3, as well as DSN1 and CHEK1. The second cluster was significantly enriched for "macrophage colony-stimulating factor signaling pathway" (GO:0038145) due to the presence of CSF1R, likely upregulated on chromosome 5 due to viral integration (as discussed above). This gene was strongly co-expressed with TCAM1P (testicular cell adhesion molecule 1, pseudogene). The remaining three clusters were significantly enriched for cellular division processes. Enrichment and up-regulation of these biological processes is consistent with the proliferative and tumourigenic phenotype in AAE6.

Table 2.S4 – Top-ten most significant down-regulated genes in AAE6 compared to NIKS.

Gene Symbol	Gene Name	Fold Change	Adjusted *P*-value
CYFIP2	cytoplasmic FMR1 interacting protein 2	0.15	6.92×10^{-21}
AJAP1	adherens junctions associated protein 1	0.14	7.37×10^{-19}
CSMD3	CUB and Sushi multiple domains 3	0.21	4.19×10^{-15}
ANTXR2	anthrax toxin receptor 2	0.31	1.74×10^{-12}
EFNB3	ephrin-B3	0.22	2.46×10^{-12}
SDK2	sidekick cell adhesion molecule 2	0.16	5.97×10^{-12}
SMTN	smoothelin	0.35	2.92×10^{-11}
EDA2R	ectodysplasin A2 receptor	0.19	2.94×10^{-11}
POU2F3	POU class 2 homeobox 3	0.28	1.84×10^{-10}
FLRT2	fibronectin leucine rich transmembrane protein 2	0.38	3.74×10^{-10}

Table 2.S5 – Top-ten most significant up-regulated genes in AAE6 compared to NIKS.

Gene Symbol	Gene Name	Fold Change	Adjusted *P*-value
SLC26A2	solute carrier family 26 (anion exchanger), member 2	114.19	2.14×10^{-173}
CSF1R	colony stimulating factor 1 receptor	407.82	4.70×10^{-112}
GPAT2	glycerol-3-phosphate acyltransferase 2, mitochondrial	387.18	4.76×10^{-80}
IFITM1	interferon induced transmembrane protein 1	9.98	7.56×10^{-45}
MCM2	minichromosome maintenance complex component 2	6.25	4.95×10^{-34}
MEST	mesoderm specific transcript	14.63	5.97×10^{-33}
CDCA7	cell division cycle associated 7	6.40	1.21×10^{-32}
RMI2	RecQ mediated genome instability 2	8.80	2.43×10^{-29}
MCM6	minichromosome maintenance complex component 6	4.68	3.23×10^{-24}
KLHL35	kelch-like family member 35	33.26	1.39×10^{-23}

Table 2.S6 – Top-ten most significant down-regulated genes in AAE6 compared to EPE6.

Gene Symbol	Gene Name	Fold Change	Adjusted *P*-value
EFNB3	ephrin-B3	0.21	3.57×10^{-13}
AJAP1	adherens junctions associated protein 1	0.20	7.32×10^{-13}
CYFIP2	cytoplasmic FMR1 interacting protein 2	0.24	6.01×10^{-12}
PAQR5	progestin and adipoQ receptor family member V	0.32	6.59×10^{-11}
CSMD3	CUB and Sushi multiple domains 3	0.26	7.60×10^{-11}
DAPK1	death-associated protein kinase 1	0.24	1.02×10^{-10}
ANTXR2	anthrax toxin receptor 2	0.36	9.30×10^{-10}
SMTN	smoothelin	0.40	5.63×10^{-09}
LOC100216001	long intergenic non-protein coding RNA 704	0.22	9.07×10^{-08}
MFAP5	microfibrillar associated protein 5	0.35	1.40×10^{-07}

Table 2.S7 – Top-ten most significant up-regulated genes in AAE6 compared to EPE6.

Gene Symbol	Gene Name	Fold Change	Adjusted *P*-value
SLC26A2	solute carrier family 26 (anion exchanger), member 2	118.42	1.98×10^{-173}
CSF1R	colony stimulating factor 1 receptor	551.09	1.28×10^{-113}
GPAT2	glycerol-3-phosphate acyltransferase 2, mitochondrial	290.62	6.59×10^{-78}
IFITM1	interferon induced transmembrane protein 1	6.95	7.74×10^{-34}
MCM2	minichromosome maintenance complex component 2	5.20	4.83×10^{-28}
CDCA7	cell division cycle associated 7	5.25	2.40×10^{-26}
RMI2	RecQ mediated genome instability 2	5.84	1.18×10^{-20}
MEST	mesoderm specific transcript	7.06	1.19×10^{-20}
LOC254559	long intergenic non-protein coding RNA 925	12.34	2.50×10^{-20}
C1R	complement component 1, r subcomponent	7.27	1.76×10^{-19}

2.7.5 – Additional file 5: GO output

Significantly enriched GO terms (biological processes) for NIKS and AAE6 contrast as well as EPE6 and AAE6 contrast; available for download (XLSX file, 28 kb): https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-016-3203-3/MediaObjects/12864_2016_3203_MOESM5_ESM.xlsx

2.7.6 – Additional file 6: Pearson correlations

Pearson correlation coefficients for gene-gene pairwise comparisons of down- and up-regulated genes for EPE6 and AAE6 contrast; available for download (XLSX file, 298 kb): https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-016-3203-3/MediaObjects/12864_2016_3203_MOESM6_ESM.xlsx

CHAPTER 3A – EPITHELIAL ORGANOID MODEL

This chapter was accepted for publication as a commentary article in *Philosophical Transactions of the Royal Society B: Biological Sciences* on 27 Nov 2018 in the special theme issue *Silent cancer agents: multi-disciplinary modelling of human DNA oncoviruses* ("Tissue models" section; DOI: 10.1098/rstb.2018.0288) [Jackson *et al*., 2019]. It has been adapted with permission for re-use within this dissertation (non-commercial purpose), as per the Royal Society's 2018 Licence to Publish document.

An epithelial organoid model with Langerhans cells for assessing virus-host interactions

Robert Jackson[1,2], Statton Eade[1], Ingeborg Zehbe[1,3]

[1]Probe Development and Biomarker Exploration, Thunder Bay Regional Health Research Institute, Thunder Bay, Ontario, Canada
[2]Biotechnology Program, Lakehead University, Thunder Bay, Ontario, Canada
[3]Department of Biology, Lakehead University, Thunder Bay, Ontario, Canada

Keywords: Human papillomavirus (HPV), Host immune surveillance, Keratinocytes and Langerhans cells (LCs), Organoids, Pathogen-host interaction, Next-generation sequencing

3A.1 – Abstract

Persistent infection with oncogenic human papillomavirus (HPV) may lead to cancer in mucosal and skin tissue. Consequently, HPV must have developed strategies to escape host immune surveillance. Nevertheless, most HPV infections are cleared by the infected host. Our laboratory investigates Langerhans cells (LCs), acting at the interface between innate and adaptive immunity. We hypothesize that this first line of defence is vital for potential HPV elimination. As an alternative to animal models, we use smaller-scale epithelial organoids grown from human primary keratinocytes derived from various anatomical sites. This approach is amenable to large sample sizes—an essential aspect for scientific rigour and statistical power. To evaluate LCs phenotypically and molecularly during the viral life cycle and onset of carcinogenesis, we have included an engineered myeloid cell line with the ability to acquire an LC phenotype. This model

is accurately tailored for the crucial time-window of early virus elimination in a complex organism and will shed more light on our long-standing research question of how naturally-occurring HPV variants influence disease development. It may also be applied to other microorganism-host interaction research or enquiries of epithelium immunobiology. Finally, our continuously-updated pathogen-host analysis tool enables state-of-the-art bioinformatics analyses of next-generation sequencing data.

3A.2 – Introduction

Since the late 1970's, there has been an evolution of growing human keratinocytes in a three-dimensional (3D) rather than 2D manner for "the reconstitution of living skin" [Bell *et al*., 1983] in a cell culture dish with nomenclatures such as "organotypic culture" [Merrick *et al*., 1992], "keratinocyte raft cultures" [Southern *et al*., 2001], "organotypic raft cultures" [Anacker and Moody, 2012], or most recently skin "organoids" [Fatehullah *et al*., 2016] [**Figure 3A.1**]. Initially, the main driver for *in vitro* skin cultivation was grafting. To grow keratinocytes as stratified epithelium was first reported by Rheinwald and Green [Rheinwald and Green, 1975], followed by Bell's "full thickness skin equivalent" [Bell *et al*., 1979; 1981; 1983]. In the early 1990's primary keratinocytes infected with high-risk (HR) human papillomavirus (HPV) were found to have premalignant characteristics in organotypic culture [Blanton *et al*., 1991] and various models capable of reproducing the HPV infectious cycle were established by several independent groups [Dollard *et al*., 1992; Meyers *et al*., 1992; Flores *et al*., 1999]. Globally, HPV—a DNA virus, is the most common sexually transmitted infectious agent with an ~10% prevalence in healthy women [Crow, 2012] and more than 300 types identified to date (the majority of the 500+ types of human and non-human PVs identified [Van Doorslaer *et al*., 2017a], based on the Papillomavirus Episteme, PaVE: https://pave.niaid.nih.gov [Van Doorslaer *et al*., 2013; 2017b]). A subset of 12 HR types, with HPV16 being the most common, cause cancer (*e.g.* oropharyngeal and gynaecological) in humans [Bouvard *et al*., 2009; Crow, 2012]. The malignant potential of HR HPVs is largely due to the E6 and E7 oncogenes encoding their corresponding proteins which interfere with cellular integrity [Hoppe-Seyler *et al*., 2018 and references therein]. HPV16 variants implicated in a higher risk for cervical cancer [Zehbe *et al*., 1998b] showed a higher degree of dysplasia in 3D raft cultures compared to lower risk variants [Richard *et al*., 2010; Jackson *et al*., 2014; 2016]. The next generation of rafts included immune components to study the role (or lack

thereof) of innate and adaptive immunity in combatting HPV. For instance, models including lymphocyte infiltration [Jacobs *et al*., 1998] and the microenvironment of Langerhans cells (LCs) [Hubert *et al*., 1999]—epithelium-specific dendritic cells (DCs), were developed. More recently, a bone marrow-derived cell line called MUTZ-3 [Masterson *et al*., 2002] has been used to produce DCs and LCs [Kosten *et al*., 2015b]. Such a cell line allows far better reproducibility in a research context where many biological replicates are necessary for robust research results with high statistical power.

HR HPV such as type 16 is the main risk factor for cervical cancer provided it can persist in the host [zur Hausen, 2002]. Variant designations are based on their geographical region of origin [Mirabello *et al*., 2018 and references therein]. The European Prototype (EP) was the first HPV16 genome published more than 30 years ago [Seedorf *et al*., 1985] and at present, four (formerly five) lineages are known: A (European sub-lineages A1-A3 and Asian sub-lineage A4), B (African-1 sub-lineages B1-B4), C (African-2 sub-lineages C1-C4), and D (North American sub-lineages D1 and D4, and Asian-American sub-lineages D2 and D3) [de Araujo Souza *et al*., 2009; Burk *et al*., 2013; Mirabello *et al*., 2018]. Recently, we have put our efforts on the common EP and AA variants which differ in only 3 amino acid changes at residues 14 (Q>H), 78 (H>Y) and 83 (L>V) in the major transforming protein E6 [Richard *et al*., 2010; Niccoli et al., 2012; Jackson *et al*., 2014; 2016; Cuninghame *et al*., 2017]. Epidemiological studies revealed that the AAE6 variant is a higher risk factor for dysplasia as well as an earlier onset of invasive tumours than EPE6 [Xi *et al*., 1997; Berumen *et al*., 2001]. AAE6 has a greater transforming, migratory, and invasive potential than EPE6 when retrovirally transduced into primary human keratinocytes during recent long-term *in vitro* immortalization studies [Richard *et al*., 2010; Niccoli et al., 2012]. Further, AAE6 is more prone to integrate into the host cell genome [Jackson *et al*., 2016] and demonstrated an altered metabolic phenotype reminiscent of the Warburg effect [Cuninghame *et al*., 2017]. These results suggest that coding changes in E6 have strong mechanistic and functional consequences for infection and thus contribute to marked differences in cancer risk.

To decipher the fundamental biology of HPVs and their tumourigenic features in a model system, the organotypic 3D infection model (raft culture, or organoid, the latter will be used henceforth) has the advantage of allowing reproducible and simultaneous epithelial differentiation and therefore the occurrence of an active viral life cycle. Our approach is a joint venture between the biological, clinical, and computer sciences, with like implications for clinical and basic

research. In addition to how E6 variants are implicated in the development of HPV-related diseases, this model has been developed to adapt to new enquiries regarding how the early cell-based innate immune system fights a common virus such as HPV. Here, we discuss an organoid epithelial model drawing on previous research and experience by us [Richard *et al*., 2010; Jackson *et al*., 2014; 2016] and others [Rheinwald and Green, 1975; Bell *et al*., 1979; 1981; 1983; Blanton *et al*., 1991; Dollard *et al*., 1992; Merrick *et al*., 1992; Meyers *et al*., 1992; Jacobs *et al*., 1998; Flores *et al*., 1999; Hubert *et al*., 1999; Southern *et al*., 2001; Masterson *et al*., 2002; Anacker and Moody, 2012; Kosten *et al*., 2015b; Fatehullah *et al*., 2016]. Its strength lies in the unique combination of components essential for our application: primary, HPV-permissive host cells from various anatomical sites, the recapitulation of the HPV infectious cycle using full-length HPV16 genomes, naturally existing HPV16 variants throughout its genome, controllable copy numbers of HPV-positive keratinocytes, tissue-residing LCs, phenotypical and immunological characterization as well as refined bioinformatics tools for next generation sequencing (NGS).

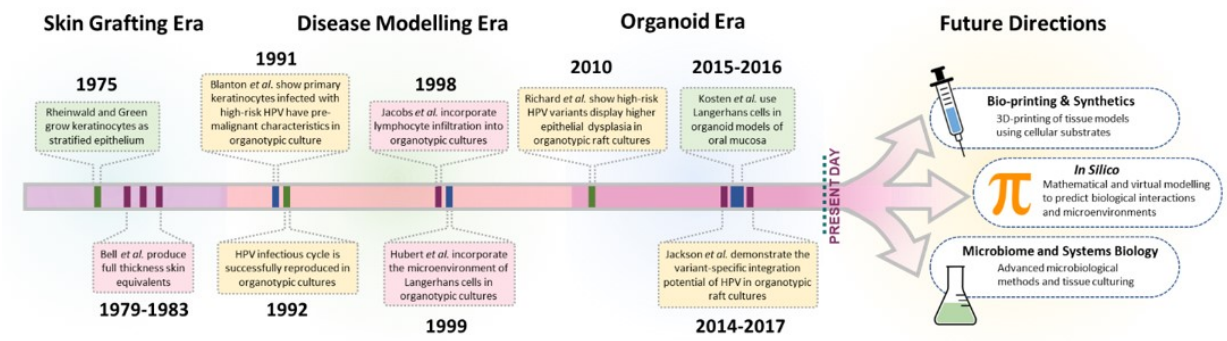


Figure 3A.1 – Historical overview of epithelial models. Significant developments in the evolution of three-dimensional epithelial models (focusing on the incorporation of Langerhans cells and HPV16 variant research) in the last four decades are depicted along with potential future directions. Due to space constraints we are unfortunately unable to showcase all the excellent model studies that exist.

3A.3 – The "Silent Killer": How HPV Evades Host Immune Recognition

HPV, the "silent killer", is a master in evading the host immune response, causing illness and in some cases death of the infected host. The Indigenous, supernatural monster Windigo comes to mind serving as a metaphor when our group explained the danger of HPV to First Nations women in Northwest Ontario, Canada [Zehbe *et al*., 2016b]. Here, we outline an eclectic, experimental approach appropriate to study HPV infection. We will employ healthy mucosal or skin tissue including epidermis and dermis either from the uterine cervix (non-keratinizing, cervical keratinocytes), the oropharyngeal area (non-keratinizing, gingival keratinocytes) or the skin (keratinizing epithelial keratinocytes) with matching fibroblasts. Fortunately, these cells can now be bought, avoiding the barrier of lengthy procurement processes [Villa *et al*., 2018]. Epidermis on its own or combined with dermis models are also commercially available. However, our context precludes the use of such models since donor background cannot be controlled and to "infect" with HPV, we need to grow 3D cultures to allow the viral life cycle to take place for the production of infection-competent virions. Our lab has the necessary experience with organoids, as we can rely on a decade or so of in-house experience in 3D culturing [Zehbe *et al*., 2009; Richard *et al*., 2010; Niccoli *et al*., 2012; Jackson *et al*., 2014; 2016; Villa *et al*., 2018].

Under normal physiological conditions, infiltrating T-cells from the underlying blood vessels migrating into the dermis and even the epidermis has not been observed. However, immature and mature LCs belong to the normal mucosa and skin landscape and hence only their integration in our current model has been considered. Rodrigues Neves & Gibbs [2018] pointed out that a 3D keratinocyte model with both immune components is still lacking in the scientific literature. This is clearly a draw-back when researching various allergens where inflammation is the biggest obstacle to tackle. However, in the context of persistent HPV in immune-comprised organs such as the cervix, the sheer lack of these components may provide major viral immune evasion strategies to escape host immune surveillance. Indeed, HPV is thought to keep a low inflammatory profile—meaning that immune cells are not attracted to the site. Without causing epithelial inflammation, HPV has developed many strategies to go undetected, both during innate and adaptive phases [Grabowska and Riemer, 2012].

3A.3.1 – The HPV clearance hypothesis: the role of LCs

LCs are a distinct DC subset comprising about 2–3% of epidermal cells [Kissenpfennig and Malissen, 2006] that are at the interface between innate and adaptive immunity. They reside in the supra-basal part of the epithelium and their adhesion is mediated by E-cadherin [Tang *et al*., 1993]. Keratinocytes are important for LC activation and maturation as they modulate immunity through their production of chemokines and pro-inflammatory cytokines. We hypothesize that in hosts clearing their HPV infection (the majority) [Stanley, 2012], keratinocytes stimulate LC activation, *e.g.* via TNF-α and TGF-β secretion which promotes LC migration from the epidermis to dermis and lymph nodes. Melief [2005] depicts such a scenario where "danger signals" and "activated dendritic cells" lead to the activation of cytotoxic T-cells—the first step towards tumour elimination. While we generally agree with this notion, further details were added to illustrate our model and adapt it to the viral infectious cycle and early carcinogenesis [**Figure 3A.2**].

LCs are patrolling antigen-presenting cells (APCs) acting as immunological sentinels in mucosal and skin tissues. They likely engulf and phagocytose viral particles and/or dead HPV+ keratinocytes before a lesion develops and may be the first encounter of an infected host to fight an HPV infection. In turn, keratinocytes secrete pro-inflammatory cytokines and dermal fibroblasts secrete chemokines/chemo-attractants, helping LCs to mature and migrate to lymph nodes for cross-presenting HPV peptides to CD4+ and CD8+ T-cells. LCs, however, seem to be less frequent in the transformation zone [Deligeoroglou *et al*., 2013] where most cervical cancers develop, and E6 negatively interferes not only with E-cadherin but also with the expression of LC chemo-attractants [Iijima *et al*., 2013 and references therein]. It is this phase of the immune system that we think is one of the key interfaces between pathogen and host, and which the virus must escape to persist in its host. Interestingly, the oncoprotein E6 of HPV16 down-regulates E-cadherin [Matthews *et al*., 2003; Togtema *et al*., 2015; references therein], and an E6 variant with only one amino acid change of AA at residue 83, showed this to be increased compared to EP [Togtema *et al*., 2015]. This, and the potential role of LCs in HPV clearance, prompted us to investigate LC immune functions in the AAE6 variant context.

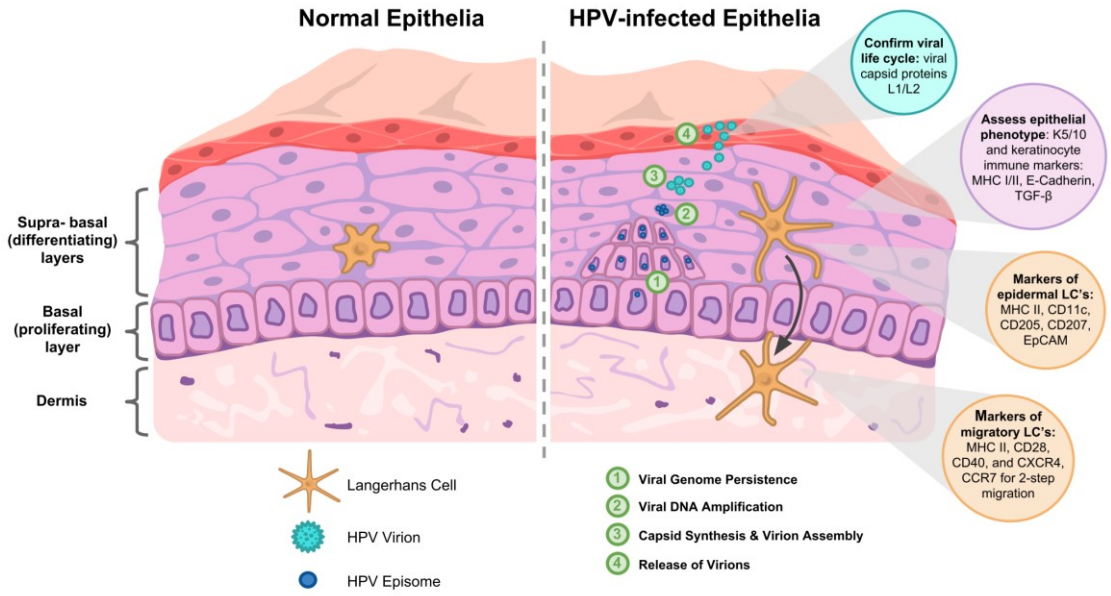


Figure 3A.2 – Epithelial microenvironment and immune landscape. HPVs infect epithelial tissues lining the upper digestive or genital tract most likely by infecting the stem cells of the proliferating basal layer. If the viral genome persists in the basal layer, then it can be amplified in the differentiating layers. Thereafter new virions are released without causing an inflammation. Langerhans cells (LCs) reside in the epithelia as immunosurveillance cells. Many parameters (markers) to investigate LC patterns of activation, differentiation and migration from epidermis to dermis (and potentially to draining lymph nodes) are known. They will be closely investigated in the depicted model in the context of the two HPV variants under study. Viral life cycle will be confirmed, *e.g.* with viral capsid proteins L1/L2. The epithelial phenotype will be assessed via appropriate keratin (K) markers in the various anatomical sites, *e.g.* K5/10 in skin, and keratinocyte immune markers for an environment commensal for LC epidermis to dermis migration, *e.g.* MHC I/II, E-cadherin, TNF-α, and TGF-β. We will further test markers of epidermal LCs, *e.g.* MHC II, CD11c, CD207 (langerin), EpCAM and markers of LC migration ability, *e.g.* MHC II, CD28, CD40, and CXCR4 and CCR7 for a 2-step migration. The above markers will all be tested *in situ* by immunofluorescence. Monitoring early HPV infection by the host needs danger signals (cytokines and chemokines) to attract LCs traveling to sentinel lymph nodes. Such LC attractants, *e.g.* CCL2/5/20, and CXCL12 (fibroblast-derived) and CCL27/28, IL-18 and type I IFNs (KC-derived) will be characterized via supernatant. In collaboration with the Alizon group in Montpellier, we use mathematical models to discover novelties/unknowns, test new hypotheses and research questions *in silico* [Murall *et al.*, 2019].

3A.3.2 – Implication of HPV16 variants in modulating the host immune system

In the past decennium and outlined in the introduction, our lab has provided ample evidence that E6 variants within the HPV16 genome strongly promote functional changes in the mammalian host. More recently, we discovered that the highly oncogenic AAE6 variant may also be involved in immune escape like the L83V variant [see *3A.3.1 – The HPV clearance hypothesis: the role of LCs*]. In our given context, other than promoting the differentiation of LCs, and in contrast to keratinocytes, TGF-β stimulates stromal fibroblasts to proliferate and to synthesize matrix proteins [Buschke *et al*., 2011]. An interesting finding from the Jackson *et al*. [2016] RNA-seq data is worth exploring further: in addition to one of the significantly down-regulated clusters of genes being involved in the "negative regulation by host of viral transcription", we also found that the TGF-β innate immune pathway signature seems to be down-regulated in the AAE6 but not in the EPE6 organoids. This is intriguing since TGF-β is the key cytokine for LC maturation. TGF-β was expressed ~1/3 lower in AAE6 vs HPV-negative and EPE6-containing organoids. Although this is not significant when considering the differential expression results which use a conservative filtering technique (given the high-throughput nature), the similarly listed "TGF-β induced" (TGFBI) is significantly lower. TGFBI, an extra-cellular matrix protein, promotes inflammation, integrin-mediated monocyte adhesion, migration and chemotaxis [Kim and Kim, 2008]. TGFBI down-regulation by AAE6 may be a means to further tone down inflammation in the HPV environment. Moreover, innate immune receptor signalling [Li *et al*., 2016] may also be affected by the AAE6 variant. In a recent protein-protein interaction screen, we found that AAE6 but not EPE6 binds to various E3 ligases (other than E6AP) indicating that innate immune-regulated transcription factors (*e.g.* NF-kβ and IRF7) as well as anti-microbial peptides (*e.g.* defensins and cathelicidins) and pro-inflammatory cytokines (*e.g.* IL-1β) may be down-regulated [Mehran Masoom, unpublished observations]. Lastly, our finding that AAE6 interferes with host cell metabolism by shifting to a Warburg effect [Cuninghame *et al*., 2017] has also been looked at from the perspective of tumour-associated macrophages (TAMs) in the stroma [Colegio *et al*., 2014]. Interestingly, it was reported that tumour-derived lactic acid caused the M2-like polarization in TAMs. Altogether, we conclude that in an immunological context, the AAE6 variant may indirectly modify LC biology as well as the tumour microenvironment in the stroma.

3A.4 – An Eclectic Methodological Approach for an Epithelial Organoid

We will perform systematic organotypic epithelial culturing [Jackson *et al*., 2014; 2016] in a step-wise manner through three primary experimental phases: Preparation, Cultivation, and Characterization. The first phase provides a foundation focused on *a priori* experimental design for reproducibly answering biological questions in a "life-like" model of human epithelium. The second phase continues with the process of growing [**Figure 3A.3**] and harvesting these lab-grown tissues. Finally, the third phase concludes with thoroughly characterizing the tissues via biological analysis and interpretation. While additional components and complexities (*e.g.* beyond LCs) could be included in an organoid model to make it ever more "life-like", we strive to present a simplistic yet useful approach with a focus on studying HPV biology (and specifically, the molecular underpinnings of increased tumourigenic risk due to viral variants, such as via LC interactions). The overall objective is to establish an immune-competent 3D *in vitro* organoid to study two commonly found HPV16 E6 variants undergoing their active HPV life cycle. We will determine an initial host-pathogen interaction, *i.e*. the suppression by HPV at the innate immunity level—for which the life cycle needs to be activated. Our previous model, even without LCs, fulfilled closely the needs of *in vivo* but prevented us from controlling the number of HPV+ cells after transfection [Jackson *et al*., 2014; 2016]. The new model, on the other hand, can be adapted to such needs as we are able to use various ratios of HPV+ and – cells.

