Genomic and Epigenomic Analysis of HPV positive and HPV negative Head and Neck Squamous Cell Cancer

Matthias Alexander Lechner, MD MRCS DOHNS

UCL Cancer Institute, 72 Huntley Street, London, WC1E 6BT

A thesis submitted to
University College London

for the degree of
Doctor of Philosophy

2012

Declaration

The dissertation is the result of my own work and contains nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

I also declare that this dissertation is not substantially the same as any that I submitted for a degree or diploma or other qualification at any other University, and that no part has already been, or is concurrently being, submitted for any degree, diploma, or other qualification.

This dissertation comprises a total of 136 pages, excluding references and appendices. The dissertation does not exceed the word limit of 100,000 words, excluding tables, figures, references, and appendices.

Matthias Alexander Lechner

Acknowledgements

I would first like to thank my supervisors Prof. Stephan Beck and Prof. Chris Boshoff for their invaluable guidance, great enthusiasm, support, time and continuous feedback during my research work at UCL, writing this thesis and publishing the papers contained within.

I would like to thank members of the Medical Genomics Group (in particular Dr. Lee Butcher, Dr. Andrew Feber, Dr. Chrissie Thirlwell, Dr. Helena Caren and Paul Guilhamon), Viral Oncology Group (in particular Dr. Tim Fenton), and Experimental Cancer Medicine Centre (ECMC) for teaching me various lab techniques which I used in my work. I would like to thank UCL Genomics and UCL Advanced Diagnostics for their technical support.

My grateful thanks are due to Dr. Andrew Teschendorff and Dr. Gareth Wilson for their guidance, help and advice on the biostatistical planning and analysis of the methylation data. I would also like to thank James West for his advice on R/Bioconductor related issues and Dr. Stephen Henderson for his guidance on the multidimensional data analysis.

I would also like to thank my collaborators at Foundation Medicine, in particular Dr. Garrett Frampton, Dr. Gary Palmer, Dr. Maureen Cronin, Dr. Doron Lipson, Dr. Vincent Miller, Dr. Roman Yelensky, and Dr. Philip Stephens.

I would like to thank Dr. Juan Martin-Serrano (Department of Infectious Diseases, King's College London School of Medicine, London), Dr. Susanne Gollin and Dr. Theresa Whiteside (University of Pittsburgh Cancer Institute, US), Dr. Hans Joenje (VU Medical Centre, Netherlands) and Dr. Thomas Carey (University of Michigan, US) for the provision of plasmids and HNSCC cell lines.

I appreciate the invaluable help of Dr. Amrita Jay and the team at the Department of Histopathology at UCLH who helped me to select the FFPE tumour samples used for the methylation and genomic analysis. I owe a vote of thanks to Dr. Nischalan Pillay, who interpreted some of the immunohistochemistry results together with Dr. Amrita Jay. I also owe a great vote of thanks to the Head and Neck Cancer MDT coordinator Ciaran Mariner, who greatly facilitated the gathering of the relevant clinical data. Moreover, I would like to thank the team of head and neck surgeons at UCLH who

continuously supported the sample collection and for their enthusiasm for this project and lastly all the patients at UCLH who participated in this study.

Furthermore, I would like to thank my clinical supervisor Mr. Francis Vaz (University College London Hospital) for his invaluable clinical teaching and guidance during the years of my PhD and for having acted as a role model of a very successful ENT surgeon. He has greatly supported me to stay involved in clinical work at UCLH along with his colleagues Mr. Paul O'Flynn Mr. Nicholas Kalavrezos, Mr. Colin Liew, Mr. Colin Hopper at UCLH together with Prof. Martin Birchall at the Royal National Throat, Nose and Ear Hospital, who I would like to thank for their teaching, their hints when approaching a clinical problem and their support during this busy time.

I owe a great vote of thanks to Dr. Martin Forster (UCL/UCLH) and once again to my supervisors, who supported me to get involved in graduate student teaching at UCL, as a lecturer and tutor on the MSc Cancer programme (on the modules 'Bed-to-Bedside' and 'Cancer Epigenetics'). I also owe a great vote of thanks to two of my UCL course tutors, Angela Cooper and Jackie Watson, and to Dr. Julie Olszewski (graduate tutor at the UCL Cancer Institute).

I would like to very much thank the funding bodies which made it possible for me to conduct the presented work, namely the Wellcome Trust for the award of a Clinical Research Fellowship (WT093855MA), the UCLH/UCL Comprehensive Biomedical Research Centre and the Austrian Science Fund for their invaluable initial support by awarding me an Erwin-Schroedinger Fellowship (J2856).

Finally, I owe a huge debt of gratitude to my family who have supported me throughout my studies and career to date.

I dedicate my work to my late grandfather August Steinleitner.

The most beautiful thing we can experience is the mysterious.

It is the source of all true art and all science.

He to whom this emotion is a stranger,

who can no longer pause to wonder and stand rapt in awe,

is as good as dead: his eyes are closed.

Albert Einstein

Contents

Abstract

Human Papillomavirus positive (HPV+) head and neck squamous cell carcinoma (HNSCC) represents a distinct clinical and epidemiological entity compared with HPV negative (HPV-) HNSCC. In this thesis I conducted both an epigenomic and genomic analysis to test the possible involvement of epigenetic modulation by HPV in HNSCC and associated genetic changes.

Using laser-capture microdissection of formalin-fixed paraffin-embedded (FFPE) HNSCCs, I generated both DNA methylation and genetic profiles of HPV+ and HPV– samples. I then used an independent clinical sample set and HPV+ and HPV- HNSCC cell lines for the validation of the obtained methylation data by two independent methods (Infinium 450k BeadArray and MeDIP-seq). Paired end sequencing of captured DNA, representing 3,230 exons in 182 genes often mutated in cancer was applied for mutation profiling. I validated the latter findings by Infinium copy number variation (CNV) profiling, Sequenom MassArray sequencing and immunohistochemistry.

Significant differences in the methylation and genomic profiles between HPV+ and HPV- HNSCC were observed. Methylation analysis revealed a hypermethylation signature with involvement of Cadherins of Polycomb group target genes in HPV+ HNSCC samples. Integration with independent expression data showed strong negative correlation, especially for the Cadherin gene family members. Combinatorial ectopic expression of the two HPV oncogenes (*E6* and *E7*) in an HPV– HNSCC cell line partially phenocopied the hypermethylation signature observed in HPV+ HNSCC tumours and established *E6* as the main viral effector gene. Moreover, MeDIP-Seq data revealed methylation sites within integrated HPV genomes. These methylation sites were confirmed by bisulfite sequencing, both in HNSCC samples and HPV+ HNSCC cell lines. Validated genomic changes clustered HPV+ and HPV- oropharyngeal carcinomas into two distinct subgroups with *TP53* mutations detected in 100% of HPV- cases. Abrogation of the G1/S checkpoint by *CCND1* amplification and *CDKN2A* loss occurred

in the majority of HPV- tumours, indicating that trials with CDK inhibitors in this disease subtype may be warranted.

My data establish archival FFPE tissue to be highly suitable for these types of methylation and mutation analysis and suggest that HPV modulates the HNSCC epigenome through hypermethylation of Polycomb repressive complex 2 target genes such as Cadherins which are implicated in tumour progression and metastasis. Moreover, my findings reinforce the causal role of HPV in oropharyngeal cancer and indicate that therapeutic stratification according to somatic genomic changes, in addition to HPV status, could be the most appropriate future approach for these cancers.

Publications

The majority of the presented work in this thesis has been published, submitted for publication or is in preparation. The following publications are based on or related to work in this thesis:

1.) Genome-wide DNA methylation analysis of archival formalin-fixed paraffin-embedded tissue using the Illumina Infinium HumanMethylation27 BeadChip. C. Thirlwell, M. Eymard, A. Feber, A.E. Teschendorff, K. Pearce, M. Lechner, M. Widschwendter, S. Beck; *Methods*. 2010, 52(3):248-54.

2.) Cancer Epigenome. M. Lechner et al.; *Advances in Genetics.* 2010, Vol. 70, Chapter 9; pages 247-276.

3.) Identification and Functional Validation of HPV-mediated Hypermethylation in Head and Neck Squamous Cell Carcinoma. M. Lechner et al.; submitted for publication.

4.) A Beta-Mixture Quantile Normalisation method for correcting probe design bias in Illumina Infinium 450k DNA methylation data**.** A.E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero and S. Beck; Accepted for publication in *Bioinformatics*.

5.) Targeted Next-generation Sequencing of Head and Neck Squamous Cell Carcinoma identifies novel genetic alterations in HPV+ and HPV- tumours. M. Lechner, G. Frampton, et al.; submitted for publication.

6.) MeDIP-Seq Analysis of the HPV-16 Methylome in Head and Neck Squamous Cell Cancer. M. Lechner et al.; in preparation.

7.) Invited Mini-Review entitled 'The Genomics and Epigenomics of HPV+ and HPV- Head and Neck Cancer – Taking a broader look'; M. Lechner et S. Beck; in preparation for publication in *Biochemical Society Transactions*.

In many instances, I use the same figures as used in the publications, or a re-formatted version to make use of the greater space allowed in this PhD thesis, compared with a journal research publication, or to include new developments in the field.

Facilities

1.) UCL Cancer Institute, University College London, 72 Huntley Street, London, WC1E 6BT, UK.

2.) Department of Histopathology, University College London Hospitals NHS Trust, Rockefeller Building, University Street, London, WC1E 6JJ, UK (Honorary appointment).

3.) Head and Neck Centre, University College London Hospitals NHS Trust, Euston Road, London, NW1 2PG, UK (Honorary appointment).

4.) Foundation Medicine, One Kendall Square, Suite B3501, Cambridge, MA 02139, US.

Keywords

human papillomavirus, HPV; head and neck squamous cell cancer, HNSCC; DNA methylation; methylome; epigenome; genome; HPV methylome.

Abbreviations

5-AZA … 5-Aza-2'-deoxycytidine

BS-Seq … bisulfite sequencing

CBS … circular binary segmentation

CDKN1A … cyclin-dependent kinase inhibitor 1A (p21, Cip1)

CDKN2A … cyclin-dependent kinase inhibitor 2A (p16)

CGP … Cancer Genome Project

CNA … copy number alteration

CNV … copy number variation

CIMP … CpG Island Methylator Phenotype

DMR … differentially methylated region

DNMT … DNA methyltransferase

EMT … epithelial to mesenchymal transition

FDR … false discovery rate

FF … fresh-frozen

FFPE … formalin-fixed paraffin-embedded

GEO … Gene Expression Omnibus

GSEA … Gene Set Enrichment Analysis

HEP … Human Epigenome Project

HPV … human papillomavirus

HPV+ … HPV positive

HPV- … HPV negative

HNSCC … head and neck squamous cell cancer

ICGC … International Cancer Genome Consortium

IHEC … International Human Epigenome Consortium

LCM … laser-captured microdissection

LCR … long-control region

MAD estimator … median absolute deviation estimator

MALDI-TOF MS … Matrix Assisted Laser Desorption Ionization Time of Flight Mass Spectrometry

MeDIP-Seq … methylated DNA immunoprecipitation coupled with NGS

MSP … methylation-specific PCR

MVP … methylation variable position (hyper-MVP or hypo-MVPs used wherever directionality towards differential hyper- or hypomethylation has been ascertained)

NGS … next-generation sequencing

PcG … Polycomb group protein

PRC2 … Polycomb repressive complex 2

PCA … principal component analysis

RB1 … Retinoblastoma Protein 1

RMT … Random Matrix Theory

SEER database … surveillance, epidemiology, and end results database

TCGA … Cancer Genome Atlas

WGA … whole genome amplification

List of Figures

List of Tables

Table 6.1: Percentage of somatic mutations (and genomic alterations in the cases of HNSCC) found within the listed genes.

Chapter 1

General Introduction

This chapter provides an introduction into cancer genomics and epigenomics, the background to head and neck cancer and gives an overview of the literature that has bearing on the work in this thesis, having led me to the formulations of the hypotheses and objectives of the presented work.

1.1 The role of HPV in Head and Neck Cancer.

Head and neck cancer is the sixth most common cancer worldwide with an incidence of around 600,000 cases per annum, with rising trends particularly in the young [1, 2]. Despite recent advances in the treatment and in the understanding of its biology, the 5-year survival rate of 50% for patients with head and neck cancer on the whole has remained largely unchanged for the past three decades with relatively few advances since the 1990s [3]. The most common type of head and neck cancer is squamous cell carcinoma (HNSCC).

As an umbrella term, HNSCC includes cancers found at several locations, e.g. in the oral cavity, pharynx and larynx. Although these cancers have different aetiologies and prognoses, they share similar risk factors. These major risk factors include the smoking and chewing of tobacco as well as alcohol consumption [1]. Heavy smokers under the age of 46 have a 20-fold increased risk of oral and pharyngeal cancer, whereas heavy drinkers are reported to have a 5-fold increased risk [4]. A combination of heavy smoking and drinking leads to an almost 50-fold increased risk of oral and pharyngeal cancer (ibid). However, a proportion of HNSCC patients (up to 20%) have

little or no tobacco or alcohol exposure, indicating other risk factors for the disease.

Over recent years human papillomavirus (henceforth HPV) has been shown to represent a major independent risk factor for HNSCC [5, 6]. HPV is particularly associated with oropharyngeal carcinoma (HNSCC of the tonsils and base of tongue) of which 20-50% test positive for the HPV-16 subtype [5-7]. HPV-positive HNSCC appears distinct from HPV-negative HNSCC with regard to the genetic mutation and gene expression status as well as to epidemiological factors and clinical features, leading to the conclusion that HPV-positive HNSCC represents a distinct molecular, epidemiologic and clinical entity [8, 9].

Epidemiological differences include the fact that HPV+ HNSCC is also increasing in the developed Western World and especially in younger age groups. Both HPV+ and HPV- HNSCCs occur more frequently in men at a rate of 3 to 1 compared to women [1, 2]. The increasing rate of squamous cell cancer of the tonsils and of the tongue base is particularly evident in patients under the age of 45 and this group has experienced a 2% per year increase in base of tongue cancer and a 4% per year increase in tonsillar cancer between 1973 and 2004. In contrast, the rates for cancers at all other head and neck sites during the same period remained constant or decreased [10]. This trend is also underscored by a further study investigating archived tonsil cancer specimens: HPV DNA was isolated in 23% of the specimens from the 1970s, 28% of the specimens of the 1980s, 57% from the 1990s, and 68% of specimens since 2000, showing an almost 3-fold increase [11]. A recently published study by the same research group suggests that the incidence of HPV-associated tonsillar carcinoma has even increased further from 2003 to 2007 (with 68% 2000-2002 to even 93% 2006-2007) with a parallel decline of HPV-negative tumours. Taking into account the previously presented data the incidence of HPV positive tonsillar tumours has almost doubled each decade between 1970 and 2007, suggesting an epidemic of viral-induced carcinoma [12].

HPV related head and neck cancer more frequently occurs in younger patients (age <50 years) with minimal tobacco and/or alcohol exposure but greater exposure to oral sex, multiple sexual partners, or marijuana [13]. Compared to non-HPV associated head and neck cancer patients, HPV-positive head and neck cancer patients are often of higher socioeconomic status with better nutritional status, dentition and overall health compared to patients suffering from HPV-negative head and neck cancer. Racial differences for HPV related head and neck cancer have also been demonstrated. In their study, Settle et al. found that 35% of Caucasian patients developed HPV-positive tumours compared with only 4% of patients of African origin, explaining the differences in overall survival observed between the two groups [14].

In summary, the epidemiological trend suggests that HPV-related oropharyngeal squamous cell cancer is becoming more widespread at a time when the overall incidence of HNSCC is falling and this subtype typically occurs in younger patients (age <50 years).

Moreover, this has major clinical implications. HPV+ HNSCC responds better to chemotherapy and radiotherapy (82% versus 55% response rate for HPV– cases) and has a better disease-free and overall survival (95% versus 62% at 2 years) [15]. Individuals with HPV+ HNSCC have a lower rate of second primary tumours, as well as a decreased cumulative incidence of relapse [16, 17]. Thus, knowledge of a patient's HPV status offers the possibility to stratify such patients for treatment and to elucidate the mechanisms underlying the virus-associated advantage in drug response and survival in HNSCC.

A summary of the distinct molecular, epidemiological and clinical features is provided in Table 1.1.

The causes responsible for the different clinical behaviour between HPV+ and HPV– tumours remain poorly understood. Numerous studies comparing gene expression patterns of HPV+ and HPV– cancers have shown different profiles between the two groups [18-22]. It is therefore likely that virus-

mediated changes in both the genome and epigenome account for the differing clinical behaviour.

	HPV+ HNSCC	HPV- HNSCC
Molecular factors	• TP53 wild-type present in most cases (if TP53 mutations are present, they are usually not disruptive [23]) • TP53 loss of function (rendered by the viral oncoprotein E6) [24]. • RB1 loss of function (rendered by the viral oncoprotein E7) [25] • CDKN2A(p16) overexpression [26] • HPV DNA (type 16 in up to 85% of cases) [5, 27, 28]	• TP53 mutational loss (disruptive) common [23] • CDKN2A(p16) underexpression [26] or mutational loss [26, 29] • No HPV DNA/RNA
Possible epidemiological factors	• Often non-smokers [5, 13] • Mild/moderate alcohol intake [5, 13] • High marijuana exposure [13] • Intact dentition [13] • High oral sex exposure [5, 13] • Younger age (<45 years) [30, 31] • Increasing incidence of the disease [12, 32]	• Heavy smokers [5, 13] • Heavy alcohol intake [5, 13] • Low marijuana exposure [13] • Poor dentition [13] • No association with sexual behaviours [5, 13] • Older age (>50 years) [30, 31] • Decreasing incidence of the disease [12, 32]
Clinical factors	• Predominantly oropharynx (tonsil and tongue base) [5-7] • Better survival [15, 33] • More radiosensitive [15]	• All head and neck sites [5-7] • Worse survival [15, 33] • Radiation response unpredictable [15]

Table 1.1: Distinct molecular, epidemiological and clinical features of HPV+ HNSCC, compared with HPV- HNSCC (adapted from [9]).

1.2 HPV-driven carcinogenesis – Current models in head and neck cancer.

Human papillomavirus is a ubiquitous, nonenveloped, circular, double-stranded DNA virus. It is believed that HPV represents the most common cause of cancers induced by an infectious agent [34]. Of the over 30 - 40 known types of HPV that are typically transmitted by sexual contact (of more than 100 types of human papillomavirus; see Appendix Figure A.1), mainly 'high risk' types (at least 15), such as HPV type 16 and 18, have been shown to promote malignant transformation. In HPV related head and neck cancer HPV type 16 is most frequently found (in over 85% of cases), thus representing the most important HPV subtype in this form of cancer [41], [28].

The viral genome (Figure 1.1) which uses the host DNA machinery to replicate and proliferate is around 8000 base pairs in length and encodes for only a few gene products: 6 early proteins (incl. the oncoproteins E6 and E7, proteins required for DNA amplification, such as E4 and E5 and the two regulatory proteins E1 and E2, which also play a role in DNA replication) as well as two late proteins (L1 and L2, representing the two structural capsid proteins) [35]. The HPV life cycle is complex and coupled to the cellular differentiation programme that occurs in the epithelium (Figure 1.1).

Through microabrasions in the skin and mucous membranes, HPV infects cells in the basal layer of the stratified epithelium where they establish their double-stranded DNA genome as a circular extrachromosomal element in the nucleus of infected cells (nuclear episome). In cervical lesions DNA is thought to be present at low copy numbers with 50-100 copies per infected cell during the early stage of infection (required to sustain the viral life cycle and its associated pathologies) [36]. In the upper layers of the epithelium HPV DNA is then amplified to greater than 1000 copies per cell [37], having been tested in raft cultures of human keratinocytes. In the terminally differentiated epithelial cells the capsid proteins are expressed and progeny virions are assembled. These virions are then shed [25].