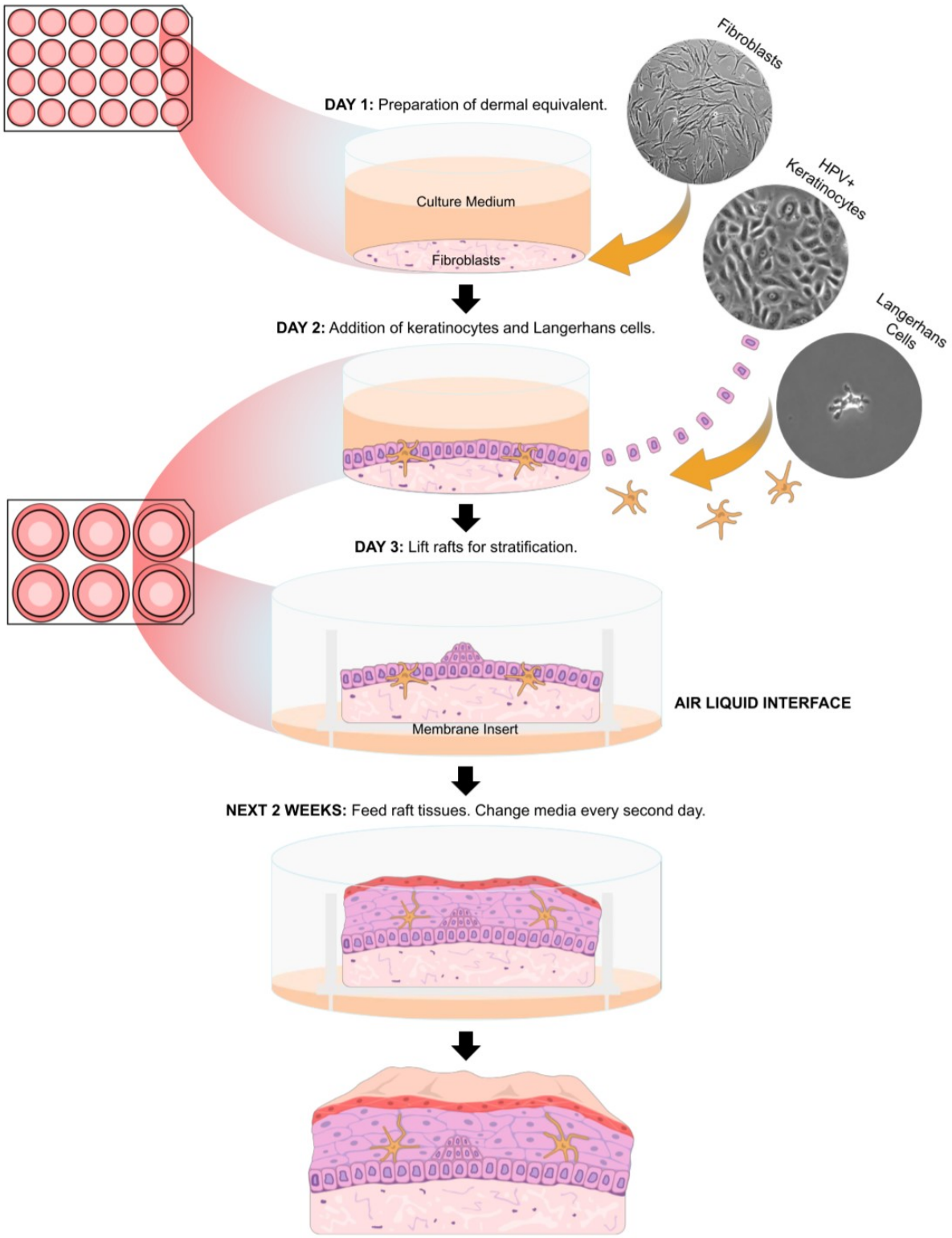


Figure 3A.3 – Flow diagram of the epithelial model. To culture immunocompetent organoids: on day 1, in a 24-well plate, fibroblasts are embedded in a collagen matrix as a dermal equivalent that supports keratinocyte growth. On day 2, keratinocytes (containing HPV episomes) and

Langerhans cells are seeded on top of the dermal equivalent at a ratio of 1:1. On day 3, the rafts are lifted and transferred to a 6-well plate with a membrane insert to create an air liquid interface exposing only the dermal equivalent to media. Subsequently, the rafts are fed for two weeks with a media change every second day. This environment encourages differentiation and stratification. With our experimental model, we strive to mimic what happens *in vivo*. With this model, the viral life cycle can be propagated, and this is only possible when cells are grown in 3 dimensions as organoid cultures. The fully-grown epithelium depicts mucosal, non-keratinizing epidermis and dermis.

3A.4.1 – Preparation of the revised model

Here, we will use the entire HPV16 AA variant genome's SNPs (*i.e.* altogether ~150 not just the 3 SNPs within the E6 gene as done previously) [Jackson *et al*., 2014; 2016]. Using the Cre-Lox recombination system [Lee *et al*., 2004; Wang *et al*., 2009] with subsequent selection potentially all transfected cells will carry the full HPV16 genome, which allows us to control the number of HPV+ cells within the 3D tissue. An alternative approach will also be used to increase DNA transfection efficiency three-fold, *i.e.* from ~10% using chemical transfection to ~30% with electroporation [Potter and Heller, 2003] prior to selection. We have found a way to create smaller 3D cultures enabling us to have more biological replicates. This will increase the effect and sample size for more robust and reliable statistics. For LCs, we will use the MUTZ-3 cell line [Masterson *et al*., 2002] as this allows experimental reproducibility and avoids donor variability when using DCs derived from peripheral blood monocytes [Chau *et al*., 2013]. We performed experiments to assess the suitability of this line using flow cytometry to identify a "differentiated", double positive population (langerin/CD1a) [**Figure 3A.4a,b**]. Another group has used this approach successively, albeit in a context of skin allergens and irritants [Kosten *et al*., 2015b].

The first step of our epithelial organoid approach will be to prepare the experimental model by carefully designing experiments with the research question at the forefront. This includes first establishing the hypotheses to be tested, defining the experimental variables (dependent and independent), and performing sample size estimations through power calculations based on expected effect size and variability (which can be informed via previous work or preliminary "pathfinder" experiments). Biostatistics are one facet of our eclectic approach, and while under-represented in past literature, may be applied to help resolve reproducibility concerns in cancer biology. When estimating sample sizes to appropriately test hypotheses, the concept of "biological independence" is important to consider. While truly independent replicates would be altogether uniquely derived biological samples (such as different donor individuals), it can be helpful to think of biological independence as a spectrum ranging from these uniquely derived specimens on one end, to increasingly more feasible options, albeit with a proportional loss in true biological independence. For epithelial cultures, we suggest that the "level of interrogation" should be the main consideration when determining the level of biological independence required. For example, with our research question focused on viral variants and their differential tumourigenic risk to the host, the level of interrogation would be at the interface between virus and host, such as the

introduction of viral genomes into keratinocytes via transfection (chemical or physical). So long as the research question is focused on the difference between the viral variants, a simple yet effective design would be to replicate the experiment with independent transfections using the same donor pool of cells (controlling this background if possible, to some extent, given inherent variability in passaging cells over time). If the research question was broader, by including donor/host variability, then the level of interrogation would correspondingly be at the host-level, requiring unique donors such as via patient-derived samples [Villa *et al*., 2018]. Once the experimental design has been established, the next stage of the approach is to perform material calculations and establish the organoid cultivation time-course.

This includes growing the required host cells (in a humidified incubator at 37°C and 5% CO2) either with the intention of incorporating into organotypic epithelia (such as keratinocytes, fibroblasts, and Langerhans cells differentiated from the MUTZ-3 myeloid leukaemia cells), or in support of those cells (such as 5637 bladder carcinoma cells and J2/3T3 mouse embryonic fibroblasts). While these cells vary in their media requirements, adherence (monolayer vs suspension cultures for the immune cells), and culture technique, an important consideration is the tissue origin and donor-background of the cells used, and whether they are primary or immortalized, and whether they will be matched for an experiment (*e.g.* gingival fibroblasts with gingival keratinocytes to model the oral epithelia). We have tried near-diploid immortalized keratinocytes (NIKS, [Allen-Hoffmann *et al*., 2000]), primary human foreskin/epidermal keratinocytes (PHFKs or HEKs) and primary human oral/gingival keratinocytes (HGKs). Also, we have generated epithelial organoids from patient-derived cervical biopsies (via suspected lesions at colposcopy, [Villa *et al*., 2018]). Primary cervical cells (both keratinocytes and fibroblasts from the uterine cervix) are now commercially available. The decision on which cells to use may be relevant due to tissue region susceptibility differences [Deng *et al*., 2018], tissue microenvironments, and variations in signalling affecting differentiation and possibly the viral life cycle and immune environment. Components of the dermal equivalent (typically collagen and fibroblasts), making up the extra-cellular matrix (ECM) may also be relevant factors.

Finally, beyond the host components, the sourcing and preparation of the viral genomes is at the crux of our experimental design and research question. Viral genomes have previously been from isolates, where a small number of modifications (*e.g.* in the E6 gene) can be introduced via mutagenesis, but now it is also possible to synthesize whole genomes, including all desired SNPs,

or intentional modifications/deletions, based on reference genomes. Using gene synthesis and cloning, it is also possible to introduce LoxP sites which are useful for Cre-LoxP-mediated transfections as a method of introducing viral genomes along with selection genes into host keratinocytes [**Figure 3A.4c**]. Using this technique, we performed a proof of principle study where we achieved an active viral life cycle using synthesized HPV16 EP whole genomes with LoxP sites in the SphI restriction site (previously used by others [Lee *et al*., 2004; Wang *et al*., 2009]) as well the alternatively chosen PmlI site (with a previously unknown effect on the viral life cycle), further upstream in the upper regulatory region (URR) and outside of potential transcription factor binding sites [**Figure 3A.4d**]. Synthesized HPV16 EP whole genomes both yielded an active life cycle, with the PmlI LoxP site performing as good if not better than the default SphI location.

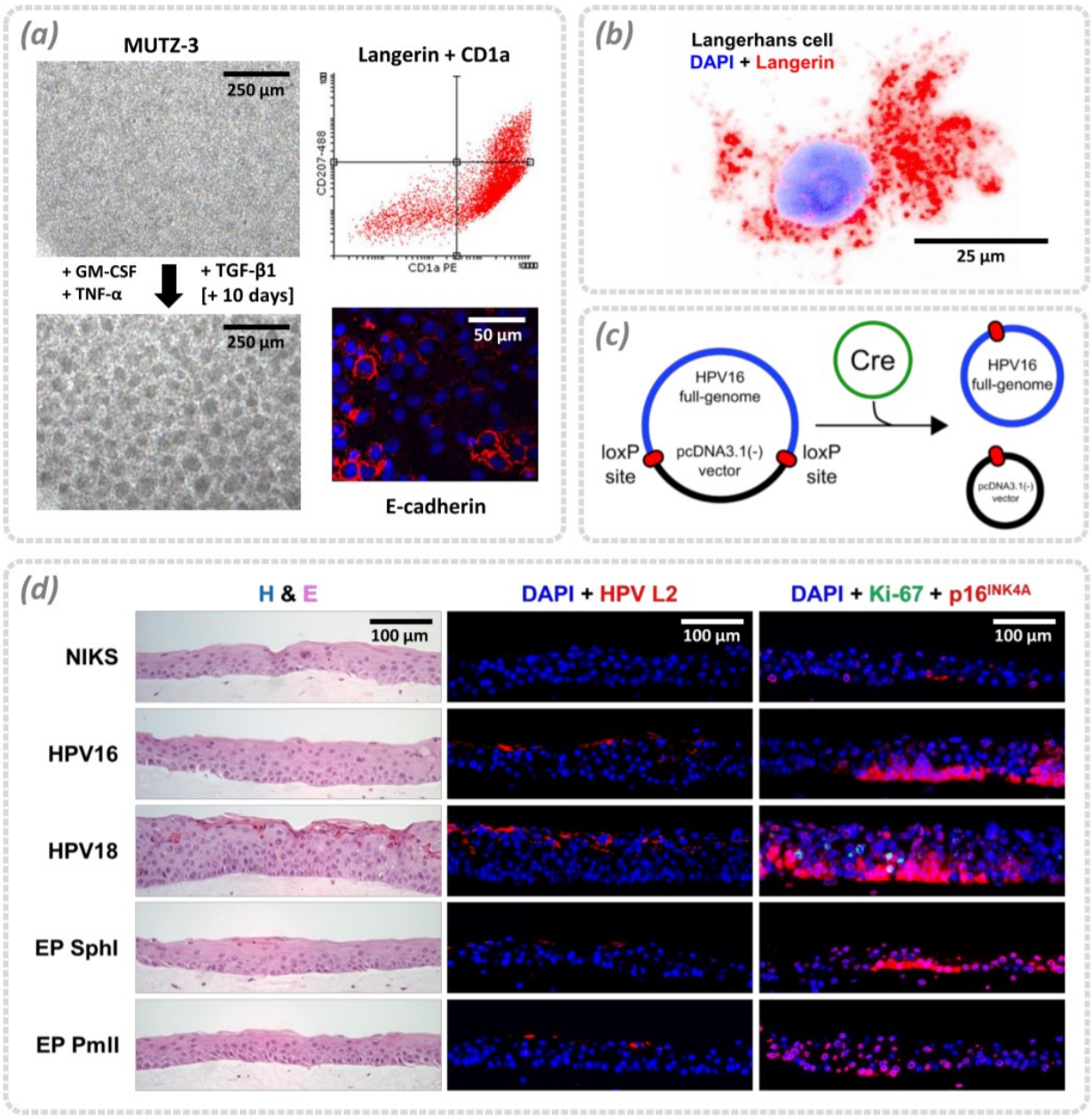


Figure 3A.4 – Characterization of the epithelial model. Langerhans cells (LCs) can be derived from MUTZ-3 cells via a cytokine-mediated differentiation over approximately 10 days **[a]**. Morphologically they appear more irregular and clustered together than their progenitors and differentiation success can be verified via flow cytometry for cell-surface markers CD1a and CD207 (langerin) as well as immunofluorescence for CD207 **[b]** and adhesion molecule E-cadherin. In addition to LCs, a major focus of our research is to study HPV16 genome variants, using full-genomes which contain all naturally-occurring polymorphisms, and to introduce these into our epithelial model. In this example case, near-diploid immortalized keratinocytes (NIKS,

[Allen-Hoffmann *et al*., 2000]) are co-transfected with LoxP-flanked HPV genomes and a Cre-recombinase expression plasmid to yield selected populations of keratinocytes containing viral DNA **[c]** and able to undergo a viral life cycle when grown as epithelial organoids (evidenced by capsid protein production, L2). As well, epithelium can be further characterized for proliferation markers (such as Ki-67) and cell-cycle dysregulation markers due to E7 expression ($p16^{INK4A}$) **[d]**. European Prototype (EP) genomes were designed to contain a LoxP site in either the SphI or PmlI restriction site, both in the non-coding upper regulatory region (URR) of the viral genome.

3A.4.2 – Organoid cultivation and characterization

While many past studies have focused on the cultivation of epithelial organoids [**Figure 3A.1**], including detailed methodology papers that serve as a good foundation and provide more thorough descriptions of the process, as well as video tutorials [Anacker and Moody, 2012], it will now be important to expand the perspective to different experimental design considerations (such as multi-component rafts) and a variety of end-points and time-courses. Recently, novel methods for studying the HPV life cycle have been described elsewhere [Bienkowska-Haba *et al*., 2018].

Different size organoids can be grown, and we have tried small (96-well sized), medium (48-well sized), and large (24-well sized) rafts, where smaller rafts allow for increased sample size for a variety of downstream applications. While being cost effective (with the exception of plate inserts and per sample materials), there is increased difficulty manually handling smaller rafts and the extraction yields of tissues are lower. Hence, we settled for medium-sized rafts as an optimal balance. As well, different amounts of fibroblasts can be seeded into the dermal equivalent, where we have found that an increased number helps with epithelial differentiation. Varying the number of infected cells within the epithelium could be a way to model varying stages of disease, from low-grade to high-grade lesions (where higher-grade lesions have a greater number of HPV+ cells [Algeciras-Schimnich *et al*., 2007]).

The duration of culturing (typically 1 to 21 days) can be assessed using time-series rafts, as we have done previously to aid in mathematical modelling [Murall *et al*., 2019]. While 14 days is typically used as the peak of the viral life cycle, shorter or longer durations (to a limited extent) may be relevant for assessing changes over time as well as the interaction between differentiation and viral replication, genome amplification, and transcription as they relate to persistence and integration. Prior to harvesting (typically <24 hours before), the thymidine analogue BrdU can be added to culture media to assess proliferation (where suprabasal proliferation is indicative of keratinocytes infected with HPV). When culturing and harvesting, biosafety precautions are required (*e.g.* biosafety level II in Canada), as active viral particles may be produced. Harvesting can be done to preserve structure, via fixation (*e.g.* formalin), or through tissue dissection (*e.g.* manual separation of the epidermis from dermis) followed by homogenization of the relevant compartment and molecular extractions (*e.g.* DNA, RNA, protein), or processing for single-cell suspensions. For formalin-fixed and paraffin-embedded tissues, typical histological assessment is performed using Haematoxylin & Eosin staining, whereas *in situ* techniques such as

immunohistochemistry and immunofluorescence can be used for qualitative and semi-quantitative characterization of host markers (*e.g.* Ki-67 proliferation marker, BrdU-incorporation proliferation marker, $p16^{INK4A}$ cell cycle dysregulation and surrogate E7 marker, cytokeratin 5 and 10 differentiation pattern markers, and viral markers (L1 and L2 capsid proteins)). Data and statistical analysis can be performed using open-source software. Extracted and purified molecules, such as nucleic acids (DNA or RNA), can be used for common assays such as viral copy number or viral/host gene expression or used for more high-throughput techniques such as microarray or NGS.

3A.4.3 – Immune-competent component

The feasibility of incorporating MUTZ-3-derived LCs into 3D epithelial cultures has been reported previously by the Gibbs group [Rodrigues Neves and Gibbs, 2018]. While LCs normally make up ~2-3% of the epithelium, this group used a ratio of 1:1 or even 2:1 in 3D cultures compared to the number of keratinocytes. This high number was deemed necessary because only a subset of cytokine-treated MUTZ-3 cells differentiate to double-positive langerin/CD1a LCs: *e.g.* 30-70% (S.W. Spiekstra, Gibbs group, personal communication) or even lower at 10-20% in our hands. Moreover, a proportion of differentiated LCs die or do not attach. Nevertheless, in our preliminary results from monolayer attachability experiments with half, equal or double ratios of cytokine-differentiated LCs *versus* keratinocytes, we yielded similar proportions of attached LCs (average of ~13%). The efficacy may be less in a 3D scenario due to LCs maturing and migrating out of the epithelium too soon. Instead of compensating this loss with an increased ratio of LCs, our populations will be further enriched for an increased number of differentiated LCs using anti-langerin conjugated microbeads. The starting population before differentiation will also have to be monitored carefully so that CD14+ and CD34+ proportions are close to equal [Santegoets *et al*., 2006]. Likewise, timely LC maturation, which should be triggered by cytokine production of surrounding keratinocytes and chemokines from dermal fibroblasts will have to be considered. LC maturation may not happen through these "natural" means and instead require a cytokine boost, *e.g.* of TNF-α and IL-1β and prostaglandin E2, or an allergen or irritant [Kosten *et al*., 2016].

So far, the outlined LC incorporation approach only allows a qualitative assessment of LC characteristics in the tissue. Hence, a quantitative assay is also needed for our research question to detect any measurable effect between the two variants under study. Therefore, we will use a

modified transwell migration assay adding artificial extracellular matrix (ECM) on top of the upper transwell membrane before adding various ratios of LCs and keratinocytes and delay LC maturation for at least two days to measure any effect due to HPV. Fibroblasts will be seeded into the underlying well. With this system, cell chemotaxis of LCs can be measured like *in* or *ex vivo*. Because the transfected HPV+ cells are not yet expected to be tumourigenic, only a few will spontaneously pass through the ECM but LCs are expected to do so if they receive the appropriate signals to "mature" from keratinocytes and fibroblasts.

Finally, it is tempting to use two different methods in parallel: to include LCs in one and omit them altogether in the other. Indeed, even without LCs, we can still test keratinocytes and fibroblasts for markers that render a milieu suitable for LC migration from epidermis to dermis [summarized in Kosten *et al*., 2015a]. Consequently, both approaches will be attempted using immunofluorescence and bead-based multiplex assays to detect immunological markers (or lack thereof) in keratinocytes and culture supernatant.

3A.4.4 – NGS and the Pathogen-Host Analysis Tool (PHAT)

NGS of extracted nucleic acids (DNA and RNA) allows for a comprehensive analysis of the molecular background of our organoids and the ability to determine differences in HPV16 variant interactions with host tissue [Jackson *et al*., 2016]. Important considerations include library preparation (whether or not a sequence capture or enrichment step will be used to enrich the low abundance viral sequences relative to the host), sequencing (platform used, short vs long reads, read depth anticipated), as well as bioinformatics analysis pipelines to be used (custom-scripts, high-performance computing cluster access, or desktop tools). Ultimately, next-generation or high-throughput sequencing (HTS) analysis can enable hypothesis-testing as well as hypothesis-generation via exploration, and possibly spur future research questions and additional cycles of organoid culturing.

To aid researchers with these data, we developed a platform to analyse pathogen-host relationships in next-generation sequencing data by using industry standard methods while reducing barrier to entry (https://github.com/chgibb/PHAT, [Gibb *et al*., 2019]). For the PHAT "toolbar", or graphical user interface (GUI), obtained sequence data is added (Input), quality-controlled (QC) and aligned with the appropriate reference sequence; this platform gives the user access to the alignment summary of obtained NGS data (DNA and/or RNA-seq) compared to a

pathogen or host reference sequence. Number and percentage of reads that have aligned to the reference are provided, and the user has access to an interactive (scrolling and zooming) linear visualization of viral reads and coverage across a reference genome, which also highlights single-nucleotide polymorphisms (SNPs) that are within reads, compared to the reference, that can be helpful for pathogen genotyping. The Output button gives access to customizable data tables that can be output and saved in preferred spreadsheet format (such as .csv or Excel) for use in other software or for publication/reports. This allows the user to select exactly which samples and columns of data they would like to output, such as quality control information, alignment statistics, SNPs/genotyping, etc. The Genome Builder is a tool within PHAT that can be used to create circular visualizations of pathogen reference genomes (rather than just linear maps for the round HPV plasmids). Viral genes, SNPs within the genes, and circular read coverage (which is important for assessing episomal/integrated forms in the case of HPV) can be added and annotated. We recently added the ability to work with pre-aligned data from large datasets, where users can take output from high-performance computing clusters and work with it on their desktop or laptop computer, and we are currently adding enhanced viral integration detection features (given one of our most pertinent research questions is the pattern of viral integration into host DNA between HPV16 variants). Overall, PHAT is under continued development and users are automatically notified of any updates, which can then be downloaded and installed seamlessly.

3A.5 – Limitations of Existing Research Models

While we advocate for a simple yet useful model of human epithelium, we appreciate that achieving this goal has limitations. George E. P. Box's wisdom on model utility can be extended to biological models: "Since all models are wrong the scientist cannot obtain a 'correct' one by excessive elaboration" [Box, 1976]. Hence, avoiding excessive elaboration may be important for keeping experiments manageable and focused, but an overly simplistic model may generalize or altogether lack essential components or characteristics that are found naturally to provide insight for the biological phenomenon studied. Box continues "… following William of Occam, one should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist, over-elaboration and over-parameterization is often the mark of mediocrity" [Box, 1976]. This limitation can be overcome by adapting the model to include only the essential components to answer the actual research

question. In this focused approach, we knowingly exclude parts in our organoid model, such as lymphoid cells, that *eventually* may become involved in the host immune process but that *currently* are outside our study frame. With continued advances in high-throughput assays as well as computing power and analytics, it may be possible to simulate human epithelia *in silico* using mathematical modelling [Murall *et al*., 2019]. Such a technique could be a powerful tool in combination with biological models to provide dataset training and verification, for answering a variety of questions pertaining to epithelial responses (*e.g.* due to viral infection) and testing hypotheses by modulating key parameters. Finally, advances in single-cell sequencing and *in situ* analysis may uncover new information about tissue heterogeneity and molecular signatures during a viral life cycle.

3A.6 – Conclusion

This commentary describes our concept accurately tailored to the crucial time-window of early virus infection, *e.g*. HPV elimination (or not) in a complex organism such as humans. We describe a unique, robust and reproducible alternative to animal models to study HPV immune biology during the full viral life cycle. While most other investigations centred around skin, we and a select few others [Kosten *et al*., 2015a; Kosten *et al*., 2016] also consider anatomical areas lined by mucosal tissue mostly affected by HPV: the uterine cervix and the oral cavity. Consequently, immune molecules in skin are well characterized [Rodrigues Neves & Gibbs, 2018] including a global approach [Spurgeon *et al*., 2017], while this is still largely lacking for both mucosal sites. Most importantly, HPV variants have not been addressed at all. Our approach will shed new light on host immune evasion by HPV in the context of HPV variants focusing on HPV and its interactions with epithelial LCs. In particular, we investigate the molecular signature of LCs surrounding keratinocytes in the epidermis and underlying fibroblasts in the dermis in the context of a high-risk HPV, *i.e.* type 16 and what role HPV variants, differently implicated in cervical disease have [Jackson *et al*., 2014; 2016 and references therein]. Here, for the sake of creating a reproducible and complex model to study two common, naturally occurring HPV16 variants, we do not wish to include genetic variables (in addition to those found in the viral genomes under study). However, the outlined model is amenable to be expanded to an epidemiologic study with multiple host genome backgrounds involving individual or all three anatomical sites with an appropriate sample size. To bear in mind: preparing organoids the way

we describe here requires an interdisciplinary mindset, meticulous preparedness coupled with a good pair of lab hands, a large portion of patience while maintaining these cultures, and a great deal of understanding and appreciation for bioinformatics. The lucky, successful candidate will be hugely rewarded with new discoveries.

3A.7 – Declarations

3A.7.1 – Acknowledgements

We are thankful to Kathlyn Alexander and Josee Bernard for initial experimental assistance as well as Dr. Melissa Togtema and Peter Villa for model discussions. Plasmids were kind gifts from Dr. Nagy (Cre), Dr. Lee (HPV16), and Drs. Chow and Broker (HPV18).

3A.7.2 – Data accessibility

The data and materials supporting this article are in text or available upon request.

3A.7.3 – Authors' contributions

RJ and IZ have contributed equally to conception and design, acquisition of data and literature, and overall interpretation. They were both instrumental in drafting and critically revising the article. SE created the final figures, assisted with acquisition and interpretation of data and formatted the manuscript for submission. All authors have contributed to revisions and approve the final version to be published.

3A.7.4 – Competing interests

We have no competing interests other than a co-authorship with one of the special issue guest editors, Dr. Alizon [Murall *et al*., 2019].

3A.7.5 – Funding

This work was supported NSERC grants to IZ (#355858-2008, #435891-2013, #RGPIN-2015-03855), an NSERC Alexander Graham Bell Canada Graduate Scholarship-Doctoral (CGS-D) to RJ (#454402-2014), and a Northern Ontario Heritage Fund Corporation (NOHFC)-sponsored internship for SE. The funding bodies had no role in study design, data collection, data analysis and interpretation, or preparation of the manuscript.

CHAPTER 3B – THEORETICAL APPLICATION OF THE MODEL

This chapter is an interdisciplinary application of our epithelial organoid model, specifically for the study of epithelial stratification and infection dynamics *in silico* using mathematical models. Below are excerpts relevant to this dissertation, focused on my primary contributions, from the full-text of a collaborative research article published in *PLoS Computational Biology* on 23 Jan 2019 (DOI: 10.1371/journal.pcbi.1006646, originally pre-printed in *bioRxiv* on 10 Dec 2017, DOI: 10.1101/231985) [Murall *et al*., 2019]. It has been adapted with permission from the authors for re-use within this dissertation.

Excerpts from: Epithelial stratification shapes infection dynamics

Carmen Lía Murall[1], **Robert Jackson**[2,3], Ingeborg Zehbe[2,4], Nathalie Boulle[5], Michel Segondy[5], Samuel Alizon[1]

[1]Laboratoire MIVEGEC (UMR CNRS 5290 UM), Montpellier, France
[2]Probe Development and Biomarker Exploration, Thunder Bay Regional Health Research Institute, Thunder Bay, Ontario, Canada
[3]Biotechnology Program, Lakehead University, Thunder Bay, Ontario, Canada
[4]Department of Biology, Lakehead University, Thunder Bay, Ontario, Canada
[5]Pathogenesis and Control of Chronic Infections, INSERM, EFS, Université de Montpellier, Montpellier, France

Keywords: Stage-structure populations, Human papillomaviruses, *Chlamydia trachomatis*, Three-dimensional organotypic culture, Mathematical models, Within-host dynamics

3B.1 – Abstract

Infections of stratified epithelia contribute to a large group of common diseases, such as dermatological conditions and sexually transmitted diseases. To investigate how epithelial structure affects infection dynamics, we develop a general ecology-inspired model for stratified epithelia. Our model allows us to simulate infections, explore new hypotheses and estimate parameters that are difficult to measure with tissue cell cultures. We focus on two contrasting

pathogens: *Chlamydia trachomatis* and Human papillomaviruses. Using cervicovaginal parameter estimates, we find that key infection symptoms can be explained by differential interactions with the layers, while clearance and pathogen burden appear to be bottom-up processes. Cell protective responses to infections (*e.g*. mucus trapping) generally lowered pathogen load but there were specific effects based on infection strategies. Our modeling approach opens new perspectives for 3D tissue culture experimental systems of infections and, more generally, for developing and testing hypotheses related to infections of stratified epithelia.

3B.1.1 – Author summary

Many epithelia are stratified in layers of cells and their infection can result in many pathologies, from rashes to cancer. It is important to understand to what extent the epithelial structure determines infection dynamics and outcomes. To aid experimental and clinical studies, we develop a mathematical model that recreates epithelial and infection dynamics. By applying it to a virus, human papillomavirus, and a bacterium, chlamydia, we show that considering stratification improves our general understanding of disease patterns. For instance, the duration of infection can be driven by the rate at which the stem cells of the epithelium divide. Having a general model also allows us to investigate and compare hypotheses. This ecological framework can be modified to study specific pathogens or to estimate parameters from data generated in 3D skin cell culture experiments.

3B.2 - Introduction

We address to what extent epithelium dynamics affect infection dynamics and as a result determine infection outcomes. First, we introduce a general epithelium model, which we calibrate using existing data, as well as original cell culture data from a spontaneously immortalized human cell line (NIKS) [Allen-Hoffmann *et al*., 2000]. With this data we infer parameters that are difficult to measure, such as the fraction of symmetric cell divisions. We then 'infect' this epithelial model with chlamydia, wart-associated HPVs and oncogenic (high-risk, HR) HPVs to investigate how protective measures by the epithelium affect infection load and duration, while identifying the parameters that control key infection traits. We find that epithelium stratification plays a key role in the dynamics and outcomes of these infections.

3B.3 – Results

3B.3.1 – Uninfected epithelial dynamics

To obtain experimentally relevant parameter estimates, we used our model and the known parameters as priors to estimate values using original data from raft cultures of NIKS (Normal Immortal KeratinocyteS) cells. The NIKS cell-line grows into a 3D epithelium structure and is commonly used as a model of cervicovaginal tissue, though they are known to differ from *in vivo* tissue [Allen-Hoffmann *et al*., 2000]. **Figure 2A** and **B** show an example of NIKS cell growth into stratified form. **Figure 2C** shows the dynamics of the number of basal and suprabasal (non-keratinized and keratinized) cells, along with the inferred dynamics from the model.

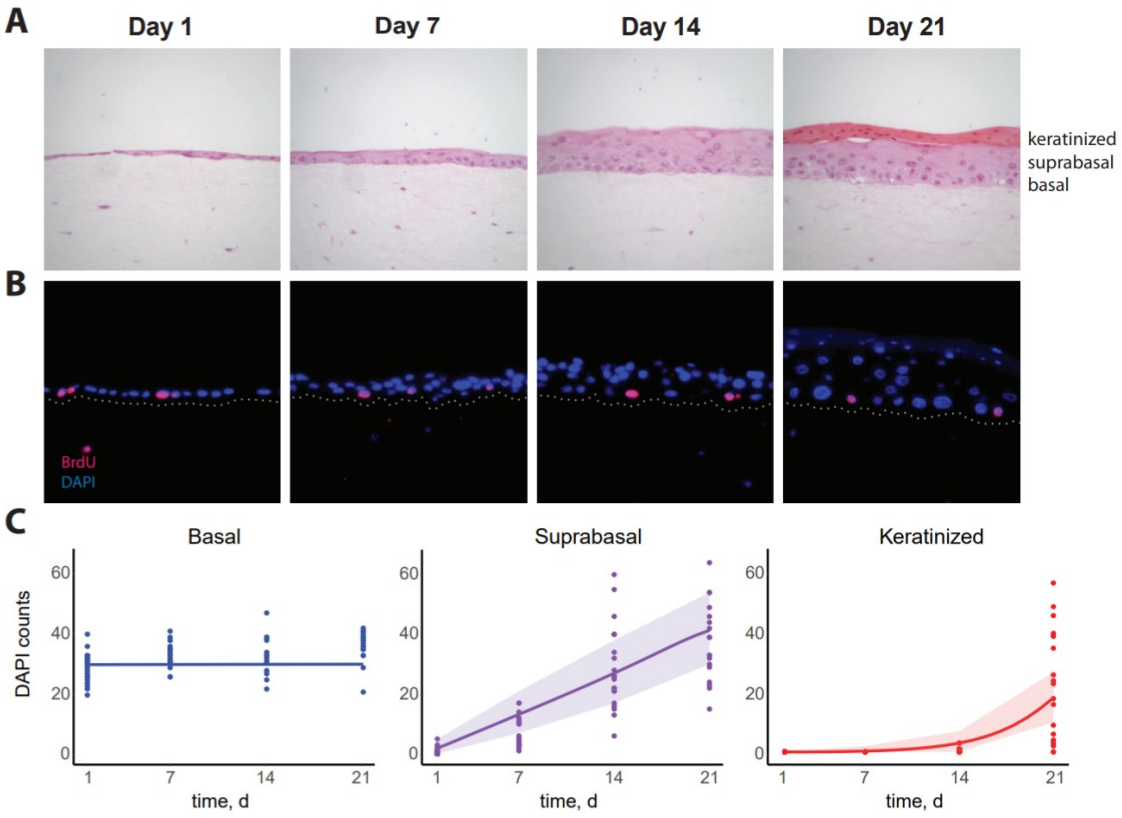


Figure 3B.1 – Epithelial cell growth in 3D raft cultures. [a] NIKS grown from a single layer over a period of three weeks. Dark pink layer in week 3 consist of cornified cells that accumulate on the surface. **[b]** Immunofluorescence staining: DAPI (blue) is nuclear staining for cell counting and BrdU (red) is for identifying cells undergoing division; white dots are added to delineate basal lines. **[c]** Data of NIKS growth over time with model fitting. Shading corresponds to 95% prediction interval, assuming the data follows a Poisson distribution.

3B.4 – Discussion

Epithelial infections are a major public health burden, and, in particular, STIs are on the rise causing a worldwide concern [Hay *et al*., 2014; Carmona-Gutierrez *et al*., 2016; WHO, 2016]. Quantitative models, both experimental and mathematical, are essential in developing our understanding of these infections. As for systemic (and virulent) infections such as HIV and HCV, mathematical models have been developed to predict and analyze the kinetics of epithelial infections. Here, we show that to understand the kinetics of epithelial infections, it is essential to account for the stratified structure of the epithelium, a property that is absent from most models. We illustrated how such a general framework can be combined with 3D cell culture data to estimate key parameters and how it can generate relevant insights regarding the course of epithelial infections.

3B.4.1 – Dynamical implications of ecological features

The rate of basal cell proliferation had a strong effect on the homeostasis of both uninfected and infected epithelia, which suggests an ecological 'bottom-up controlled' system [Lindeman, 1942; Gruner *et al*., 2008], analogous to those found in free-living food webs. These bottom-up effects are more apparent if we consider that basal cell replication is strongly determined by the resources that are available in the basal lamina, such as growth factor. While we did not explicitly model the resources of the basal layer (it is implicit in the basal proliferation rates), the growth of the cells in the experimental set-up does depend on concentration and temporo-spatial distribution of growth factors, impacting epithelial thickness and proliferation rates. Therefore, this ecological insight of bottom-up driven systems, could be tested more formally in experimental systems by monitoring resource concentrations.

3B.4.2 – Perspectives

Finally, opening a dialogue between mathematical modeling and experimental data generates new hypotheses to test. One of the clearest illustrations of this is our result that burst size differences appear as the most parsimonious explanation to explain symptom differences between wart-causing and lesion-causing HPV infections. Technological improvements in clinical and experimental techniques also allow us to test more subtle predictions. Testing hypotheses generated by the model will allow us to move forward by validating the model assumptions that

are consistent with the data and rejecting the others. This will allow us to increase the model complexity and test more elaborate predictions. We hope to inspire experimental studies on infections of stratified epithelia to focus more on dynamics and time series approaches (including mathematics) to better understand these varied and broadly impacting pathogens.

3B.5 – Materials and Methods

3B.5.1 – Ethics statement

The Thunder Bay Regional Health Research Institute's Biosafety Committee approved all research involving NIKS cell line cultures. The NIKS cell line [Allen-Hoffmann *et al*., 2000] was obtained from Dr. Paul Lambert, McArdle Laboratory for Cancer Research, University of Wisconsin.

3B.5.2 – Cell culture data

Organotypic culture growing techniques used here have already been described in detail elsewhere [Jackson *et al*., 2014; 2016]. Original experiments were performed to obtain time series data with sufficient replicates for model fitting. Three independent experiments were performed, with rafts harvested at one-week intervals (0, 1, 2, and 3 weeks) starting the day after lifting them to an air-liquid interface. From a total of 12 formalin-fixed, paraffin-embedded (FFPE) rafts, 48 tissue slices were imaged using fluorescence microscopy (DAPI staining for cell nuclei) and resulted in 3 Fields of View (FOV) per slice (n = 144). Counts in each FOV were done semi-automatized using ImageJ cell counting software.

3B.6 – Declarations

3B.6.1 – Acknowledgements

We would like to thank the anonymous reviewers for significantly improving the manuscript. We would like to thank Kathlyn Alexander for assistance with the raft culture experiments and also to Drs. Ignacio G. Bravo, Jessie L. Abbate and Jérémie Guedj for helpful discussions. We also thank Dr. Paul Lambert (McArdle Laboratory for Cancer Research, University of Wisconsin) for providing the NIKS cell line.

3B.6.2 – Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 648963). CLM and SA also received funding from the CNRS and the IRD. RJ and IZ received funding from the Natural Sciences and Engineering Research Council of Canada (NSERC), with a Discovery Grant to IZ (#RGPIN-2015-03855) and a NSERC Alexander Graham Bell Canada Graduate Scholarship-Doctoral (CGS-D) to RJ (#454402-2014). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors have declared that no competing interests exist.

3B.6.3 – Data availability

All relevant data are within the paper and its Supporting Information files.

CHAPTER 3C – ENHANCING THE MODEL'S BIOLOGICAL RELEVANCE

This chapter contains recent progress on enhancing aspects of our epithelial organoid model for studying the unique pathogen-host relationships of HPV16 sub-lineages. I begin by describing improvements to the experimental viral genomes [**3C.1 – Full-Length HPV16 Sub-Lineage Genomes**], followed by methodologies for incorporating them into host cells [**3C.2 – Strategies for Introducing HPV16 Genomes to Host Keratinocytes**], and conclude on our continued efforts to model innate immunity, such as including Langerhans cells, in epithelial organoids [**3C.3 – Adapting the Organoid Model to Study Innate Immunity**].

3C.1 – Full-Length HPV16 Sub-Lineage Genomes

Studying the viral life cycle of HPV16 in a controlled experimental setting typically requires the use of full-length circular viral genomes. Our previous studies on HPV16 sub-lineages in an organoid epithelial model used full-length genomes [Jackson *et al*., 2014; 2016; see **CHAPTER 2**], but did not include all EP (A1) and AA (D2 or D3) variations across the whole viral genome. Rather, these were based on a mutagenized HPV16 W12E genome [GenBank # AF125673], as a common genetic backbone, to isolate the effects of only the non-synonymous variant SNPs in the E6 gene. Beyond E6, there may be relevant SNPs in other genes and genomic regions [see **Figure 3C.1**] such as E2 (with 23 SNPs), which regulates E6 expression via *cis*-binding to the non-coding viral upper regulatory region (URR, with 15 SNPs of its own) [Hubert *et al*., 2005; Lace *et al*., 2009]. To address this limitation, we have sought full-length HPV16 sub-lineage genomes (A1, D2, and D3) containing all the natural SNPs present. Overall, variant lineages are considered to have 1.0 to 10.0% difference in their whole genome (90.0 to 99.0% similarity), whereas sub-lineages have 0.5 to 1.0% difference in their whole genome (99.0 to 99.5% similarity) [Burk *et al*., 2013]. These differences include SNPs, but can also include indels, leading to gaps in alignment and different genome lengths (which is also why reference genomes differ slightly over time, as sequencing errors have been identified and corrected). It is worth keeping in mind that these reference genomes are useful for researchers as they can be agreed upon representatives of a sub-lineage branch based on phylogenetic analysis, but ultimately, they are individual isolates of heterogenous groups. There is an ~98% sequence homology between the entire HPV16 genomes of the EP reference sequence (sub-lineage A1, GenBank # K02718) and the AA variant (sub-lineage D3 as an example, GenBank # AY686579.1) [**Figure 3C.1**].

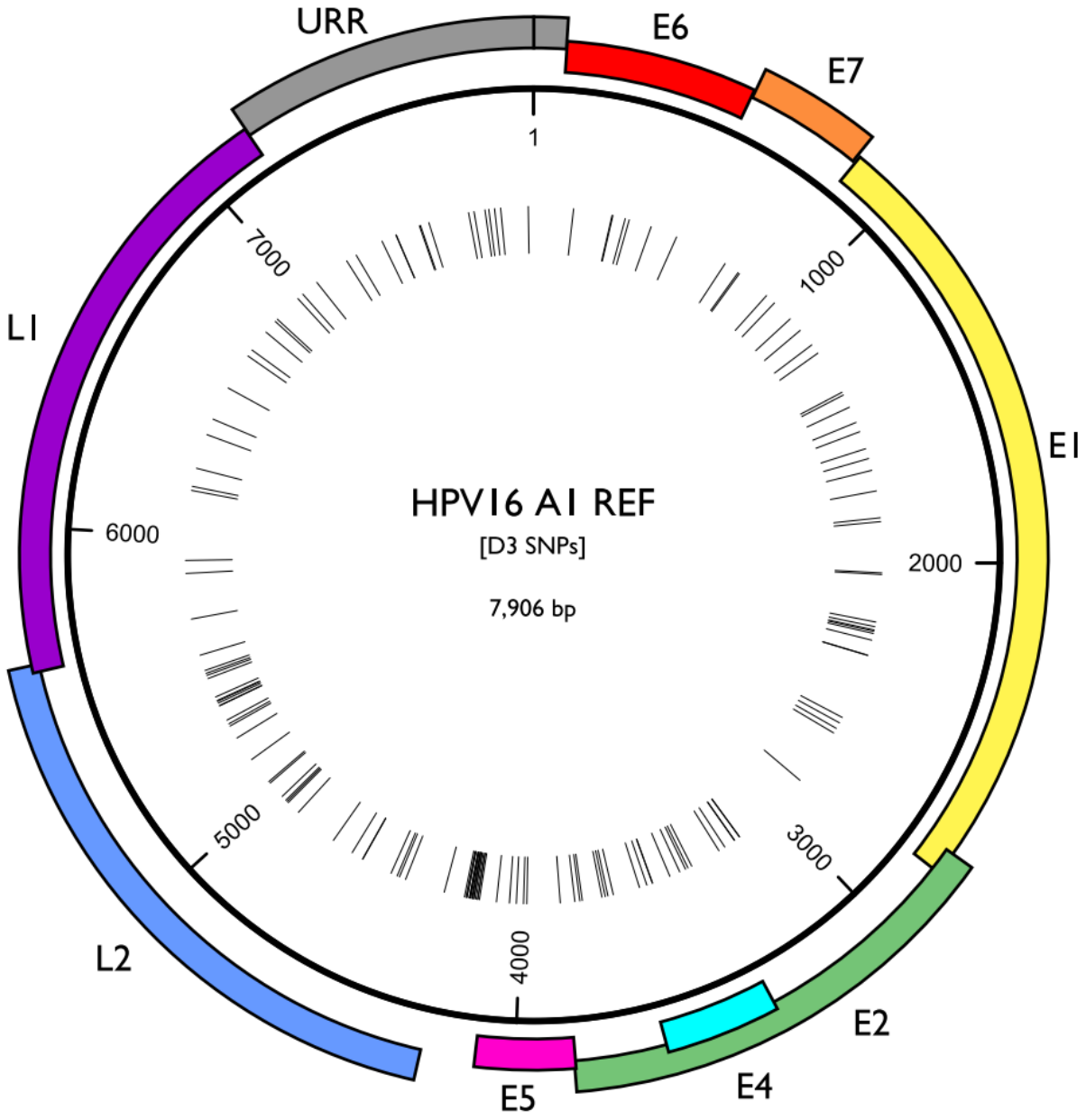


Figure 3C.1 – HPV16 A1 vs D3 single-nucleotide polymorphisms (SNPs). The circular genome of HPV16 A1 (EP reference sequence, GenBank # K02718) is annotated with genomic features. Black lines in the inner ring represent the 151 SNPs present in the D3 variant sub-lineage (AA, GenBank # AF402678) throughout the entire genome and present in genes (E6 [7 SNPs], E7 [3], E1 [31], E2 and E4 [23], E5 [6], L2 [32], L1 [17]), as well as non-coding regions (short non-coding region [17], URR [15]). Circular visualization was created using Circos [Krzywinski *et al*., 2009].

While D2 (Asian-American 2, GenBank # AY686579) and D3 (Asian-American 1, GenBank # AF402678) are considered separate sub-lineages, I often refer to them together (*i.e.*, D2/D3) as they have identical E6 proteins, with no additional non-synonymous SNP differences between their E6 reference sequences. Meanwhile, an additional synonymous SNP is present at the gene-level: D2 vs D3 E6 gene (nt 83-559) sequences = 476/477 (99% match). Compared to the A1 reference sequence (GenBank # K02718), D2 has six E6 SNPs: G145T, T286A, A289G, C335T, T350G, A532G. When comparing D3 to A1, there are seven E6 SNPs (with the additional synonymous one underlined): G145T, T286A, A289G, C335T, T350G, G433A, A532G. The three non-synonymous SNPs that are shared between D2/D3, relative to the A1 E6 reference sequence, lead to amino acid residue changes and potential (albeit subtle) structure differences that could have a functional effect [**Figure 3C.2**].

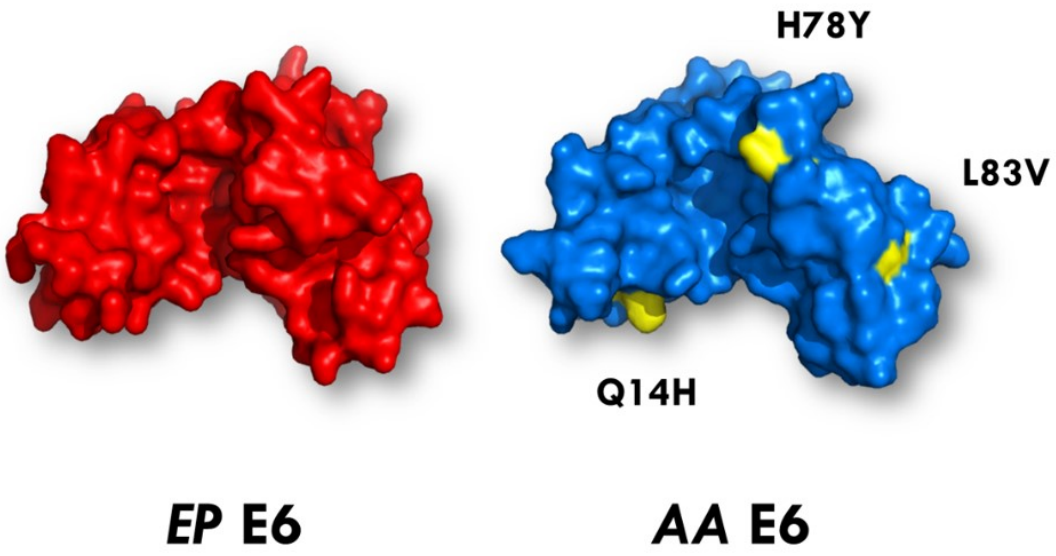


Figure 3C.2 – Predicted 3D protein structures for EP (A1, red) and AA (D2/D3, blue). Phyre2 (intensive mode) [Kelley and Sternberg, 2009; Kelley *et al*., 2015] was used for *in silico* prediction of the 3D protein structures for EP E6 (GenBank #AAA46939.1) and AA E6 (GenBank #AAV91644.1). The sequences were used to generate a 3D model file (.pdb) based on the 151-residue version of E6, starting at the 2nd Met residue. The 3D models were visualized using PyMOL [The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC]. Amino acid residue differences numbered based on the 151-residue version of E6 are highlighted in yellow (Q14H/H78Y/L83V), causing subtle changes to the protein structure and potential functions.

While D2/D3 E6 proteins may be identical, when considering the full-length genomes there are 37 SNPs between D2 and D3 (spanning the entire genome), as well as two indels. To properly address our research question, considering the differences between full-length A1, D2, and D3 HPV16 genomes, we acquired the necessary genomes as plasmids [see **APPENDIX A, Table A.1**]. Generally, the full-length viral genomes were inserted reverse-orientation into LoxP-flanked (floxed) multiple-cloning site plasmids also containing a neomycin (G418) resistance gene followed by a promoter [as described in **CHAPTER 3A**; Lee *et al*., 2004; Wang *et al*., 2009; Bodily *et al*., 2011]. With this strategy, after Cre-recombination (enabled due to the LoxP sites surrounding the viral genome and co-transfection with a Cre-recombinase-expressing plasmid) the viral genomes are cut out, recircularized, and have a single LoxP sequence (34 bp long) in a region of the viral genome that has been reported not to affect its function (at nt 7,465, the *SphI* restriction site) [Lee *et al*., 2004]. In addition, the recombination yields a second recircularized plasmid containing the remainder of the original plasmid: a neomycin resistance gene (conferring G418 resistance in mammalian cells), but with the promoter now positioned in front. This design ensures that cells transfected with plasmids having been recombined by Cre-recombinase express a resistance gene as well as have an intact and active viral episome. While useful, a limitation of the Cre-LoxP system which is important to consider for our experiments is that illegitimate DNA recombination and damage could occur in cells expressing Cre recombinase, due to cryptic or pseudo LoxP sites present within the mammalian genome [Thyagarajan *et al*., 2000].

While working on adapting this design to our HPV16 sub-lineage genomes we considered whether the SphI restriction site is the best location for an interruption in the genome. Given that the URR has many binding sites for viral as well as host factors, and that there are SNPs in this region between EP and AA variants [Hubert *et al*., 2005], it was possible that the exact site of any interruption could have a significant functional effect on the viral life cycle. The final location of the LoxP site should be in a non-coding region of the viral genome, with no known function or binding sites, and with no sequence variation between EP and AA. The alternative site that we selected to compare to the traditional SphI site was the PmlI site [**Figure 3C.3**]. Given its location upstream of major transcription factor binding sites, we hypothesized that the viral life cycle could be more productive with this alternative location for a LoxP sequence interruption. On the other hand, the PmlI site is located near a late polyadenylation signal/site of the late genes, which could possible interfere with capsid gene transcription.

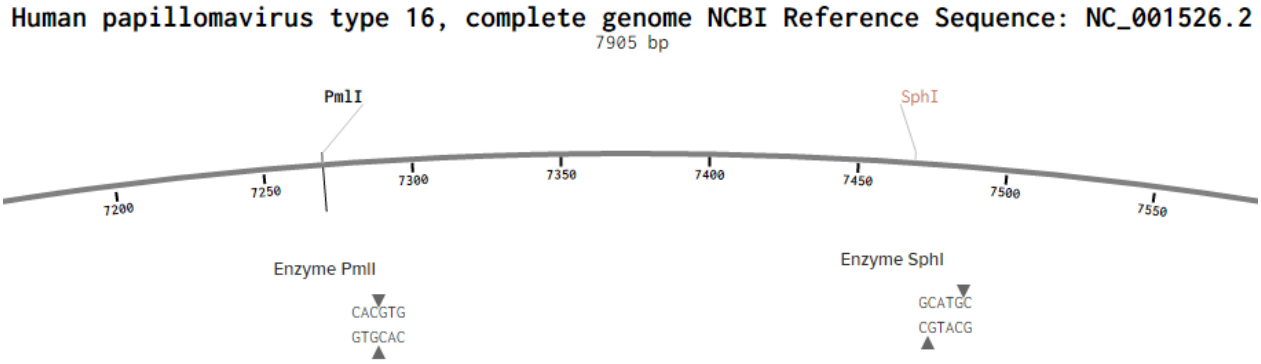


Figure 3C.3 – Location of preferred LoxP sites in HPV16's URR. PmlI and SphI restriction sites are shown in the non-coding URR of HPV16's genomes. Image was constructed using Benchling Biology Software.

We had the A1 genome synthesized by GenScript, using the latest version of the A1 reference sequence available at the time (GenBank # NC_001526.2). We custom ordered three plasmids via GenScript's gene synthesis and cloning service as 4 µg lyophilized DNA: Full Length_pcDNA3.1(-), hpv16-ep-sphi-linearized (same site as previous studies, close to E2 binding site and within host factor binding sites, start 7,469 - end 7,468, flanked by LoxP sites in the same orientation, 8 bp spacer = GCATACAT, to permit excision by Cre recombinase), and hpv16-ep-pmli-linearized (new site chosen by us, follows the late polyA signal, start 7,270 - end 7,269, flanked by LoxP sites in the same orientation, 8 bp spacer = GCATACAT, to permit excision by Cre recombinase). Full Length_pcDNA3.1(-) is the non-functional precursor plasmid used to construct the two alternate LoxP site placements in the natural EP genome (this was stored at -20°C as 4 µg lyophilized DNA in the sealed tube as there was no need to transform it into bacteria in our hands). The functional plasmids (stored at -20°C until re-suspended) were created via mutagenesis from the precursor plasmid to result in a LoxP site at the endogenous viral SphI or PmlI restriction site (both contained within the HPV16 URR). These were all cloned into pcDNA3.1(-) vectors, containing Amp^R (bacterial) and Neo (mammalian) resistance. Target sequences were inserted in reverse orientation using XhoI and XbaI to prevent CMV-driven expression of viral genes. *In silico* Cre-lox recombination of these sequences using Cre-ACEMBLER [Christian Becke and Imre Berger, University of Bristol] resulted in the desired products. From Dr. Nagy we requested pCAGGS-NLS-cre to serve as Cre expression plasmid to

permit Cre-Lox recombination when co-transfected with floxed plasmids. From Dr. John Lee we requested pEGFP Ni HPV16 to use as an HPV16 control for the Cre-Lox system [Lee *et al*., 2004; Bodily *et al*., 2011]. From Dr. Louise Chow and Dr. Tom Broker we requested pNeo-loxP HPV-18 to serve as a highly productive positive control for this system in our hands. The HPV18 LoxP by the Chow group gives high amplification and viral production in 3D culture [Wang *et al*., 2009; see **CHAPTER 3A, Figure 3A.4**] whereas Dr. Lee's HPV16 clone, which contains the W12 genome, gives low amplification. We suspected originally that the low viral life cycle production may be due to the SphI LoxP placement, but using the PmlI site, we found a similar level of capsid protein [**CHAPTER 3A, Figure 3A.4**]. This phenomenon, of HPV16 having low amplification and viral life cycle production *in vitro* compared to HPV18 and HPV31 [personal communication with Dr. Bodily], even though HPV16 is the more prevalent type in human cancers, could be a topic for future research.

Once we confirmed that both synthetic A1 genomes yielded a productive life cycle, and that both SphI and PmlI LoxP site placements were feasible (albeit, without substantial enhancement in productivity using the alternate PmlI site), we sought full-length D2 and D3 genomes to incorporate into a Cre-LoxP system. While synthesis was an option, we were able to receive D2 and D3 isolate plasmids from Dr. Michael Dean [**Appendix A, Table A.1**] and instead used these to incorporate as floxed inserts into pcDNA3.1(-) vectors, as described in the paragraph above. D2 and D3 sequences are currently being sub-cloned by GenScript and will be used in future experiments (after being fully sequence verified as well) to rigorously test the variant-specific integration hypothesis. Addressing the limitations of the epithelial organoid model also required enhancements to our strategy for introducing the genomes into host keratinocytes.

3C.2 – Strategies for Introducing HPV16 Genomes to Host Keratinocytes

Previously, we did not use Cre/LoxP-mediated recombination (as described in the section above), but instead our original chemical co-transfection used a re-circularized HPV16 genome, along with a GFP selection plasmid, but without any selection gene of its own [Jackson *et al*., 2014; 2016]. These experiments were performed in immortalized cells, NIKS [Allen-Hoffman *et al*., 2000], and with this prior technique we would expect at most 50% of the cells after selection to contain HPV DNA due to the nature of the co-transfection. We have sought to address the limitations of our prior model from the following perspectives: *i*) cells, *ii*) plasmids (discussed in

3C.1), and *iii*) transfection technique. First: *i*) the host cells used should be more biologically relevant: mucosal cervical/oral keratinocytes, rather than primary or immortalized foreskin-derived [non-mucosal] keratinocytes [as previously used in Jackson *et al*., 2014; 2016]. Next: *ii*) plasmids previously used had only the changes in HPV16 E6 which lead to amino acid changes, so full-length variant genomes (A1, D2, D3) should be used. Finally: *iii*) the overall transfection technique required enhancements: the use of Cre-Lox co-transfection based on Dr. Lee (HPV16, although they used an adenoviral vector) [Lee *et al*., 2004] and Dr. Chow (HPV18, but similar methods to ours) [Wang *et al*., 2009], where each requires LoxP sites in the HPV genome in the URR. Another aspect of the transfection technique considered was the transfection method, where we have tried electroporation in the hopes to improve overall transfection efficiency versus chemical transfection efficiency (which is notoriously low in keratinocytes). To achieve efficient HPV-DNA transfection, maintenance, and amplification, the full-length HPV16 genomes are transfected into monolayer keratinocytes (*e.g*., previously NIKS and primary human foreskin keratinocytes, but more relevant are mucosal keratinocytes: primary oral/gingival keratinocytes or primary cervical keratinocytes) employing an approach based on Cre/LoxP-mediated recombination [Lee *et al*., 2004; Wang *et al*., 2009, Bodily *et al*., 2011]. Our current strategy [see **Figure 3A.4** in **CHAPTER 3A**] involves co-transfection of a floxed viral genome with the Cre-recombinase expression plasmid. Transfected cells could then be selected (to increase the proportion of HPV-containing cells) or expanded and used immediately as the basal keratinocytes for generating a multi-layered epithelium.