Figure 1.1: Genome of HPV type 16 (A) and illustration of the virus life cycle (B). The above figure was obtained from [25]. (A) Early (p97) and late (p670) promoters are marked by arrows within the HPV type 16 genome (7904 bp in length). The six early ORFs (E1, E2, E4 and E5 are shown in green and E6 and E7 in red) are expressed from either p97 or p670 at different stages during epithelial cell differentiation, whereas the late ORFs (L1 and L2, shown in yellow) are only expressed from p670 (B). Within the long control region (7156–7184 bp) E2-binding sites and the TATA element of the p97 promoter are illustrated. (B) Key events following infection are shown diagrammatically within the epidermis (both in cutaneous tissue and within mucous membranes). Virus infection, and in particular, the expression of the viral oncogenes, E6 and E7, lead to the appearance of cells expressing cell cycle markers in the basal layer of the epidermis (shown with red nuclei). In these cells the expression of viral proteins necessary for genome replication then follows the activation of p670 in the upper epithelial layers (cells illustrated in green with red nuclei). Expression of the L1 and L2 genes (shown in yellow) occur only in a subset of the cells that contain amplified viral DNA in the upper epithelial layer (Figure Legend adapted from [25])

Integration of the virus is frequently observed in HPV-associated cancers [38, 39]. Integration is thought to disrupt the regulation of E6 and E7 expression and to subsequently lead to higher expression levels of these oncoproteins with the effects described above [25].

HPV+ HNSCC is characterized by distinct molecular pathways which are different from those involved in HPV- HNSCC [8]. The important role of high-risk HPV oncoproteins, in particular E6 and E7, in malignant transformation and the maintenance of this malignant phenotype is well recognized, having been described in cervical cancer first [25].

Current model of HPV+ and HPV- HNSCC carcinogenesis:

As described above, integration of the HPV into the host genome is found in HNSCC and the HPV genome is interrupted by the integration event. Upon integration E5 function is frequently lost and thus, E5 is not regarded as being a mediator in late stage carcinogenesis [40]. However, the HPV E5 protein is thought to play an important role during the early stage of infection by binding epidermal growth factor receptor (EGFR), platelet derived growth factor receptor beta polypeptide (PDGFRB) and colony stimulating factor 1 receptor (CSF1R), thereby promoting cell proliferation [41]. The same applies to the E2 gene which lies in the same region and is a potent inhibitor of E6 and E7 viral gene expression. E2-mediated inhibition is also frequently lost upon integration of HPV into the host genome, as described in cervical cancer [42].

The main mechanism by which HPV causes cancer is the expression of the two viral oncoproteins E6 and E7. These oncoproteins degrade and destabilize two major tumour suppressor proteins of the host, tumour protein p53 (TP53) and retinoblastoma 1 (RB1), through ubiquitination. The HPV E6 protein forms a complex with an E3 ubiquitin ligase (UBE3A, also referred to as E6AP), and ubiquitinates TP53 [43-45]. This ubiquitination leads to rapid degradation of TP53, which results in deregulation of both the G1/S and G2/M cell cycle checkpoints during cell damage and other cellular stress, which leads to genomic instability [46]. The HPV E7 oncoprotein ubiquinates RB1 by binding to the cullin 2 ubiquitin ligase complex [47-49]. Ubiquination

of RB1 results in its degradation leading to deregulation of the G1/S phase of the cell cycle. The absence of RB1 then causes the release of the E2F family of transcription factors with the transcription of S-phase genes, resulting in cell proliferation [50].

A summary of the cell cycle deregulation by human papillomavirus is illustrated in Figure 1.2 (from [51]).

Figure 1.2: Cell cycle deregulation by human papillomavirus (adapted from [51]). Effects of the two main viral oncoproteins on TP53 and RB1 function and cell-cycle regulation are illustrated.

Moreover, studies have shown that E6 and E7 can directly bind to numerous host proteins other than TP53 and RB1. These include, for example, BCL2-antagonist/killer 1 (BAK1) and cyclin-dependent kinase inhibitor 1A (CDKN1A; also referred to as p21 or Cip1) [51]. This further contributes to genetic instability. However, many more genetic alterations are necessary for malignant progression in the setting of virus-induced genomic instability. Most of these are currently unknown [52].

Looking at the most well studied example of HPV-induced cancer, cervical cancer, the period of time from the initial infection with a high-risk type of HPV to malignant transformation and the development of cervical cancer is typically between 10 and 20 years [53].

In contrast to this, HPV- HNSCC takes around 30 – 40 years to evolve, with multiple genomic changes and epigenomic changes in key pathways thought to be acquired by the exposure of the respective epithelium to carcinogens, instigated by smoking or alcohol abuse, and chronic inflammation. These genomic and epigenomic alterations are discussed in detail in section 1.3.

A current model of the molecular carcinogenesis of HNSCC is illustrated in Figure 1.3 (from [51]).

Figure 1.3: Genetic and epigenetic changes currently regarded key steps in the molecular carcinogenesis of HNSCC (obtained from [51]). This figure illustrates a multi-step model of head and neck carcinogenesis via precursor lesions. The progression from normal tissue to metastatic cancer is associated with an accumulation of specific genetic and epigenetic changes.

1.3 Genetics and Epigenetics of Head and Neck Cancer.

Cancer is a heterogeneous disease caused largely by the acquisition of biological capabilities, rendered by changes in the genome and the epigenome in a partly stochastic and partly directed multi-step process, with underlying genomic instability. These capabilities were referred to as hallmark capabilities by Hanahan and Weinberg [54] and are outlined in Figure 1.4.

Typically, genetic changes such as mutations and other genomic rearrangements or epigenetic changes such as aberrant DNA methylation and histone modifications underlie such abnormalities and are mediated by exposure to environmental or life-style factors.

Therefore, comprehensive genetic and epigenetic analysis of cancer genomes is the most effective way to identify causative changes involved in tumourigenesis, irrespective of whether they are inherited or acquired.

Figure 1.4: The 10 hallmark capabilities of cancer (obtained from [54]).

1.3.1 The Cancer Epigenome of Head and Neck Cancer

Oncogenic loci in HNSCC encompass genes which are mutated (and amplified/deleted) or epigenetically compromised (epimutation), for instance, by promoter methylation. These epigenetic changes at oncogenic loci associated with progression to carcinoma, particularly in HPV+ HNSCC, are not well characterized.

Recent progress in technology now enables the study of these epigenetic variations at the genome level. The obtained profiles are called epigenomes and in the context of cancer are referred to as cancer epigenomes.

’Epigenome’ is defined as the complete set of chemical modifications of chromatin constituents and ‘methylome’ as the complete set of DNA methylation modifications of the genome. [55] Additional non-covalent modifications such as micro RNAs and chromatin remodelling complexes also modulate the epigenome (Figure 1.5).

Compared to the genome, the epigenome is much more dynamic, reflecting many different functional states in time and space. These dynamics are governed in part by reversible covalent modifications of the epigenome such as DNA methylation and histone tail modifications.

Figure 1.5: Illustration of chemical modifications of chromatin constituents (Adapted from [56])

Epigenetic processes may affect both the viral genome itself and the host cell methylome. Modification of normal epigenetic processes in the host occurs during viral infection: for example, compared with HPV-negative HNSCC, HPV+ tumours show less genome-wide hypomethylation [57]. These epigenetic changes may occur through direct interactions between HPV proteins (e.g. E2, E7) and proteins involved in DNA methylation. Viral oncoproteins (in particular E7) can bind and regulate DNMT1 enzymatic

activity [58]. Moreover, interactions between E7 and components of the chromatin remodelling machinery, e.g. p300, are also described [59]. It has also been shown that E7 induces KDM6A and KDM6B histone demethylase expression and causes epigenetic reprogramming [35].

This research work is set in an environment in which the analysis of cancer genomes and epigenomes is well underway. This era of cancer genome and epigenome projects as part of multiple national and international efforts [56], [60] represent an important endeavour to understand the genetic and epigenetic basis of malignant disease. The recent endeavours and achievements, which have been massively boosted by decreasing sequencing costs, are summarized in Figure 1.6.

Figure 1.6: Landmark initiatives towards comprehensive analysis of cancer (epi)genomes. The emergence of next-generation sequencing (indicated by blue shading) since 2004 has clearly galvanized the field (adapted from [55])

Past and present projects include the Human Epigenome Project (HEP), Cancer Genome Project (CGP), The Cancer Genome Atlas (TCGA), NIH Roadmap Epigenomics Program, International Cancer Genome Consortium

(ICGC), and International Human Epigenome Consortium (IHEC) (recently summarized in [55]). Head and neck cancer is now also one of the featured cancers in the Cancer Genome Atlas (TCGA) project, a main focus in extensive epidemiological studies, i.e. International Head and Neck Cancer Epidemiology (INHANCE) consortium [61] and future collaborative projects on head and neck cancer, such as the 'Enabling Head and Neck Cancer Therapy via Novel Biomarkers (ENHANCER)' project.

Not only will these projects identify a whole-genome map of genetic and epigenetic changes that determine the phenotype of the cancer investigated, but also lead the development of novel diagnostic and prognostic markers, of drugs specifically directed to the described genetic and epigenetic changes (targeted and combinatorial therapies) and gain the necessary information to direct the therapeutics effectively and safely to the identified targets.

The reversible nature of epigenetic aberrations has already led to the emergence of the promising field of epigenetic therapy (with the two groups of drugs targeting the epigenome, DNA methylation inhibitors and Histone deacetylase (HDAC) inhibitors). The aim of this new treatment option is to reverse epigenetic changes which contribute to carcinogenesis (driver epimutations) and to restore a 'normal epigenome'. Some agents of this class of therapeutics (although not yet target-specific) have recently been approved by the FDA for cancer treatment and can be expected to play a major role in epigenetic cancer therapy in the future [62].

Analysis of the epigenome is more complex than that of the genome and the majority of studies have therefore focused on the methylome as DNA methylation is the most accessible epigenetic modification in clinical samples [63].

DNA methylation refers to the addition of a methyl-group to cytosine (Figure 1.7), catalysed by DNA (cytosine-5-)-methyltransferases (DNMTs). In mammals, it plays a key role in regulating gene expression and other biological processes such as X-chromosome inactivation in women [64], parent-of-origin-specific gene expression (imprinting) [65] and epigenetic reprogramming during mammalian development [66]. It occurs on cytosines

not only at CpG sites (mostly clustered in ‘islands’ that are predominantly found at gene promoters or in nearby ‘shores’ – regions up to several kilobases up- or downstream of the promoter) but also at non-CpG sites (as recently described [67]) across the genome.

Figure 1.7: The catalysis of 5-methylcytosine from cytosine by DNA methyltransferase (DNMT); adapted from [68].

To date most studies have focused on DNA methylation at these gene promoters and CpG islands only, as many technologies covered these efficiently, such as restriction landmark genomic scanning (RLGS) [69, 70] and CpG island microarrays [71]. However, recent evidence suggests that areas outside these regions – such as the methylation of CpG island shores and shelves [72, 73] or of cytosines in non-CpG sites play key roles in cellular differentiation and thus malignant transformation. It remains unclear whether methylation within CpG sites in the ‘open sea’ has any significance. These areas of the genome were not covered by former methods, either due to the fact that they were not global enough or biased toward CpG islands. Only fairly recently have efficient technologies become available for genome-wide methylation analysis with the introduction of the microarray and later second-generation sequencing, exemplified by the Roche *454* platform [74],

the Illumina *GA* platform [75] and the Applied Biosystems *SOLiD* platform [76]. These next-generation sequencing techniques have been successfully combined with different capture techniques for methylated DNA, resulting e.g. in AutoMeDIP-Seq [77] and MBD-seq [78]. These methods are now capable of covering around 70% of the >28 million CpG sites across the haploid human genome (at 50-100 bp resolution). Another method based on bisulfite sequencing allows methylome analysis at single base resolution [67]. In this thesis I applied MeDIP-Seq [77] which allowed me to interrogate approximately 70% of the host methylome of HNSCC and 100% of the HPV methylome. One of the main limitations of these techniques was that until recently these technologies were very expensive.

Thus, until recently microarrays have been the most widely used tool for comprehensive DNA methylation analysis. Advantages of using microarrays include ease of use, high levels of multiplexing and parallel processing to facilitate high-throughput and cost-effective analysis as well as the possibility of capturing specific target regions for subsequent array- or sequencing based analysis [79]. Being essentially an analogue technology, disadvantages include lower sensitivity and specificity than those achievable by digital technologies such as sequencing. Nevertheless, the first methylomes to be reported were generated using microarray-based approaches [80], [81]. However, the introduction of next-generation sequencing is increasingly replacing and will continue to replace microarrays in the coming years. A likely exception is so-called beadarrays (e.g. in conjunction with the Illumina Infinium assay) which allow highly sensitive and quantitative methylation analysis at the single base-pair level [82]. In 2009 Illumina announced the release of its Infinium HumanMethylation450 BeadChip which was a big step towards its application for more comprehensive methylome analysis, based on Infinium® Assay used for analysis of CpG methylation using bisulfite-converted genomic DNA. This technology which was used for the host methylome analysis of HNSCC samples and for the validation of the obtained results, allowed me to interrogate > 485,000 CpG (methylation) sites per sample at a single nucleotide resolution. The content of CpG sites probed was selected

according to the guidance of a consortium of methylation experts, covering 99% of RefSeq genes, with multiple probes per gene region distributed across the transcription start site, 5'UTR, first exon, gene body, and 3'UTR. It covers 96% of CpG islands, with additional coverage in island shores, shelves and the open sea. It also covers a proportion of non-CpG methylated sites identified in human stem cells [83].

The main mechanism of epigenome-induced carcinogenesis is the silencing of tumour suppressor genes mediated by site-specific DNA hypermethylation. Such genes are heavily involved in cell cycle control, apoptosis, cell adhesion, DNA repair, and angiogenesis (e.g. *TP53*, *RB1, P16, PTEN, MLH1,* and *BRCA1*). Their loss of function leads to the gain of hallmark capabilities of cancer in human cells [84]. In addition to direct effects by DNA hypermethylation of genes, various indirect effects exist. These include the silencing of effector molecules, such as transcription factors that mainly effect downstream targets and thereby contribute to the process of malignant transformation. An example includes the silencing of the transcription factor *RUNX3* in oesophageal cancer [85]. The Knudson two-hit model in the context of DNA methylation describes a model in which one allele of a tumour suppressor gene is silenced, e.g. by somatic mutation, and the second allele is then hypermethylated, leading to a complete loss of gene function [86].

DNA hyopmethylation often results in the activation of oncogenes, thus initiating carcinogenesis. This mechanism has been described in various cancers, e.g. *R-RAS* and *MAPSIN* in gastric cancer, *S-100* in colon cancer and *MAGE* in melanoma [87]. In conjunction with histone modifications (e.g. the methylation at lysine residues 9 and 27 of histone H3), these epigenetic changes affect the chromatin structure through a cascade of chemical reactions and interactions with additional proteins and protein complexes.

In summary, DNA hypermethylation can result in the silencing of tumour suppressor genes and DNA hypomethylation in the activation of oncogenes, thereby significantly contributing to carcinogenesis.

Studies, describing the methylome analysis on a selected range of cancers, have been reviewed in a book chapter recently [88] and I will now mainly focus on the current state of affairs with regard to methylome analyses of head and neck cancer in the following section.

The role of DNA methylation in HNSCC:

As with other cancer types, numerous DNA methylation changes are detected in head and neck cancer, suggesting an important role of epigenetics in head and neck carcinogenesis.

Most studies reported to date, however, only assessed single methylation sites (mostly promoter regions) or covered tiny proportions of the cancer methylome. Studies which assessed the methylation status of single genes in HNSCC samples identified numerous tumour suppressor genes to be differentially methylated, e.g. APC [89], CDKN2A(p16) [89], MLH1 [90], and MSH2 [90]. Some of these, such as CDKN2A, were also shown to potentially have a diagnostic value when examined as a panel including further relevant genes in blood samples [91] and in oral and salivary rinses [91], [92]. However, I have identified only a few studies which investigated DNA methylation profiles in head and neck cancer samples. Using the Illumina Golden Gate assay (covering 1,536 CpG sites), Marsit et al. [93] showed that specific DNA methylation profiles in HNSCC are associated with certain risk factors such as tobacco smoking and alcohol abuse as well as tumour stage. Using combined bisulfite restriction analysis of the LRE1 sequence, it was shown that DNA hypomethylation in whole blood is associated with a 1.6-fold increased risk of HNSCC [94]. Only a very few studies can be identified which assess the methylation status of samples in the context of the HPV status of HNSCC. Using pyrosequencing, LINE hypomethylation was found to be more pronounced in HPV- than in HPV+ tumours. Genomic instability was found to be greater in HNSCC samples with more pronounced LINE hypomethylation [57]. A recent study [95] comparing two HPV+ with two HPV– HNSCC cell lines showed that HPV infection is associated with changes in methylation of various host genes. Their study identified five Polycomb Repressive Complex 2 (PRC2) targets among hypermethylated

promoters. Polycomb group (PcG) proteins are transcriptional repressors, modifying histone tails to reversibly suppress genes required for differentiation [96]. These proteins play a major role in neoplasia [97] and their oncogenic function is associated with a well-established role in stem cell maintenance. Stem cell PcG targets were shown to be 12-fold more likely to have cancer-specific promoter hypermethylation than non-targets [98-100], supporting the theory of a stem cell origin of cancer [101].

These described epigenetic changes that occur both in the host [102] and viral methylomes [103, 104], may occur through direct interactions between HPV proteins (e.g. E2, E7) and proteins involved in DNA methylation. Viral oncoproteins (in particular E7) can bind and regulate DNMT1 enzymatic activity [58]. Moreover, interactions between viral oncoproteins and components of the chromatin remodelling machinery, e.g. p300, are also described [59, 105].

In summary, the analysis of cancer epigenomes in the context of HNSCC is still at an early stage. Most of the studies published so far have only assessed methylation changes in the context of CpG methylation, predominantly investigating the promoter regions of genes.

Thus, the work presented in this thesis represents a significant step forward in the understanding of the epigenetic basis of head and neck cancer.

1.3.2 The Cancer Genome of HNSCC – Current State of Affairs

Early neoplastic events in up to 80% of HPV- HNSCC include the mutational loss of Cyclin-dependent kinase inhibitor 2A (CDKN2A; frequently referred to as p16) and TP53 (mainly mediated by significant alcohol and tobacco exposure), leading to uncontrolled cellular growth [106]. In most cases of HPV+ HNSCC, TP53 mutations are not found at all [23]. In these, however, unmutated wild-type TP53 is degraded by the E6 protein and CDKN2A (p16) is overexpressed [107]. Previous studies reported that HPV- tumours show losses at chromosomal regions 3p, 9p, and 17p, which are absent in HPV+ tumours [1]. It is noteworthy that 18q12.1-23 is gained in HPV+ tumours and lost in HPV- tumours [108]. Taking a broader view, HPV associated HNSCC show fewer genome-wide DNA copy number alterations [108], less genome-

wide hypomethylation [57], and lower expression of EGFR [109]. Two recent Science papers investigated the genomic profile of HNSCC, but both only included a very small number of HPV+ samples. Stransky et al. [110] looked at 74 samples (final number for analysis) of which 11 were from the oropharynx (7 HPV+ and 4 HPV-), not checking for copy number alterations. Agrawal et al. [111] tested 32 samples for HPV (4 HPV+ samples and 28 HPV- samples). The average number of mutations in HPV+ samples were 4.8 and in HPV- 20.6. The main copy number alterations observed in their study are illustrated in Table 1.2.

Chromosome location	Cytoband	Change	Number of samples
	3p, 8p, 13q, 18q, 21q	loss	>5
	3q, 5p, 8q, 18p, 20	gain	>5
chr9:21970150-21983344	9p21.3	loss	12
chr11:68903267-70541010	11q13.3-11q13.4	gain	7-11
chr3:167285555-181968731	3q26.1 - 3q26.33	gain	5-6
chr7:54959633-56069782	7p11.2	gain	3-4

Table 1.2: Copy number alterations observed in HNSCC (Adapted from Supplemental Table S.7 [111])

In summary, these studies showed that over 80% of tumours contain *TP53* mutations (a potential genomic stratifier for HPV status) and strikingly up to 20% have loss-of-function *NOTCH1* mutations. However, in these two studies, only seven and four HPV+ samples were included respectively. Both studies confirmed the lack of *TP53* mutations compared to HPV- samples, and overall, a lower mutational burden in HPV+ disease.

1.3.3 Gene expression changes between HPV+ and HPV- head and neck cancer

Five studies are identified which evaluated gene expression changes in HPV+, compared with HPV- head and neck cancers [18, 20-22, 26]. In their gene expression analysis, Jung et al. [26] found the most significantly differentially expressed gene to be cyclin-dependant kinase inhibitor 2A (CDKN2A or often referred to as p16), a known biomarker for HPV+ HNSCC

[112]. All studies show clear differences in the gene expression between HPV+ and HPV- HNSCC.

1.4 The Epigenome of HPV in HNSCC.

As described in section 1.1, HPV type 16 is most frequently found (in over 85% of cases) in HPV+ HNSCC and thus the focus shall lie on this high-risk type.

Not only do epigenetic processes affect the host genome, but also the viral genome itself. The HPV genome does not encode for any DNA methyltransferases (DNMT) and thus it is believed that the viral genome is methylated by the human host cell DNMTs [104]. This may either be directed by the virus through cross-talk between viral proteins and DNMTs (in order to regulate intrinsic cell-differentiation dependent and temporal viral gene expression) or it may represent a potential host defence mechanism, with the host aiming to alter viral gene expression once the virus integrated into the host genome. By micro-dissecting cell fractions and isolating DNA from the basal, intermediate, and superficial epithelial cell layers of HPV type 16 positive cervical biopsy sections, S. Vinokurova et al. detected differential methylation of transcription factor binding sites within the HPV 16 upstream regulatory region which were associated with the squamous epithelial differentiation stage. This is an indicator that methylation of the HPV type 16 genome is linked to differentiation dependent transcription and replication control, thus potentially triggering neoplastic transformation [113].

In cervical epithelium, HPV DNA hypermethylation has been observed to be more closely associated with cervical carcinoma, than with asymptomatic infection or with dysplasia [103, 114]. Methylation of the L1 gene of both HPV-16 and of HPV-18 has previously been reported in cervical cancer and in penile cancer [114-117].

Methylation changes within HPV-16 in HNSCC have been demonstrated by Balderas-Loaeza et al. [118]. Methylation of the viral genome near the 3’ end

of the L1 ORF (3 CpGs) and in the LCR (5 CpGs) was observed. The most extensive study of methylation of the HPV-16 genome in HNSCC was performed by Il-Seok Park et al. [119]. Using BS-Seq, they determined the methylation status of all 110 CpG sites within the viral genome in advanced stage HPV+ HNSCC. In most cases they found that the viral methylome was not methylated within the LCR, which controls the expression of the viral oncogenes E6 and E7. Methylation in other regions was detected to varying degrees [119].

1.5 Aims of my thesis and hypotheses tested in this work.

As outlined, the analysis of the cancer genome and epigenome of HPV+, compared with HPV- HNSCC, is still at an early stage. Most of the studies published to date have only assessed small proportions of the genome and methylome or included only a very small number of HPV+ cases which makes it impossible to draw definitive conclusions.

Thus, I have formulated the following aims of this work to make a significant step forward in the understanding of the genetic and epigenetic basis of head and neck cancer

Aims of the project:

The overall aim of this project was to understand the genetic basis of HNSCC and epigenetic alterations induced by HPV, contributing to HNSCC. I investigated the effects of HPV infection on the host methylome (human HNSCC), the viral methylome (i.e. HPV) and the specific genetic changes associated with both subtypes. In order to achieve this I aimed to perform:

1.) Comprehensive methylome analysis of HPV(+) and HPV(-) HNSCC, which informed me about detailed epigenetic changes unique to HPV associated head and neck cancer (both detected in the human genome as well as in the virus itself) and the validation of these results.

2.) Genomic profiling of the same samples which informed on differences in the mutational profiles between these two patient groups.

3.) *In vitro* experiments to confirm the obtained results, and to induce the observed changes in *in vitro* models of HPV(+) and HPV(-) cancer.

Results were linked to public gene expression array and whole-genome somatic mutation databases. Having integrated these data with public databases and having integrated my obtained data on HNSCC with those from other HPV-induced and smoking-induced cancers, it has advanced the current understanding of HPV+ HNSCC, compared with HPV- HNSCC, and helped to partly explain the better outcome and survival of patients suffering from this distinct subtype. In the longer-term, these data will be used to identify potential diagnostic and prognostic markers, as well as putative therapeutic targets.

My workplan is outlined in Figure 1.8: The chapters of my thesis are indicated which are dealing with the experimental approaches set out to answer the following research question and to test my hypotheses.

Research question and hypothesis:

What is the role of HPV infection in oropharyngeal carcinoma? In particular, my thesis investigated the genetic and epigenetic alterations in HPV+, compared with HPV-HNSCC. My project also investigated the regulatory role of such changes, and the methylation status of HPV open reading frames (henceforth ORFs) in positive tumours. My thesis addressed whether such genetic and epigenetic differences could contribute to the better overall responses to chemo-radiotherapy, and better overall outcome, of HPV+ HNSCC.

My hypothesis was that HPV+ HNSCC, compared with HPV- HNSCC, has a distinct genomic and epigenomic profile and that HPV modulates the host epigenome.

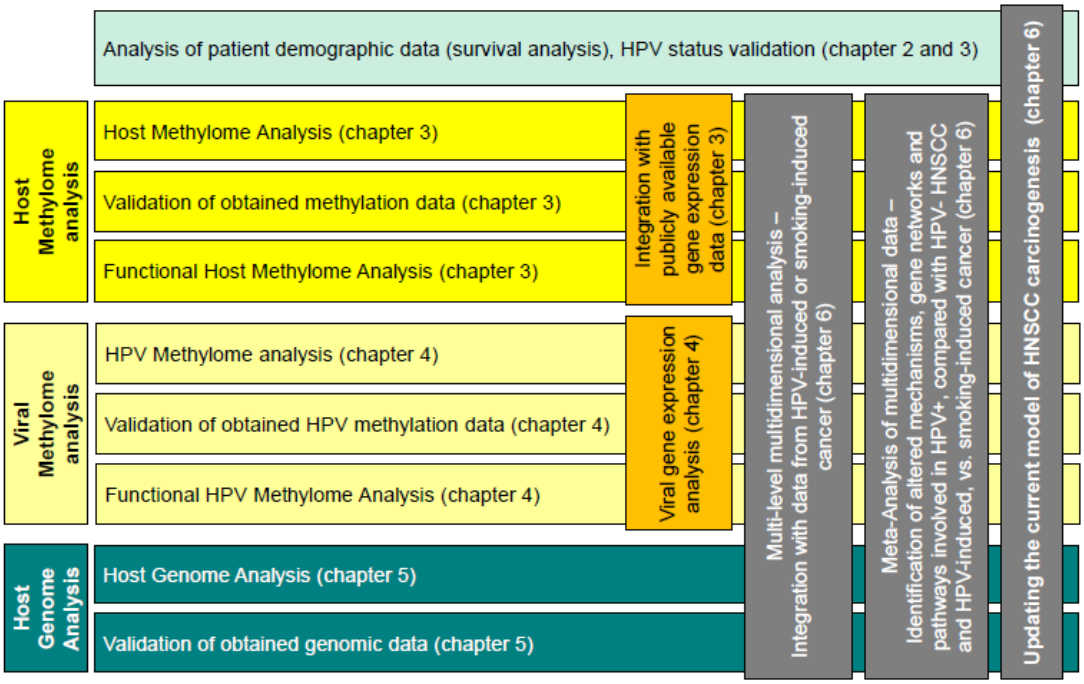


Figure 1.8: Outline of the work presented in this thesis.

Chapter 2

Materials, Methodology and Experimental Procedures

This chapter describes the materials, various methodologies and experimental procedures which were used in carrying out this research.

2.1 Materials.

2.1.1 Human Tissue Samples and Clinical Data:

Ethical approval for the study of human tissue samples was granted by the UCL/UCLH Ethics Committee (Reference number 04/Q0505/59). 107 archival formalin fixed paraffin embedded tissue (FFPE) oropharyngeal cancer samples were obtained from the Department of Histopathology at University College London Hospitals (UCLH). 82 of these samples were considered suitable for p16 staining after careful histopathological review. 82 samples (77%) were stained for p16 (CDKN2A). 28 oropharyngeal cancer samples stained positive (34.2 %). 46 oropharyngeal cancer samples stained negative (56.1 %). 8 oropharyngeal cancer samples showed a mixed staining – a tumour sample showed both positively and negatively stained areas (9.7 %). Laser capture microdissection was performed on 74 samples (90 %). The dissection and DNA extraction from these samples (yielding sufficient amounts of good DNA quality) was successful in 66 samples (89.2 %). E6 qPCR was performed on 66 samples. This method was assumed to be the gold standard for further calculation. 24 samples were positive for E6

amplification (36.4 %) and 40 samples were negative for E6 amplification (60.6 %). 2 samples showed inconsistent or borderline results in repeat measurements (3 %). 64 samples were used for the final analysis. Considering p16 staining and confirming the results by E6 qPCR (assuming that E6 qPCR is the gold standard) in a total of 64 samples: true positives: 22 (34.4 %); true negatives: 34 (53.1 %); false positives: 6 (9.4 %); false negatives: 2 (3.1 %). In my set of samples p16 staining had a sensitivity of 91.6 % and a specificity of 85 % (Table 2.1). The 22 p16 and E6 qPCR positive and 34 p16 and qPCR negative samples were selected for further analysis, according to the quality and the amount of the yielded DNA and gender and age-matching requirements (Workflow illustrated in Figure 2.1).

Assuming E6 qPCR to be the gold standard:
- True positives: 22 (34.4 %)
- True negatives: 34 (53.1 %)
- False positives: 6 (9.4 %)
- False negatives: 2 (3.1 %)

2x2 table

	E6 qPCR positive	E6 qPCR negative
p16 positive	22	6
p16 negative	2	34

p16 staining: Sensitivity: 91.6%; Specificity: 85%

Table 2.1: Calculation of sensitivity and specificity of p16 staining in the set of tested samples.

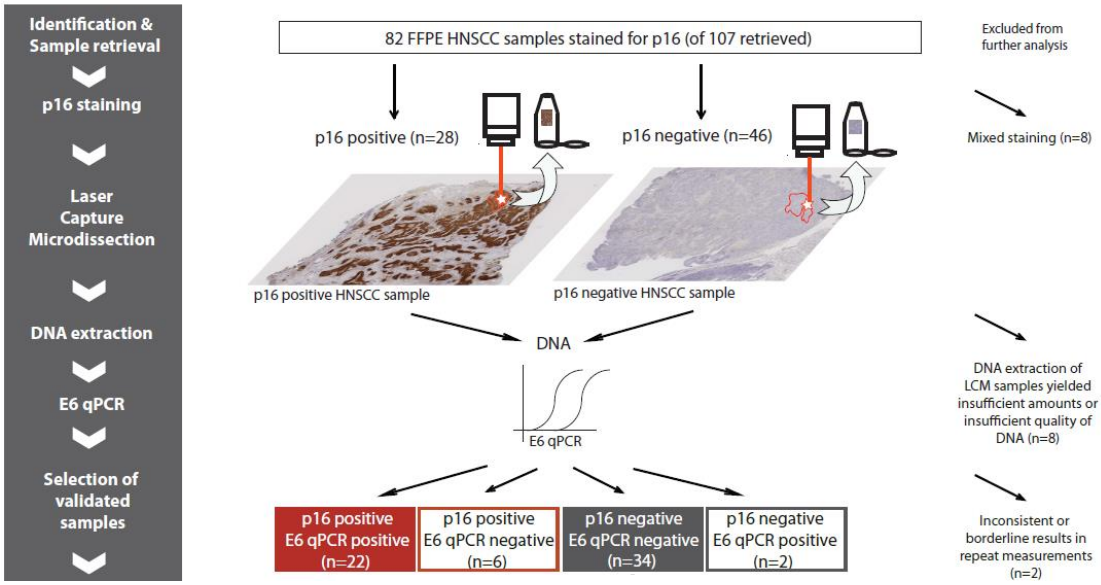


Figure 2.1: Workflow of FFPE sample preparation and selection. The 22 p16 and E6 qPCR positive and 34 p16 and qPCR negative samples were available for further analysis.

Furthermore, three fresh-frozen (FF) HPV+ and three FF HPV– HNSCC samples (Table 2.2) were obtained from the UCLH Head and Neck Tumour

Bank. These three fresh-frozen HPV+ and HPV- HNSCC samples (Table 2.3) were used for MeDIP-Seq analysis (incl. Illumina HumanOmni1-Quad BeadChip genotyping array for normalization of MeDIP-Seq data) as well as for Illumina Infinium 450k array analysis to validate FFPE findings. They were also used for the analysis and the validation of results of the viral methylome experiments.

	HPV+ (n=3)	HPV- (n=3)
Age	58.3 years (53-62)	73.6 years (66-80)
Gender	M: 2 F: 1	M: 3 F: 0
Tumour site	Oropharynx: 3	Oropharynx: 3
Tumour grade	Well diff: 0 Mod diff: 0 Poorly diff: 3	Well diff: 0 Mod diff: 1 Poorly diff: 2
Tumour stage (T)	T1: 0 T2: 3 T3: 0 T4: 0 N/a: 0	T1: 0 T2: 0 T3: 0 T4: 2 N/a: 1
Cervical lymph node involvement (N)	Yes: 3 No: 0 N/a: 0	Yes: 0 No: 2 N/a: 1
Smoking	Ever: 3 Never: 0 N/a: 0	Ever: 2 Never: 0 N/a: 1
Alcohol	Heavy drinker: 1 Occ. alcohol: 2 Never: 0 N/a: 0	Heavy drinker: 0 Occ. alcohol: 2 Never: 1 N/a: 0

Table 2.2: Patient characteristics of 3 HPV+ and 3 HPV- fresh frozen HNSCC samples.

Human Tissue samples used for the host methylation analysis (chapter 3):

According to the quality and the amount of the yielded DNA and gender and age-matching requirements 21 HPV+ and 21 HPV– age-matched samples were selected for methylation analysis. Histological diagnosis was confirmed by an experienced histopathologist and correlated with clinical findings (Table 2.3). Furthermore, the described three FF HPV+ and three FF HPV– HNSCC

samples (Table 2.2) were used for MeDIP-Seq analysis as well as for Illumina Infinium 450k array analysis to validate FFPE findings.

	HPV+ (n=21)	HPV- (n=21)
Median Age (range)	56.9 years (42-81)	58.4 years (43-77)
Gender	M: 14 F: 7	M: 17 F: 4
Tumour site	Oropharynx: 21	Oropharynx: 21
Tumour grade	Well diff: 1 Mod diff: 10 Poorly diff: 10	Well diff: 0 Mod diff: 15 Poorly diff: 6
Tumour stage (T)	T1: 5 T2: 7 T3: 4 T4: 3 N/a: 2	T1: 1 T2: 3 T3: 8 T4: 9 N/a: 0
Cervical lymph node involvement (N)	Yes: 16 No: 2 N/a: 3	Yes:16 No: 5 N/a: 0
Smoking	Ever: 8 Never: 10 N/a: 3	Ever: 17 Never: 0 N/a: 4
Alcohol	Heavy drinker: 2 Occ. alcohol: 6 Never: 4 N/a: 9	Heavy drinker: 13 Occ. alcohol: 3 Never: 1 N/a: 4

Table 2.3: Patient characteristics of 21 HPV+ and 21 HPV- FFPE HNSCC cases used for the host methylome analysis.

Human Tissue samples used for the viral methylome analysis (chapter 4):

The above-mentioned three FF HPV+ and three FF HPV– HNSCC samples (Table 2.2) were obtained from the UCLH Head and Neck Tumour Bank.

Human Tissue samples used for the genomic analysis (chapter 5):

According to the quality and the amount of the yielded DNA and gender and age-matching requirements, 20 HPV+ and 20 HPV- oropharyngeal carcinomas were selected, all formalin fixed paraffin-embedded, for paired-end sequencing of hybrid-captured DNA, targeting 3,230 exons in 182 genes often mutated in cancer. These data were validated by Sequenom MassArray sequencing, Infinium copy number variation (CNV) profiling, and immunohistochemistry (Table 2.4).

	HPV+ (n=20)	HPV- (n=20)
Median Age (range)	56.5 years (42-81)	58 years (45-77)
Gender	M: 14 F: 6	M: 14 F: 6
Tumour site	Oropharynx: 20	Oropharynx: 20
Tumour grade	Well diff: 1 Mod diff: 9 Poorly diff: 10	Well diff: 0 Mod diff: 16 Poorly diff: 4
Tumour stage (T)	T1: 5 T2: 8 T3: 3 T4: 3 N/a: 1	T1: 1 T2: 4 T3: 5 T4: 10 N/a: 0
Cervical lymph node involvement (N)	Yes: 16 No: 2 N/a: 2	Yes:13 No: 6 N/a: 1
Smoking	Ever: 9 Never: 8 N/a: 3	Ever: 15 Never: 0 N/a: 5
Alcohol	Heavy drinker (>20U/w): 2 Occ. alcohol: 5 Never: 4 N/a: 9	Heavy drinker (>20U/w): 12 Occ. alcohol: 3 Never: 0 N/a: 5

Table 2.4: Patient characteristics of selected HPV+ and HPV- FFPE HNSCC samples for NGS.

2.1.2 Cell lines used for the experiments:

HNSCC cell lines UPCI:SCC090 (HPV+), UPCI:SCC003 (HPV–), UPCI:SCC036 (HPV–) and PCI-30 (HPV–) were generous gifts from Dr. Susanne Gollin and Dr. Theresa Whiteside (University of Pittsburgh Cancer Institute, US). 93VU-147T (HPV+) was a generous gift from Dr. Hans Joenje (VU Medical Center, Netherlands) and UM:SCC047 (HPV+) was from Dr. Thomas Carey (University of Michigan, US). All cell lines were maintained in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% foetal bovine serum (FBS) and penicillin/streptomycin. I also generated clones of SCC003 expressing either empty vector control, HPV-16 *E7*, HPV-16 *E6,* or both HPV-16 *E6* and *E7*, respectively.

The main features of these cell lines are described in Table 2.5. HPV status was confirmed by E6 and E7 qPCR as described in section 2.2.8.

Cell line	Origin (site)	HPV status
UPCI:SCC90	oropharynx	HPV+ (integrated)
UM:SCC47	oropharynx	HPV+ (integrated)
93-VU-147T	oral cavity	HPV+ (integrated)
UPCI:SCC003	oropharynx	HPV-
UPCI:SCC036	oropharynx	HPV-
PCI-30	oral cavity	HPV-

Table 2.5: HPV+ and HPV- HNSCC cell lines used for experiments.

2.2 Core experimental procedures.

All general purpose chemicals and media were purchased from Sigma-Aldrich (Sigma-Aldrich Company Ltd, Fancy Road, Poole, BH12 4QH) or BDH (BDH Laboratory Supplies, Poole, BH15 1TD) unless otherwise stated.

2.2.1 DNA extraction:

DNA from the fresh frozen tumour samples and HNSCC cell lines was extracted using the QIAamp DNA Blood Mini Kit (QIAGEN). The QIAamp DNA FFPE Tissue Kit (QIAGEN) was used for the DNA extraction of laser dissected FFPE samples.

2.2.2 RNA extraction:

RNA from cell lines and infected cell line clones was extracted using the RNeasy Plus Mini Kit (QIAGEN).

2.2.3 DNA and RNA quantification:

DNA and RNA were quantified using the ND-1000 Spectrophotometer (NanoDrop®). For sequencing experiments DNA was quantified using the standardized PicoGreen fluorescence assay (LifeTechnologies, Carlsbad, CA).

2.2.4 Assessment of HPV status:

HPV status was determined by CDKN2A (p16) immunostaining (a biomarker for HPV+ HNSCC) of the corresponding FFPE blocks (in collaboration with UCL Advanced Diagnostics) and confirmation by E6 qPCR on extracted DNA from both fresh-frozen samples, FFPE samples, and cell line samples [120]. This has been shown to have 97% sensitivity, 94% specificity and to be the best discriminator of favourable outcome [121]. Sequencing demonstrated 100% concordance of HPV status.

2.2.5 p16 staining:

p16 staining was performed under the guidance of Philippa Jones (UCL Advanced Diagnostics) using the Bond™-III system, a fully-automated immunohistochemistry staining system (Leica Microsystems, Inc., Buffalo Grove, US). 3 µm sections were cut from a total of 82 FFPE blocks from HNSCC samples (see section 2.1.1) and were prepared for p16 staining. Using the Bond™-III system, slides were dewaxed using the Bond Dewax Solution (Leica Microsystems, Inc., Buffalo Grove, US) according to the manufacturer's recommendation (protocol '*D'). Antigen retrieval was conducted using the Bond ER1 solution for 30 minutes (Leica Microsystems, Inc., Buffalo Grove, US) according to manufacturer's protocol '*H1(30)'. Staining was then performed using the Bond Polymer Refine Kit and employing the manufacturer's protocol '15,8,8' using the pre-diluted p16 antibody clone E6H4™ and a negative reagent control (CINtec, Ventana Medical Systems, Inc., US). The stained slides were examined and scored as described by Thavaraj et al. [122], by two experienced histopathologists.