From most recent experiments, we noticed that primary cells are more sensitive to HPV DNA transfection, yielding higher GFP-positive cells, but with increased cytotoxicity. There is also a trade-off between selecting cells and immediately using them for rafts. In terms of our research question, the longer we have them in culture the greater the chance of spontaneous integration; *i.e.* selection allows the depletion of the negative cells but with the trade-off that the cells are in culture longer, which is especially problematic with primary cells of limited proliferative capacity [Villa *et al*., 2018]. We calculated transfection (electroporation) efficiency with a GFP Cre-Lox HPV16 plasmid (control from Dr. Lee) in primary cervical keratinocytes. We found the following conditions were optimal for primary cells: 10 µg/mL of DN, 1 x 10^6 cells per cuvette (4 x 10^5 cells in 400 µL, for each well of a 6-well plate), 100% hypotonic electroporation buffer, room temperature, and 1x 100 µs pulse at 280 V. GFP fluorescence images were overlaid

with phase-contrast images and GFP signals were only counted in healthy-looking, non-apoptotic cells. Smaller, rounded up cells expected to be apoptotic, were not included in our counting. Based on these criteria a mean of 6.5% keratinocytes showed GFP expression after 24 hours. After 48 hours, these numbers increased to a mean ranging from 9.6% (only those strongly expressing GFP) to 17.3% (both weakly and strongly expressing GFP). Notably, approximately 50% of the electroporated cultures subdued due to cytotoxicity caused by the procedure itself, which could not be remedied, *e.g.,* using salmon sperm carrier DNA and/or diluting the plasmid DNA in molecular-grade water rather than TE buffer. Hence, these electroporated cultures take four days to grow up to ~75% confluence, where they would be used for rafting. We recommend in future experiments with adult primary cells to exchange the culture media following attachment of cells (to remove hypotonic buffer and decrease toxicity) and to carefully consider selection (and opt not to include it) as this may compromise the research question by introducing HPV DNA integration. It would also slow down proliferation and take at least twice (or greater) the time to reach the same confluence. Altogether, this enhanced procedure based on electroporation is a more gentle and straightforward approach with the best possible transfection efficiency (*i.e.*, higher than the <10% that we typically obtained in an earlier study) [Jackson *et al*., 2016] and provides a means to grow organoids from mucosal keratinocytes with HPV DNA introduced.

Improving the full-length viral genomes, the relevance of the host cells, and the transfection strategies used for introducing the viral genomes to the host cells are all important factors for enhancing our epithelial organoid model to address our research questions. Beyond these improvements, future experiments should also assess *in situ* amplification of HPV DNA as well as the replication competence of HPV episomes introduced into host keratinocytes. This can be accomplished using DNA qPCR copy number assays and comparing DpnI-treated and untreated DNA from experimental cultures [Mori *et al*., 2014], as DpnI digests bacterially-synthesized plasmid DNA.

3C.3 – Adapting the Organoid Model to Study Innate Immunity

The next aspect of the organoid epithelial model which we sought to enhance was the ability to study innate immunity, specifically to test the hypothesis that HPV16 sub-lineage have differential innate immune evasion abilities [Jackson *et al*., 2014]. A framework for incorporating Langerhans cells into the model is discussed in **CHAPTER 3A**, and our recent progress has

revealed that incorporating these cells into the model is challenging. We have optimized cytokine-induced differentiation of MUTZ-3 cells into Langerhans cells ($CD1a^+/CD207^+$, assessed via flow cytometry) and have attempted incorporating them into the organoid model. Preliminary experiments were performed using magnetic-enrichment to purify the differentiated MUTZ-3 (10-30% population of $CD1a^+/CD207^+$ when unpurified) and seeding different ratios of these cells relative to the primary human keratinocytes (2%, a realistic amount found in epithelia, assuming all embed vs 20%, 10x greater, in case there is loss) [**Figure 3C.4**]. While optimal epithelial stratification was associated with increased purity and a lower ratio of LCs, as expected, no CD207 or even CD1a positive cells were detected. Media collections were performed throughout the rafting process, followed by membrane washes, and few cells were detected and therefore could not account for the undetected cells. It could be possible that the LCs did not survive through the 2-week rafting period (perhaps due to lack of optimal media and growth factors), or that they are very rare and further analysis of serial sections along with optimization of the immunostainings will yield their detection. These experiments are currently in progress. Future experiments could also include time-course harvests of the rafts [such as described in **CHAPTER 3B**; Murall *et al*., 2019] to check the status of the LCs earlier in the process.

Finally, beyond incorporating LCs to study innate immunity in the organoid model, it is worth emphasizing that innate immunity can be assessed without LCs, using endogenous keratinocyte pathways and markers [as discussed in **CHAPTER 3A**]. For example, keratinocytes express a variety of relevant biomarkers that could provide insight on the innate immune environment and whether changes due to viral activity could be immuno-evasive (*e.g*., innate sensing molecules such as TLRs, chemokines such as CCL27/28 and IL-18; pro-inflammatory cytokine expression such as TNF-α, TGF-β, and type I IFNs; cell-to-cell adhesion molecules, such as E-cadherin; and cell-surface receptors such as MHC I/II). An important consideration for future experiments, especially for next-generation sequencing, is that the abundance of these keratinocyte immune markers may be very low and require sensitive methods (*i.e*., deeper sequencing, as was the case in Jackson *et al*. [2016], where the average read depth of ~40 million/sample was not enough to detect desired keratinocyte markers via RNA-Seq).

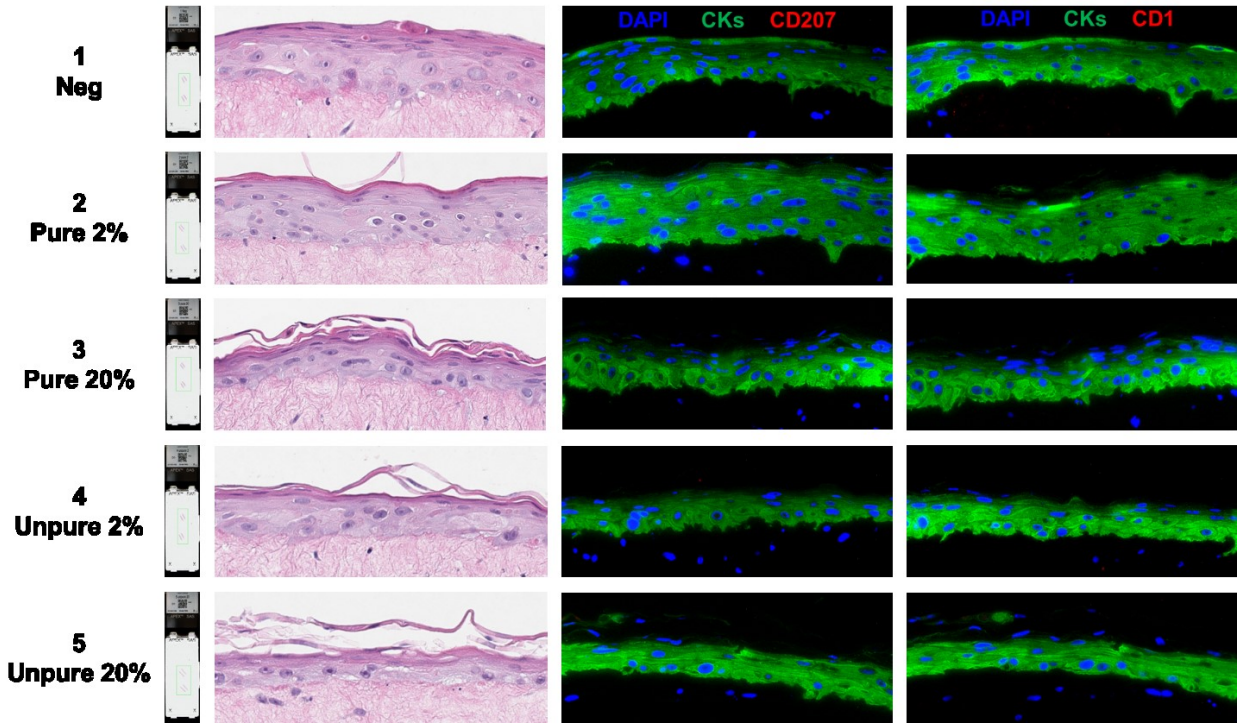


Figure 3C.4 – Immunocompetent epithelia trials. H&E and immunofluorescence micrographs (200X magnification) of five epithelial organoids with differential purity and ratios of Langerhans cells to keratinocytes. Blue (DAPI) staining represents nuclei, green (CKs) represents a pan-cytokeratin immunostain, and red (CD207 or CD1) represents Langerhans cell immunostains (none detected).

CHAPTER 4A – PATHOGEN-HOST ANALYSIS TOOL

This chapter was accepted for publication as an applications note in *Bioinformatics* on 12 Dec 2018 (DOI: 10.1093/bioinformatics/bty1003) [Gibb *et al*., 2019], having been originally pre-printed in *bioRxiv* on 18 Aug 2017 (DOI: 10.1101/178327). It has been included with permission for re-use within this dissertation as per the Oxford University Press (License Number 4511960249027). While the applications note included below is a brief contribution, the project as a whole is more expansive, including GitHub repositories for the software and its documentation.

Pathogen-Host Analysis Tool (PHAT): an integrative platform to analyze next-generation sequencing data

Christopher M. Gibb[1,2,†], **Robert Jackson**[1,3,†], Sabah Mohammed[2], Jinan Fiaidhi[2], Ingeborg Zehbe[1,4,5]

[1]Probe Development and Biomarker Exploration, Thunder Bay Regional Health Research Institute, Lakehead University, Thunder Bay, Ontario, Canada
[2]Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada
[3]Biotechnology Program, Lakehead University, Thunder Bay, Ontario, Canada
[4]Department of Biology, Lakehead University, Thunder Bay, Ontario, Canada
[5]Northern Ontario School of Medicine, Lakehead University, Thunder Bay, Ontario, Canada

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Keywords: Bioinformatics, Virology, Visualization, Software engineering, Programming, Next-generation sequencing

4A.1 – Abstract

4A.1.1 – Summary

The Pathogen-Host Analysis Tool (PHAT) is an application for processing and analyzing next-generation sequencing (NGS) data as it relates to relationships between pathogens and their hosts. Unlike custom scripts and tedious pipeline programming, PHAT provides an integrative platform encompassing raw and aligned sequence and reference file input, quality control (QC) reporting, alignment and variant calling, linear and circular alignment viewing, and graphical and tabular output. This novel tool aims to be user-friendly for life scientists studying diverse pathogen–host relationships.

4A.1.2 – Availability and implementation

The project is available on GitHub (https://github.com/chgibb/PHAT) and includes convenient installers, as well as portable and source versions, for both Windows and Linux (Debian and RedHat). Up-to-date documentation for PHAT, including user guides and development notes, can be found at https://chgibb.github.io/PHATDocs/. We encourage users and developers to provide feedback (error reporting, suggestions and comments).

4A.2 – Introduction

Analysis of pathogen data, especially of their genomes [Xiang *et al*., 2007] via high-throughput or next-generation sequencing (NGS), is an essential endeavour to understanding intricate pathogen–host relationships. While the ease of producing NGS data has grown significantly, bottlenecks still exist in its processing and analysis. In particular, short-read alignment algorithms and the tools that implement them have matured to the point that they no longer represent the major hurdle in the data analysis process [Li and Homer, 2010]. Instead, the availability of fast and user-friendly tools has become the limiting factor [Milne *et al*., 2010]. While there are excellent tools which perform one or several discrete functions in the same domain, *e.g.* Bowtie2 [Langmead and Salzberg, 2012] and SAMtools [Li *et al*., 2009], all-in-one type platforms can offer a breadth of features that help address barrier-to-entry (*i.e.* the ease in which users can setup and perform analyses). Integrative multi-tool platforms such as Comparative Genomics (CoGe) [Lyons and Freeling, 2008], VirBase [Li *et al*., 2015], Pathogen-Host Interaction Data Integration and Analysis System (PHIDIAS) [Xiang *et al*., 2007], Galaxy [Afgan

et al., 2016] and Unipro UGENE [Okonechnikov *et al*., 2012] exist, but they are often server or cloud-based. The infrastructure behind some of these projects, and their cloud-based nature, introduce roadblocks in the transfer of data to and from their servers [Li and Homer, 2010]. One solution to such a limitation is to establish an onsite computational cluster. However, technical and infrastructure requirements may pose further barrier-to-entry for data analysis.

We sought to develop the Pathogen-Host Analysis Tool (PHAT) to alleviate these issues by presenting an easy-to-setup and easy-to-use platform for life scientists conducting pathogen-host NGS analysis on common desktop computing hardware and operating systems (*e.g.* Windows).

4A.3 – Features

Pathogen-host NGS analysis typically begins with high-throughput sequencing output files: experimentally relevant nucleic acid read information. PHAT is a platform for analyzing these data, with a focus on pathogen sequences within NGS data [**Figure 4A.1**]. Reads are entered into PHAT as FASTQ files [Cock *et al*., 2010], comprised of sequence reads with per base nucleotide identities and quality scores, or pre-aligned SAM/BAM files [Li *et al*., 2009] generated via powerful cloud-based tools such as Galaxy [Afgan *et al*., 2016]. Quality control can be performed on individual files, with graphical reports generated. Reference genomes, recorded as FASTA files, must be indexed before they can be visualized or used for analysis. Once a pair of forward and reverse reads (paired FASTQ files) and a reference have been input, alignment can occur. PHAT also supports unpaired alignment and visualization of pre-aligned sequences.

The core functions of the PHAT platform as well as FASTQ quality control, sequence alignment, visualization, and its automated analyses are performed through well-known, established implementations. FastQC (https://www.bioinformatics.babraham. ac.uk/projects/fastqc) is used for quality control scoring. Sequence alignment is done by Bowtie2 [Langmead and Salzberg, 2012] or HISAT2 [Kim *et al*., 2015], while linear alignment visualization is via pileup.js [Vanderkam *et al*., 2016]. Circular genomes are viewed with our enhancements to AngularPlasmid (http://angularplasmid. vixis.com/) which we make available as a new project called ngPlasmid (https://github.com/chgibb/ngPlasmid). Automated variant calling of single-nucleotide polymorphisms (SNPs) is by VarScan2 [Koboldt *et al*., 2012].

The graphical user interface, based on GitHub's Electron project (https://electronjs.org/docs), operates in a client-server-based architecture. Each window acts as a client, communicating with a background server process. The server manages the saving and propagation of workspace data, as well as the generation of additional processes such as sequence alignment and quality control. This mechanism allows processes to act as threads, allowing the flow of data to and from the application window that invoked it and the created process itself. On systems with limited power, the server process limits the number of concurrently running processes and the amount of data propagated between windows to reduce memory and central processing unit (CPU) usage. We utilize an internal pipeline, spawning new processes as others end, passing data from one application window to another (*e.g.* alignment output). The server process, as well as the application windows themselves are implemented in Typescript. These windows can be conveniently undocked from the main toolbar.

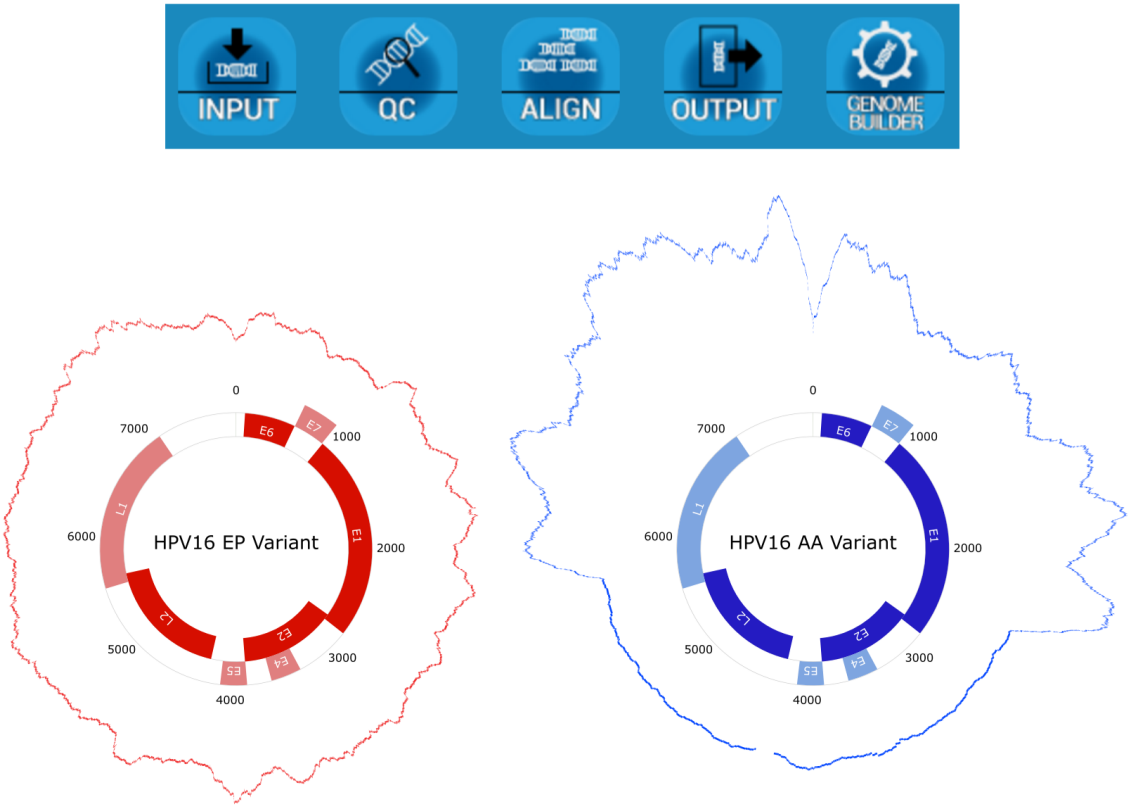


Figure 4A.1 – Pathogen-Host Analysis Tool (PHAT) and visualization. The toolbar allows input of pathogen-containing NGS data and reference sequences, quality control with FastQC, alignment with Bowtie2 or HISAT2, SNP detection with VarScan2, visualization with pileup.js and the genome builder, as well as tabular output. Example HPV16 genome maps and coverage plots were generated to contrast viral variants: the European Prototype (EP) variant is episomal, whereas the Asian-American (AA) variant's coverage is disrupted by integration into host DNA [Jackson *et al*., 2016].

4A.4 – Future Work

With the development of PHAT, we aim to bring simple-to-use cross-platform NGS analysis to off-the-shelf hardware for life scientists studying pathogen–host relationships. In our own lab, we study human papillomavirus type 16 (HPV16) variants and their tumourigenicity in epithelia using NGS [Jackson *et al*., 2016], but PHAT can be applied to a wide variety of pathogen–host relationships (*e.g.* genotyping of microbes such as viruses, bacteria, fungi and protozoans) from host NGS samples. To aid in our own experimental work, including analysis of HPV sequences within curated datasets (*e.g.* The Cancer Genome Atlas, TCGA), we are currently testing a viral-host integration detection feature in PHAT, with linkage to sequence databases. Additional features could include advanced alignment options as well as tools for further exploring pathogen-host interactions. We plan to actively develop, update and support PHAT based on user feedback and needs, with auto-updating features already included, in anticipation of building an active user and developer community.

4A.5 – Declarations

4A.5.1 – Acknowledgements

The need for PHAT was conceived by RJ and IZ. Interface was designed by CMG and RJ, with programming by CMG. Manuscript writing was carried out by CMG, RJ, SM, JF and IZ. Intellectual property considerations made by SM and JF. Thanks to Zehbe Lab members for user testing, Dr. M. Togtema for feedback, as well as students M. Pynn, J. Braun, S. Liu, Z. Moorman and N. Catanzaro for improvements. We are thankful to the developers of open-source tools that are used in PHAT as well as GitHub and Reddit communities for helpful discussions.

4A.5.2 – Funding

This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) grants to IZ (#RGPIN-2015-03855) and RJ (CGS-D #454402-2014). The funding body had no role in study design, data collection, data analysis and interpretation, or preparation of the manuscript. *Conflict of Interest:* none declared.

CHAPTER 4B – HPV16 SUB-LINEAGE AND INTEGRATION IN BIG DATA

This chapter contains an unpublished manuscript in preparation, with additional analyses in progress, for pre-print submission to *bioRxiv* (with the intention to submit for future peer-review). The purpose of this ongoing bioinformatics study is to further test our hypothesis, using large datasets of cancer genome data (along with our experimental work), that the HPV16 D2/D3 sub-lineage (AA variants) increase the chance (or pattern) of host genome integration compared to the A1 sub-lineage (EP variant).

Human papillomavirus type 16 sub-lineages and their host genome integration capability

Robert Jackson[1,2], Dallas Nygard[3], Christopher M. Gibb[2], Ingeborg Zehbe[2,4,5]

[1]Biotechnology Program, Lakehead University, Thunder Bay, Ontario, Canada
[2]Probe Development and Biomarker Exploration, Thunder Bay Regional Health Research Institute, Lakehead University, Thunder Bay, Ontario, Canada
[3]Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, Canada
[4]Department of Biology, Lakehead University, Thunder Bay, Ontario, Canada
[5]Northern Ontario School of Medicine, Lakehead University, Thunder Bay, Ontario, Canada

Keywords: The Cancer Genome Atlas (TCGA), Human papillomavirus type 16 (HPV16), Sub-lineage, Variants, Integration, Cervical cancer, Head and neck cancer

4B.1 – Abstract

Our lab has been intrigued by the fact that viruses often take on the role of mobile elements to perpetuate their existence in a complex organism's genome. Multiple DNA viruses such as Epstein-Barr virus, hepatitis B virus, and human papillomavirus (HPV) can invade the human genome, as "genomic parasites". Here we investigate the HPV family which is a widespread group of tumour viruses in humans. In our recent *in vitro* work using 3D organoids, a common variant of HPV16's coding region elicited early integration into the host genome. Next-generation sequencing (NGS) data confirmed a phenotype of active proliferation and chromosomal

instability—both hallmarks of cancer. Epidemiologically, this variant is associated with a high cervical cancer incidence. To substantiate our *in vitro* data and confirm viral variant-specific integration patterns we used bioinformatics analyses of NGS data from population-derived clinical samples in The Cancer Genome Atlas (TCGA) curated database. We are particularly interested to investigate whether the HPV integration mechanism is sequence-specific (*e.g*. due to microhomologies) or functionally-related (*e.g*. chromosomal instability, hypomethylation). We will align obtained integration data with previously identified integration patterns, conclude with an evolutionary perspective on integration mechanisms of mobile elements and highlight the clinical utility in developing prognostic personalized biomarkers for cancer treatment efficiency, *e.g*. to detect residual disease.

4B.2 – Introduction

The co-evolutionary interplay between pathogens and their hosts involves a plethora of intriguing biological phenomena, including genomic integration of pathogen DNA into host cells. While cellular integration of viral genomic material is a means of propagation for families such as *Retroviridae* (*e.g*., human immunodeficiency virus, HIV), it can also occur during persistent infections with DNA viruses and is often associated with virally-induced cancers. One such family of common DNA tumour viruses, *Papillomaviridae*, is responsible for causing ~5% of all human cancers worldwide [Ghittoni *et al*., 2015]. There are 500+ types of papillomaviruses (PVs) discovered to date, including the 83 recently added in Dec 2018 [Pastrana *et al*., 2018], bringing the total to 513 as of Feb 2019 based on the PaVE reference genome database [PaVE: https://pave.niaid.nih.gov, Van Doorslaer *et al*., 2013; 2017b]. The number of non-human PVs (currently 183) is also expected to rise using high-throughput genomics [Van Doorslaer and Dillner, 2019]. However, most PVs are identified as human papillomaviruses (HPVs) [Van Doorslaer *et al*., 2017a] and these types (330 to date) differ in their propensity for inducing cancer as well as their epithelial tropism, with "high-risk" types being the primary cause: predominantly via sexually-transmitted persistent HPV type 16 (HPV16) infection of anogenital and oropharyngeal mucosa [Ndiaye *et al*., 2014]. Although HPV16's genome is small, only ~7.9 kb containing nine open reading frames (ORFs) and a variety of alternatively spliced and multicistronic variant transcripts [Graham & Faizo, 2017], it encodes potent oncoproteins such as E6 and E7 which drive host cells to acquire the hallmarks of cancer [Hanahan and Weinberg, 2000;

2011; Mesri *et al*., 2014]. Their combined actions disable protective apoptotic mechanisms and promotes host cell proliferation as outlined by Hoppe-Seyler and colleagues [2018]; a phenomenon that could potentially be exploited therapeutically at an RNA [Togtema *et al*., 2018b] or protein-level [Togtema *et al*., 2018a]. This augmented environment can permit chromosomal instability and eventual integration of HPV16 sequences into human DNA via double-strand break (DSB) repair events [Winder *et al*., 2007], yielding unique integration signatures found in next-generation sequencing (NGS) data from human carcinomas [Holmes *et al*., 2016].

HPV16 sub-lineages (variants) via epidemiological and lab-based studies have been reported to infer differing cancer risk [Mirabello *et al*., 2018 and references within; Clifford *et al*., 2019]. HPV16 variant designations were originally based on their geographical region of origin. The European "prototype" sub-lineage (EP; A1 according to new nomenclature) was the first HPV16 genome published [Seedorf *et al*., 1985] followed by two additional European (E; A2 and A3 according to new nomenclature), one Asian (As; A4 according to new nomenclature), eight African (Af-1a, -1b and -2; B1-4 and C1-4 according to new nomenclature), two North-American (NA; D1 and D4 according to new nomenclature) and two Asian-American (AA-1 and -2; D3 and D2, respectively, according to new nomenclature) [Burk *et al*., 2013; Mirabello *et al*., 2018]. Recently, we have put our efforts on the variants EP and AA, which differ in only three amino acid changes at residues 14 (Q to H), 78 (H to Y) and 83 (L to V) in the major transforming protein E6 [Richard *et al*., 2010; Niccoli *et al*., 2012; Jackson *et al*., 2014; 2016; Togtema *et al*., 2015; Cuninghame *et al*., 2017]. Epidemiological studies revealed that the AA sub-lineage infers a higher risk factor for dysplasia and an earlier onset of invasive tumours than EP [Xi *et al*., 1997; Berumen *et al*., 2001]. Our functional assays showed that AAE6 has a greater transforming, migratory, and invasive potential than EPE6 when the respective E6 gene was retrovirally transduced into primary human keratinocytes in recent long-term *in vitro* immortalization studies [Richard *et al*., 2010; Niccoli *et al*., 2012; Togtema *et al*., 2015]. These observations may be due, at least in part to an AAE6-mediated altered, metabolic phenotype reminiscent of the Warburg effect [Richard *et al*., 2010; Cuninghame *et al*., 2017]. Consistent with the hypothesis that AAE6 is more oncogenic than EPE6, our recent experimental work revealed that AAE6 in the context of the full-length viral genome is more prone to host genome integration than is EPE6 [Jackson *et al*., 2014; 2016]. Another study has mentioned HPV16 sub-lineage and integration, but did not have sufficient sample size to address this with respect to the AA sub-lineage (from Jackson *et al*., 2016): "A

previous study of HPV16 integration propensity with respect to the variants did not demonstrate a statistically significant difference (*P*-value = 0.28, two-tailed Fisher's exact test) between EPE6 (3 episomal and 20 integrated cases) and the E-T350G variant (6 episomal and 16 integrated cases, responsible for one of the residue changes also found in AAE6: L83V) [Xu *et al*., 2013]. Only one tumour sample in their set contained the AA variant, therefore precluding a formal analysis of its propensity to integrate, but notably it was in integrated form".

To re-conciliate our lab-based finding on integration patterns, we set out to do a search of clinical samples with full genome sequence analyses reported on available data bases. Two landmark publications of The Cancer Genome Atlas (TCGA) reported on cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) [TCGA Research Network, 2017] and head and neck squamous cell carcinoma (HNSC) [TCGA Research Network, 2015]. More recently, data reported by Cantalupo *et al*. [2018] became available and is of high-interest for our study. Prior to the primary TCGA data analysis papers, there was a publication about DNA viruses in TCGA RNA-Seq data across a variety of human cancers [Khoury *et al*., 2013]. Additional literature search found several more related papers: HPV and HNSC comprehensive analysis [Castellsagué *et al*., 2016] and an HPV16-focused follow-up TCGA-HNSC analysis [Nulton *et al*., 2017]. Here, they report three different groupings of integration status: episomal, integrated, and human-viral episomal hybrids, challenging integration data by Parfenov *et al*. [2014]. The Parfenov *et al*. study disclosed HPV typing and sub-lineage analyses of the 35 HPV+ HNSC cases with 29 containing HPV16 of which just one belonging to the AA variant. Interestingly, integration with 16 breakpoints in the human genome (*i.e.* ~ 4-fold more than in EP) was reported. In contrast, in the Nulton *et al*. study, for the same case, it was concluded that the human-viral hybrids after initial integration were in fact episomal again [Nulton *et al*., 2017]. Recently, Zapatka *et al*. [2018] further analyzed TCGA data to determine viral associations with cancer, including the impact of integrations on host copy number variations.

Our objective is to data mine TCGA datasets for all HPV16-containing samples, verify their sub-lineage genotypes, then characterize their integration status. The most pertinent research question for us is whether there are qualitative and quantitative relationships between HPV16 sub-lineage genotypes and integration status. We hypothesize that HPV16 sub-lineages D2 and D3 (AA variants) would be more strongly associated with integration than A1 (EP variant), either via proportion of episomal vs integrated samples, or more complex characteristics such as total

number of integrations (breakpoints) or unique patterns with potential functional relevance. Such a finding would confirm our variant-specific integration hypothesis [Jackson *et al*., 2014; 2016], as well as provide a novel mechanism for differing variant tumourigenesis in host tissue. Our approach of HPV16 sub-lineage genotyping to detect differential integration patterns correlated to A1 and D2/D3 sub-lineages, in a genome-wide analysis, has not been previously addressed.

4B.3 – Methodology

4B.3.1 – Data acquisition and storage

Access to controlled sample data from TCGA, using the database of Genotypes and Phenotypes (dbGaP), was acquired via application to the electronic Research Administration (eRA) commons. We registered our host institution, Lakehead University, as a new organization and received a Data Universal Numbering System (DUNS) identifier for access. Our proposal title was "Characterization of human papillomavirus type 16 integration sites in cervical and head and neck tumour biopsies", with approval allowing access to TCGA Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (TCGA-CESC) as well as Head and Neck Squamous Cell Carcinoma (TCGA-HNSC) files for analysis. As the focus of this study was on HPV16, only files for TCGA cases that had been previously been marked as HPV16 positive [TCGA Research Network, 2015; 2017; Cantalupo *et al*., 2018] were downloaded. After receiving access to the data (which was a lengthy process, with steps outlined in **Figure 4B.1**), the files were stored in an encrypted fashion on a local file server and analyzed on Compute Canada's Graham cluster. Additional work with the data was performed with an in-house server and laptops, all password protected and stored securely behind multiple levels of locked doors (as physical barriers). See **Figure 4B.2** for a graphical summary of our analytical workflow. Additional data may be required to further address our research question, such as corresponding RNA-Seq data to compare to DNA-Seq (WGS and WXS) data used in our primary analysis, as well as external data from the Sequence Read Archive (SRA).

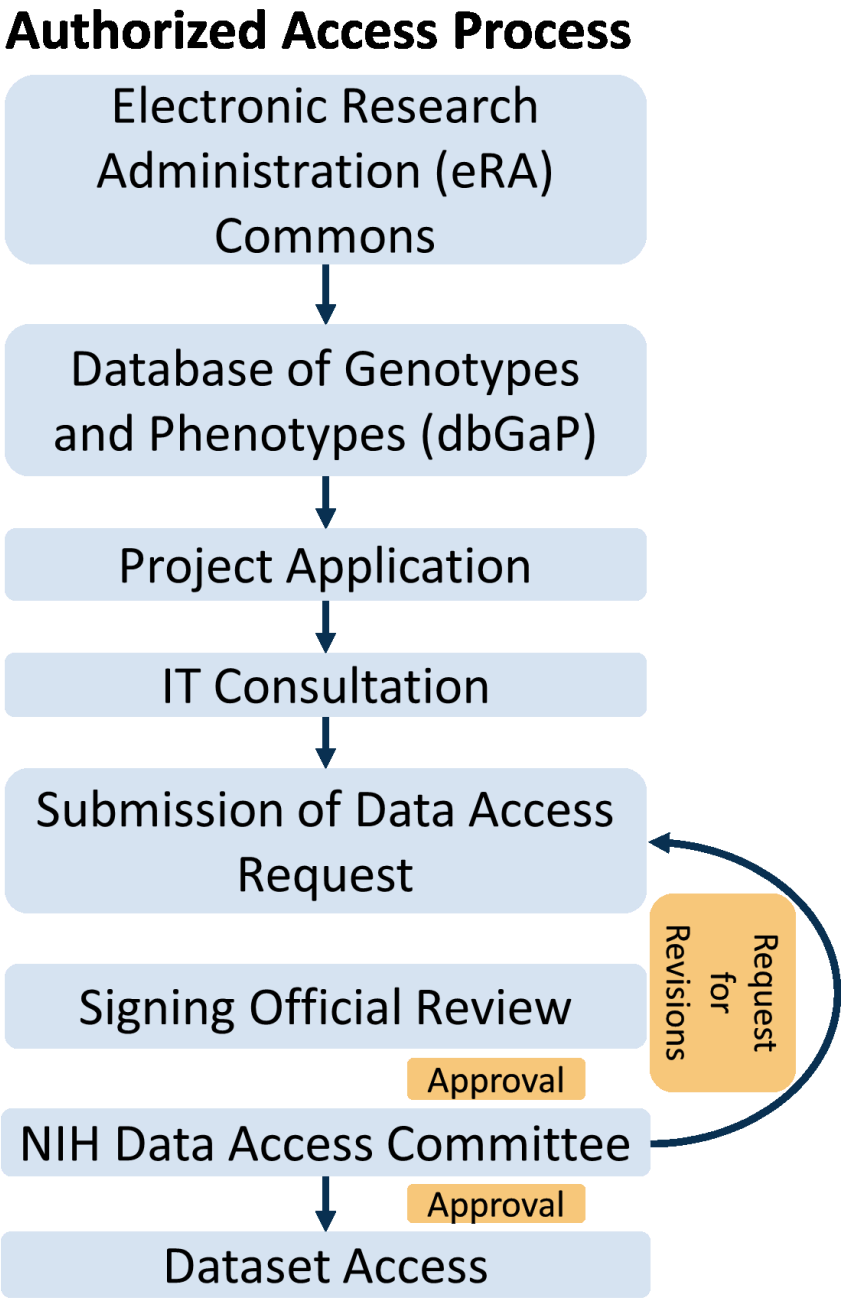


Figure 4B.1 – Authorized access process for TCGA data. Access to controlled sample genomic data from TCGA required a multi-step authorization process before we could access relevant datasets.

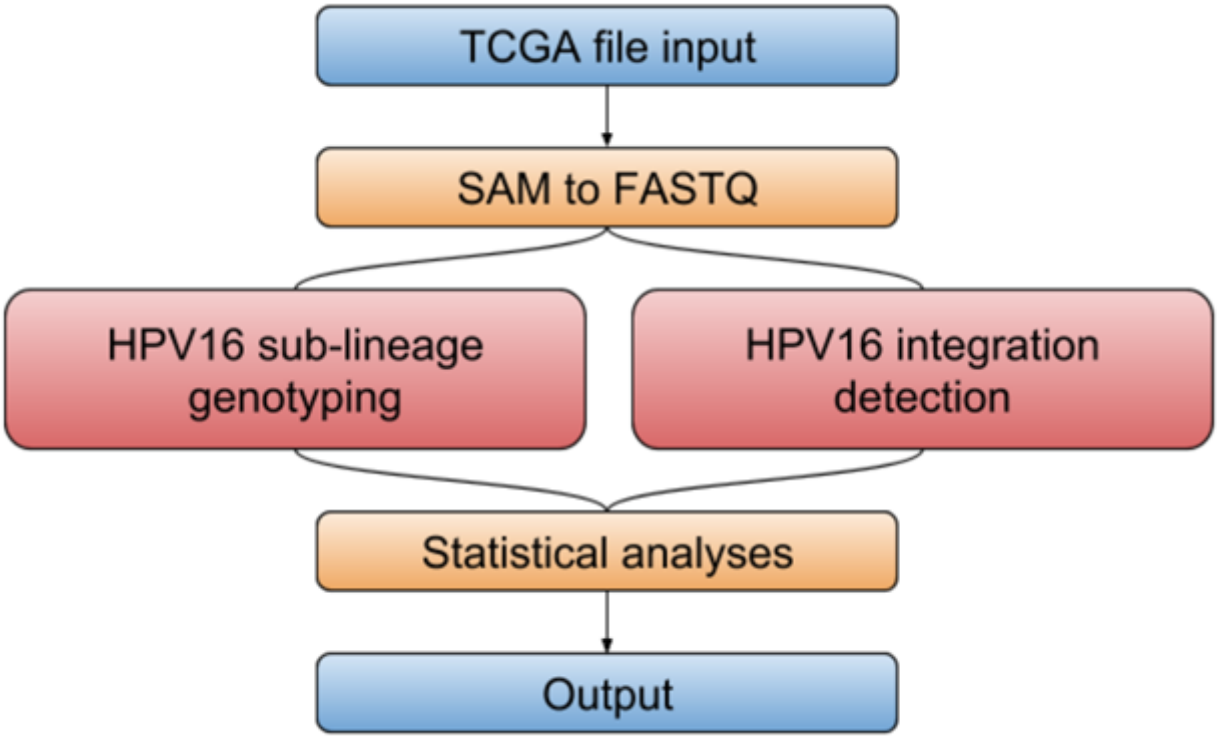


Figure 4B.2 – Summary of the analytical workflow. First, TCGA high-throughout sequence read data (WGS, WXS, or RNA-Seq) is accessed and converted from pre-aligned BAM/SAM format to FASTQ read format. This data is then subjected to our HPV16 sub-lineage genotyping analysis as well as our HPV16 integration detection analysis, followed by statistical analyses, and output as visualizations.

4B.3.2 – HPV16 sub-lineage genotyping analysis

The samples chosen for this analysis were those identified as HPV16 positive in the landmark TCGA Research Network studies [2015; 2017] as well as Cantalupo *et al*. [2018]. While HPV16-positive cervical cancer (TCGA-CESC) samples had associated sub-lineages already identified [TCGA Research Network, 2017], sub-lineage genotyping was not immediately available for the head and neck (TCGA-HNSC) cases [TCGA Research Network, 2015]. It was thus necessary to develop a pipeline for ascertaining the sub-lineages of unknown identified samples (*i.e*., those that were identified as HPV16 positive by Cantalupo *et al*., [2018] as well). Our pipeline utilized the raw Binary Alignment Map (BAM) files from TCGA. First, Picard's SamToFastq function [http://broadinstitute.github.io/picard/] converted the BAM file to a format that could be utilized by HISAT2 [Kim *et al*., 2015], which then aligned the reads to our HPV16 A1 reference (GenBank # K02718). Once aligned by HISAT2, SAMtools [Li *et al*., 2009] was used to sort and prepare the file for conversion to mpileup format, which was accomplished using SAMtools' mpileup command. VarScan [Koboldt *et al*., 2012] was used to extract single nucleotide polymorphism (SNP) and indel (insertion and deletion of bases) information from the mpileup file, which was then run through a custom R script [*4B.8.1 – Supplementary Script 1*]. The R script altered the reference A1 sequence with the detected SNPs and indels to create a new sequence that resembled the true integrated HPV sequence. The final step was to perform a multiple sequence alignment (MSA) on the new sequence and references for each sub-lineage of HPV16 (A1, A2, A3, A4, B1, B2, B3, B4, C1, C2, C3, C4, D1, D2, D3, D4) using ClustalW [McWilliam *et al*., 2013]. The reference sequences used for each sub-lineage are listed in **Table 4B.1**. From Clustal's Newick tree output, sub-lineages were confirmed visually based on their proximity to the sub-lineage references on the phylogenetic tree (example reference tree visualized in **Figure 4B.3**). We are further improving our sub-lineage genotyping pipeline, especially for difficult to analyze low-coverage samples (and samples that may contain multiple types of HPV), by testing alternative tools such as rkmh [Dawson *et al*., 2018] and RAxML [Stamatakis, 2014]. As well, genotyping could be performed following integration analysis to further verify that integrated viral reads belong to their particular sub-lineages (based on representative SNPs and appropriate coverage).

Table 4B.1 – HPV16 sub-lineage reference genomes used for genotyping. Derived from reference data summarized in Burk *et al*. [2013], Chen *et al*. [2018a], and Mirabello *et al*. [2018].

HPV16 sub-lineage	GenBank #	Alternate ID	Alternative name
A1	K02718	NC_001526.4 (latest)	European Prototype (EP)
A2	AF536179	w0122	European (E)
A3	HQ644236	AS411	European (E)
A4	AF534061	w0724	Asian (As)
B1	AF536180	w0236	African-1a (Af-1a)
B2	KU053915	IARC1100085NI	African-1 (Af-1)
B3	HQ644298	Z109	African-1b (Af-1b)
B4	KU053914	IARC907912AL	African-1 (Af-1)
C1	AF472509	R460	African-2 (Af-2)
C2	HQ644244	IARC240211GU	African-2 (Af-2)
C3	KU053920	IARC040105BE	African-2 (Af-2)
C4	KU053925	IARC304612MO	African-2 (Af-2)
D1	HQ644257	Qv00512	North-American-1 (NA1)
D2	AY686579	Qv15321	Asian-American (AA) 2
D3	AF402678	Qv00995	Asian-American (AA) 1
D4	KU053931	IARC903812AL	North-American-2 (NA2)

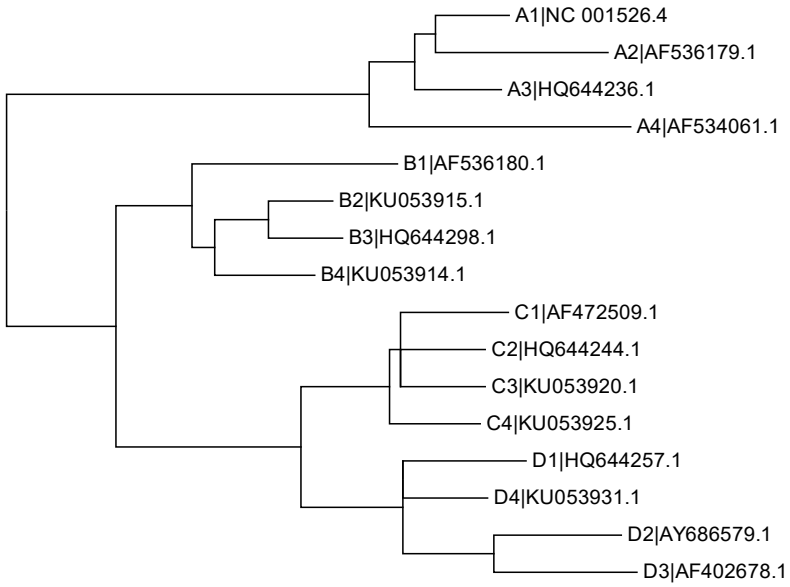


Figure 4B.3 – HPV16 sub-lineage reference tree for phylogenetic analyses. Branches correspond to four lineages (A, B, C, and D), each containing four sub-lineage genomes (1 to 4).

4B.3.3 – Viral-human integration detection and characterization

Integration analysis was performed using ViFi [Nguyen *et al*., 2018] with default settings and the included genome references for human (Hg19) and HPV16 A1 (GenBank # K02718). ViFi was able to determine viral breakpoints and the number of chimeric reads found in each sample. It was difficult to confirm integration status from the ViFi output alone, so to obtain a better visualization of the data, our developed Pathogen-Host Analysis Tool (PHAT) [Gibb *et al*., 2019] and a custom R script [*4B.8.2 – Supplementary Script 2*] were used to create circular visualizations of the viral genome and histograms of chimeric reads respectively. Further improvements to our integration detection and characterization pipelines are in progress. Full characterization of the genomic landscape surrounding detected HPV16 integration sites within human chromosomes will be required to comprehensively address our research question, including viral and host nucleotide positions, sequence overlap/microhomologies, functional annotations of viral and host features, including gene proximity (inside/outside, exon/intron), promoters, enhancers, transcription factor binding sites, repeat elements, and proximity to fragile sites. "Virtual ChIP-seq" [Karimzadeh and Hoffman, 2018] could be a useful tool to predict nearby transcription factor binding sites (*e.g*., CTCF [Doolittle-Hall *et al*., 2015; Paris *et al*., 2015]). As described previously [Jackson *et al*., 2016], the region surrounding a viral integration site can be scanned for repeat elements, which may be prone to rearrangements, using the UCSC human genome browser RepeatMasker track. The current version of the human genome is GRCh38/hg38 (hg19 was GRCh37, and this was iterated to "hg38" rather than hg20 to be on par with the GRCh nomenclature, https://genome.ucsc.edu/cgi-bin/hgGateway). Another important aspect will be to visualize and compare the A1 vs D2/D3 cases; examples exist in the literature [Akagi *et al*., 2014; Tang *et al*., 2013; Holmes *et al*., 2016; Jackson *et al*., 2016; Warburton *et al*., 2018; Lagström *et al*., 2019].

4B.3.4 – Statistical analyses

Statistical analyses will be performed using the open-source programming language R, version 3.5.0 [R Core Team, 2018], within the integrated development environment RStudio, version 1.1.453 [RStudio Team, 2015]. Inferences on count data, such as with X^2 or Fisher's exact tests, will use an *a priori* significance level of $P < 0.05$. Beyond these basic statistical analyses, future work could explore using deep/machine learning (artificial intelligence-based approaches) to aid biological analysis as well as create predictive models [Ching *et al*., 2018].

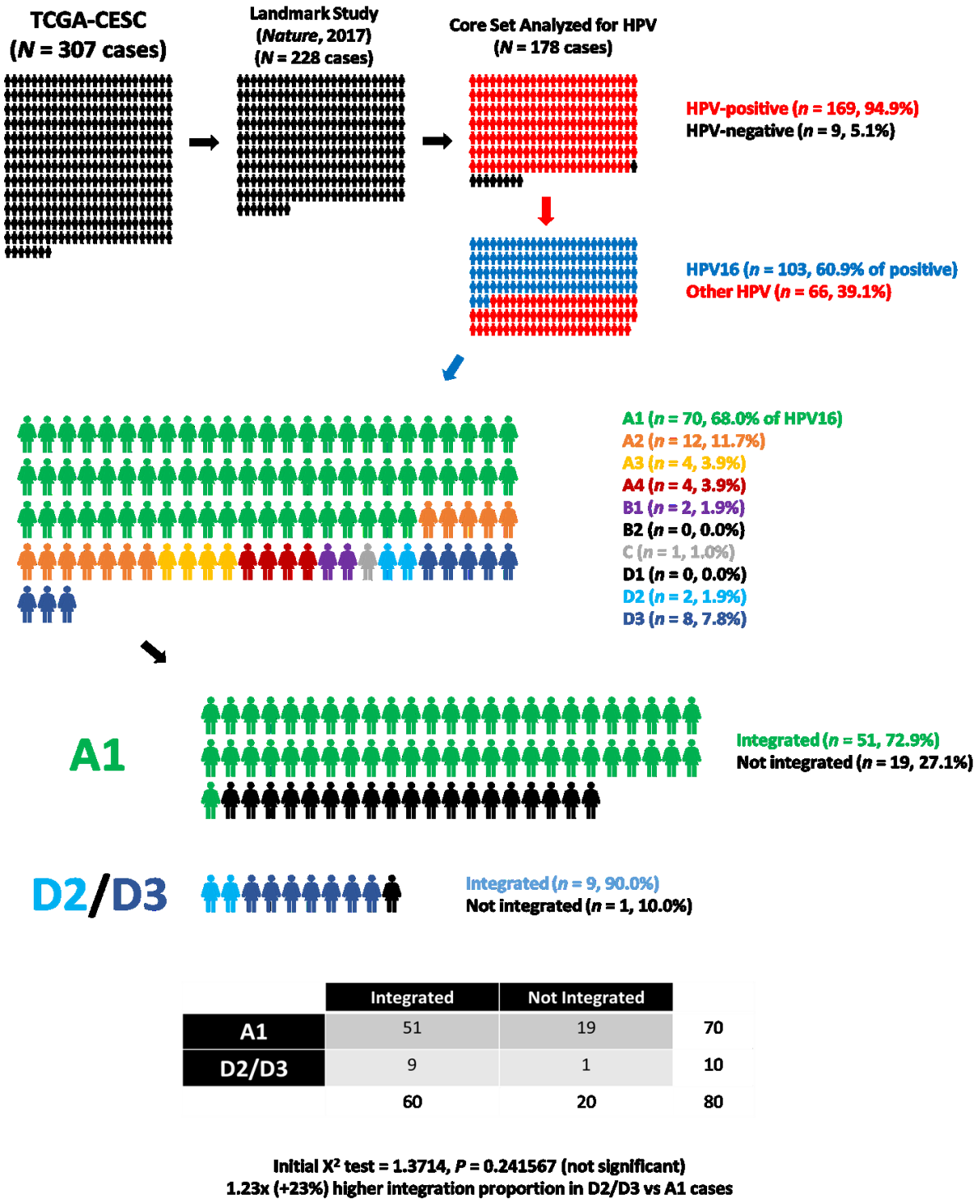
4B.4 – Preliminary Results

4B.4.1 – HPV16 sub-lineages in cervical and head and neck cancer datasets

The number of CESC and HNSC samples that were previously marked as being positive for HPV16 differs between studies, likely due to differences in analysis pipelines (*e.g.*, data types analyzed, thresholding, and filtering for contamination) [TCGA Research Network, 2017; Cantalupo *et al*., 2018] and poses the major challenge to a valid analysis of sub-lineage association with integration. A preliminary sample summary is reported in **Table 4B.2** along with a preliminary sub-lineage and integration analysis for available CESC sample data in **Figure 4B.4**. Overall, while the proportion of integrated cases was 1.23x higher in D2/D3-containing samples, the sample size was very low and this finding was not statistically significant ($P = 0.242$, X^2 test). To expand upon these preliminary results, we will use our own HPV16 genotyping and sub-lineage analyses to verify sub-lineage calls by prior studies, resolve discrepancies, and further analyze the "unknown" samples. Based on these prior studies, it is expected that the majority of CESC samples are HPV positive, whereas a lower proportion of HNSC samples are HPV positive. While HPV16 is expected to be the predominant type, we must also consider multi-type infections as well as co-infections with other pathogens (such as human herpesviruses). As well, cross-contamination between samples should be assessed, especially for adjacent cases where sequence libraries could be contaminated. From the GDC Data Portal, there are currently 835 relevant CESC and HNSC cases for analysis, including 2,568 BAM files (27.84 TB total, average of 33.34 GB/case, including all WXS and RNA-Seq sequencing read files and excluding miRNA-Seq files as they are not primarily relevant for HPV16 genotyping).There are 307 CESC cases (924 BAM files, 10.44 TB) and 528 HNSC cases (1,644 BAM files, 17.41 TB), with both sets including samples from primary tumours, metastases, normal blood, and normal solid tissues. Beyond these GDC Data Portal cases and files, there appears to be additional files available via the legacy data portal (which includes WGS files). Our analysis could also further extend to additional HPV16-positive cancers, such as the bladder cancers samples identified by Cantalupo *et al*. [2018]. As well, additional sample data from non-TCGA datasets could be included.

Table 4B.2 – Sample summary breakdown for each TCGA case set. Sample genotypes for cancer cases from the Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC) as well as Head and Neck Squamous Cell Carcinoma (HNSC) datasets. CESC HPV16 data is based on TCGA Research Network [2017] (using the sub-lineages identified at the time, which were only 10 of the current 16) and HNSC HPV16 data is based on Parfenov *et al*. [2014], where they performed lineage genotyping (resulting in some sub-lineages being grouped, such as A1-A3, as "EUR" genotypes).

Sample genotype	**CESC (*n* = 178)**	**HNSC (*n* = 279)**
Total HPV16 cases	103 (57.9% of CESC cases)	29 (10.4% of HNSC cases)
A1	70 (68.0% of HPV16 cases)	
A2	12 (11.7%)	22 (75.9% of HPV16 cases)
A3	4 (3.9%)	
A4	4 (3.9%)	2 (6.9%)
B1	2 (1.9%)	3 (10.3%)
B2	-	-
C	1 (1.0%)	1 (3.4%)
D1	-	-
D2	2 (1.9%)	-
D3	8 (7.8%)	1 (3.4%)



	Integrated	Not Integrated	
A1	51	19	70
D2/D3	9	1	10
	60	20	80

Initial X^2 test = 1.3714, P = 0.241567 (not significant)
1.23x (+23%) higher integration proportion in D2/D3 vs A1 cases

Figure 4B.4 – Preliminary analysis of CESC samples. HPV16 sub-lineage (10 of the 16 considered in TCGA Research Network [2017]) and integration data was analyzed from the CESC dataset. While the proportion of integrated samples was 1.23x higher in D2/D3, this finding was not statistically significant (P = 0.242, X^2 test).

4B.4.2 – Integration histograms reveal different patterns of integration

Histograms were created using the integration data for all cases [example shown in **Figure 4B.5**]. These histograms provide insight into the integration pattern across chromosomes and revealed cases that had clear integration points and others that were more convoluted. Additional characterization, both qualitative and quantitative, are required to clarify these patterns and if they differ between sub-lineages. To gain additional insight we scanned prior integration data from three previous studies, and while these should be mainly assessed in a "qualitative" manner, due to sample size restriction, they do not provide evidence against our state hypothesis. The fusion transcript for the single AA case in a previous study was revealed to be with ERBB4 near the fragile site (FRA) 21 (2q33) with 3Mb distance from the break point [Schmitz *et al*., 2012]. This study used an in-house DNA-Seq approach with DNA fragmentation and adapter targeting alongside PCR enrichment of HPV16's early region and Illumina NGS. Fusion transcripts were then confirmed with the APOT assay. The Holmes *et al*. [2016] study used the Capture-NGS approach and describes five integration patterns vs the four patterns in the former study. In the Holmes *et al*., study, they found 33 different SNPs in their pooled sample. Based on these signatures 14 SNPs are identical with AA. One SNP was only found in cases with episomal HPV but this SNP is not found in the AA. There was no evidence of any of the detected AA SNPs being associated with episomal HPV DNA. In fact, one of 33 SNPs was found in cases with predominantly episomal HPV (>50%) while 14/33 SNPs are found in cases with <50% episomal HPV suggesting that in the Holmes study, there is no evidence to disprove our hypothesis. The only HPV16 AA case in the 30 HPV+ head and neck cancers of the TCGA data showed 16 break points and was considered integrated in a study by Parfenov *et al*. but was believed to be episomal albeit after initial integration events as shown by Nulton *et al*. Numerous questions arise from these data, such as what would happen to these hybrid HPV DNA "plasmids" containing human sequences? Are they replication-competent? How do they impact on HPV evolution? How can they form dimers and trimers? This represents a kind of reverse "hijacking" scenario, where HPV can pickup pieces of host DNA and incorporate it into its episome. Beside the classical interrupted E2 HPV gene, other genes have now also been reported to be interrupted, *e.g.* E1, E5 and L1 [Akagi *et al*., 2013; Warburton *et al*., 2018]. These findings have implications in how to define HPV integration into the human genome, and while breakpoints could possibly arise at any location throughout the HPV genomes, particular regions could be more advantageous to viral

expression (and ultimately, host cell proliferation, and clonal expansion). Ideally, to classify HPV integration events, we will have to analyze both DNA and RNA-Seq data to clarify functional effects of DNA integration.

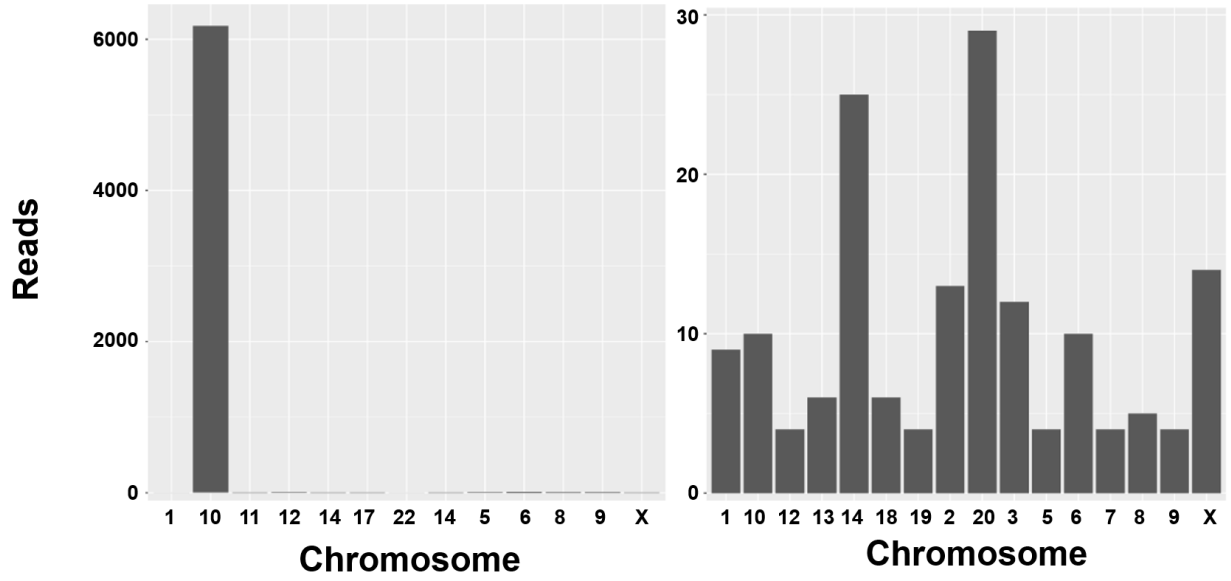


Figure 4B.5 – Integration histograms. On the left, a clear case of viral integration, for TCGA-IR-A3LK (A1 sub-lineage), into human chromosome 10 (supported by > 6000 reads). On the right, a "convoluted" integration status for TCGA-IR-A3LL (A2 sub-lineage), into multiple chromosomes (158 reads total), with a similar number of reads supporting different chromosomal integrations. These could represent true breakpoints or be due to non-specific read alignments.