2.2.6 Laser-capture microdissection:

Samples were laser-captured microdissected (LCM) to separate tumour epithelial from surrounding stromal tissues, enriching tumour DNA for further analyses. These were processed as 10 µm thick unstained slides which were reviewed by an expert pathologist who had marked the slides for tumour subtype enrichment in a corresponding H&E stained section. LCM was

carried out on P.A.L.M. MembraneSlide 1.0 PEN slides (*Zeiss* Microimaging, Munich, Germany) using the Zeiss Palm MicrobeamTM system. Tissue was collected into extraction tubes and processed using the QIAamp DNA FFPE Tissue Kit (Qiagen, Hilden, Germany). Extracted DNA was quantified using a standardized PicoGreen fluorescence assay (LifeTechnologies, Carlsbad, CA).

2.2.7 Reverse transcription of extracted RNA:

cDNA was obtained from the extracted RNA by using the SuperScript™ First-Strand Synthesis System (Invitrogen Life Technologies**,** Paisley UK).

2.2.8 E6, E7, L1 and L2 qPCR:

Both DNA and cDNA (reverse transcribed from the extracted RNA samples) were used for qPCR. After having checked the accordance of the respective DNA and cDNA sequences with the primer sequences used, E6, E7, L1 and L2 qPCR and RT-qPCR were optimized using the following primers and TaqMan probes (and using GAPDH as a house-keeping control) to test for the DNA and cDNA regions of interest:

Name of primer/probe	**Primer/probe Sequence (5'-3')**
HPV type 16 GAPDH forward primer	5'- GGAGTCAACGGATTTGGTCGTA -3'
HPV type 16 GAPDH forward primer	5'- GGCAACAATATCCACTTTACCAGAGT -3'
HPV 16 GAPDH TaqMan probe	5'-(FAM)- CGCCTGGTCACCAGGGCTGC -(TAMRA)-3'

Table 2.6: GAPDH primers and probe used for qPCR.

Name of primer/probe	**Primer/probe Sequence (5'-3')**
HPV type 16 E6 forward primer	5'-TCAGGACCCACAGGAGCG-3'
HPV type 16 E6 reverse primer	5'-CCTCACGTCGCAGTAACTGTTG-3'
HPV 16 E6 TaqMan probe	5'-(FAM)-CCCAGAAAGTTACCACAGTT ATGCACAGAGCT-(TAMRA)-3'

Table 2.7: E6 primers and probe used for qPCR.

Name of primer/probe	Primer/probe Sequence (5'-3')
HPV type 16 E7 forward primer	5'-CCGGACAGAGCCCATTACAA-3'
HPV type 16 E7 reverse primer	5'-CGAATGTCTACGTGTGTGCTTTG-3'
HPV 16 E7 TaqMan probe	5'-(FAM)-CGCACAACCGAAGCGTAGAGTCACACT-(TAMRA)-3'

Table 2.8: E7 primers and probe used for qPCR.

Name of primer/probe	Primer/probe Sequence (5'-3')
HPV type 16 L1 forward primer	5'- ATAATCCAGATACACAGCGGC -3'
HPV type 16 L1 reverse primer	5'- CCACTAATGCCCACACCTAATG -3'
HPV 16 L1 TaqMan probe	5'-(FAM)-AACACCTACACAGGCCCAAACCA-(TAMRA)-3'

Table 2.9: L1 primers and probe used for qPCR.

Name of primer/probe	Primer/probe Sequence (5'-3')
HPV type 16 L2 forward primer	5'- ACGTGCATCGGCTACCCAACTTTA -3'
HPV type 16 L2 reverse primer	5'- ACCCGACCCTGTTCCAATTCCTAA -3'
HPV 16 L2 TaqMan probe	5'-(FAM)-AAACAGGCAGGTACATGTCCACCTGA -(TAMRA)-3'

Table 2.10: L2 primers and probe used for qPCR.

DNA and cDNA were amplified using qPCR with 25 µl 2x Abgene Buffer A (Abgene, UK), 0.3 µM forward primer, 0.3 µM reverse primer, 0.15 µM TaqMan probe, in a total volume of 50 µl. Standards, a house-keeping control (GAPDH), and one water control sample were included in each PCR setup. qPCR and RT-qPCR was performed using the Eppendorf Realplex Mastercycler (Eppendorf, UK) applying the following qPCR programme: Reactions were denatured at 95°C for 15 minutes, followed by 40 cycles of 95°C for 15 sec and 60°C for 60 sec, with no extension. All reactions were run in duplicate (two reactions at 1x concentrations and two reactions of 1:10 dilutions).

Primer efficiency for each primer set was calculated using the formula '[10^(-1/-slope)]-1', as previously described [123, 124]. Sufficient primer efficiency was confirmed for each primer set used (Table 2.11). Standard curves for each primer set used are illustrated in the Appendix Figure A.2-A.6.

Primers tested	Slope (standard curves; Appendix Fig.)	Primer efficiency
GAPDH	-3.156	**107.42%**
E6	-3.517	**92.46%**
E7	-3.3683	**98.10%**
L1	-3.2537	**102.93%**
L2	-3.264	**102.48%**

Table 2.11: Calculated primer efficiencies for each primer set tested.

Comparison of viral L1 and L2 gene expression between control states and 5-AZA treated states of the respective cell lines was calculated using the delta-delta (ΔΔ) Ct method [124], according to the following formula:

$$\Delta\Delta Ct = [(Ct_{G.O.I.(treated)} - Ct_{HK.G.C.(treated)}) - (Ct_{G.O.I.(control)} - Ct_{HK.G.C.(control)})]$$

Genes of interest (G.O.I.): viral L1 and L2 gene. House-keeping gene control (HK.G.C): GAPDH

A positive ΔΔCt value indicates an increase in expression of the tested gene, whereas a negative ΔΔCt value indicates a decrease in expression of the tested gene. The fold variation was calculated using the following formula:

Fold change = 2^(-ΔΔCt)

Standard deviation of the Delta Ct values of the tested replicates was estimated as the sum of squares of the standard deviations of the gene of interest and the housekeeping gene. This was added to, and subtracted from, the respective ΔΔCt values, yielding corrected ΔΔCt values. The obtained values were then used to calculate the corrected fold change, according to the following formula:

Corrected fold change = 2^(-corrΔΔCt)

The obtained corrected fold change value then subtracted from the fold change value, yielding the quantity of correction (margins of error). Mean fold change values were then log-transformed and plotted, with the calculated margins of error (error bars). Two-tailed unpaired T-tests were applied to determine the significance of differences between untreated and treated cell lines.

2.2.9 Gel electrophoresis:

The electrophoretic mobility of DNA molecules is dependent on the size of a specific region and the concentration of agarose gel used. 2% Agarose gels were generally employed, although 1.5% gels were used for analysis of DNA fragments less than 500 bp in size, and 0.8 % gels for DNA fragments larger than 4 kb. The appropriate weight of agarose was added to either TBE (from a 10x stock of 1 litre; 108 g Tris, 55 g Boric acid, 40 ml 0.5 M Na2EDTA at pH 8.0, dissolved in 900 ml distilled water, or TAE buffer (40 mM Tris-Acetate/1 mM EDTA), depending on the application and the mixture heated to allow the agarose to dissolve. The solution was then immediately cooled to approximately 60°C and ethidium bromide added to a final concentration of 1 µg/ml. The melted agarose solution was poured into a mould with a well-forming comb inserted and incubated at RT to allow the gel to harden. The DNA samples were then mixed with 6 X Gel Loading Dye Blue (New England Biolabs, US), loaded into the wells of the prepared gel and fragments were separated by electrophoresis in TBE/TAE buffer at a constant current of 270 mA. Standard molecular weight markers (50 bp DNA ladder and 1 kb DNA ladder, New England Biolabs, US) were used alongside the samples. The DNA was visualized and photographed under a long-wave UV light.

2.3 Experimental procedures used to obtain results in chapter 3.

2.3.1 Whole-genome methylation analysis with MeDIP-seq:

Methylated DNA immunoprecipitation (MeDIP) was developed by Weber et al. in 2005 [125]. It is a whole genome DNA methylation assay that enriches the methylated fraction using a versatile immunocapturing approach and by that allows for unbiased detection of methylated DNA. This allowed me to interrogate approximately 70% of the host methylome and 100% of the HPV methylome.

In my experiments genomic DNA was randomly sheared by sonication (DNA preparation step) and the length of obtained fragments was validated using gel electrophoresis. The NEBNext DNA Sample Prep Mastermix (New England Biolabs) was then used for the library preparation step. In test runs I performed both manual MeDIP [126] and automated MeDIP (AutoMeDIP) [127]. For the analysis of fresh-frozen tumour samples AutoMeDIP was conducted using the IP Star (Diagenode) as previously described [127]. During this process the prepped DNA was immunoprecipitated with a monoclonal antibody that specifically recognizes 5-methylcytidine and subsequently the immunoprecipitated (methylated) fraction was recovered. Enrichment of the methylated DNA fraction (compared to input) was assessed using qPCR. Thereafter the methylated portion was amplified by adapter mediated PCR. After the adapter mediated PCR step, the library was subjected to size selection (300-350 bp) from low melting-point agarose gels. The excised fraction was quality controlled by qPCR. The work-flow is also illustrated in Figure 2.2.

There are some pitfalls which should be mentioned here and which I have also encountered. Gel excision is a laborious and time-consuming step and a lot of DNA is lost in its course both during the gel excision itself and during the gel purification step. Thus, this has to be taken into account when deciding on the amount of DNA to start with in order to end up with a sufficient amount for cluster generation. Moreover, there is currently no consensus on the best methods for quantifying library concentration. Although I aimed to sonicate the sample DNA to have a similar size distribution, invariably slight variation has been observed between samples.

MeDIP-Seq describes the combination of MeDIP with Next Generation sequencing. Cluster generation and 36 base paired-end sequencing was performed by the UCL Genomics Core Facility according to the manufacturer’s recommendation, using an Illumina GAIIX platform.

Figure 2.2: Illustrated Auto-MeDIP work-flow; e, electrophorese, p, purify, q, quantitate, qPCR, quantitative polymerase chain reaction (obtained from [127]).

2.3.2 Data analysis of MeDIP-Seq data:

The sequencing analysis was performed with the guidance and help of Dr. Gareth Wilson (Medical Genomics Group, UCL Cancer Institute). The data were analysed using the MeDUSA pipeline [128]. Reads were aligned to the reference genome (Human assembly GRCh37) using the alignment software BWA (v0.5.8) [129], with default parameters. Filtering was performed using SAMtools, (v0.1.9) [130] to remove erroneously mapped and low quality (score of <10) reads. Only those forming a correctly aligned pair were kept. A

final filtering step removed potential PCR artefacts by discarding all but one read-pair within groups of non-unique fragments (Table 2.12). Read quality was ascertained using FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) and the Bioconductor package MEDIPS (v1.0.0) [131]. Probes from the 450k BeadChips located within CpG island regions were isolated and these sites were extended to create 500bp windows. Absolute methylation scores for each of these regions were calculated from my MeDIP read files using MEDIPS. Methylation scores were calculated for each extended probe site using default values.

Sample	HPV	Cycles	Total Reads (pre-alignment)	Unique fragments (paired reads)	Mean Insert Size	CpG Enrichment Score	CpG cover-age	% GC
HN_29	HPV-	36	52655678	17002647	174.43	4.06	48%	52
HN_32	HPV-	36	58152760	19332002	189.9	3.78	54%	52
HN_39	HPV+	36	40889864	11829705	196.16	3.45	45%	50
HN_96	HPV-	36	41538032	12474190	185.06	3.19	47%	49
HN_105	HPV+	36	36259072	11385250	173.75	3.68	44%	52
HN_125	HPV+	36	47522856	15101872	178.68	3.96	47%	52

Table 2.12: Read counts obtained from the methylome analysis on MeDIP-Seq data.

2.3.3 Genome-wide methylation analysis with Illumina 450k BeadChips:

DNAs were prepared in a total volume of 20 µl (1 µg for FF and cell line DNAs and 2 µg for FFPE DNA per sample) using a previously optimized protocol [132] in conjunction with the REPLIg FFPE kit (Qiagen GmbH D-40724 Hilden, cat No. 150243) and EZ DNA Methylation kit (Zymo Research Corp, Orange, CA 92867, USA, cat No. D5001). The latter was modified to improve bisulfite conversion efficiency by inclusion of a cyclic denaturation step as described previously [132]. The Infinium HumanMethylation450 BeadChips were purchased from Illumina and processed by the UCL Genomics core facility according to the manufacturer's recommendation. The scanned data and image output files were managed with the Genomestudio software (version 1.9.0) (Illumina®, Inc., San Diego, CA 92121, USA).

2.3.4 Data analysis of Illumina 450k BeadChips data:

The statistical analysis was performed under the guidance, and with the assistance, of Dr. Andrew Teschendorff (Statistical Genomics Group, UCL Cancer Institute). I initially tested 3 approaches in addition to the Bayesian regularised T-statistics model described in detail (see below). The work-flow/parameters used for each analysis approach are outlined in Figure 2.3. Using the Illumina software package 'Genome Studio' v1.9.0 (Illumina®, Inc., San Diego, CA 92121, USA), designed for the analysis of Illumina 450k BeadChips methylation data, 77917 MVPs were identified (adjusted $P < 0.01$). Employing the Lumi pipeline [133] and an in-house method developed at the time of analysis, 'Methyl-T' (script kindly provided by P. Guilhamon, Medical Genomics Group, for testing), 394 MVPs and 306 MVPs were identified, respectively. Using a Bayesian regularised T-statistics model which is described in detail in the following section, 2757 MVPs were identified. The overlap of obtained significant MVPs is illustrated in a Venn diagram (Figure 2.4).

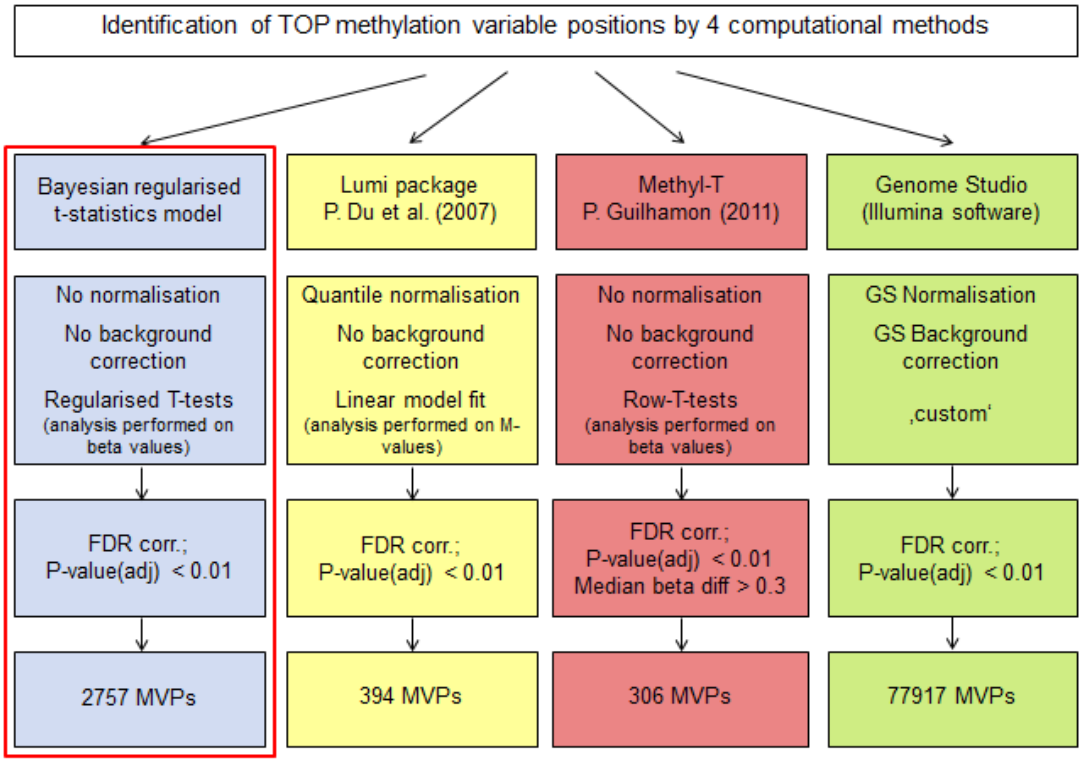


Figure 2.3: Evaluation of 4 methods tested for the methylation analysis of Illumina 450k BeadChip data (the one chosen is highlighted in red)

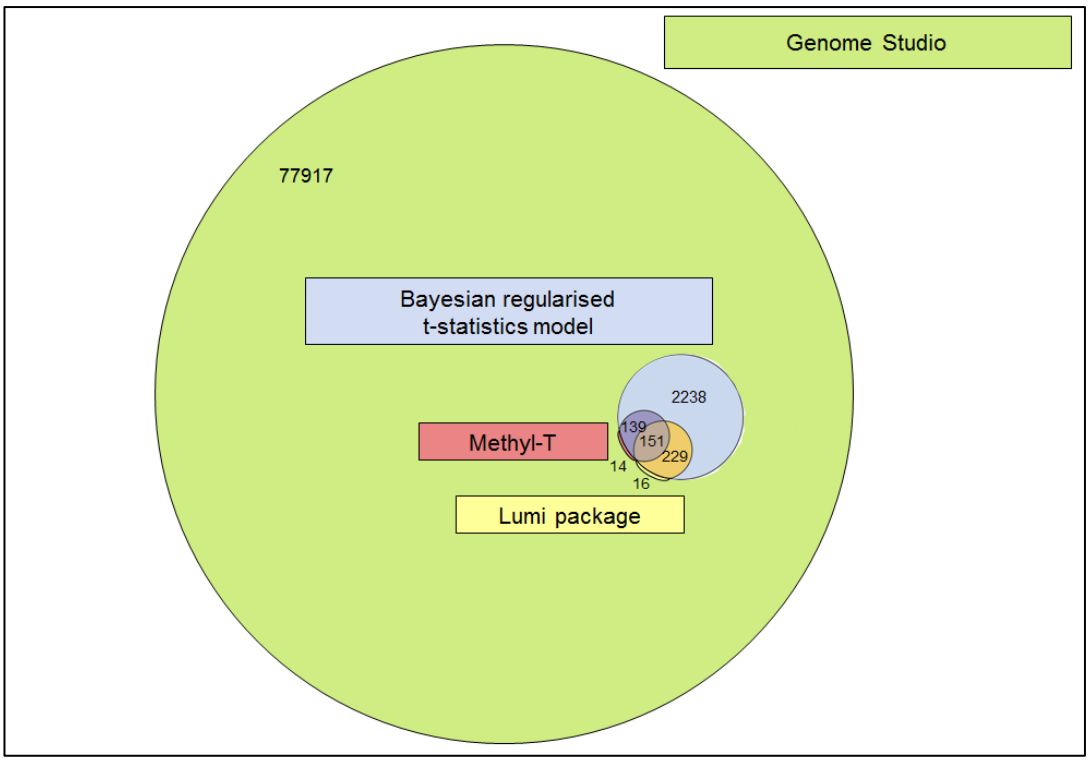


Figure 2.4: Venn diagram based on numbers of significant methylation variable positions (adjusted *P* < 0.01) by applying the 4 different methods outlined in Figure 2.3.

The Bayesian regularised T-statistics model was selected on the basis of its superior power, compared with the Methyl-T approach and the approach used by the Lumi package. I disregarded the results of the Genome Studio algorithm due to the implausibly large number of MVPs found and doubts over the proprietary methods employed. R statistical software v2.14.0 (http://www.r-project.org) was used for the subsequent data analysis. Raw data were subjected to a stringent quality control analysis as follows: samples were removed which showed reduced coverage and only probes detected above background across all samples were kept (detection p-value < 0.01), resulting in a raw data matrix of 439,385 probes and 32 samples with 18 HPV+ and 14 HPV-. This raw data matrix was then subjected to a principal component analysis to determine the nature of the largest components of variation. Random Matrix Theory was used to estimate the number of significant components of variation [134, 135].

450k BeadChips contain two types of probes (type 1 and 2) which have slightly different profiles. While there have been attempts to normalize for that difference [136], it was observed that proposed normalization

methods and in-house methods that were developed overcorrected the data leading to worse performance as evaluated using a rigorous training-test set partition strategy. Thus, in the supervised analysis type1 and type2 design probes were treated equally and tested a posteriori for a potential skew favouring type1 probes. Although there were only 1075 type 1 probes among the top 2757, this amounted to an over-enrichment with an odds ratio of 1.48 (P < 1e - 16). However, after correcting for differences in CpG density between type1 and type2 probes, the enrichment odds ratio favouring type1 probes was significantly reduced to 1.13 (P ~ 0.03). Thus, there was no substantial skew favouring type1 probes and it was observed that normalising for the design using the Peak-Based Correction method of Dedeurwader only led to overcorrection and increased technical variability (Figure 2.5).