4B.4.3 – Circular visualizations reveal sequence differences between sub-lineages

Using the same set of HPV16-positive samples from the TCGA-CESC dataset, circular visualizations were created using PHAT [**Figure 4B.6**]. A pattern of peaks and valleys is seen in each visualization, representing genomic coverage and viral reads corresponding to regions of the viral genome. Some valleys were present in all visualizations (at *ori*, near 4,200 bp and near 6,900 bp), but two valleys in particular (near 2,300 bp and 5,350 bp) were only found in the D3 cases. As expected, these valleys were very close to the breakpoints in simple integration cases from our ViFi analysis. A valley is representative of few reads mapping to the reference sequence in an area and typically could be associated missing regions due to integration events. However, the valleys that were present in all visualizations, and not unique to either A1 or D2/D3 cases, can be attributed

to technical artifacts: the *ori* valley is due to sequence alignment being performed on a linear sequence which starts and ends at this location, and when circularized is seen as an artificial gap in coverage. Circular visualizations and integration histograms will be used to make conclusions regarding integration for each sample.

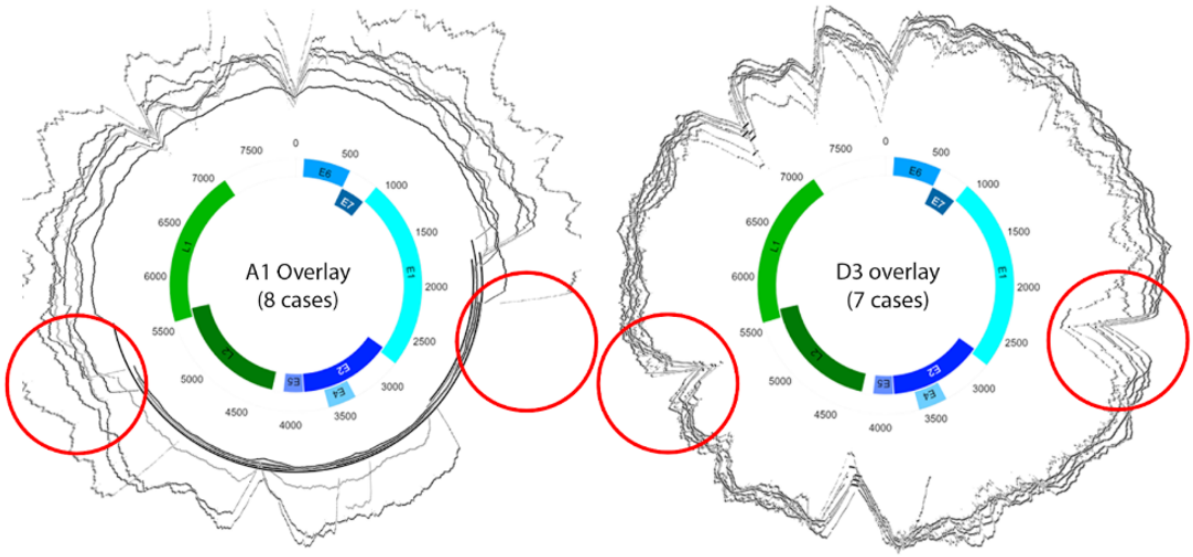


Figure 4B.6 – Circular genome visualizations of HPV16. A1 vs D3 coverage plots were created using PHAT [Gibb *et al*., 2019] and then overlaid to emphasize the consistency of the pattern valleys. D3 cases, when overlaid, showed consistent valleys (circled in red) at the 2100 – 2400 bp range and the 5300 – 5500 bp range.

4B.4.4 – TCGA data used for functional analysis other than integration

Eckardt *et al*. [2018] mapped a global network of virus-host protein interactions by purification of the complete set of HPV proteins in multiple cell lines followed by mass spectrometry analysis. Their goal was to determine which HPV-human interactions were most directly involved in cancer. They described an integrative strategy combining the complete HPV-human interactome with the genomic mutational landscape of tumours. Genomic data were obtained from 177 CESCs and endocervical adenocarcinomas and 505 HNSCs from the TCGA along with 118 HNSC samples from the University of Chicago (http://gdac.broadinstitute.org) [Seiwert *et al*., 2015], yielding a combined cohort of 295 HPV-positive and 505 HPV-negative

tumour samples. The genes impacted by either single nucleotide alterations or copy number alterations in each sample were determined based on subtractive analysis of paired tumour and normal whole-exome sequences or SNP arrays, respectively. This combined proteomic and genetic approach provided a systematic means to study the cellular mechanisms hijacked by virally induced cancers. Like Eckardt *et al*. [2018], the Seiwert *et al*. [2015] study also used the HNSC samples from the University of Chicago showing that the mutational makeup of HPV-positive and HPV-negative cancers differs. Generally, HNSC harbours multiple therapeutically important genetic aberrations, including frequent aberrations in the FGFR and PI3K pathway genes. A different database, *i.e.* from the University of Texas Medical Branch (UTMB Health) in Galveston, Texas was used by LeConte *et al*. [2018] to investigate SNPs in the HPV16 genome and compared cervical with head and neck. Notably, no full genome sequence data or information regarding population demographics (*e.g*., ethnicity) were available. This study was an archival tissue cohort assessment comprising 226 HNSCs and 154 CESCs formalin-fixed and paraffin-embedded samples. Interestingly, compared to the EP sub-lineage, AA was less frequent in HNSC vs CESC.

4B.5 – Discussion

While presently we have performed literature review, acquired access to relevant TCGA data, setup our analysis infrastructure, and carried out preliminary analyses for HPV16 sub-lineage genotype and integration, continued efforts are required to comprehensively address our research aims. First, what insight do we glean from the sub-lineage genotyping of both head and neck and cervical cancer datasets? From initial results, it seems there are only a small number of cases with D2/3 (AA) sub-lineage compared to A (European) sub-lineages, which presumably is representative of the prevalence of these sub-lineages in the study/sample population. Additionally, is there a difference in sub-lineage distribution between head and neck and cervical cancer samples, and how do unique anatomical sites factor into the natural history of these viral infections, including risk of integration [LeConte *et al*., 2018; Combes and Franceschi, 2018; Eckhardt *et al*., 2018; Funk *et al*., 2018]? Next, what insight do we glean from integration analysis of both head and neck and cervical cancer datasets, with respect to sub-lineages present? Do we reject or fail to reject our hypothesis that D2/D3 is more strongly associated with integration, and do we have sufficient statistical power for this conclusion, or is sample size too low? Besides the general hypothesis (based on counts/proportion data), are there any differences in integration

patterns that are unique? Even with TCGA datasets, a major limitation in answering these questions are the small number of D2/D3 samples, and it may be necessary to extend these analyses beyond the well-curated TCGA database, but more broadly into other databases (*e.g.*, the SRA). While additional samples could possibly be found and incorporated into our analyses, these data may lack sufficient metadata to be easily searchable and screened (*e.g.*, experimental vs clinical data). As well, since the needs of our specific HPV-based analyses may not have been considered *a priori* when data were collected for those samples, there may be significant variability in data formats, read lengths, depths, and quality. Specifically, whether sequence capture or enrichment for HPV sequences was performed would greatly affect the probability of detecting HPV sequences [Holmes *et al*., 2016; Jackson *et al*., 2016]. Additionally, contaminant sequences, such as HPV18 sequences from the seemingly ubiquitous HeLa cell line, could be present and require filtering [Cantalupo *et al*., 2015]. It will be important to determine the role of co-infections with multiple PV types, as well as other microbes common to the epithelia. Sequencing errors could further confound variant identification, requiring careful mitigation [Stewart *et al*., 2018]. Overall, while these challenges exist, potentially relevant datasets deserve analysis and could be useful contributions. There have been efforts to curate reference viral databases for helping detect novel viruses and aide analysis of high-throughput sequencing data [Goodacre *et al*., 2018]. As well, recent perspectives on optimizing sample preparation, sequencing, and analysis for viral detection have been discussed elsewhere [Lambert *et al*., 2018; Ng *et al*., 2018; Chen *et al*., 2018b]. Overcoming these challenges can help address the rapidly increasing number of viruses discovered via high-throughput sequencing [Tisza *et al*., 2019] and provide support for the recent call to classify viruses [Kuhn *et al*., 2019].

Additional work is also required to determine the factors involved in a HPV16 sub-lineage-specific genomic instability and integration risk, including the role of defective DNA damage repair pathways [Seiwert *et al*., 2015; Ratnaparkhe *et al*., 2018], and whether there are also connections to altered innate immune pathways [Bakhoum *et al*., 2018]. It is also worthwhile to explore relationships between E6 splice variants and genomic stability [Olmedo-Nieva *et al*., 2018]. Researchers continue to explore novel mechanisms surrounding HPV16 integration, such as "super-enhancer-like elements" [Warburton *et al*., 2018] and interruption of tumour suppressor genes [Zhao *et al*., 2016]. Future research could continue exploring these relationships, such as determining the role of non-coding genome elements in genomic instability (*e.g.*, lncRNAs)

[Munschauer *et al*., 2018], the similarities of viral-human integration with mobile elements and human fusion genes [Imielinski and Ladanyi, 2018], as well as evolutionary perspectives [Petrie *et al*., 2018]. Finally, it remains to be seen how clinically-relevant HPV16 sub-lineage genotyping could be, and whether this could lead to improved care and outcome for patients. One strategy could be to compute individual risk scores based on the co-factors affecting risk of disease progression [Bastarache *et al*., 2018]. New screening techniques could also be useful, such as urine sampling [Van Keer *et al*., 2018] followed by high-throughput sequencing and HPV genotyping, as well as assessing methylation signatures [Cook *et al*., 2018].

4B.6 – Conclusion

From a basic science perspective, analysis of these big data helps us understand more about the mechanisms by which pathogens can become integrated into their hosts, akin to mobile elements, and by consequence lead to persistence of those sequences as well as disease progression. In the case of HPV16, a small number of changes in its genome may be responsible for increased host genetic instability and integration propensity. Additional bioinformatics analysis of these and other *ex vivo* data, coupled with *in vitro* experimental work, is required to further test this hypothesis. Hypothesis-free exploration of these data will also be important to clarify the gaps in our existing knowledge and identify fascinating phenomena to test in future experiments [Pipas, 2019]. From an applied science and clinical perspective, HPV16 sub-lineage genotyping can be used for preventative screening and routine monitoring, and next-generation sequencing specifically could be used on blood to detect integrated HPV sequences in circulating-tumour DNA (ctDNA) [Holmes *et al*., 2016] as a tumour-specific diagnostic biomarker and for monitoring residual disease and recurrence after treatments.

4B.7 – Declarations

4B.7.1 – Acknowledgements

The results reported here are based upon data generated by The Cancer Genome Atlas Research Network (cancergenome.nih.gov). This research was enabled in part by computational support provided by Compute Ontario (www.computeontario.ca) and Compute Canada (www.computecanada.ca). Thank you to Darryl Willick and Dr. Wely Floriano for their assistance in helping set up and maintain required bioinformatics tools (such as the Galaxy platform) hosted

at the Lakehead University High Performance Computing Centre (LUHPCC). We are also thankful for Bruce Bogacki's technical support at the Thunder Bay Regional Health Research Institute (TBRHRI). Additional thanks to Kathleen Roulston and Vanessa Masters for assisting with initial data access and analysis, Anirudh Shahi for helping with literature review on integration and transposable elements, and Alejandro Ortigas Vásquez for bioinformatics support. Preliminary work from this study was presented as a poster in Heidelberg, Germany (11-14 Oct 2017) at The Mobile Genome: Genetic and Physiological Impacts of Transposable Elements conference and a talk in Thunder Bay, Canada (12 Oct 2018) at the 1st Health & Information Technology (HIT) Research Group Workshop. This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) grants to IZ (#RGPIN-2015-03855) and RJ (CGS-D#454402-2014). The funding body had no role in study design, data collection, data analysis and interpretation, or preparation of the manuscript. We declare no conflicts of interest.

4B.7.2 – Authors' contributions

The study was initially conceived of and designed by IZ and RJ. Data access was managed by IZ and DN. Methods were developed, optimized, and analyses performed by DN and CG, with input from RJ and IZ. Biological interpretation and manuscript writing were led by RJ and IZ, while all authors were involved in providing critical revision and feedback.

4B.8 – Supplemental Information

4B.8.1 – Supplementary Script 1: Custom R script for sub-lineage pipeline

Script "vcf2clustal.R" takes file output from VarScan and creates a multiple sequence FASTA file that is ready for multiple sequence alignment by Clustal. The script alters an HPVA1 reference sequence (K02718) with the SNPs and indels found by VarScan to create a sequence that attempts to replicate the true sequence found in the case sample.

4B.8.2 – Supplementary Script 2: Custom R script for histogram creation

Script "HistoViFi.R" takes full ViFi output file as an argument and produces a histogram of integration hits, with each chromosome represented as a bin.

CHAPTER 5 – CONCLUSIONS

This dissertation addresses the unique pathogen-host relationship of HPV16 sub-lineages by combining novel techniques in human epithelial organoid modelling and "-omics"-based bioinformatics analyses. By combining these approaches, I sought to provide additional insight into how small changes in this tumourigenic virus could lead to increased risk for cancer progression, specifically via genomic instability and host genome integration. Although the HPV genome is small, and much about its natural history and molecular interactions within human epithelia has been learned in these past decades, there is clearly more to uncover regarding the complexity of its interactions and its multipurpose gene products [Mirabello *et al*., 2018]. While exploring this biological theme I also aimed to make original contributions in the field of biotechnology, specifically for human epithelial organoids and informatics tools.

Chapter-specific findings, challenges, and future directions are discussed throughout the dissertation, but their significance is summarized here. **CHAPTER 2**'s significance is primarily in the domain of basic natural science, in that we have potentially uncovered a natural phenomenon (a human papillomaviral variant with a predisposition for genomic integration under experimental conditions) using bio-engineered human tissue. We hypothesize that this integration phenomenon is related to an increased cancer risk. This work is impactful due to its collaborative and interdisciplinary nature, using innovative technologies such as three-dimensional organotypic epithelial cultures, viral sequence-capture, and next-generation sequencing coupled with bioinformatics. The use of "-omics" techniques allowed us to gain an understanding of widespread changes in the human (and viral) genome and transcriptome. The main limitation of this work is that the integration phenomenon may be attributable to artificial experimental conditions (*e.g.*, monolayer cell culture transfection and selection of viral genomes prior to rafting; including only E6 SNPs and not all those naturally found throughout the entire viral genome), which is why we sought to further enhance the model and confirm these findings using clinically-derived data.

Following the initial findings, the next three chapters focus on our epithelial organoid model. **CHAPTER 3A** is a significant contribution as it provides a historical overview of epithelial culturing, our perspectives on experimental designs for studying HPV16 viral life cycle in a natural setting, as well discussion on how to include innate immune components for assessing the HPV16 innate immune evasion hypothesis. As discussed in its chapter, one of the main limitations of epithelial organoids is that they are a simplification of the actual tissue complexity and

heterogeneity that is present *in vivo*, lacking other cell types, vasculature, and microbiota. While these factors likely play a role in HPV-induced carcinogenesis [Łaniewski *et al*., 2018], modelling them all *in vitro* becomes a complex task. **CHAPTER 3B** applies our organoid model to *in silico* applications (*i.e*., the stratification of epithelia for studying infection dynamics), and represents a significant cross-pollination of our efforts with evolutionary ecologists studying infectious diseases. This is a timely contribution, as cross-talk between biologists and mathematical modellers can lead to improved biological and mathematical models, such as the recent example of factors affecting epidermal thickness being used to predict optimal shapes ("sinusoidal undulations", in this case) for epithelial growth [Kumamoto *et al*., 2018]. **CHAPTER 3C** is a significant work in progress as it chronicles our recent progress and challenges of enhancing the epithelial organoid model. A limitation of our controlled experimental approach is that we introduce viral genomes into monolayer keratinocytes (which make up the basal layer of epithelium at the start of 3D culturing) using either chemical transfection or electroporation, rather than a natural infectious process via exposure of viral particles to naïve epithelia. An alternative approach could potentially allow us to study factors affecting viral entry and the earliest stages of the viral life cycle. As well, a significant challenge encountered when improving the model was that introducing immune cells, such as Langerhans cells, can negatively affect stratification (and as a result, could impact a viral life cycle). Alternative methods for testing innate immune evasion may be necessary (*e.g*., using a transwell culture assay to investigate the role of LC and assessing keratinocyte innate immune markers like TLRs and IFNs) [Jackson *et al*., 2014]. Future studies should also consider taking advantage of time-course experiments as well as single-cell [Lukowski *et al*., 2018] and *in situ* techniques, as the epithelia organoid is heterogenous due to the various cells, stratified layers, and viral activities at different times and spatial locations.

A computational "-omics" approach, offered by high-throughput sequencing and bioinformatics analyses, provides an ideal method for addressing and generating basic research questions. However, such analyses require advanced computational skills, which can be a barrier to researchers [Blankenberg *et al*., 2011]. The software described in **CHAPTER 4A**, Pathogen-Host Analysis Tool (PHAT), is a significant contribution aimed at reducing this barrier-to-entry. This project involved developing custom software, in collaboration with computer scientists, allowing for high-throughput data analysis and visualization for biologists without extensive computational knowledge. This was accomplished by integrating open-source tools into a user-

friendly platform. Our intention was to provide a tool with simplified workflows and a means of engaging with data, but as a result the main limitations are that the software is not fully customizable from a user perspective (*e.g.*, default parameters for alignments) and is not available for all possible operating systems and configurations. As well, since new tools frequently emerge in the bioinformatics space, it would become an intractable problem to integrate every potentially useful tool into PHAT. Researchers comfortable with programming and scripting, either standalone or using the cloud, or that have dedicated infrastructure and support for bioinformatics, would likely prefer to use their own custom-workflows allowing greater customization, flexibility, and access to processing power. To alleviate these concerns, we provide PHAT as an open-source project and invite users and developers to provide feedback and further improvements. Beyond this informatics tool contribution, **CHAPTER 4B** contains recent progress in testing our HPV16 variant-specific integration hypothesis using curated clinical data from TCGA datasets. This work is significant as it utilizes our software tool (PHAT) and could provide *ex vivo* support of our *in vitro* findings, making good use of pre-existing data. The main limitation is the small number of D2/D3 sub-lineage samples and that a more comprehensive analysis of their specific integration patterns is required for both cervical and head and neck cancer datasets. Beyond these "-omics" analyses, computational techniques continue to be utilized for HPV-associated cancers, with a recent report on deep learning to detect cervical cancer via image analysis [Hu *et al*., 2019].

There is fundamental scientific value in understanding how subtle changes in a pathogen's genome can lead to host consequences, as is the case with HPVs promoting tumourigenesis within human epithelia. Not only does this extend our knowledge of the molecular basis of these afflictions (*i.e*., anogenital and oropharyngeal cancers), and their heterogeneity, but it provides fodder for developing biomarkers, diagnostics, prognostics, and therapeutics to relieve disease burden and ultimately improve the quality of human life. Furthermore, the resulting biotechnologies that are developed, enhanced, and implemented can be extended to study diverse pathogen-host relationships (*i.e.*, other oncoviruses and infectious microbes) and used in other fields of research. To maximize the benefit of these tools for understanding the variability of complex biological systems researchers should continue to integrate biological and computational sciences (along with a healthy dose of philosophy [Laplane *et al*., 2019]), in an open-source and accessible manner, to further improve our ability to design meaningful experiments as well as generate, analyze, and interpret biological data.

LITERATURE CITED

Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*. **2016**; 44: W3-10. DOI: 10.1093/nar/gkw343

Akagi K, Li J, Broutian TR, Padilla-Nash H, Xiao W, Jiang B, *et al*. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res*. **2014**; 24: 185-99. DOI: 10.1101/gr.164806.113

Algeciras-Schimnich A, Policht F, Sitailo S, Song M, Morrison L, Sokolova I. Evaluation of quantity and staining pattern of human papillomavirus (HPV)-infected epithelial cells in thin-layer cervical specimens using optimized HPV-CARD assay. *Cancer Cytopathol*. **2007**; 111: 330-8. DOI: 10.1002/cncr.22946

Allen-Hoffmann BL, Schlosser SJ, Ivarie CA, Sattler CA, Meisner LF, O'Connor SL. Normal growth and differentiation in a spontaneously immortalized near-diploid human keratinocyte cell line, NIKS. *J Invest Dermatol*. **2000**; 114: 444-55. DOI: 10.1046/j.1523-1747.2000.00869.x

Amieva MR. Stanley Falkow (1934-2018). *Nature*. **2018**; 558: 190. DOI: 10.1038/d41586-018-05377-6

Anacker D, Moody C. Generation of organotypic raft cultures from primary human keratinocytes. *J Vis Exp*. **2012**; 60: e3668. DOI: 10.3791/3668

Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**; 11: R106. DOI: 10.1186/gb-2010-11-10-r106

Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*. **2014**; 31: 166-9. DOI: 10.1093/bioinformatics/btu638

Baba Y, Watanabe M, Murata A, Shigaki H, Miyake K, Ishimoto T, *et al*. LINE-1 hypomethylation, DNA copy number alterations, and CDK6 amplification in esophageal squamous cell carcinoma. *Clin Cancer Res*. **2014**; 20: 1114-24. DOI: 10.1158/1078-0432.CCR-13-1645

Bahnassy AA, Zekri AR, Alam El-Din HM, Aboubakr AA, Kamel K, El-Sabah MT, *et al*. The role of cyclins and cyclins inhibitors in the multistep process of HPV-associated cervical carcinoma. *J Egypt Natl Canc Inst*. **2006**; 18: 292-302. DOI: N/A

Bakhoum SF, Ngo B, Laughney AM, Cavallo JA, Murphy CJ, Ly P, *et al*. Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature*. **2018**; 553: 467-72. DOI: 10.1038/nature25432

Banzai C, Nishino K, Quan J, Yoshihara K, Sekine M, Yahata T, *et al*. Promoter methylation of DAPK1, FHIT, MGMT, and CDKN2A genes in cervical carcinoma. *Int J Clin Oncol*. **2014**; 19: 127-32. DOI: 10.1007/s10147-013-0530-0

Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, Ho WT, *et al*. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science*. **2018**; 359: 1233-9. DOI: 10.1126/science.aal4043

Bell E, Ehrlich HP, Buttle DJ, Nakatsuji T. Living tissue formed *in vitro* and accepted as skin-equivalent tissue of full thickness. *Science*. **1981**; 211: 1052-4. DOI: 10.1126/science.7008197

Bell E, Ivarsson B, Merrill C. Production of a tissue-like structure by contraction of collagen lattices by human fibroblasts of different proliferative potential *in vitro*. *Proc Natl Acad USA*. **1979**; 76: 1274-8. DOI: 10.1073/pnas.76.3.1274

Bell E, Sher S, Hull B, Merrill C, Rosen S, Chamson A, *et al*. The reconstitution of living skin. *J Invest Dermatol*. **1983**; 81: S2-10. DOI: 10.1111/1523-1747.ep12539993

Bernard HU. The clinical importance of the nomenclature, evolution and taxonomy of human papillomaviruses. *J Clin Virol*. **2005**; 32: S1-6. DOI: 10.1016/j.jcv.2004.10.021

Berumen J, Ordoñez RM, Lazcano E, Salmeron J, Galvan SC, Estrada RA, *et al*. Asian American variant of human papillomavirus 16 and risk for cervical cancer: a case–control study. *J Natl Cancer Inst*. **2001**; 93: 1325-30. DOI: 10.1093/jnci/93.17.1325

Bester AC, Roniger M, Oren YS, Im MM, Sarni D, Chaoat M, *et al*. Nucleotide deficiency promotes genomic instability in early stages of cancer development. *Cell*. **2011**; 145: 435-46. DOI: 10.1016/j.cell.2011.03.044

Bienkowska-Haba M, Luszczek W, Myers JE, Keiffer TR, DiGiuseppe S, Polk P, *et al*. A new cell culture model to genetically dissect the complete human papillomavirus life cycle. *PLoS Pathog*. **2018**; 14: e1006846. DOI: 10.1371/journal.ppat.1006846

Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A. Manipulation of FASTQ data with Galaxy. *Bioinformatics*. **2010a**; 26: 1783-5. DOI: 10.1093/bioinformatics/btq281

Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, *et al*. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*. **2010b**; Chapter 19: Unit 19.10.1-21. DOI: 10.1002/0471142727.mb1910s89

Blankenberg D, Taylor J, Nekrutenko A. Making whole genome multiple alignments usable for biologists. *Bioinformatics*. **2011**; 27: 2426-8. DOI: 10.1093/bioinformatics/btr398

Blanton RA, Perez-Reyes N, Merrick DT, McDougall JK. Epithelial cells immortalized by human papillomaviruses have premalignant characteristics in organotypic culture. *Am J Pathol*. **1991**; 138: 673-85. DOI: N/A

Bodelon C, Vinokurova S, Sampson JN, den Boon JA, Walker JL, Horswill MA, *et al*. Chromosomal copy number alterations and HPV integration in cervical precancer and invasive cancer. *Carcinogenesis*. **2016**; 37: 188-96. DOI: 10.1093/carcin/bgv171

Bodily JM, Mehta KP, Cruz L, Meyers C, Laimins LA. The E7 open reading frame acts in cis and in trans to mediate differentiation-dependent activities in the human papillomavirus type 16 life cycle. *J Virol*. **2011**; 85: 8852-62. DOI: 10.1128/JVI.00664-11

Bonfert T, Csaba G, Zimmer R, Friedel CC. Mining RNA-Seq data for infections and contaminations. *PLoS One*. **2013**; 8: e73071. DOI: 10.1371/journal.pone.0073071

Bouvard V, Baan R, Straif K, Grosse Y, Secretan B, El Ghissassi F, *et al*. A review of human carcinogens—Part B: biological agents. *Lancet Oncol*. **2009**; 10: 321-2. DOI: 10.1016/S1470-2045(09)70096-8

Box GEP. Science and statistics. *J Am Stat Assoc*. **1976**; 71: 791-9. DOI: 10.2307/2286841

Bryant D, Onions T, Raybould R, Flynn Á, Tristram A, Meyrick S, *et al*. mRNA sequencing of novel cell lines from human papillomavirus type-16 related vulval intraepithelial neoplasia: Consequences of expression of HPV16 E4 and E5. *J Med Virol*. **2014**; 86: 1534-41. DOI: 10.1002/jmv.23994

Burk RD, Harari A, Chen Z. Human papillomavirus genome variants. *Virology*. **2013**; 445: 232-43. DOI: 10.1016/j.virol.2013.07.018

Buschke S, Stark H-J, Cerezo A, Prätzel-Wunder S, Boehnke K, Kollar J, *et al*. A decisive function of transforming growth factor-β/Smad signaling in tissue morphogenesis and differentiation of human HaCaT keratinocytes. *Mol Biol Cell*. **2011**; 22: 782-94. DOI: 10.1091/mbc.E10-11-0879

Cantalupo PG, Katz JP, Pipas JM. HeLa nucleic acid contamination in The Cancer Genome Atlas leads to the misidentification of HPV18. *J Virol*. **2015**; 89: 4051-7. DOI: 10.1128/JVI.03365-14

Cantalupo PG, Katz JP, Pipas JM. Viral sequences in human cancer. *Virology*. **2018**; 513: 208-16. DOI: 10.1016/j.virol.2017.10.017

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. *Bioinformatics*. **2009**; 25: 288-9. DOI: 10.1093/bioinformatics/btn615

Carmona-Gutierrez D, Kainz K, Madeo F. Sexually transmitted infections: old foes on the rise. *Microbial Cell*. **2016**; 3: 361-2. DOI: 10.15698/mic2016.09.522

Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat Genet*. **2006**; 38: 1043-8. DOI: 10.1038/ng1861

Castellsagué X, Alemany L, Quer M, Halec G, Quirós B, Tous S, *et al*. HPV involvement in head and neck cancers: comprehensive assessment of biomarkers in 3680 patients. *J Natl Cancer Inst*. **2016**; 108: djv403. DOI: 10.1093/jnci/djv403

Chandrani P, Kulkarni V, Iyer P, Upadhyay P, Chaubal R, Das P, *et al*. NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. *Br J Cancer*. **2015**; 112: 1958-65. DOI: 10.1038/bjc.2015.121