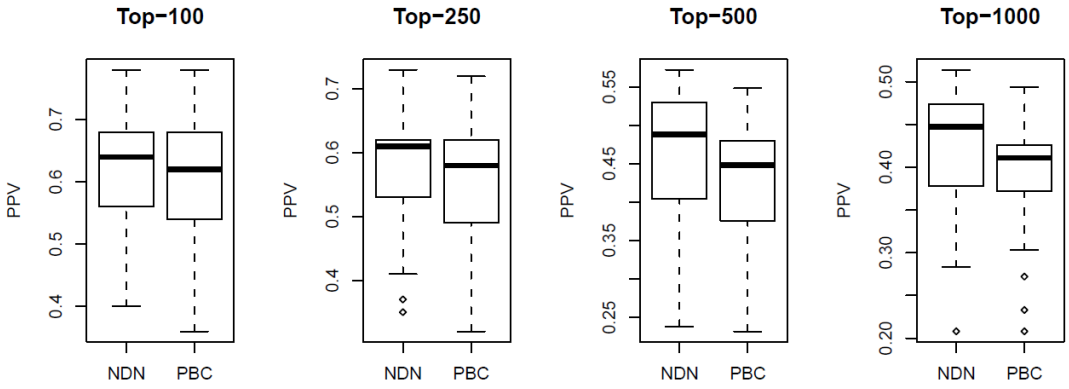


Figure 2.5: Peak-Based Correction (PBC) does not improve predictive power in FFPE HNSCC samples. Positive predictive value (PPV, y-axis) for the case where no design normalisation is performed (NDN) against PBC normalisation. The PPV is estimated in a test set, where features have been ranked and selected from a mutually exclusive training set. The boxplots are over 25 distinct training/test set choices. This shows that PBC increases variability within phenotypes resulting in worse performance. The PPV is estimated as the fraction of the selected features that have a P-value < 0.05 in the test and preserving the same directionality as in the training set.

To quantify the strength of the association and to adjust for multiple testing the false discovery rate (FDR) was estimated using the q-value procedure [137]. Since the analytical q-value estimates assume independence of the underlying tests, an assumption which does not necessarily apply to neighbouring probes that are spatially correlated, the

FDR was also estimated using a permutation approach which preserves the potential correlation structure of proximal probes. However, empirical and analytical FDR estimates were in close agreement (Figure 2.6). Both procedures estimated approximately 2750 methylation variable positions (MVPs) at an FDR< 0.01, i.e less than 1% of the 2757 probes are expected to be false positives.

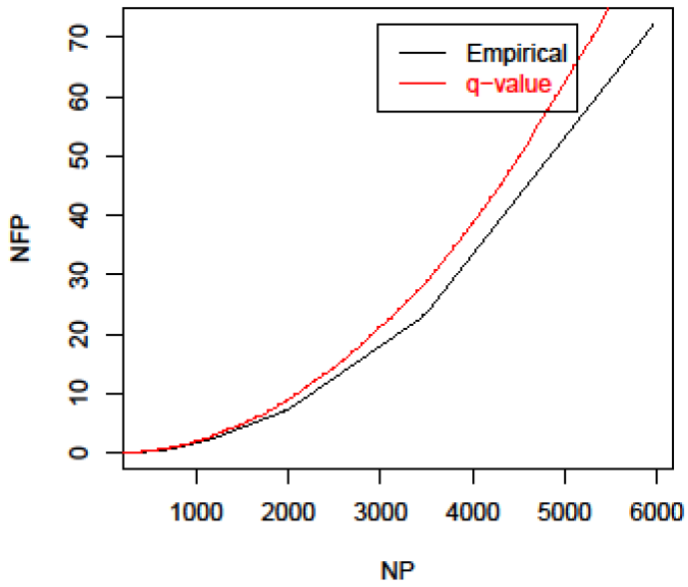


Figure 2.6: Close agreement of analytic and permutation-based FDR estimates. False discovery rate (FDR) of FFPE probes estimated by (1) red: analytic *q*-value procedure of Storey-Tibshirani and (2) black: using a permutation approach preserving the potential correlative structure of proximal probes (empirical). NP = number of probes, NFP = Number of False Positives.

2.3.5 Copy number variation (CNV) analysis:

CNV analysis was performed on the DNA of the three FF HPV+ and three FF HPV–HNSCC samples using the Illumina HumanOmni1-Quad BeadChip genotyping array. This analysis was required for the normalization of the MeDIP-seq data. CNV data were analysed using Illumina Genome Studio.

2.3.6 Generation of SCC003 clones expressing HPV-16 oncogenes:

Clones of SCC003 expressing either empty vector control, HPV-16 *E7*, HPV-16 *E6,* or both HPV-16 *E6* and *E7*, respectively, were generated by infection of the SCC003 cell line with retroviruses and single cell cloning as follows.

Viruses were produced by transfecting Human Embryonic Kidney (HEK) 293T cells with pLXSN (empty vector or containing the HPV-16 *E6*, *E7* or *E6&E7* cDNAs, kind gifts from Dr. David Beach) together with pHIT- VSVG and MLV-gag/pol (kind gifts from Dr. Juan Martin-Serrano) using polyethylenimine (Polysciences, Inc.). 72 hours post-transfection, viruses were harvested by removal of the medium and filtration through 0.45um surfactant-free filters (Nalgene). The filtered virus stocks were either frozen at -80C or diluted 1:2 in DMEM/10%FBS containing 8ug/ml Polybrene (to give a final conc. of 4ug/ml) and added to SCC003 cells grown to a confluence of 40-50%. Following overnight incubation, the cells were washed to remove virus and medium was replaced with DMEM/10%FBS. At 48-72hrs post-infection cells were passaged at a ratio of 1:5 into selection medium containing 400ug/ml G418. Following the death of all mock-infected cells (approx. 1 – 2 weeks), cells were removed from selection and plated at limiting dilution in 96 well plates to generate single cell clones. HPV-16 *E6* and *E7* qPCR were conducted as described [138, 139]. To assess *E6* and *E7* expression levels in *E6*-transduced, *E7*-transduced and *E6&E7*-transduced SCC003 cell line clones (and empty vector controls), qPCR was performed on cDNA following reverse transcription (Superscript-II, Invitrogen) of total RNA purified from cells using the miRNeasy kit (QIAGEN) as per manufacturers' recommendations.

2.3.7 Survival Analysis:

Survival Analysis on the selected 21 FFPE HPV+ and 21 FFPE HPV- HNSCC samples was performed using the IBM SPSS Statistics Package Version 20, after having retrieved the relevant clinical information. Overall survival was used as the outcome measure. The Log Rank test was used to test for significant differences and Kaplan-Meier curves plotted.

2.4 Methodology used and data analysis applied to obtain results in chapter 4

2.4.1 MeDIP-Seq data analysis of the HPV genome:

The analysis of the MeDIP-Seq data was performed with the help of Dr. Gareth Wilson (Medical Genomics, UCL Cancer Institute). For this the MeDIP sequence data from the 3 HPV+ FF HNSCC samples were used, allowing the interrogation of 100% of the HPV methylome. The 36 base paired-end sequence data were aligned to the human reference genome (Human assembly GRCh37) using BWA (v0.5.8) (Li and Durbin 2009), with default parameters. Those that failed to align to the human genome were mapped, using BWA, to 1776 viral genomes, including 29 human papilloma viruses (Table 2.13), obtained from NCBI Genomes (http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239).

Filtering was performed using SAMtools (v0.1.9) (Li, Handsaker et al. 2009) to remove erroneously mapped and low quality (score of <10) reads. In order to compare viral aligned reads between samples, the read counts were normalized for viral load by dividing the number of reads in 10bp windows by the viral load as estimated by E6 qPCR. Further normalization to account for total library size led to read counts being transformed to reads per million (RPM) as described in Chavez et al. 2010 [131].

Additionally, the MeDIP-seq data were analysed to identify potential viral integration sites. For this, reads that failed to form a correctly aligned pair in the human genome were stored. For each HPV+ sample, these stored reads were aligned to the HPV-16 (NC_001526) genome using BWA. Pairs in which one read mapped to the viral genome with an alignment score >= 10 and the corresponding mate failed to map, were identified (n = 182) using SAMtools and custom perl scripts. These reads were compared to the relevant human alignment to find pairs. In order to be classed as a pair, the human read also required an alignment score >= 10 and to have an unmapped mate (n = 660428). Therefore the resulting pair comprises of one read aligning with high confidence to HPV-16 and the other to human,

indicating that a potential integration site is located within the fragment (n=33).

HPV type	Reference Sequence Accession Code (NCBI)
Human_papillomavirus_type_18_uid15506	NC_001357
Human_papillomavirus_type_1_uid15491	NC_001356
Human_papillomavirus_type_2_uid15512	NC_001352
Human_papillomavirus_type_54_uid15466	NC_001676
Human_papillomavirus_type_101_uid17121	NC_008189
Human_papillomavirus_type_103_uid17119	NC_008188
Human_papillomavirus_type_108_uid34847	NC_012213
Human_papillomavirus_type_10_uid15504	NC_001576
Human_papillomavirus_type_16_uid15505	NC_001526
Human_papillomavirus_type_26_uid15507	NC_001583
Human_papillomavirus_type_32_uid15508	NC_001586
Human_papillomavirus_type_34_uid15509	NC_001587
Human_papillomavirus_type_41_uid15485	NC_001354
Human_papillomavirus_type_48_uid14027	NC_001690
Human_papillomavirus_type_49_uid15455	NC_001591
Human_papillomavirus_type_4_uid15492	NC_001457
Human_papillomavirus_type_50_uid14327	NC_001691
Human_papillomavirus_type_53_uid15510	NC_001593
Human_papillomavirus_type_5_uid15511	NC_001531
Human_papillomavirus_type_60_uid14028	NC_001693
Human_papillomavirus_type_63_uid15486	NC_001458
Human_papillomavirus_type_6b_uid15454	NC_001355
Human_papillomavirus_type_7_uid15450	NC_001595
Human_papillomavirus_type_88_uid28737	NC_010329
Human_papillomavirus_type_90_uid15424	NC_004104
Human_papillomavirus_type_92_uid14406	NC_004500
Human_papillomavirus_type_96_uid15488	NC_005134
Human_papillomavirus_type_9_uid15456	NC_001596

Table 2.13: 29 human papilloma viruses, including Reference Sequence Accession Codes, obtained from NCBI Genomes (http://www.ncbi.nlm.nih.gov/genomes)

To determine if there was a significant enrichment of these reads within a specific location on the viral genome, counts of the aligned ‘integration site’ reads were obtained for 500bp windows with 10bp slide. Permutation analysis was performed (n=10000). The probability of an ‘integration site’ read aligning to a specific genomic region was weighted according to the total number of reads aligning in that window. The resulting empirical p-values were log transformed (-log10) and plotted along the HPV16 genome.

2.4.2 Validation of obtained results by bisulfite sequencing (BS-Seq):

In vitro methylation of HPV+ HNSCC DNA samples:

In vitro methylation of HPV+ HNSCC DNA samples was performed using CpG Methyltransferase (M.Sssi) (New England Biolabs Inc., Ipswich, US), according to the protocol of the supplier, in order to obtain a positive control for subsequent BS-Seq analysis.

Whole-genome amplification of HPV+ HNSCC DNA samples:

Whole-genome amplification of HPV+ HNSCC DNA samples was performed using the GenomePlex® Complete Whole Genome Amplification (WGA) kit (Sigma-Aldrich, Missouri, US) and the REPLI-g Mini Kit (QIAGEN), according to the protocol of the supplier, in order to obtain a negative control for subsequent BS-Seq analysis.

Bisulfite modification:

It is not possible to detect methylated cytosines by using conventional DNA sequencing. By applying the bisulfite conversion technique [140], methylation can be analysed by means of sequencing (BS-Seq). Bisulfite deaminates unmethylated cytosines and leads to the generation of the base uracil. Methylated cytosines are not converted and remain cytosines. During the subsequent PCR step, the modified template is amplified by PCR (methylated cytosines remain unchanged and unmethylated cytosines which are converted to uracils are amplified, leading to their replacement by thymines).

One µg of genomic DNA from 3 FF HPV+ HNSCC tissue samples (analysed by MeDIP-Seq previously) and from 3 HPV+ HNSCC cell lines, was bisulfite converted using the Zymo EZ DNA Methylation kit (Zymo Research Corp, Orange, CA 92867, USA, cat No. D5001). The protocol was modified to improve bisulfite conversion efficiency by inclusion of a cyclic denaturation step as described previously [132]. Bisulfite conversion efficiency was

confirmed by methylation-specific PCR (MSP) for a methylated and unmethylated DNA fragment.

Targeted HPV methylome analysis:

In order to obtain the DNA segment of interest (the region at the boundary of the L1 and L2 gene within the HPV-16 genome), the targeted modified DNA fragment was amplified using touchdown PCR with 1x Reaction Buffer, 0.5 mM dNTPs, 2.0-3.0 mM MgCl2, 0.4 µM of forward and reverse primers respectively and 1 unit of HotStar Taq (Qiagen, Crawley, UK), in a total volume of 20 µl (Qiagen, Crawley, UK). Reagents were denatured at 95°C for 10 min, followed by 20 cycles of 95°C for 45 sec, 60°C with a decrease of half a degree per cycle for 45 sec, 68°C for 60 sec and 30 cycles of 95°C for 45 sec, 50°C for 45 sec, 68°C for 60 sec and ending with a seven-minute extension at 68°C. One *in vitro* methylated control sample (positive control) and one unmethylated (whole-genome amplified) control sample (negative control), as well as a 50/50 mixture of both, were included in each PCR to ensure that methylated and unmethylated templates were both equally amplified, as previously described [141].

Initially primers were designed with BiSearch (http://bisearch.enzim.hu/) for the generation of a PCR product of 292 bp in length, covering the 6 CpG sites of interest in a region at the boundary of the L1 and L2 gene within the HPV-16 genome. Sequences of the initial primers are summarized in Table 2.14.

Name of primer	**Primer Sequence (5'-3')**
forward L1/2 PCR primer vol. 1 (5548-5567)	AATTATTGTTGATGTAGGTG
reverse L1/2 PCR primer vol. 1 (5819-5839)	AGTTTTTAAAGTATTAGGATT

Table 2.14: L1 and L2 PCR primers tested initially.

The specificity of the reaction products was inspected using 1.5 % agarose gel electrophoresis before being purified using the Illustra ExoStar 1-Step kit (GE Healthcare Life Sciences, Little Chalfont, UK).

One problem encountered was that only the methylated template was amplified (with primers vol. 1 used; Table 2.14). The BS-Seq trace of the

negative control consistently showed cytosine peaks at all 6 CpG sites of interest within this region, although peaks of the base thymine in the sequencing data at this location were expected. It was hypothesised that the GenomePlex® Complete Whole Genome Amplification (WGA) kit (Sigma-Aldrich, Missouri, US) would not sufficiently amplify viral DNA and the experiment was repeated using the REPLI-g Mini Kit (QIAGEN). The result did not change following that. To check that the viral DNA was amplified as well as the human DNA, an E6 qPCR was run on 3 HPV+ HNSCC samples pre and post WGA. The delta Ct between GAPDH and E6 showed no difference between samples pre and post WGA, suggesting that both the viral and the human DNA were amplified equally well.

Using the same PCR programme as before with a gradient of 3 temperatures for amplification (55°C, 60°C, 65°C), MSP primers were tested, covering 4 of the 6 CpG sites of interest (primer sequences obtained from [119] and outlined in Table 2.15), on *in vitro* methylated viral DNA, whole-genome amplified DNA and a 50/50 mixture of both, in 3 HPV+ HNSCC cell line DNA samples.

Name of primer	**Primer Sequence (5'-3')**
M-SP F (5595-5619)	TGTTACGAAAACGACGTAAACG
M-SP R (5731-5707)	CGTACAACATATTCATCCG
UnM-SP F (5595-5619)	TGTTATGAAAATGATGTAAATG
UnM-SP R (5724-5707)	CATACAACATATTCATCCA

Table 2.15: PCR primers used for MSP.

In contrast to the results published by Park et al. [119], amplification of both the methylated and unmethylated fraction, using both of these primers, was found, not allowing the validation of the methylation status within this region.

PCR primers used for BS-Seq by Park et al., were then tested, covering the 6 CpG sites of interest in a region at the boundary of the L1 and L2 gene within the HPV-16 genome, in order to validate my viral methylome analysis employing MeDIP-Seq. Sequences of these primers are summarized in Table 2.16. A product of 421 bp (see Figure 2.7) was generated by PCR

using the PCR programme outlined above and purified using the Illustra ExoStar 1-Step kit (GE Healthcare Life Sciences, Little Chalfont, UK).

Name of primer	Primer Sequence (5'-3')
forward L1/2 PCR primer vol. 2 (5541-5569)	AATATATAATTATTGTTGATGTAGGTGAT
reverse L1/2 PCR primer vol. 2 (5984-5962)	ACCCACACCTAATAACTAACCAC

Table 2.16: L1 and L2 PCR primers used for validation of MeDIP-Seq results by BS-Seq.

```
5521                        aatatacaat tattgctgat gcaggtgact tttatttaca
5581 tcctagttat tacatgttac gaaaacgacg taaacgttta ccatattttt tttcagatgt
5641 ctctttggct gcctagtgag gccactgtct acttgcctcc tgtcccagta tctaaggttg
5701 taagcacgga tgaatatgtt gcacgcacaa acatatatta tcatgcagga acatccagac
5761 tacttgcagt tggacatccc tattttccta ttaaaaaacc taacaataac aaaatattag
5821 ttcctaaagt atcaggatta caatacaggg tatttagaat acatttacct gaccccaata
5881 agtttggttt tcctgacacc tcattttata atccagatac acagcggctg gtttgggcct
5941 gtgtaggtgt tgaggtaggc cgtggtcagc cattaggtgt gggc
```

Figure 2.7: 421 bp viral DNA segment of interest.

Bisulfite sequencing:

Methylation analysis was performed with bisulfite genomic sequencing using the bisulfite modified and PCR amplified and purified DNA segment of interest (421 bp segment at the border of the L1 and L2 gene in the HPV-16 genome, harbouring the 6 CpG sites of interest in genomic positions 5600, 5606, 5609, 5615, 5707, and 5724 (and two additional CpG sites). The same primers (Table 2.16) were used for BS-Seq.

It ought to be mentioned that, after bisulphite treatment and PCR amplification, a subcloning step is frequently included before sequencing. This is done to avoid a molecular weight shift, detected during the sequencing reaction, when CpG rich regions are analysed. This shift is a result of the analysis of a mixture of methylated (+ CH_3) and unmethylated (- CH_3) CpG sites within the segment of interest in a heterogeneous DNA population. However, in my case, the segment of interest only contained 8 CpG sites (within a 421 bp viral DNA segment). Thus, no significant molecular weight shift was expected. In order to cause a molecular weight shift of one base, the DNA segment would have to contain 22 CpG sites which are differentially methylated in a cell population. Hence, a subcloning step was not included. Although Sanger sequencing is not a quantitative method, looking at the results obtained from the positive (in vitro methylated)

and negative (whole-genome amplified) control, one can be confident that it is a good indicator of methylation of this locus in the HPV type 16 genome.
BS-Seq was performed by the UCL Genomics Core Facility according to the manufacturer's recommendation using Applied Biosystems Bigdye3.1 chemistry. Sequencing reactions were cleaned up using in-house-made sephadex filtration plates and then run on Applied Biosystems 3730XL Genetic Analyser.

2.4.3 5-AZA treatment of cell lines:

Three HPV+ HNSCC cell lines (UPCI:SCC90, UM:SCC47, 93VU-147T) were plated in 6-well plates (1.5 x 10 cells/well) with 1 ml of culture medium and incubated with increasing amounts of 5-Aza-2'-deoxycytidine 98+% (5-AZA; Fisher Scientific, Pittsburgh, US), dissolved in phosphate-buffered saline (PBS), for 72 hours. Culture medium was changed every 24 hours and freshly prepared 5-AZA was added. Concentrations of 5-AZA (1 µM, 2 µM, 5 µM and 10 µM) used were chosen according to available literature, describing the treatment of HNSCC cell lines [142, 143], cervical cancer cell lines [144, 145], and cell lines from other cancers, such as colorectal cancer [146] with 5-AZA.

2.5 Methodology used and data analysis applied to obtain results in chapter 5.

2.5.1 Exon-sequencing of 182 genes often mutated in cancer:

Paired-end sequencing of hybrid-captured DNA, targeting 3,230 exons in 182 cancer-related genes plus 37 introns from 14 genes often rearranged in cancer was done in collaboration with Foundation Medicine (Cambridge, MA, US). These genes are known to be somatically altered in human solid cancers based on recent scientific and clinical literature.

DNA Library Construction and Hybrid Capture

Molecular barcode-indexed ligation-based sequencing libraries were constructed using 50-200ng of total genomic DNA recovered from the sample. Libraries were hybridization captured with custom biotinylated RNA oligo pools (custom SureSelect kit, Agilent) representing 3,230 exons in 182 cancer-related genes (most commonly altered in cancer, from http://www.sanger.ac.uk/genetics/CGP/cosmic/) plus 37 introns from 14 genes often rearranged in cancer (Appendix Table A.1).

Sequencing and Analysis

Paired end sequencing (49 x 49 cycles) was performed using the HiSeq2000 (Illumina). Sequence data from gDNA, available from 18 HPV+ and 16 HPV- samples, were mapped to the reference human genome (human assembly GRCh37) using the BWA aligner [129] and processed using publicly available SAMtools [130], Picard (http://picard.sourceforge.net) and GATK [147]. Genomic base substitutions and indels were detected using custom tools optimised for mutation calling in heterogeneous tumour samples, based on statistical modelling of sequence quality scores and local sequence assembly. Variations were filtered using dbSNP_135_ENREF_1 (http://www.ncbi.nlm.nih.gov/projects/SNP/) and a custom artifact database, then annotated for known and likely somatic mutations using COSMIC [148]. Copy number alterations were detected by comparing targeted genomic DNA sequence coverage with a process-matched normal control sample. Genomic rearrangements were detected by clustering chimeric reads mapping to targeted introns.

2.5.2 Validation of selected mutations by Sequenom OncoCarta:

DNA extracted from FFPE samples were sent to Sequenom (Hamburg, Germany) for blind testing and analysis, using Sequenom OncoCarta panels v1.0 and v3.0, as previously described [149]. In brief, the Sequenom OncoCarta Panel™ represents a high-throughput panel for somatic mutation profiling [150], combining PCR and Matrix Assisted Laser Desorption

Ionization Time Of Flight Mass Spectrometry (MALDI-TOF MS). Mutations covered by panel v1.0 and v3.0 are outlined in Appendix Table A.2.

2.5.3 Confirmation of copy number changes by Infinium CNV Profiling:

Data analysis was performed under the guidance of Dr. Andrew Feber (Medical Genomics, UCL Cancer Institute). CNA data was generated from un-normalised signal intensities from the methylation data from HPV+ and HPV- HNSCC. Signal intensities were extracted for each sample using Genome Studio. Intensities were subsequently quantile normalised and a log2 ratio to normal reference DNA generated. Intensities from Type I and Type II probes were normalised independently as well as in combination. GC content normalisation was carried out using loess (Marioni et al., 2007). Circular Binary Segmentation (CBS), from the R package *DNAcopy*, was then performed to define chromosomal segments with differing copy number states. Thresholds for the identification of single copy CNAs were derived from the difference in log ratio between normal reference DNA from male and female samples (cut-off +/- 0.33) denoting a single copy change in the X chromosome, high-level amplifications and homozygous deletions were defined incrementally from this threshold.

2.5.4 Validation of exon-sequencing data by Immunohistochemistry and interpretation of results:

Immunohistochemistry was done in collaboration with UCL Advanced Diagnostics. All sequenced HNSCC samples were stained for PTEN and for Cyclin D1 using the Bond™-III system (Leica Microsystems, Inc., Buffalo Grove, US) according to the manufacturer's recommendation, as described in section 2.2.5. Antibody 04-409 (Millipore-Merck KGaA, Darmstadt, Germany) was used for PTEN staining and antibody P2D11F11 (Novocastra) was used for Cyclin D1 staining of 3 µm thick sections. For antigen retrieval method and primary antibody dilution please refer to Table 2.17. The stained

slides were examined and scored as previously described [151, 152], by two experienced histopathologists.

	Antibody dilution	Antigen retrieval method
CCND1	1:20	Bond solution 'ER2' for 30 minutes
PTEN	1:200	Bond solution 'ER2' for 20 minutes

Table 2.17: Antigen retrieval method and primary antibody dilution for CCND1 and PTEN staining.

2.5.5 Survival Analysis:

R statistical software v2.14.0 (http://www.r-project.org) was used for survival analysis of 9 HNSCC patients with a CCND1 amplification in their tumour tissue (detected by next generation sequencing; see method section 2.5.1) and 25 HNSCC patients who did not harbour a CCND1 amplification within their tumours. A Cox regression model (proportional hazard model) was fitted with a single CCND1 predictor (after having removed non-significant factors). A log rank test was performed and Kaplan-Meier curves were plotted.

2.6 Methodology used and data analysis applied to obtain results in chapter 6

2.6.1 Multi-dimensional scaling (MDS; principal coordinates analysis):

R statistical software v2.15.1 (http://www.r-project.org) was used for pre-processing of data and for classical multi-dimensional scaling (MDS) (principal coordinates analysis). MDS was used to visualise HPV+ and HPV- HNSCC methylation signatures within methylation datasets obtained from an HPV-induced cancer type (cervical cancer) and an smoking-induced cancer type (lung cancer).

The following Illumina Infinium 27k BeadChip datasets were identified on the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/), a public functional genomics data repository:

- ‘GSE32861 : Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression [Homo sapiens]’ from [153]
- ‘GSE30759: Epigenome analysis of normal and cancer tissue from the uterine cervix [Homo sapiens]’ from [154]

The respected datasets were downloaded and data matrices were extracted. Phenotype information was checked and the relevant methylation data (beta values) were extracted for cancer samples only (data on normal tissue/adjacent tissue were removed).

In detail, for the lung cancer dataset 27578 probe IDs of 59 lung cancer samples (of a total of 118 with 59 lung cancer and 59 adjacent tissue samples) were selected. For the cervical cancer dataset, 27578 probe IDs of 48 cervical cancer samples (of 63 samples in total) were selected.

The relevant methylation data from my own processed dataset of 18 HPV+ and 14 HPV- HNSCCs were extracted. 439385 probe IDs of 32 HNSCC, 18 HPV+ and 14 HPV- samples, respectively, were used for further filtering. Having ensured the same ordering of respective probe IDs, NA’s were removed, culminating in data matrices of 27300 probe IDs (59 lung cancer samples), 26871 probe IDs (48 cervical cancer samples), 439385 probe IDs (18 HPV+ HNSCC samples), and 439385 probe IDs (14 HPV- HNSCC samples). All the datasets were restricted to the common probe IDs (24145 probe IDs), by the ‘intersect’ function, resulting in the following datasets:

- beta values from intersected 24145 CpGs from 48 cervical cancer samples
- beta values from intersected 24145 CpGs from 59 lung cancer samples
- beta values from intersected 24145 CpGs from 18 HPV+ HNSCC samples
- beta values from intersected 24145 CpGs from 14 HPV- HNSCC samples

These were then used for multi-dimensional analysis.

In detail, the previously obtained methylation differences identified by comparing HPV+ with HPV- HNSCC samples served as a starting point, a semi-supervised selection of HPV-associated vs. smoking associated features. As outlined in chapter 3, 2757 MVPs were identified (using a Bayesian regularised t-statistics model) with a false discovery rate < 0.01 in HPV+ HNSCC, compared with HPV- HNSCC. The respective probe IDs were intersected with the common probe IDs identified in each dataset (lung cancer, cervical cancer, HPV+ HNSCC, HPV- HNSCC) using the 'intersect' function in R. This resulted in 90 common probe IDs identified across all the datasets.

In order to check whether I observe the same pattern in cervical cancer vs. lung cancer, a multidimensional scaling of the samples was created using a Euclidean distance measure after scaling all common probe ID features, with the R 'cmdscale' function. A Wilcoxon rank sum test was employed to test intersample distances between the evaluated datasets (the set of distances between HPV+ HNSCC and cervical cancer samples was compared with the set of distances between HPV+ HNSCC and lung cancer samples). Distances were graphically illustrated using the 'ggplot2' function.

2.6.2 Hierarchical clustering:

In order to obtain the presented results in chapter 6 (section 6.1.5), R statistical software v2.15.1 (http://www.r-project.org) was used and the 'hclust' function (http://svn.r-project.org/R/trunk/src/library/stats/R/hclust.R) was applied to perform hierarchical clustering (Euclidean distance, complete linkage) of HPV+ and HPV- HNSCC samples, using panels of genetic (chapter 5) and epigenetic candidate markers (chapter 3). 4 samples had to be removed as for these only methylation data, but no mutation data were available or vice versa (P40_neg, P70_neg, P95_neg, and P105_pos). A cut-off of a beta value of 0.5 was used to define a CpG site as either methylated (1) or unmethylated (0) for the purpose of the clustering.

Chapter 3

Host Methylome Analysis of HPV+ and HPV- HNSCC

The following chapter evaluates changes in the host methylome of HPV+ HNSCC, compared with HPV- HNSCC. I test my hypothesis that HPV modulates the epigenome in HNSCC by comprehensive methylome analysis of HPV+ and HPV- primary tumours and cell lines. In this chapter the results of my experiment are outlined, aiming to phenocopy the observed HPV-mediated DNA methylation signature by ectopic expression of HPV oncogenes in HPV- HNSCC cell lines.

3.1 Results.

3.1.1 Patient demographic data:

The median age is similar, but slightly higher in the HPV- group (58.4 vs. 56.9 years) (Table 2.3). Male to female ratio is similar between the groups, and the majority of cases show moderately or poorly differentiated histology with evidence of lymph nodal involvement at presentation. In the evaluated cohort, as predicted, the vast majority of HPV- cases are amongst active smokers and/or heavy alcohol users. The survival analysis which I performed on these 21 HPV+ and 21 HPV- samples, showed that the overall survival, is significantly higher (P = 0.036; Log Rank test) in patients who have HPV+ HNSCC, compared with those who have HPV- HNSCC (Figure 3.1).

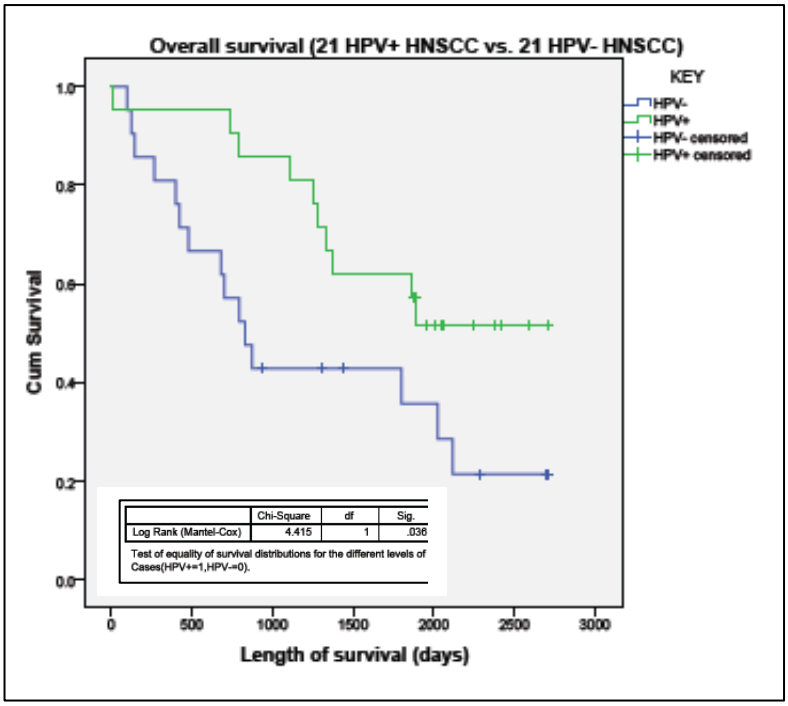


Figure 3.1: Survival analysis on the 21 HPV+ and 21 HPV- samples, selected for methylation analysis, showing that the overall survival is significantly higher in patients who have HPV+ HNSCC, compared with those who have HPV- HNSCC (*P* = 0.036; Log Rank test). One patient died of a stroke 3 days after the diagnosis with HPV+ HNSCC and this explains the initial drop of the Kaplan-Meier curve of the HPV+ HNSCC patient group.

3.1.2 HPV+ tumours have a distinct DNA methylation signature:

To investigate whether HPV+ and HPV– tumours have distinct epigenetic signatures, genome-wide DNA methylation profiling was performed using the 450k Illumina Infinium platform [83], which allows the methylation state of over 480,000 cytosine sites (mostly CpG sites) to be interrogated. Formalin Fixed Paraffin Embedded (FFPE) samples from 21 HPV+ and 21 HPV– tumours were analysed. Raw data were subjected to a stringent quality control analysis (Methods). This resulted in a raw data matrix of 439,385 probes and 32 samples (18 HPV+ and 14 HPV–). This raw data matrix was then subjected to a principal component analysis (PCA) to determine the nature of the largest components of variation. Using Random Matrix Theory (RMT) [134, 135] it was estimated that a total of 9 significant components of variation which were mainly correlated with biological factors. The first two

components correlated with HPV status (Figure 3.2) and confirming this, a scatterplot along these showed that samples segregated according to HPV status (Figure 3.3).

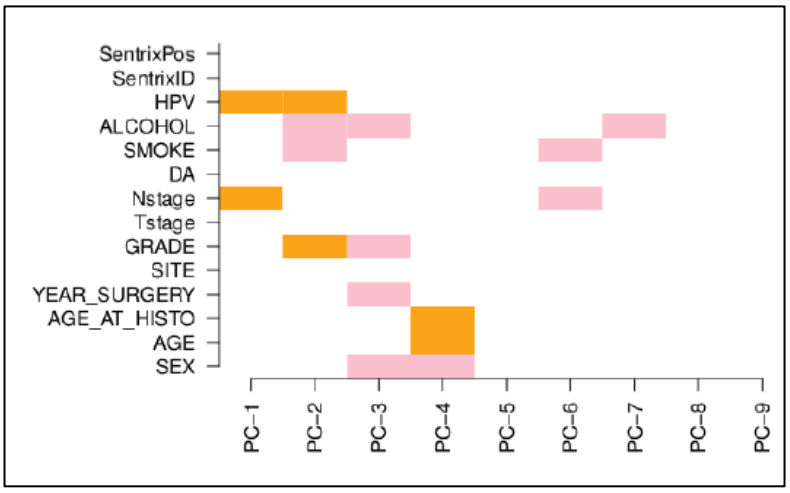


Figure 3.2: Singular value decomposition: PC-k denotes the kth principal component, DA denotes survival at censoring date. The first two principal components PC-1 and PC-2 most strongly correlate with HPV status, whilst the remaining significant components associate to clinical parameters including alcohol consumption, smoking, age, sex, tumour stage and grade. No association was found with technical factors (such as Sentrix position and Sentrix ID) on the array.

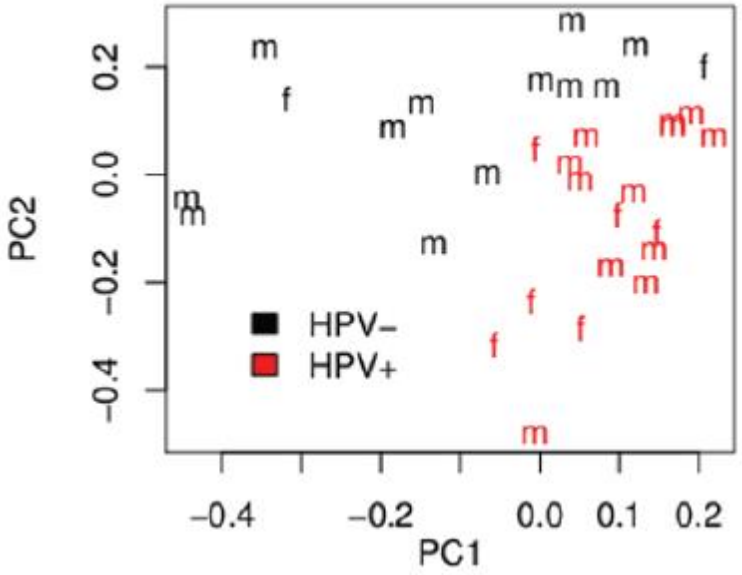


Figure 3.3: Singular value decomposition: The first two principal components clearly distinguish HPV status: m/f indicates male/female, HPV+ samples are plotted in black, HPV– samples are in red.

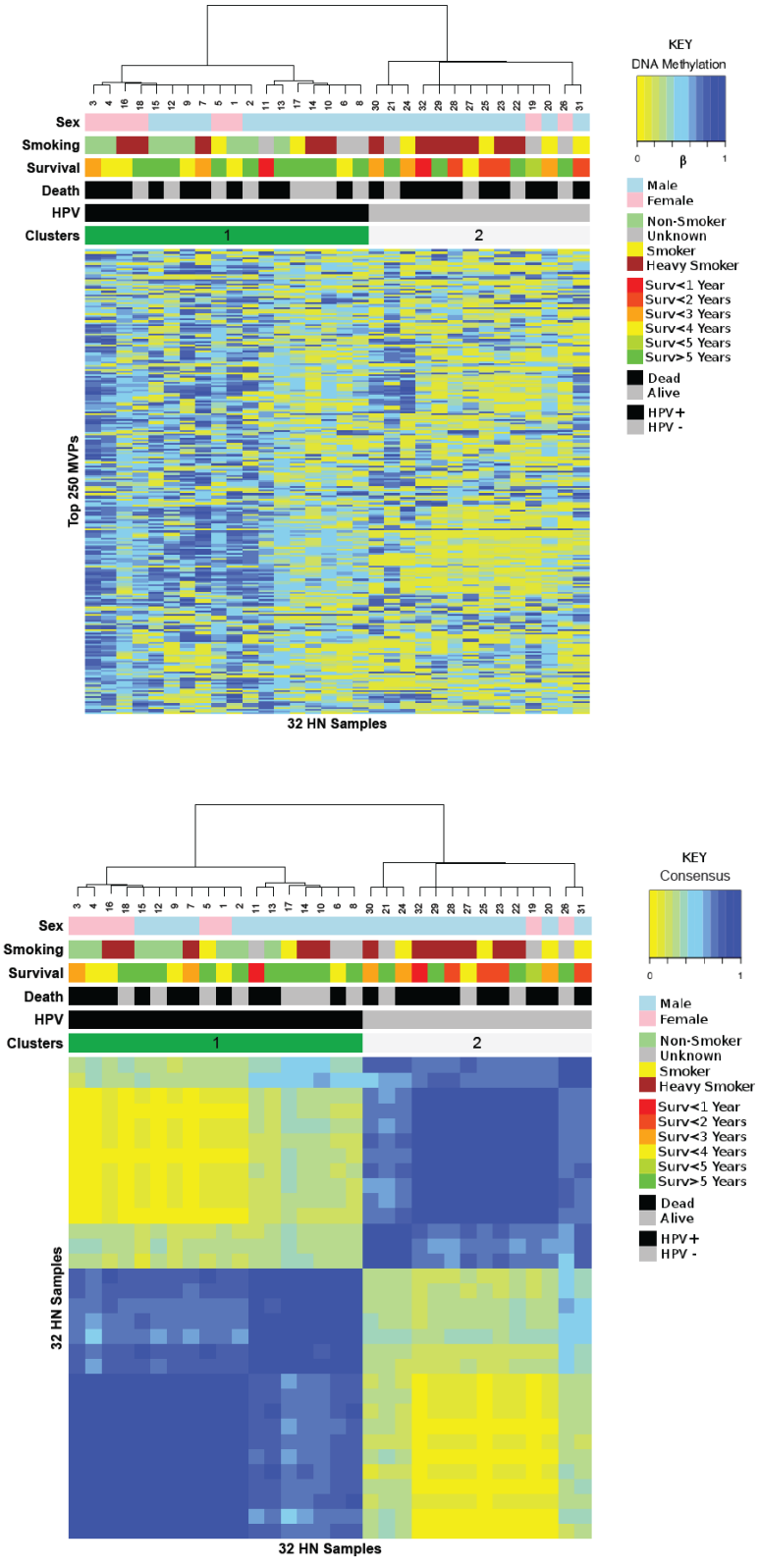


Figure 3.4: Unsupervised analysis of the top 250 MVPs in FFPE HPV+ and HPV- tumour samples.

Clusters inferred by the unsupervised consensus clustering algorithm for the top 250 methylation variable positions (MVPs) as found using the median absolute deviation (MAD) estimator.