Chang Y, Moore PS, Weiss RA. Human oncogenic viruses: nature and discovery. *Philos Trans Royal Soc B*. **2017**; 372: 20160264. DOI: 10.1098/rstb.2016.0264

Chau DYS, Johnson C, MacNeil S, Haycock JW, Ghaemmaghami AM. The development of a 3D immunocompetent model of human skin. *Biofabrication*. **2013**; 5: 035011. DOI: 10.1088/1758-5082/5/3/03511

Chen X, Kost J, Li D. Comprehensive comparative analysis of methods and software for identifying viral integrations. *Brief Bioinform*. **2018b:** 8 Aug. 10.1093/bib/bby070

Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics*. **2013**; 29: 266-7. DOI: 10.1093/bioinformatics/bts665

Chen YC, Huang RL, Huang YK, Liao YP, Su PH, Wang HC, *et al*. Methylomics analysis identifies epigenetically silenced genes and implies an activation of β-catenin signaling in cervical cancer. *Int J Cancer*. **2014**; 135: 117-27. DOI: 10.1002/ijc.28658

Chen Z, DeSalle R, Schiffman M, Herrero R, Wood CE, Ruiz JC, *et al*. Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. *PLoS Pathog*. **2018a**; 14: e1007352. DOI: 10.1371/journal.ppat.1007352

Chen Z, Terai M, Fu L, Herrero R, DeSalle R, Burk RD. Diversifying selection in human papillomavirus type 16 lineages based on complete genome analyses. *J Virol*. **2005**; 79: 7014-23. DOI: 10.1128/JVI.79.11.7014-7023.2005

Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, *et al*. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. **2018**; 15: 20170387. DOI: 10.1098/rsif.2017.0387

Clifford GM, Tenet V, Georges D, Alemany L, Pavón MA, Chen Z. Human Papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: whole viral genome sequences from 7116 HPV16-positive women. *Papillomavirus Res*. **2019**; 7: 67-74. DOI: 10.1016/j.pvr.2019.02.001

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*. **2010**; 38: 1767-71. DOI: 10.1093/nar/gkp1137

Colegio OR, Chu N-Q, Szabo AL, Chu T, Rhebergen AM, Jairam V, *et al*. Functional polarization of tumour-associated macrophages by tumour-derived lactic acid. *Nature*. **2014**; 513: 559-63. DOI: 10.1038/nature13490

Combes JD, Franceschi S. Human papillomavirus genome variants and head and neck cancers: a perspective. *Infect Agent Cancer*. **2018**; 13: 13. DOI: 10.1186/s13027-018-0185-6

Conway MJ, Meyers C. Replication and assembly of human papillomaviruses. *J Dent Res*. **2009**; 88: 307-17. DOI: 10.1177/0022034509333446

Cook DA, Krajden M, Brentnall AR, Gondara L, Chan T, Law JH, *et al*. Evaluation of a validated methylation triage signature for human papillomavirus positive women in the HPV FOCAL cervical cancer screening trial. *Int J Cancer*. **2018**: 9 Nov 2018. DOI: 10.1002/ijc.31976

Cornet I, Gheit T, Franceschi S, Vignat J, Burk RD, Sylla BS, *et al*. Human papillomavirus type 16 genetic variants: phylogeny and classification based on E6 and LCR. *J Virol*. **2012**; 86: 6855-61. DOI: 10.1128/JVI.00483-12

Crow JM. HPV: The global burden. *Nature*. **2012**; 488: S2-3. DOI: 10.1038/488S2a

Cullen M, Boland J, Schiffman M, Zhang X, Wentzensen N, Yang Q, *et al*. Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. *Papillomavirus Res*. **2015**; 1: 3-11. DOI: 10.1016/j.pvr.2015.05.004

Cuninghame S, Jackson R, Lees SJ, Zehbe I. Two common variants of human papillomavirus type 16 E6 differentially deregulate sugar metabolism and hypoxia signalling in permissive human keratinocytes. *J Gen Virol*. **2017**; 98: 2310-9. DOI: 10.1099/jgv.0.000905

Cuninghame S, Jackson R, Zehbe I. Hypoxia-inducible factor 1 and its role in viral carcinogenesis. *Virology*. **2014**; 456: 370-83. DOI: 10.1016/j.virol.2014.02.027

Daniel B, Rangarajan A, Geetasree M, Elizabeth V, Krishna S. The link between integration and expression of human papillomavirus type 16 genomes and cellular changes in the evolution of cervical intraepithelial neoplastic lesions. *J Gen Virol*. **1997**; 78: 1095-101. DOI: 10.1099/0022-1317-78-5-1095

Dawson ET, Wagner S, Roberson D, Yeager M, Boland J, Garrison E, *et al*. rkmh: a MinHash toolbox for analyzing HPV coinfections. *Cancer Res.* **2018**; 78; abstract 3273. DOI: 0.1158/1538-7445.AM2018-3273

de Araujo Souza PS, Sichero L, Maciag PC. HPV variants and HLA polymorphisms: the role of variability on the risk of cervical cancer. *Future Oncol*. **2009**; 5: 359-70. DOI: 10.2217/fon.09.8

Deans AJ, West SC. DNA interstrand crosslink repair and cancer. *Nat Rev Cancer*. **2011**; 11: 467-80. DOI: 10.1038/nrc3088

DeCarlo CA, Rosa B, Jackson R, Niccoli S, Escott NG, Zehbe I. Toll-like receptor transcriptome in the HPV-positive cervical cancer microenvironment. *Clin Dev Immunol*. **2012**; 2012: 785825. DOI: 10.1155/2012/785825

del Rosario RC, Rayan NA, Prabhakar S. Noncoding origins of anthropoid traits and a new null model of transposon functionalization. *Genome Res*. **2014**; 24: 1469-84. DOI: 10.1101/gr.168963.113

Deligeoroglou E, Giannouli A, Athanasopoulos N, Karountzos V, Vatopoulou A, Dimopoulos K, *et al*. HPV infection: immunological aspects and their utility in future therapy. *Infect Dis Obst Gynecol*. **2013**; 2013: 540850. DOI: 10.1155/2013/540850

Deng H, Hillpot E, Yeboah P, Mondal S, Woodworth CD. Susceptibility of epithelial cells cultured from different regions of human cervix to HPV16-induced immortalization. *PLoS One*. **2018**; 13: e0199761. DOI: 10.1371/journal.pone.0199761

Dollard SC, Wilson JL, Demeter LM, Bonnez W, Reichman RC, Broker TR, *et al*. Production of human papillomavirus and modulation of the infectious program in epithelial raft cultures. *Genes & Dev*. **1992**; 6: 1131-42. DOI: 10.1101/gad.6.7.1131

Doolittle-Hall J, Cunningham Glasspoole D, Seaman W, Webster-Cyriaque J. Meta-analysis of DNA tumor-viral integration site selection indicates a role for repeats, gene expression and epigenetics. *Cancers*. **2015**; 7: 2217-35. DOI: 10.3390/cancers7040887

Doorbar J. The papillomavirus life cycle. *J Clin Virol*. **2005**; 32: S7-15. DOI: 10.1016/j.jcv.2004.12.006

Doorbar J, Quint W, Banks L, Bravo IG, Stoler M, Broker TR, *et al*. The biology and life-cycle of human papillomaviruses. *Vaccine*. **2012**; 30: F55-70. DOI: 10.1016/j.vaccine.2012.06.083

Duensing S, Lee LY, Duensing A, Basile J, Piboonniyom SO, Gonzalez S, *et al*. The human papillomavirus type 16 E6 and E7 oncoproteins cooperate to induce mitotic defects and genomic instability by uncoupling centrosome duplication from the cell division cycle. *Proc Natl Acad Sci USA*. **2000**; 97: 10002-7. DOI: 10.1073/pnas.170093297

Duensing S, Münger K. The human papillomavirus type 16 E6 and E7 oncoproteins independently induce numerical and structural chromosome instability. *Cancer Res*. **2002**; 62: 7075-82. DOI: N/A

Dürst M, Bosch FX, Glitz D, Schneider A, zur Hausen H. Inverse relationship between human papillomavirus (HPV) type 16 early gene expression and cell differentiation in nude mouse epithelial cysts and tumors induced by HPV-positive human cell lines. *J Virol*. **1991**; 65: 796-804. DOI: N/A

Eckhardt M, Zhang W, Gross AM, Von Dollen J, Johnson JR, Franks-Skiba KE, *et al*. Multiple routes to oncogenesis are promoted by the human papillomavirus-host protein network. *Cancer Discov*. **2018**; 8: 1474-89. DOI: 10.1158/2159-8290.CD-17-1018.

Faluhelyi Z, Rodler I, Csejtey A, Tyring SK, Ember IA, Arany I. All-trans retinoic acid (ATRA) suppresses transcription of human papillomavirus type 16 (HPV16) in a dose-dependent manner. *Anticancer Res*. **2004**; 24: 807-9. DOI: N/A

Fatehullah A, Tan SH, Barker N. Organoids as an *in vitro* model of human development and disease. *Nat Cell Biol*. **2016**; 18: 246-54. DOI: 10.1038/ncb3312

Ferber MJ, Eilers P, Schuuring E, Fenton JA, Fleuren GJ, Kenter G, *et al*. Positioning of cervical carcinoma and Burkitt lymphoma translocation breakpoints with respect to the human papillomavirus integration cluster in FRA8C at 8q24.13. *Cancer Genet Cytogenet*. **2004**; 154: 1-9. DOI: 10.1016/j.cancergencyto.2004.01.028

Fischer J. Autosomal recessive congenital ichthyosis. *J Invest Dermatol*. **2009**; 129: 1319-21. DOI: 10.1038/jid.2009.57

Flores ER, Allen-Hoffmann BL, Lee D, Sattler CA, Lambert PF. Establishment of the human papillomavirus type 16 (HPV-16) life cycle in an immortalized human foreskin keratinocyte cell line. *Virology*. **1999**; 262: 344-54. DOI: 10.1006/viro.1999.9868

Freitas LB, Chen Z, Muqui EF, Boldrini NAT, Miranda AE, Spano LC, *et al*. Human papillomavirus 16 non-European variants are preferentially associated with high-grade cervical lesions. *PLoS One*. **2014**; 9: e100746. DOI: 10.1371/journal.pone.0100746

Funk LC, Lee DL, Kimple RJ, Lambert PF, Weaver BA. HPV oncoproteins cause specific types of chromosomal instability in head and neck cancer. *Cancer Res.* **2018**; 78: S1846. DOI: 10.1158/1538-7445.AM2018-1846

gplots Package for R. http://cran.r-project.org/web/packages/gplots/gplots.pdf

Ghittoni R, Accardi R, Chiocca S, Tommasino M. Role of human papillomaviruses in carcinogenesis. *Ecancermedicalscience*. **2015**; 9: 526. DOI: 10.3332/ecancer.2015.526

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, *et al*. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*. **2005**; 15: 1451-5. DOI: 10.1101/gr.4086505

Gibb CM, Jackson R, Mohammed S, Fiaidhi J, Zehbe I. Pathogen-Host Analysis Tool (PHAT): an integrative platform to analyze next-generation sequencing data. *Bioinformatics*. **2019**; bty1003. DOI: 10.1093/bioinformatics/bty1003

Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. **2010**; 11: R86. DOI: 10.1186/gb-2010-11-8-r86

Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS. A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. *mSphere*. **2018**; 3: e00069-18. DOI: 10.1128/mSphereDirect.00069-18

Grabowska AK, Riemer AB. The invisible enemy - how human papillomaviruses avoid recognition and clearance by the host immune system. *Open Virol J*. **2012**; 6, 249-56. DOI: 10.2174/1874357901206010249

Graham SV, Faizo AA. Control of human papillomavirus gene expression by alternative splicing. *Virus Res*. **2017**; 231: 83-95. DOI: 10.1016/j.virusres.2016.11.016

Grodzki M, Besson G, Clavel C, Arslan A, Franceschi S, Birembaut P, *et al*. Increased risk for cervical disease progression of French women infected with the human papillomavirus type 16 E6-350G variant. *Cancer Epidemiol Biomarkers Prev*. **2006**; 15: 820-2. DOI: 10.1158/1055-9965.EPI-05-0864

Gruner DS, Smith JE, Seabloom EW, Sandin SA, Ngai JT, Hillebrand H, *et al*. A cross-system synthesis of consumer and nutrient resource control on producer biomass. *Ecology Letters*. **2008**; 11: 740-55. DOI: 10.1111/j.1461-0248.2008.01192.x

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. **2011**; 144: 646-74. DOI: 10.1016/j.cell.2011.02.013

Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. **2000**; 100: 57-70. DOI: 10.1016/S0092-8674(00)81683-9

Hästbacka J, Kerrebrock A, Mokkala K, Clines G, Lovett M, Kaitila I, *et al*. Identification of the Finnish founder mutation for diastrophic dysplasia (DTD). *Eur J Human Genet*. **1999**; 7: 664-7. DOI: 10.1038/sj.ejhg.5200361

Havre PA, Yuan J, Hedrick L, Cho KR, Glazer PM. p53 inactivation by HPV16 E6 results in increased mutagenesis in human cells. *Cancer Res*. **1995**; 55: 4420-4. DOI: N/A

Hawkins TB, Dantzer J, Peters B, Dinauer M, Mockaitis K, Mooney S, *et al*. Identifying viral integration sites using SeqMap 2.0. *Bioinformatics*. **2011**; 27: 720-2. DOI: 10.1093/bioinformatics/btq722

Hay RJ, Johns NE, Williams HC, Bolliger IW, Dellavalle RP, Margolis DJ, *et al*. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *J Invest Dermatol*. **2014**; 134: 1527-34. DOI: 10.1038/jid.2013.446

Hochmann J, Sobrinho JS, Villa LL, Sichero L. The Asian-American variant of human papillomavirus type 16 exhibits higher activation of MAPK and PI3K/AKT signaling pathways, transformation, migration and invasion of primary human keratinocytes. *Virology*. **2016**; 492: 145-54. DOI: 10.1016/j.virol.2016.02.015

Holmes A, Lameiras S, Jeannot E, Marie Y, Castera L, Sastre-Garau X, *et al*. Mechanistic signatures of HPV insertions in cervical carcinomas. *Genome Med*. **2016**; 1: 16004. DOI: 10.1038/npjgenmed.2016.4

Hoppe-Seyler K, Bossler F, Braun JA, Herrmann AL, Hoppe-Seyler F. The HPV E6/E7 oncogenes: key factors for viral carcinogenesis and therapeutic targets. *Trends Microbiol*. **2018**; 26: 158-68. DOI: 10.1016/j.tim.2017.07.007

How C, Bruce J, So J, Pintilie M, Haibe-Kains B, Hui A, *et al*. Chromosomal instability as a prognostic marker in cervical cancer. *BMC Cancer*. **2015**; 15: 361. DOI: 10.1186/s12885-015-1372-0

Hu L, Bell D, Antani S, Xue Z, Yu K, Horning MP, *et al*. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *J Natl Cancer Inst*. **2019**: djy225. DOI: 10.1093/jnci/djy225

Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, *et al*. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat Genet*. **2015**; 47: 158-63. DOI: 10.1038/ng.3178

Huang X, Shen Y, Liu M, Bi C, Jiang C, Iqbal J, *et al*. Quantitative proteomics reveals that miR-155 regulates the PI3K-AKT pathway in diffuse large B-cell lymphoma. *Am J Pathol*. **2012**; 181: 26-33. DOI: 10.1016/j.ajpath.2012.03.013

Hubert P, van den Brûle F, Giannini SL, Franzen-Detrooz E, Boniver J, Delvenne P. Colonization of *in vitro*-formed cervical human papillomavirus-associated (pre)neoplastic lesions with dendritic cells: role of granulocyte/macrophage colony-stimulating factor. *Am J Pathol*. **1999**; 154: 775-84. DOI: 10.1016/S0002-9440(10)65324-2

Hubert WG. Variant upstream regulatory region sequences differentially regulate human papillomavirus type 16 DNA replication throughout the viral life cycle. *J Virology*. **2005**; 79: 5914-22. DOI: 10.1128/JVI.79.10.5914-5922.2005

Iijima N, Goodwin EC, DiMaio D, Iwasaki A. High-risk human papillomavirus E6 inhibits monocyte differentiation to Langerhans cells. *Virology*. **2013**; 444: 257-62. DOI: 10.1016/j.virol.2013.06.020

Imielinski M, Ladanyi M. Fusion oncogenes—genetic musical chairs. *Science*. **2018**; 361: 848-9. DOI: 10.1126/science.aau8231

Jackson R, Eade S, Zehbe I. An epithelial organoid model with Langerhans cells for assessing virus-host interactions. *Philos Trans Royal Soc B*. **2019**; *accepted for publication*. DOI: 10.1098/rstb.2018.0288

Jackson R, Rosa BA, Lameiras S, Cuninghame S, Bernard J, Floriano WB, *et al*. Functional variants of human papillomavirus type 16 demonstrate host genome integration and transcriptional alterations corresponding to their unique cancer epidemiology. *BMC Genomics*. **2016**; 17: 851. DOI: 10.1186/s12864-016-3203-3

Jackson R, Togtema M, Lambert PF, Zehbe I. Tumourigenesis driven by the human papillomavirus type 16 Asian-American E6 variant in a three-dimensional keratinocyte model. *PLoS One*. **2014**; 9: e101540. DOI: 10.1371/journal.pone.0101540

Jackson R, Togtema M, Zehbe I. Subcellular localization and quantitation of the human papillomavirus type 16 E6 oncoprotein through immunocytochemistry detection. *Virology*. **2013**; 435: 425-32. DOI: 10.1016/j.virol.2012.09.032

Jackson II RS, Cho YJ, Stein S, Liang P. CYFIP2, a direct p53 target, is leptomycin-B sensitive. *Cell Cycle*. **2007**; 6: 95-103. DOI: 10.4161/cc.6.1.3665

Jacobs N, Moutschen MP, Franzen-Detrooz E, Boniver V, Boniver J, Delvenne P. Organotypic culture of HPV-transformed keratinocytes: a model for testing lymphocyte infiltration of (pre)neoplastic lesions of the uterine cervix. *Virchows Arch*. **1998**; 432: 323-30. DOI: 10.1007/s004280050173

Jeon S, Allen-Hoffmann BL, Lambert PF. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *J Virol*. **1995**; 69: 2989-97. DOI: N/A

Johansson C, Schwartz S. Regulation of human papillomavirus gene expression by splicing and polyadenylation. *Nat Rev Microbiol*. **2013**; 11: 239-51. DOI: 10.1038/nrmicro2984

Jones M, Dry IR, Frampton D, Singh M, Kanda RK, Yee MB, *et al*. RNA-Seq analysis of host and viral gene expression highlights interaction between varicella zoster virus and keratinocyte differentiation. *PLoS Pathog*. **2014**; 10: e1003896. DOI: 10.1371/journal.ppat.1003896

Karimzadeh M, Hoffman MM. Virtual ChIP-seq: Predicting transcription factor binding by learning from the transcriptome. *bioRxiv*. **2018**: 168419. DOI: 10.1101/168419

Katz JP, Pipas JM. SummonChimera infers integrated viral genomes with nucleotide precision from NGS data. *BMC Bioinformatics*. **2014**; 15: 348. DOI: 10.1186/s12859-014-0348-4

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*. **2015**; 10: 845-58. DOI: 10.1038/nprot.2015.053

Kelley LA, Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc*. **2009**; 4: 363-71. DOI: 10.1038/nprot.2009.2

Kelley ML, Keiger KE, Lee CJ, Huibregtse JM. The global transcriptional effects of the human papillomavirus E6 protein in cervical carcinoma cell lines are mediated by the E6AP ubiquitin ligase. *J Virol*. **2005**; 79: 3737-47. DOI: 10.1128/JVI.79.6.3737-3747.2005

Kessis TD, Connolly DC, Hedrick L, Cho KR. Expression of HPV16 E6 or E7 increases integration of foreign DNA. *Oncogene*. **1996**; 13: 427-31. DOI: N/A

Khoury JD, Tannir NM, Williams MD, Chen Y, Yao H, Zhang J, *et al*. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. *J Virol*. **2013**; 87: 8916-26. DOI: 10.1128/JVI.00340-13

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. **2015**; 12: 357-60. DOI: 10.1038/nmeth.3317

Kim H-J, Kim I-S. Transforming growth factor-β-induced gene product, as a novel ligand of integrin $\alpha_M\beta_2$, promotes monocytes adhesion, migration and chemotaxis. *Int J Biochem Cell Biol*. **2008**; 40: 991-1004. DOI: 10.1016/j.biocel.2007.11.001

Kissenpfennig A, Malissen B. Langerhans cells – revisiting the paradigm using genetically engineered mice. *Trends Immunol*. **2006**; 27: 132-9. DOI: 10.1016/j.it.2006.01.003

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. **2012**; 22: 568-76. DOI: 10.1101/gr.129684.111

Kocjan BJ, Bzhalava D, Forslund O, Dillner J, Poljak M. Molecular methods for identification and characterization of novel papillomaviruses. *Clin Microbiol Infect*. **2015**; 21: 808-16. DOI: 10.1016/j.cmi.2015.05.011

Kosten IJ, Buskermolen JK, Spiekstra SW, de Gruijl TD, Gibbs S. Gingiva equivalents secrete negligible amounts of key chemokines involved in Langerhans cell migration compared to skin equivalents. *J Immunol Res*. **2015a**; 2015: 627125. DOI: 10.1155/2015/627125

Kosten IJ, Spiekstra SW, de Gruijl TD, Gibbs S. MUTZ-3 derived Langerhans cells in human skin equivalents show differential migration and phenotypic plasticity after allergen or irritant exposure. *Toxicol Appl Pharmacol*. **2015b**; 287: 35-42. DOI: 10.1016/j.taap.2015.05.017

Kosten IJ, Spiekstra SW, de Gruijl TD, Gibbs S. MUTZ-3 Langerhans cell maturation and CXCL12 independent migration in reconstructed human gingiva. *ALTEX*. **2016**; 33: 423-34. DOI: 10.14573/altex.1510301

Kraus I, Driesch C, Vinokurova S, Hovig E, Schneider A, von Knebel Doeberitz M, *et al*. The majority of viral-cellular fusion transcripts in cervical carcinomas cotranscribe cellular sequences of known or predicted genes. *Cancer Res*. **2008**; 68: 2514-22. DOI: 10.1158/0008-5472.CAN-07-2776

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res*. **2009**; 19: 1639-45. DOI: 10.1101/gr.092759.109

Kuhn JH, Wolf YI, Krupovic M, Zhang YZ, Maes P, Dolja VV, *et al*. Classify viruses—the gain is worth the pain. *Nature.* **2019**; 566: 318-20. DOI: 10.1038/d41586-019-00599-8

Kumamoto J, Nakanishi S, Makita M, Uesaka M, Yasugahira Y, Kobayashi Y, *et al*. Mathematical-model-guided development of full-thickness epidermal equivalent. *Sci Rep*. **2018**; 8: 17999. DOI: 10.1038/s41598-018-36647-y

Lace MJ, Anson JR, Klussmann JP, Wang DH, Smith EM, Haugen TH, *et al*. Human papillomavirus type 16 (HPV-16) genomes integrated in head and neck cancers and in HPV-16-immortalized human keratinocyte clones express chimeric virus-cell mRNAs similar to those found in cervical cancers. *J Virol*. **2011**; 85: 1645-54. DOI: 10.1128/JVI.02093-10

Lace MJ, Isacson C, Anson JR, Lörincz AT, Wilczynski SP, Haugen TH, *et al*. Upstream regulatory region alterations found in human papillomavirus type 16 (HPV-16) isolates from

cervical carcinomas increase transcription, *ori* function, and HPV immortalization capacity in culture. *J Virol*. **2009**; 83: 7457-66. DOI: 10.1128/JVI.00285-09

Lagström S, Umu SU, Lepistö M, Ellonen P, Meisal R, Christiansen IK, *et al*. TaME-seq: an efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration. *Sci Rep*. **2019**; 9: 524. DOI: 10.1038/s41598-018-36669-6

Lambert C, Braxton C, Charlebois R, Deyati A, Duncan P, La Neve F, *et al*. Considerations for optimization of high-throughput sequencing bioinformatics pipelines for virus detection. *Viruses*. **2018**; 10: 528. DOI: 10.3390/v10100528

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. **2012**; 9: 357-9. DOI: 10.1038/nmeth.1923

Łaniewski P, Barnes D, Goulder A, Cui H, Roe DJ, Chase DM, *et al*. Linking cervicovaginal immune signatures, HPV and microbiota composition in cervical carcinogenesis in non-Hispanic and Hispanic women. *Sci Rep*. **2018**; 8: 7593. DOI: 10.1038/s41598-018-25879-7

Laplane L, Mantovani P, Adolphs R, Chang H, Mantovani A, McFall-Ngai M, *et al*. Opinion: Why science needs philosophy. *Proc Natl Acad USA*. **2019**; 116: 3948-52. DOI: 10.1073/pnas.1900357116

Lau CC, Sun T, Ching AKK, He M, Li JW, Wong AM, *et al*. Viral-human chimeric transcript predisposes risk to liver cancer development and progression. *Cancer Cell*. **2014**; 25: 1-15. DOI: 10.1016/j.ccr.2014.01.030

Lavezzo E, Masi G, Toppo S, Franchin E, Gazzola V, Sinigaglia A, *et al*. Characterization of intra-type variants of oncogenic human papillomaviruses by next-generation deep sequencing of the E6/E7 region. *Viruses*. **2016**; 8: 79. DOI: 10.3390/v8030079

LeConte BA, Szaniszlo P, Fennewald SM, Lou DI, Qiu S, Chen NW, *et al*. Differences in the viral genome between HPV-positive cervical and oropharyngeal cancer. *PLoS One*. **2018**; 13: e0203403. DOI: 10.1371/journal.pone.0203403

Lee JH, Yi SMP, Anderson ME, Berger KL, Welsh MJ, Klingelhutz AJ, *et al*. Propagation of infectious human papillomavirus type 16 by using an adenovirus and Cre/LoxP mechanism. *Proc Natl Acad USA*. **2004**; 101: 2094-9. DOI: 10.1073/pnas.0308615100

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics*. **2009**; 25: 2078-9. DOI: 10.1093/bioinformatics/btp352

Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform*. **2010**; 11: 473-83. DOI: 10.1093/bib/bbq015

Li J, Chai QY, Liu CH. The ubiquitin system: a critical regulator of innate immunity and pathogen-host interactions. *Cell Mol Immunol*. **2016**; 13: 560-76. DOI: 10.1038/cmi.2016.40

Li JW, Wan R, Yu CS, Wong N, Chan TF. ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics*. **2013**; 29: 649-51. DOI: 10.1093/bioinformatics/btt011

Li Y, Wang C, Miao Z, Bi X, Wu D, Jin N, *et al*. ViRBase: a resource for virus-host ncRNA-associated interactions. *Nucleic Acids Res*. **2015**; 43: D578-82. DOI: 10.1093/nar/gku903

Lindeman RL. The trophic-dynamic aspect of ecology. *Ecology*. **1942**; 23: 399-418. DOI: 10.2307/1930126

Lukowski SW, Tuong ZK, Noske K, Senabouth A, Nguyen QH, Andersen SB, *et al*. Detection of HPV E7 transcription at single-cell resolution in epidermis. *J Invest Dermatol*. **2018**; 138: 2558-67. DOI: 10.1016/j.jid.2018.06.169

Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J*. **2008**; 53: 661-73. DOI: 10.1111/j.1365-313X.2007.03326.x

Masterson AJ, Sombroek CC, de Gruijl TD, Graus YM, van der Vliet HJ, Lougheed SM, *et al*. MUTZ-3, a human cell line model for the cytokine-induced differentiation of dendritic cells from CD34+ precursors. *Blood*. **2002**; 100: 701-3. DOI: 10.1182/blood.V100.2.701

Maufort JP, Williams SM, Pitot HC, Lambert PF. Human papillomavirus 16 E5 oncogene contributes to two stages of skin carcinogenesis. *Cancer Res*. **2007**; 67: 6106-12. DOI: 10.1158/0008-5472.CAN-07-0921

Matthews K, Leong CM, Baxter L, Inglis E, Yun K, Bäckström BT, *et al*. Depletion of Langerhans cells in human papillomavirus type 16-infected skin is associated with E6-mediated down regulation of E-cadherin. *J Virol*. **2003**; 77: 8378-85. DOI: 10.1128/JVI.77.15.8378-8385.2003

McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, *et al*. Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res*. **2013**; 41: W597-600. DOI: 10.1093/nar/gkt376

Melief CJ. Cancer immunology: cat and mouse games. *Nature*. **2005**; 437: 41-2. DOI: 10.1038/437041a

Merico D, Gfeller D, Bader GD. How to visually interpret biological data using networks. *Nat Biotech*. **2009**; 27: 921-4. DOI: 10.1038/nbt.1567

Merrick DT, Blanton RA, Gown AM, McDougall JK. Altered expression of proliferation and differentiation markers in human papillomavirus 16 and 18 immortalized epithelial cells grown in organotypic culture. *Am J Pathol*. **1992**; 140: 167-77. DOI: N/A

Mesri EA, Feitelson MA, Münger K. Human viral oncogenesis: a cancer hallmarks analysis. *Cell Host Microbe*. **2014**; 15: 266-82. DOI: 10.1016/j.chom.2014.02.011

Meyers C, Frattini MG, Hudson JB, Laimins LA. Biosynthesis of human papillomavirus from a continuous cell line upon epithelial differentiation. *Science*. **1992**; 257: 971-3. DOI: 10.1126/science.1323879

Meyerson NR, Sawyer SL. Two-stepping through time: mammals and viruses. *Trends Microbiol*. **2011**; 19: 286-94. DOI: 10.1016/j.tim.2011.03.006

Middleton K, Peh W, Southern S, Griffin H, Sotlar K, Nakahara T, *et al*. Organization of human papillomavirus productive cycle during neoplastic progression provides a basis for selection of diagnostic markers. *J Virol*. **2003**; 77: 10186-201. DOI: 10.1128/JVI.77.19.10186-10201.2003

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, *et al*. Tablet—next generation sequence assembly visualization. *Bioinformatics*. **2010**; 26: 401-2. DOI: 10.1093/bioinformatics/btp666

Mine KL, Shulzhenko N, Yambartsev A, Rochman M, Sanson GF, Lando M, *et al*. Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer. *Nat Commun*. **2013**; 4: 1806. DOI: 10.1038/ncomms2693

Mirabello L, Clarke MA, Nelson CW, Dean M, Wentzensen N, Yeager M, *et al*. The intersection of HPV epidemiology, genomics and mechanistic studies of HPV-mediated carcinogenesis. *Viruses*. **2018**; 10: 80. DOI: 10.3390/v10020080

Mori S, Kusumoto-Matsuo R, Ishii Y, Takeuchi T, Kukimoto I. Replication interference between human papillomavirus types 16 and 18 mediated by heterologous E1 helicases. *Virol J*. **2014**; 11: 11. DOI: 10.1186/1743-422X-11-11

Muller E, Brault B, Holmes A, Legros A, Jeannot E, Campitelli M, *et al*. Genetic profiles of cervical tumors by high-throughput sequencing for personalized medical care. *Cancer Med*. **2015**; 4: 1484-93. DOI: 10.1002/cam4.492

Munschauer M, Nguyen CT, Sirokman K, Hartigan CR, Hogstrom L, Engreitz JM, *et al*. The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature*. **2018**; 561: 132-6. DOI: 10.1038/s41586-018-0453-z

Murall CL, Jackson R, Zehbe I, Boulle N, Segondy M, Alizon S. Epithelial stratification shapes infection dynamics. *PLoS Comput Biol*. **2019**; 15: e1006646. DOI: 10.1371/journal.pcbi.1006646

Ndiaye C, Mena M, Alemany L, Arbyn M, Castellsagué X, Laporte L, *et al*. HPV DNA, E6/E7 mRNA, and $p16^{INK4a}$ detection in head and neck cancers: a systematic review and meta-analysis. *Lancet Oncol*. **2014**; 15: 1319-31. DOI: 10.1016/S1470-2045(14)70471-1

Ng S, Braxton C, Eloit M, Feng S, Fragnoud R, Mallet L, *et al*. Current perspectives on high-throughput sequencing (HTS) for adventitious virus detection: upstream sample processing and library preparation. *Viruses*. **2018**; 10: 566. DOI: 10.3390/v10100566

Nguyen N-PD, Deshpande V, Luebeck J, Mischel PS, Bafna V. ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. *Nucleic Acids Res*. **2018**; 46: 3309-25. DOI: 10.1093/nar/gky180

Niccoli S, Abraham S, Richard C, Zehbe I. The Asian-American E6 variant protein of human papillomavirus 16 alone is sufficient to promote immortalization, transformation, and migration of primary human foreskin keratinocytes. *J Virol*. **2012**; 86: 12384-96. DOI: 10.1128/JVI.01512-12

Nulton TJ, Olex AL, Dozmorov M, Morgan IM, Windle B. Analysis of The Cancer Genome Atlas sequencing data reveals novel properties of the human papillomavirus 16 genome in head and neck squamous cell carcinoma. *Oncotarget*. **2017**; 8: 17684-99. DOI: 10.18632/oncotarget.15179

Ojesina AI, Lichtenstein L, Freeman SS, Pedamallu CS, Imaz-Rosshandler I, Pugh TJ, *et al*. Landscape of genomic alterations in cervical carcinomas. *Nature*. **2014**; 506: 371-5. DOI: 10.1038/nature12881

Okonechnikov K, Golosova O, Fursov M. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*. **2012**; 28: 1166-7. DOI: 10.1093/bioinformatics/bts091

Olmedo-Nieva L, Muñoz-Bello J, Contreras-Paredes A, Lizano M. The role of E6 spliced isoforms (E6*) in human papillomavirus-induced carcinogenesis. *Viruses*. **2018**; 10: 45. DOI: 10.3390/v10010045

Parfenov M, Pedamallu CS, Gehlenborg N, Freeman SS, Danilova L, Bristow CA, *et al*. Characterization of HPV and host genome interactions in primary head and neck cancers. *Proc Natl Acad Sci USA*. **2014**; 111: 15544-9. DOI: 10.1073/pnas.1416074111

Paris C, Pentland I, Groves I, Roberts DC, Powis SJ, Coleman N, *et al*. CCCTC-binding factor recruitment to the early region of the human papillomavirus 18 genome regulates viral oncogene expression. *J Virol*. **2015**; 89: 4770-85. DOI: 10.1128/JVI.00097-15

Parkin DM, Bray F. The burden of HPV-related cancers. *Vaccine*. **2006;** 24: S11-25. DOI: 10.1016/j.vaccine.2006.05.111

Pastrana DV, Peretti A, Welch NL, Borgogna C, Olivero C, Badolato R, *et al*. Metagenomic discovery of 83 new human papillomavirus types in patients with immunodeficiency. *mSphere*. **2018**; 3: e00645-18. DOI: 10.1128/mSphereDirect.00645-18

Peter M, Stransky N, Couturier J, Hupé P, Barillot E, de Cremoux P, *et al*. Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma. *J Pathol*. **2010**; 221: 320-30. DOI: 10.1002/path.2713

Petrie KL, Palmer ND, Johnson DT, Medina SJ, Yan SJ, Li V, *et al*. Destabilizing mutations encode nongenetic variation that drives evolutionary innovation. *Science*. **2018**; 359: 1542-5. DOI: 10.1126/science.aar1954

Pett M, Coleman N. Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *J Pathol*. **2007**; 212: 356-67. DOI: 10.1002/path.2192

Picard Tools. http://broadinstitute.github.io/picard/. Accessed 13 May **2016**.

Pichardo S, Togtema M, Jackson R, Zehbe I, Curiel L. Influence of cell line and cell cycle phase on sonoporation transfection efficiency in cervical carcinoma cells under the same physical conditions. *IEEE Trans Ultrason Ferroelectr Freq Control*. **2013**; 60: 432-5. DOI: 10.1109/TUFFC.2013.2581

Pipas JM. DNA tumour viruses and their contributions to molecular biology. *J Virol.* **2019**: JVI.01524-18. DOI: 10.1128/JVI.01524-18

Poreba E, Broniarczyk JK, Gozdzicka-Jozefiak A. Epigenetic mechanisms in virus-induced tumorigenesis. *Clin Epigenetics*. **2011**; 2: 233-47. DOI: 10.1007/s13148-011-0026-6

Potter H, Heller R. Transfection by electroporation. *Curr Protoc Mol Biol*. **2003**; 62: 9.3.1-6. DOI: 10.1002/0471142727.mb0903s62

R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. **2018**. http://www.R-project.org

RStudio Team. RStudio: integrated development for R. RStudio, Inc. **2015**. http://www.rstudio.com

Rabbitts TH. Chromosomal translocations in human cancer. *Nature*. **1994**; 372: 143-9. DOI: 10.1038/372143a0

Ratnaparkhe M, Wong J, Wei PC, Hlevnjak M, Kolb T, Haag D, *et al*. Defective DNA damage repair leads to frequent catastrophic genomic events in murine and human tumors. *Nat Commun*. **2018**; 9: 4760. DOI: 10.1038/s41467-018-06925-4

Reeves C, Charles-Horvath P, Kitajewski J. Studies in mice reveal a role for anthrax toxin receptors in matrix metalloproteinase function and extracellular matrix homeostasis. *Toxins*. **2013**; 5: 315-26. DOI: 10.3390/toxins5020315

Rheinwald JG, Green H. Formation of a keratinizing epithelium in culture by a cloned cell line derived from a teratoma. *Cell*. **1975**; 6: 317-30. DOI: 10.1016/0092-8674(75)90183-X

Richard C, Lanner C, Naryzhny S, Sherman L, Lee H, Lambert PF, *et al*. The immortalizing and transforming ability of two common human papillomavirus 16 E6 variants with different prevalences in cervical cancer. *Oncogene*. **2010**; 29: 3435-45. DOI: 10.1038/onc.2010.93

Richards KL, Zhang B, Baggerly KA, Colella S, Lang JC, Schuller DE, *et al*. Genome-wide hypomethylation in head and neck cancer is more pronounced in HPV-negative tumors and is associated with genomic instability. *PLoS One*. **2009**; 4: e4941. DOI: 10.1371/journal.pone.0004941

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, *et al*. Integrative genomics viewer. *Nat Biotechnol*. **2011**; 29: 24-6. DOI: 10.1038/nbt.1754

Rodrigues Neves C, Gibbs S. Progress on reconstructed human skin models for allergy research and identifying contact sensitizers. *Curr Top Microbiol Immunol*. **2018**: 1-27. DOI: 10.1007/82_2018_88

Roman A, Münger K. The papillomavirus E7 proteins. *Virology*. **2013**; 445: 138-68. DOI: 10.1016/j.virol.2013.04.013

Samanta S, Dey P, Nijhawan R. Micronucleus in cervical intraepithelial lesions and carcinoma. *Acta Cytol*. **2011**; 55: 42-7. DOI: 10.1159/000320792

Santegoets SJAM, Masterson AJ, van der Sluis PC, Lougheed SM, Fluitsma DM, van den Eertwegh AJ, *et al*. A $CD34^+$ human cell line model of myeloid dendritic cell differentiation: evidence for a $CD14^+CD11b^+$ Langerhans cell precursor. *J Leukoc Biol*. **2006**; 80: 1337-44. DOI: 10.1189/jlb.0206111

Sasaki K, Akiyama M, Yanagi T, Sakai K, Miyamura Y, Sato M, *et al*. CYP4F22 is highly expressed at the site and timing of onset of keratinization during skin development. *J Dermatol Sci*. **2012**; 65: 156-8. DOI: 10.1016/j.jdermsci.2011.12.006

Schiffman M, Rodriguez AC, Chen Z, Wacholder S, Herrero R, Hildesheim A, *et al*. A population-based prospective study of carcinogenic human papillomavirus variant lineages, viral persistence, and cervical neoplasia. *Cancer Res*. **2010**; 70: 3159-69. DOI: 10.1158/0008-5472.CAN-09-4179

Schlecht NF, Burk RD, Adrien L, Dunne A, Kawachi N, Sarta C, *et al*. Gene expression profiles in HPV-infected head and neck cancer. *J Pathol*. **2007**; 213: 283-93. DOI: 10.1002/path.2227

Schmitz M, Driesch C, Jansen L, Runnebaum IB, Dürst M. Non-random integration of the HPV genome in cervical cancer. *PLoS One*. **2012**; 7: e39632. DOI: 10.1371/journal.pone.0039632

Schütze DM, Snijders PJ, Bosch L, Kramer D, Meijer CJ, Steenbergen RD. Differential *in vitro* immortalization capacity of eleven (probable) high-risk human papillomavirus types. *J Virol*. **2014**; 88: 1714-24. DOI: 10.1128/JVI.02859-13

Seedorf K, Krämmer G, Dürst M, Suhai S, Röwekamp WG. Human papillomavirus type 16 DNA sequence. *Virology*. **1985**; 145, 181-5. DOI: 10.1016/0042-6822(85)90214-4

Seiwert TY, Zuo Z, Keck MK, Khattri A, Pedamallu CS, Stricker T, *et al*. Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. *Clin Cancer Res*. **2015**; 21: 632-41. DOI: 10.1158/1078-0432.CCR-13-3310

Sharma AK, Rigby AC, Alper SL. STAS domain structure and function. *Cell Physiol Biochem*. **2011**; 28: 407-22. DOI: 10.1159/000335104

Sichero L, Sobrinho JS, Villa LL. Oncogenic potential diverge among human papillomavirus type 16 natural variants. *Virology*. **2012**; 432: 127-32. DOI: 10.1016/j.virol.2012.06.011

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. **2011**; 27: 431-2. DOI: 10.1093/bioinformatics/btq675

Southern SA, Noya F, Meyers C, Broker TR, Chow LT, Herrington CS. Tetrasomy is induced by human papillomavirus type 18 E7 gene expression in keratinocyte raft cultures. *Cancer Res*. **2001**; 61: 4858-63. DOI: N/A

Spurgeon ME, den Boon JA, Horswill M, Barthakur S, Forouzan O, Rader JS, *et al*. Human papillomavirus oncogenes reprogram the cervical cancer microenvironment independently of and synergistically with estrogen. *Proc Natl Acad USA*. **2017**; 114: E9076-85. DOI: 10.1073/pnas.171201811

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **2014**; 30: 1312-3. DOI: 10.1093/bioinformatics/btu033

Stanley MA. Epithelial cell responses to infection with human papillomavirus. *Clin Microbiol Rev*. **2012**; 25: 215-22. DOI: 10.1128/CMR.05028-11

Stewart C, Leshchiner I, Hess J, Getz G. Comment on "DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification". *Science*. **2018**; 361: eaas9824. DOI: 10.1126/science.aas9824

Tang A, Amagai M, Granger LG, Stanley JR, Uddy MC. Adhesion of epidermal Langerhans cells to keratinocytes mediated by E-cadherin. *Nature*. **1993**; 361: 82-5. DOI: 10.1038/361082a0

Tang KW, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat Commun*. **2013**; 4: 2513. DOI: 10.1038/ncomms3513

TCGA Research Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*. **2015**; 517: 576-82. DOI: 10.1038/nature14129

TCGA Research Network. Integrated genomic and molecular characterization of cervical cancer. *Nature.* **2017**; 543: 378-84. DOI: 10.1038/nature21386

Thyagarajan B, Guimaraes MJ, Growth AC, Calos MP. Mammalian genomes contain active recombinase recognition sites. *Gene*. **2000**; 244: 47-54. DOI: 10.1016/S0378-1119(00)00008-1

Tisza MJ, Pastrana DV, Welch NL, Stewart B, Peretti A, Starrett GJ, *et al*. Discovery of several thousand highly diverse circular DNA viruses. *bioRxiv*. **2019**: 555375. DOI: 10.1101/555375

Togtema M, Hussack G, Dayer G, Teghtmeyer M, Raphael S, Tanha J, *et al*. Single-domain antibodies represent novel alternatives to monoclonal antibodies as targeting agents against the human papillomavirus 16 E6 protein. *bioRxiv*. **2018a**: 388884. DOI: 10.1101/388884

Togtema M, Jackson R, Grochowski J, Villa PL, Mellerup M, Chattopadhyaya J, *et al*. Synthetic siRNA targeting human papillomavirus 16 E6: a perspective on *in vitro* nanotherapeutic approaches. *Nanomedicine*. **2018b**; 13: 455-74. DOI: 10.2217/nnm-2017-0242

Togtema M, Jackson R, Richard C, Niccoli S, Zehbe I. The human papillomavirus 16 European-T350G E6 variant can immortalize but not transform keratinocytes in the absence of E7. *Virology*. **2015**; 485: 274-82. DOI: 10.1016/j.virol.2015.07.025

Togtema M, Pichardo S, Jackson R, Lambert PF, Curiel L, Zehbe I. Sonoporation delivery of monoclonal antibodies against human papillomavirus 16 E6 restores p53 expression in transformed cervical keratinocytes. *PLoS One*. **2012**; 7: e50730. DOI: 10.1371/journal.pone.0050730

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. **2012**; 7: 562-78. DOI: 10.1038/nprot.2012.016

Van Doorslaer K, Dillner J. The launch of an international animal papillomavirus reference center. *Viruses*. **2019**; 11: 55. DOI: 10.3390/v11010055

Van Doorslaer K, Li Z, Xirasagar S, Maes P, Kaminsky D, Liou D, *et al*. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res*. **2017b**; 45: D499-506. DOI: 10.1093/nar/gkw879

Van Doorslaer K, Ruoppolo V, Schmidt A, Lescroël A, Jongsomjit D, Elrod M, *et al*. Unique genome organization of non-mammalian papillomaviruses provides insights into the evolution of viral early proteins. *Virus Evol*. **2017a**; 3: vex027. DOI: 10.1093/ve/vex027

Van Doorslaer K, Tan Q, Xirasagar S, Bandaru S, Gopalan V, Mohamoud Y, *et al*. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res*. **2013**; 41: D571-8. DOI: 10.1093/nar/gks984

Van Keer S, Tjalma WA, Pattyn J, Biesmans S, Pieters Z, Van Ostade X, *et al*. Human papillomavirus genotype and viral load agreement between paired first-void urine and clinician-collected cervical samples. *Eur J Clin Microbiol Infect Dis*. **2018**; 37: 859-69. DOI: 10.1007/s10096-017-3179-1

Vande Pol SB, Klingelhutz AJ. Papillomavirus E6 oncoproteins. *Virology*. **2013**; 445: 115-37. DOI: 10.1016/j.virol.2013.04.026

Vanderkam D, Aksoy BA, Hodes I, Perrone J, Hammerbacher J. pileup.js: a JavaScript library for interactive and in-browser visualization of genomic data. *Bioinformatics*. **2016**; 32: 2378-9. DOI: 10.1093/bioinformatics/btw167

Villa LL, Sichero L, Rahal P, Caballero O, Ferenczy A, Rohan T, *et al*. Molecular variants of human papillomavirus types 16 and 18 preferentially associated with cervical neoplasia. *J Gen Virol*. **2000**; 81: 2959-68. DOI: 10.1099/0022-1317-81-12-2959

Villa PL, Jackson R, Eade S, Escott N, Zehbe I. Isolation of biopsy-derived, human cervical keratinocytes propagated as monolayer and organoid cultures. *Sci Rep*. **2018**; 8: 17869. DOI: 10.1038/s41598-018-36150-4

Wang HK, Duffy AA, Broker TR, Chow LT. Robust production and passaging of infectious HPV in squamous epithelium of primary human keratinocytes. *Genes & Dev*. **2009**; 23: 181-94. DOI: 10.1101/gad.1735109

Wang Q, Jia P, Zhao Z. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One*. **2013**; 8: e64465. DOI: 10.1371/journal.pone.0064465

Warburton A, Redmond CJ, Dooley KE, Fu H, Gillison ML, Akagi K, *et al*. HPV integration hijacks and multimerizes a cellular enhancer to generate a viral-cellular super-enhancer that drives high viral oncogene expression. *PLoS Genet*. **2018**; 14: e1007179. DOI: 10.1371/journal.pgen.1007179

Weitzman MD, Weitzman JB. What's the damage? The impact of pathogens on pathways that maintain host genome integrity. *Cell Host Microbe*. **2014**; 15: 283-94. DOI: 10.1016/j.chom.2014.02.010

Wentzensen N, Ridder R, Klaes R, Vinokurova S, Schaefer U, von Knebel Doeberitz M. Characterization of viral-cellular fusion transcripts in a large series of HPV16 and 18 positive anogenital lesions. *Oncogene*. **2002**; 21: 419-26. DOI: 10.1038/sj.onc.1205104

Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. *Nat Rev Microbiol*. **2012**; 10: 618-30. DOI: 10.1038/nrmicro2852

White AE, Livanos EM, Tlsty TD. Differential disruption of genomic integrity and cell cycle regulation in normal human fibroblasts by the HPV oncoproteins. *Genes & Dev*. **1994**; 8: 666-77. DOI: 10.1101/gad.8.6.666

WHO. Global health sector strategy on sexually transmitted infections, 2016-2021. Geneva, Switzerland: World Health Organisation. **2016**; WHO/RHR/16.09. DOI: N/A

Williams M, Rainville IR, Nicklas JA. Use of inverse PCR to amplify and sequence breakpoints of HPRT deletion and translocation mutations. *Environ Mol Mutagen*. **2002**; 39: 22-32. DOI: 10.1002/em.10040

Winder DM, Pett MR, Foster N, Shivji MK, Herdman MT, Stanley MA, *et al*. An increase in DNA double-strand breaks, induced by Ku70 depletion, is associated with human papillomavirus 16 episome loss and de novo viral integration events. *J Pathol*. **2007**; 213: 27-34. DOI: 10.1002/path.2206

Xi LF, Koutsky LA, Galloway DA, Kiviat NB, Kuypers J, Hughes JP, *et al*. Genomic variation of human papillomavirus type 16 and risk for high grade cervical intraepithelial neoplasia. *J Natl Cancer Inst*. **1997**; 89: 796-802. DOI: 10.1093/jnci/89.11.796

Xi LF, Koutsky LA, Hildesheim A, Galloway DA, Wheeler CM, Winer RL, *et al*. Risk for high-grade cervical intraepithelial neoplasia associated with variants of human papillomavirus types 16 and 18. *Cancer Epidemiol Biomarkers Prev*. **2007**; 16: 4-10. DOI: 10.1158/1055-9965.EPI-06-0670

Xiang Z, Tian Y, He Y. PHIDIAS: a pathogen-host interaction data integration and analysis system. *Genome Biol*. **2007**; 8: R150. DOI: 10.1186/gb-2007-8-7-r150

Xu B, Chotewutmontri S, Wolf S, Klos U, Schmitz M, Dürst M, *et al*. Multiplex identification of human papillomavirus 16 DNA integration sites in cervical carcinomas. *PLoS One*. **2013**; 8: e66693. DOI: 10.1371/journal.pone.0066693

Yamada T, Manos MM, Peto J, Greer CE, Munoz N, Bosch FX, *et al*. Human papillomavirus type 16 sequence variation in cervical cancers: a worldwide perspective. *J Virol*. **1997**; 71: 2463-72. DOI: N/A

Yang X, Li M, Liu Q, Zhang Y, Qian J, Wan X, *et al*. Dr.VIS v2.0: an updated database of human disease-related viral integration sites in the era of high-throughput deep sequencing. *Nucleic Acids Res*. **2015**; 43: D887-92. DOI: 10.1093/nar/gku1074

Zacapala-Gómez AE, Del Moral-Hernández O, Villegas-Sepúlveda N, Hidalgo-Miranda A, Romero-Córdoba SL, Beltrán-Anaya FO, *et al*. Changes in global gene expression profiles induced by HPV 16 E6 oncoprotein variants in cervical carcinoma C33-A cells. *Virology*. **2016**; 488: 187-95. DOI: 10.1016/j.virol.2015.11.017

Zapatka M, Borozan I, Brewer DS, Iskar M, Grundhoff A, Alawi M, *et al*. The landscape of viral associations in human cancers. *bioRxiv*. **2018**; 465757. DOI: 10.1101/465757

Zehbe I, Jackson R, Wood B, Weaver B, Escott N, Severini A, *et al*. Community-randomised controlled trial embedded in the Anishinaabek Cervical Cancer Screening Study: human papillomavirus self-sampling versus Papanicolaou cytology. *BMJ Open*. **2016a**; 6: e011754. DOI: 10.1136/bmjopen-2016-011754

Zehbe I, Richard C, DeCarlo CA, Shai A, Lambert PF, Lichtig H, *et al*. Human papillomavirus 16 E6 variants differ in their dysregulation of human keratinocyte differentiation and apoptosis. *Virology*. **2009**; 383: 69-77. DOI: 10.1016/j.virol.2008.09.036

Zehbe I, Voglino G, Delius H, Wilander E, Tommasino M. Risk of cervical cancer and geographical variations of human papillomavirus 16 E6 polymorphisms. *Lancet*. **1998a**; 352: 1441-2. DOI: 10.1016/S0140-6736(05)61263-9

Zehbe I, Voglino G, Wilander E, Delius H, Marongiu A, Edler L, *et al*. p53 codon 72 polymorphism and various human papillomavirus 16 E6 genotypes are risk factors for cervical cancer development. *Cancer Res*. **2001**; 61: 608-11. DOI: N/A

Zehbe I, Wilander E, Delius H, Tommasino M. Human papillomavirus 16 E6 variants are more prevalent in invasive cervical carcinoma than the prototype. *Cancer Res*. **1998b**; 58: 829-33. DOI: N/A

Zehbe I, Wood B, Wakewich P, Maar M, Escott N, Jumah N, *et al*. Teaching tools to engage Anishinaabek First Nations women in cervical cancer screening: Report of an educational workshop. *Health Educ J*. **2016b**; 75: 331-42. DOI: 10.1177/0017896915580446

Zhang CZ, Spektor A, Cornils H, Francis JM, Jackson EK, Liu S, *et al*. Chromothripsis from DNA damage in micronuclei. *Nature*. **2015**; 522: 179-84. DOI: 10.1038/nature14493

Zhang Z, Huettner PC, Nguyen L, Bidder M, Funk MC, Li J, *et al*. Aberrant promoter methylation and silencing of the POU2F3 gene in cervical cancer. *Oncogene*. **2006**; 25: 5436-45. DOI: 10.1038/sj.onc.1209530

Zhao JW, Fang F, Guo Y, Zhu TL, Yu YY, Kong FF, *et al*. HPV16 integration probably contributes to cervical oncogenesis through interrupting tumor suppressor genes and inducing chromosome instability. *J Exp Clin Cancer Res*. **2016**; 35: 180. DOI: 10.1186/s13046-016-0454-4

Zhou S. Cytochrome P450 2D6: structure, function, regulation and polymorphism. *CRC Press*; **2016**. DOI: N/A

Zuna RE, Moore WE, Shanesmith RP, Dunn ST, Wang SS, Schiffman M, *et al*. Association of HPV16 E6 variants with diagnostic severity in cervical cytology samples of 354 women in a US population. *Int J Cancer*. **2009**; 125: 2609-13. DOI: 10.1002/ijc.24706

zur Hausen H. Papillomavirus infections—a major cause of human cancers. *Biochim Biophys Acta, Rev Cancer*. **1996**; 1288: F55-78. DOI: 10.1016/0304-419X(96)00020-0

zur Hausen H. Papillomaviruses and cancer: from basic studies to clinical investigations. *Nat Rev Cancer*. **2002**; 2: 342-50. DOI: 10.1038/nrc798

APPENDIX A

Table A.1 – Plasmid information. Plasmids were either provided as gifts, shared via a material transfer agreement (MTA), or custom-ordered.

Plasmid	Source	Status	Size (kbp)	Bacterial Selection	Vector	Digest Site
pCAGGS-NLS-cre	Dr. Nagy	Maxiprep stock prepared	~5.8	amp^R	pCAGGS	N/A
pNeo-loxP HPV-18	Drs. Chow & Broker	Maxiprep stock prepared	~11.3	kan^R	pNeo	XmaI XhoI
pEGFP Ni HPV16	Dr. Lee	Maxiprep stock prepared	~12.2	kan^R	peGFP-N1	NheI HindIII
HPV16 EP SphI Linearized	Our Lab	Synthesized and cloned by GenScript	~13.4	amp^R	pcDNA3.1 (-)	XhoI XbaI
HPV16 EP PmlI Linearized	Our Lab	Synthesized and cloned by GenScript	~13.4	amp^R	pcDNA3.1 (-)	XhoI XbaI
HPV16 A1/L5398k DL0116184 HPV00001	Dr. Dean (NCI)	Maxiprep stock prepared	~10.8	amp^R (suspected)	pBMH (pBSK(+))	EcoRV
HPV16 D2/L5399k DL0116196 HPV00013	Dr. Dean (NCI)	Maxiprep stock prepared	~10.8	amp^R (suspected)	pBMH (pBSK(+))	EcoRV
HPV16 D3/L5400k DL0116190 HPV00007	Dr. Dean (NCI)	Maxiprep stock prepared	~10.8	amp^R (suspected)	pBMH (pBSK(+))	EcoRV
HPV16 AA D2 SphI	Our Lab	In prep	~13.4	amp^R	pcDNA3.1 (-)	XhoI XbaI
HPV16 AA D3 SphI	Our Lab	In prep	~13.4	amp^R	pcDNA3.1 (-)	XhoI XbaI
HPV16 114/B (DNA #959)	Dr. Müller (DKFZ)	Filter paper	~10.6	amp^R (suspected)	pUC (pUC19?)	EcoRI
HPV16 114/K (DNA #960)	Dr. Müller (DKFZ)	Filter paper	~10.6	amp^R (suspected)	pUC (pUC19?)	EcoRI