Importantly, there was no substantial variation associated with technical factors including Sentrix position/Sentrix ID (Figure 3.2). Since the first two principal components of these data corresponded most strongly with HPV status, it was naturally expected that unsupervised clustering over the most variable probes would result in segregation of samples according to HPV status. As defined previously [155], such probes or CpG sites are referred to as Methylation Variable Positions (MVPs) and as hyper- or hypo-MVPs when directionality towards differential hyper- or hypomethylation has been ascertained. Segregation was confirmed by consensus clustering over the top 250 MVPs (Figure 3.4).

Next, a supervised analysis was performed in order to ascertain the association between DNA methylation and HPV status. To rank probes, a Bayesian regularised *t*-statistics model [156] was used, which has been used and validated in the context of DNA methylation data [157]. Consistent with the previous unsupervised analysis, a histogram of *P*-values from the supervised analysis showed a clear trend towards small significant *P*-values (Figure 3.5).

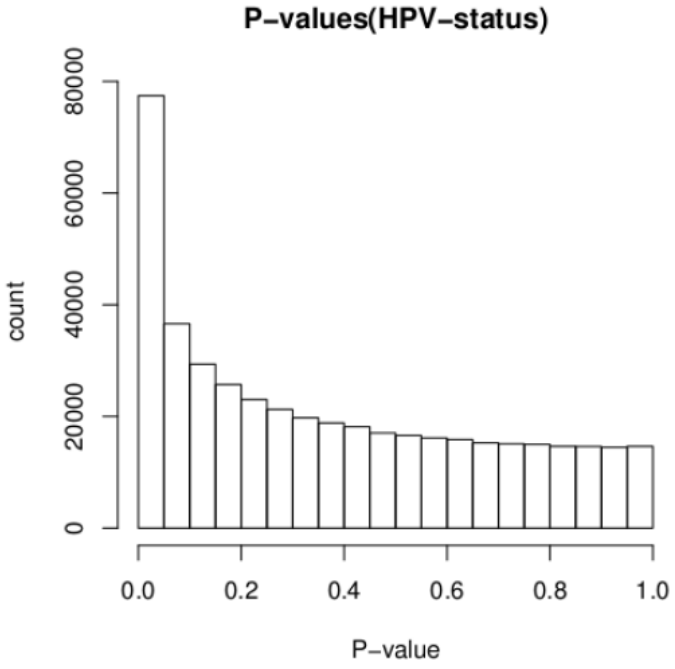


Figure 3.5: Histogram of *P*-values from the supervised analysis showed a clear trend towards small significant *P*-values

Using two alternative procedures (*q*-values [137] and a permutation approach (Methods), 2757 MVPs were found with a false discovery rate < 0.01, i.e. less than 1% of the 2757 MVPs are expected to be false positives. Of these 2757 MVPs, the overwhelming majority (2408, 87%) were hyper-MVPs in HPV+ samples compared with HPV–, indicating that HPV infection is associated with widespread gain of DNA methylation. The MVPs indicating differential methylation between HPV+ and HPV– samples did so independently of gender. Indeed, a ranked set of MVPs associated with HPV status for the 24 males and 8 females were derived separately and the resulting statistics were highly correlated (Figure 3.6).

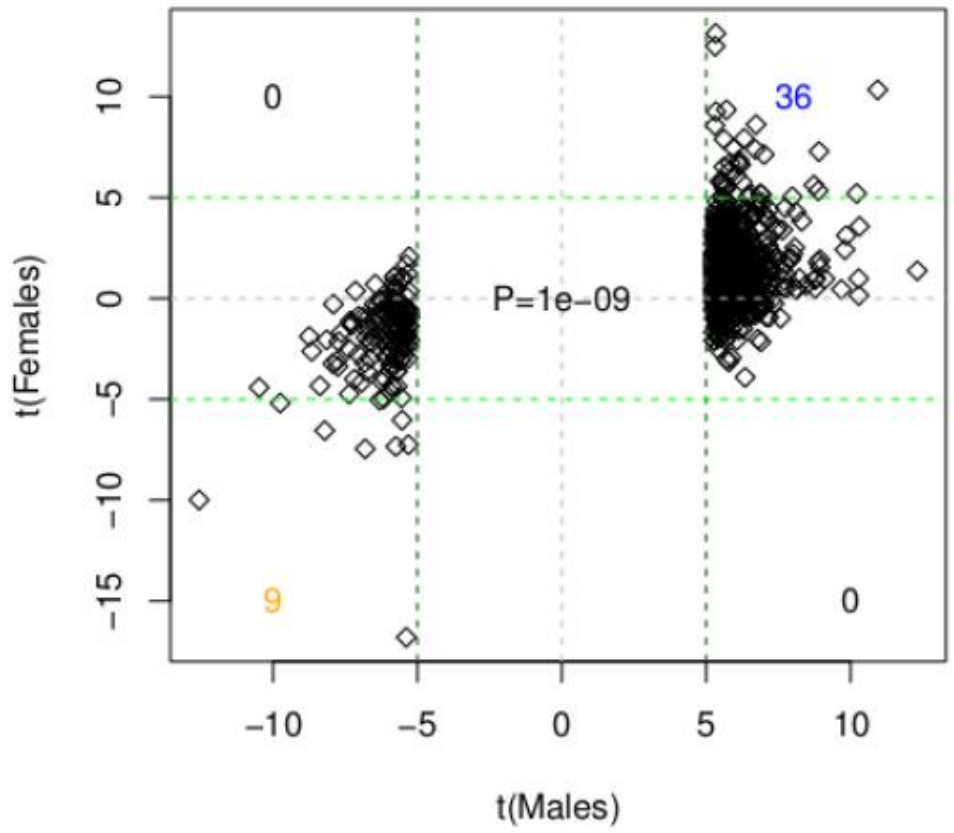


Figure 3.6: Independence of HPV status and gender: scatterplot of t-statistics of individual CpGs reflecting HPV status (positive t-statistics indicate hypermethylation in HPV infected samples). P-value computed using Wilcox rank sum test.

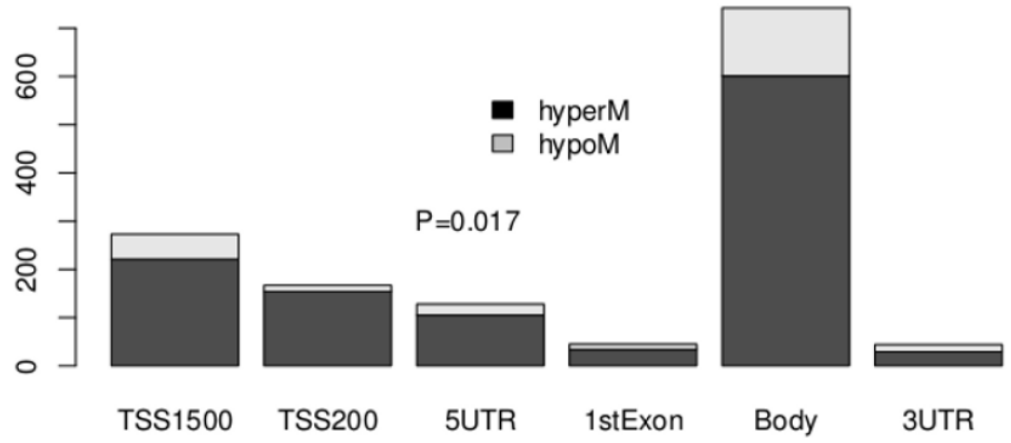


Figure 3.7: Methlyation status according to gene feature annotation, showing a clear trend towards hypermethylation (P= 0.017). Gene features: TSS1500, TSS200, 5'UTR, 1st exon, gene body, 3'UTR.

To investigate if the directional DNA methylation changes were related to the position of the MVPs relative to the corresponding genes, each MVP was first categorised into one of 6 gene feature groups (TSS1500, TSS200, 5'UTR, 1stExon, Body, 3'UTR). Hyper-MVPs in HPV+ samples were preferentially located upstream or near the TSS or in gene bodies but not in 1st exons, while hypo-MVPs in HPV+ samples were preferentially located in gene bodies (Figure 3.7). Taken together, these data clearly show that the HPV+ tumour samples have a distinct epigenetic signature which shows a significant skew towards hypermethylation (Figure 3.7).

3.1.3 HPV+ HNSCC are heterogeneous with a candidate CpG Island Methylator Phenotype (CIMP)

The enrichment of hyper-MVPs in HPV+ samples and the observation that many of these mapped to CpG islands suggested a possible association with CIMP in these samples. To investigate this further consensus clustering over the top 1000 MVPs was performed, a procedure similar to the one used previously to discover CIMP phenotypes in breast and brain cancer [158,

159]. The consensus clustering yielded 4 clusters, which still correlated with HPV status, but also revealed heterogeneity within the HPV+ and HPV- subtypes (Figure 3.8). Specifically, it was observed two main subgroups of HPV+ samples, with subtype 1a exhibiting higher methylation levels (Figure 3.8). This subtype was also characterised by higher average methylation levels when the MVPs were restricted to CpG islands (Figure 3.9), suggestive of CIMP. However, and in contrast to the CIMPs reported in breast, colon and brain cancer, there was no evidence of a stronger correlated hypermethylation pattern in this subtype than in the rest of HPV+ tumours. Interestingly, the samples in my candidate CIMP cluster 1a were all of poor outcome, exhibiting significantly shorter survival times compared with 1b which contained mostly good outcome patients (Log-rank P = 0.001, Figure 3.9; also Figure 3.8).

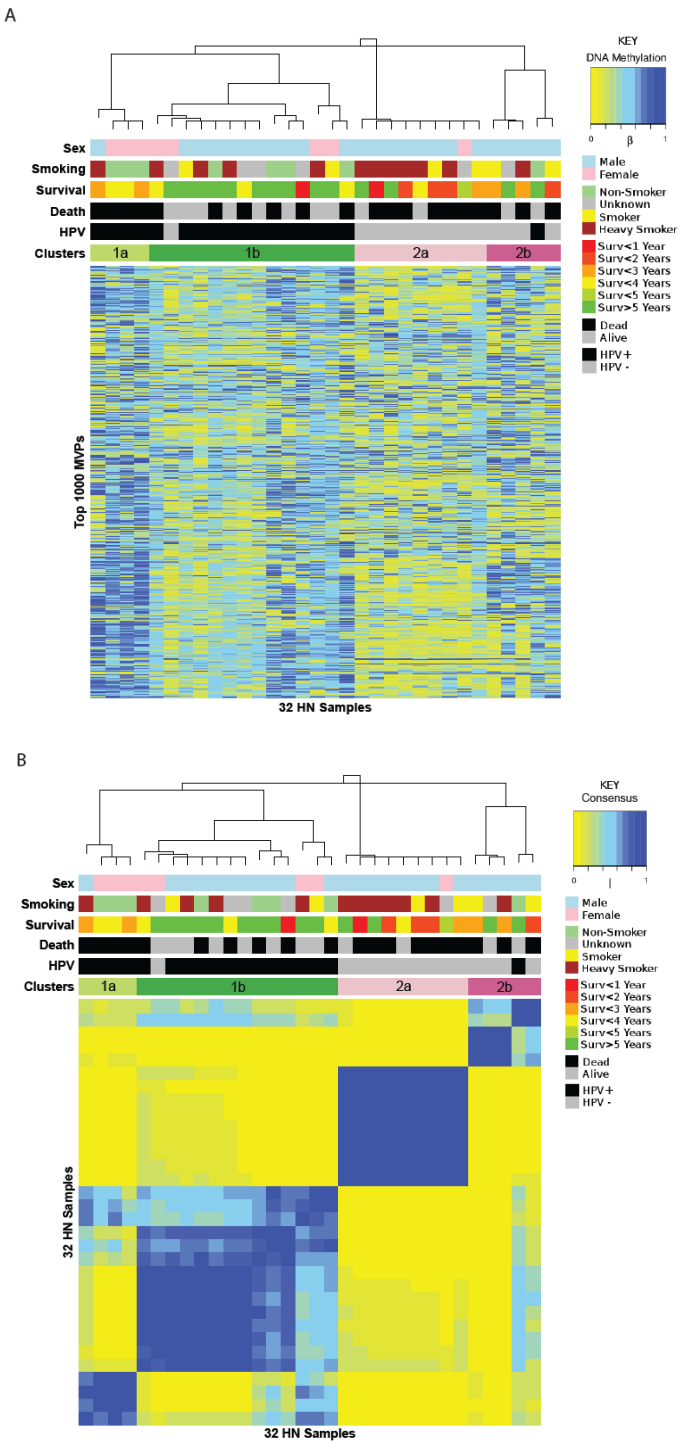


Figure 3.8: Unsupervised analysis of the top 1000 MVPs in FFPE HPV+ and HPV- tumour samples:

A) Consensus clustering identifies four sub-groups in HPV+ and HPV- groups, revealing sub-group 1a as candidate CIMP within the HPV+ group.

B) Clusters inferred by the unsupervised consensus clustering algorithm for the top 1000 MVPs as found using the MAD estimator.

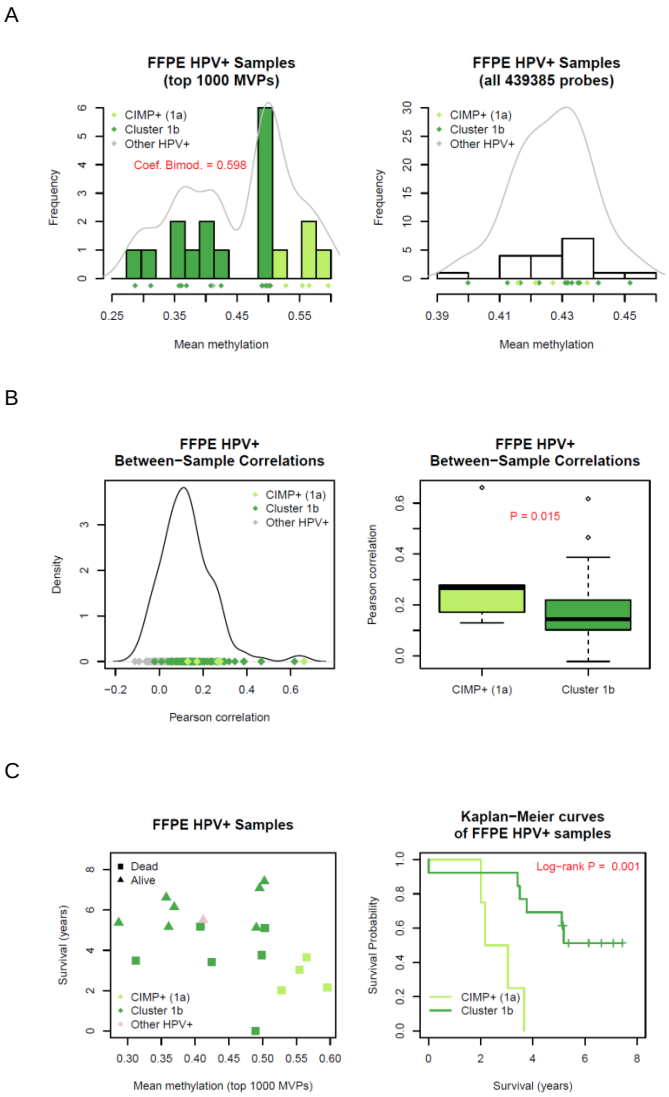


Figure 3.9: Testing for the presence of a HPV+ CIMP phenotype.

A) Distribution of mean methlyation values across the top 1000 MVPs in HPV+. As illustrated, there was a slight bimodality (coefficient of bimodality = 0.6, for which 0.555 ~ uniform and 1 ~ Bernoulli) in the distribution of mean methlyation values across the top 1000 MVPs in HPV+, with significant hypermethylation of the CIMP+ cluster (1a, light green). Across all probes, no multimodality was observed, nor any significant separation of mean methylation levels between the CIMP+ cluster and the large HPV+ positive cluster with lower methylation levels (1b, dark green). All points of the distribution are displayed below the histograms (black) and density plots (grey).

B) Significant separation (one sided *P* = 0.015) of between-sample correlation distributions of the CIMP+ cluster (1a) and cluster 1b.

C) Association of mean methylation levels with survival independent of HPV status. There was an anticorrelation between mean levels across the top 1000 MVPs and survival time. Comparing the overall survival of patients in the CIMP+ cluster (1a) with patients in HPV+ cluster 1b, the HPV+ CIMP+ patients showed a significantly worse outcome (Log-rank *P* = 0.001).

3.1.4 Validation of the hypermethylation signature in HPV+ tumours

To validate these findings, I performed 450k Infinium profiling on 6 independent fresh frozen (FF) samples (3 HPV+ and 3 HPV–). All 6 samples passed the quality control criteria (Methods). The same Bayesian supervised analysis was applied to rank MVPs according to how well they discriminated the 3 HPV+ from the 3 HPV– samples. The overwhelming majority of MVPs that were significantly hypermethylated in HPV+ FFPE samples were also hypermethylated (and many significantly hypermethylated) in the HPV+ FF samples relative to HPV– samples. Comparing the regularized *t*-statistics obtained from the 32 FFPE samples with those obtained from the 6 FF samples, a very strong agreement was observed (Wilcox *P* = 3e-35; Figure 3.10A). A control set of probes which did not differ between HPV+ and HPV– FFPE samples, also did not correlate with HPV status in the FF set (Figure 3.10B).

To further validate these findings and the 450k technology, the 450k methylation values for CpG islands to methylation scores calculated from MeDIP-seq using the MEDIPS package was compared, testing the 3 HPV+ FF and 3 HPV– FF samples. A strong agreement between the two methods was observed (Figure 3.11).

Next, it was asked if the DNA methylation changes associated with HPV status were also present in HPV infected HNSCC cell lines. HPV *t*-statistics between the FFPE and HPV cell line experiments correlated strongly (Wilcox *P* = 2e-27, Fig. 3.10c). Changes in absolute mean beta value (Δβ) observed between HPV+ and HPV– cell lines were substantially larger in contrast to the changes detected in FFPE (Paired Wilcox *P* = 7e-15) and again larger than in FF tumour samples (Paired Wilcox *P* = 3e-13, Figure 3.12).

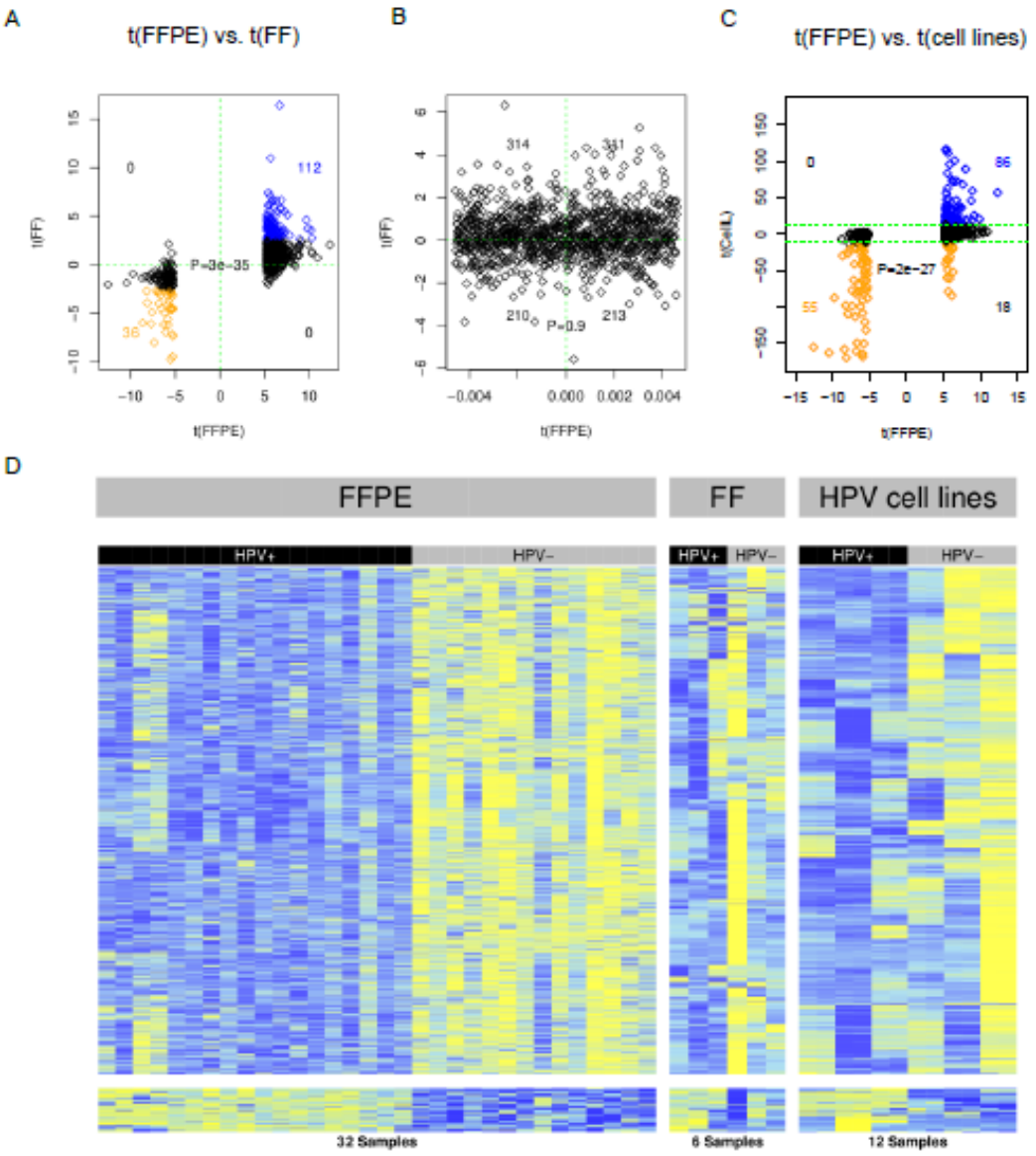


Figure 3.10: Validation of HPV+ and HPV- methylation signature.

A-C) Validation of consistency of *t*-statistics between FFPE and (A) FF samples, (B) FF control probes and (C) HPV+ against HPV- cell lines.

D) Heatmap representation of signature of consistent hyper-MVPs (top) and hypo-MVPs (bottom) in Infinium DNA methylation data. The DNA methylation (beta) values are represented using a colour scale from yellow (low DNA methylation) to blue (high DNA methylation) normalized across each MVP. The HPV+ HNSCC methylation signature contains 2194 consistent hyper-MVPs and 74 consistent hypo-MVPs across all three datasets.

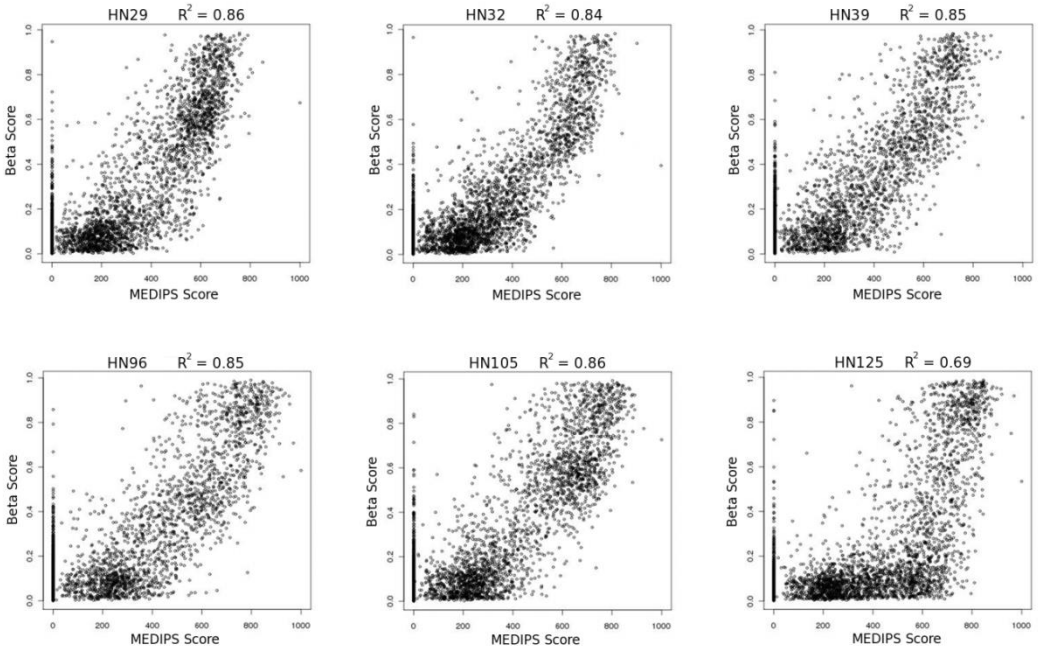


Figure 3.11: Validation of the 450k technology by MeDIP-Seq. Methylation score calculated from MeDIP-Seq using the MEDIPS package (MEDIPS Score) and beta values for CpG islands (Beta Score), testing the 3 HPV+ FF and 3 HPV– FF samples. Correlation coefficients (Pearson) illustrated above each scatterplot.

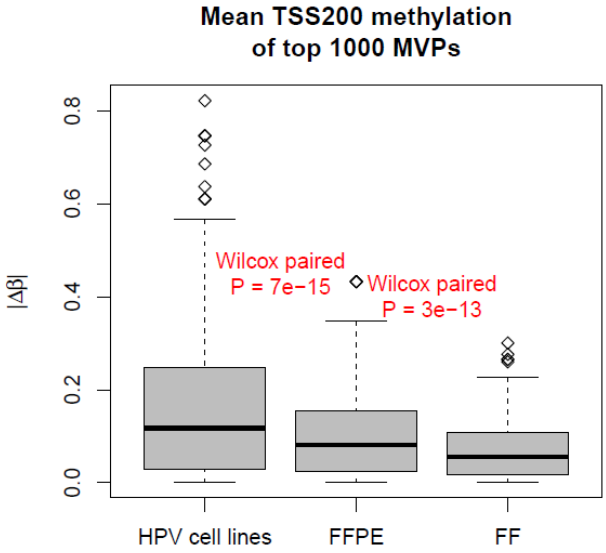


Figure 3.12: Comparison of methylation differences across all datasets. Boxplots of changes in absolute mean beta value (Δβ) of the top 1000 MVPs in HPV+ and HPV– cell lines, FFPE tumours and FF tumours. Methylation changes are substantially larger in cell lines in contrast to the changes detected in FFPE (Paired Wilcox *P* = 7e-15) and again larger than in FF tumour samples (Paired Wilcox *P* = 3e-13).

In conclusion, we identified 2194 consistent hyper-MVPs and 74 consistent hypo-MVPs across all three experiments (FFPE HNSCCs, FF HNSCCs and HNSCC cell lines, as illustrated in Figure 3.10D). This confirms that my HPV+ hypermethylation signature obtained from FFPE samples validates in an independent set of HNSCC samples as well as in HPV+ HNSCC cell lines and indicates a strong association of the observed methylation signature and HPV status.

3.1.5 Ectopic expression of HPV-16 oncogene *E6* partially phenocopies the hypermethylation signature:

In order to functionally validate my obtained HPV+ hypermethylation signature, I infected an HPV– HNSCC cell line with retroviral vectors containing either or both HPV-16 oncogenes, *E6* and *E7*. After confirmation of ectopic expression of these HPV oncogenes in 3 clones of each cell line (Table 3.1), I performed DNA methylation profiling on *E6*, *E6*&*E7* and *E7* infected clones relative to empty vector controls.

The skew towards hypermethylation (observed in the described experiment on FFPE HNSCCs) was confirmed to be highly significant in *E6* and *E6*&*E7* infected clones against the background probability of hypermethylation (there was widespread hypermethylation in *E6* and *E6*&*E7* clones) (Figure 3.13, Monte Carlo, $P = 0.007$).

The distribution of methylation changes in *E6*, *E6*&*E7* and *E7* infected clones compared with controls is illustrated in Figure 3.14.

In contrast to E6 and E6&E7, E7 was shown not to contribute to this hypermethylation signature (comparison of clones infected with *E7* compared with controls; one-sided Wilcox $P = 1$). The skew towards hypermethylation was significantly larger for *E6* than for *E6*&*E7*. This is consistent with the lower expression levels of *E6* in *E6*&*E7* co-infected clones compared with *E6* infected clones (Table 3.1). The results of the entire experiment are summarized in two graphs (Figure 3.13 A, B). In conclusion, ectopic expression of E6 in HPV– HNSCC cell line clones partially phenocopies the

hypermethylation signature observed in HPV+ HNSCC tumours. However, no final conclusion can be drawn, as E7 expression was detected at lower levels in infected cell lines, compared with those of E6 (Table 3.1).

	Normalized* E6 expression levels	Normalized* E7 expression levels
pLXSN-E6 clone 4	0.43	0.00
pLXSN-E6 clone 1	0.25	0.00
pLXSN-E6 clone 2	0.11	0.00
pLXSN-0 clone 2	0.00	0.00
pLXSN-0 clone 3	0.00	0.00
pLXSN-0 clone 4	0.00	0.00
pLXSN-E6/E7 clone 1	0.01	0.11
pLXSN-E6/E7 clone 2	0.01	0.06
pLXSN-E6/E7 clone 5	0.01	0.06
pLXSN-E7 clone 2	0.00	0.09
pLXSN-E7 clone 4	0.00	0.02
pLXSN-E7 clone 5	0.00	0.07

***Normalized to UPCI:SCC90**

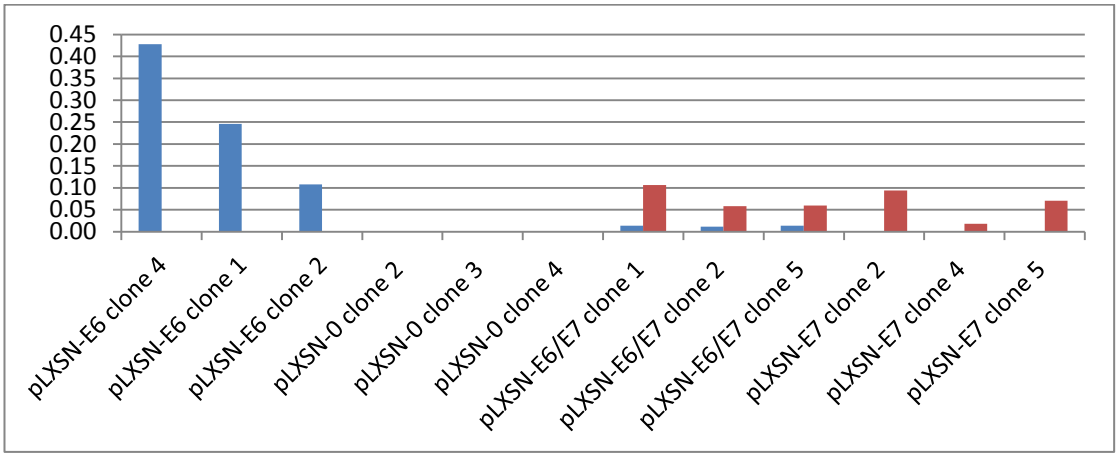


Table 3.1: Normalized expression levels of the viral oncogenes *E6* and *E7* in infected cell line clones with graphical illustration (bar chart).

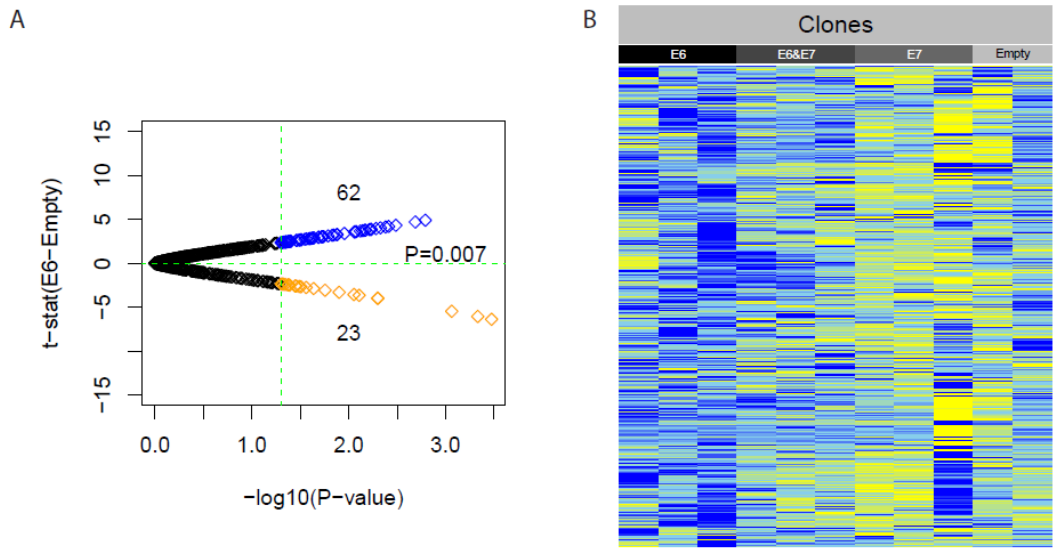


Figure 3.13: Validation of consistent hyper-MVPs in E6 and E6&E7 infected cell line clones. A) FFPE hyper-MVP signature consistent with E6 (E6 only and E6&E7 infected) vs. empty vector controls (Monte Carlo P= 0.007). The volcano plot shows t-statistics of E6 vs. empty clones plotted against log10 FFPE *P*-values.

B) Heatmap representation of consistent hyper-MVPs in E6, E6&E7, E7 infected clones and empty vector controls. Yellow indicates relative hypomethylation in HPV+ samples whilst blue indicates hypermethylation (MVPs normalized across samples).

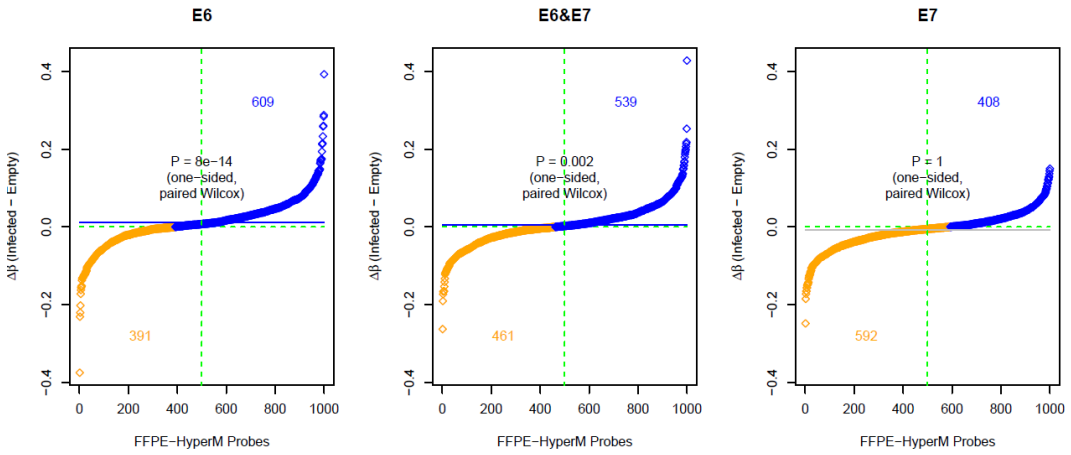


Figure 3.14: Scatterplots of change in mean methylation change Δβ (*y*-axis) against hypermethylated probes in FFPE samples (*x*-axis). A one-sided Wilcox test of top 1000 FFPE probes (annotating across all datasets) illustrating that the hypermethylation signature observed in FFPE is partially recapitulated by E6 infected and E6&E7 infected clones. The mean of the distribution is marked with a horizontal and the total number of probes in each class is annotated in the appropriate colour (hypomethylated: orange; hypermethylated: blue).

3.1.6 Enrichment of PRC2 targets, especially members of the Cadherin superfamily within the hypermethylation signature:

To find consistent targets across all the data sets (FFPEs, FFs and HPV+ cell lines), all of the consistent hyper- and hypo-MVPs identified above to genes were assigned and a Gene Set Enrichment Analysis (GSEA) was conducted. The hyper-MVPs which make up the majority (96.7 %) of MVPs identified several SUZ12 and PRC2 targets (Appendix Table A.3: Gene set enrichment analysis on consistent hyper-MVPs) including multiple members of the Cadherin superfamily. Indeed, there was significant (11 probes total, Fisher P = 4e-7) enrichment of hyper-MVPs within the Cadherin genes. This was also the case in the top 1000 MVPs (Fisher P = 0.0003). The possible biological impact of these Cadherins was tested in three separate ways: (1) by showing that their methylation states were sufficient to accurately cluster samples according to HPV status, (2) by showing that they have consistent and significant hypermethylation across their promoter regions (see DMR section) and (3) by showing that this promoter hypermethylation associates with decreased gene expression in existing data (see expression section). Using the *k*-medoids clustering algorithm (*pam* in the R package *survival*), the 11 cadherin-annotated probes within the top 1000 MVPs were found to be sufficient to detect HPV status (84% correctly classified; 27/32). These 11 MVPs mapped to CpG islands, shores or shelves of 6 Cadherin genes (*CDH8*, *CDH15*, *PCDH8*, *PCDH9*, *PCDH10* and *PCDHB3)*. The remaining 3.3% hypo-MVPs were enriched for two gene sets previously shown to display upregulation of gene expression in HPV+ head and neck cancers [20, 22]. Among the top 100 hits consistent enrichment of genes involved in DNA replication and binding, the MAPK pathway and E2F-targets was also found (Appendix Table A.4: Gene set enrichment analysis on consistent hypo-MVPs).

To assess the MVP associations in a more biologically relevant context, they were grouped into differentially methylated regions (DMRs) if at least 3 (range 3-7) had correlated differential methylation levels within the TSS200

promoter region. TSS200 was chosen because it was the most significantly positively (*P* = 2.4e-5) enriched category of the 6 tested (Figure 3.15).

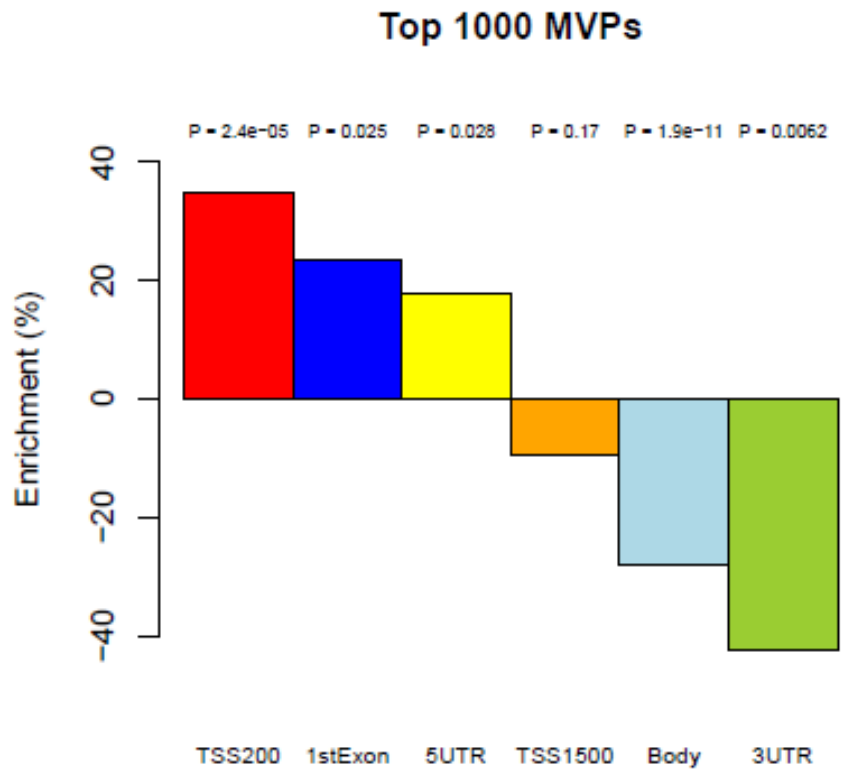


Figure 3.15: Enrichment (%) of hypermethylation of promoter regions amongst MVPs. Significant over-representation of the TSS200 (35%), 1stExon (23%) and 5'UTR (18%) regions amongst the top 1000 MVPs (*P* < 0.05 in each case) and significant under-representation in the Body (–28%) and 3'UTR (–42%)

Applying this filter, the 2194 consistent hyper-MVPs mapped to 906 distinct genes such that 416 had at least three probes in their respective TSS200 regions. Of these 416 genes, 43 hypermethylated TSS200 DMRs (delta beta > 0.1) were derived across FFPE HNSCCs, FF, HNSCCs and HNSCC cell lines. A sample permutation approach yielded an expected 4.4 false positives (Methods). HPV cell lines showed the largest changes in mean TSS200 hypermethylation, significantly larger than FFPE (Paired Wilcox *P* = 5e-05) which showed significant hypermethylation relative to FF (Paired Wilcox *P* = 1e-06, Figure 3.16).

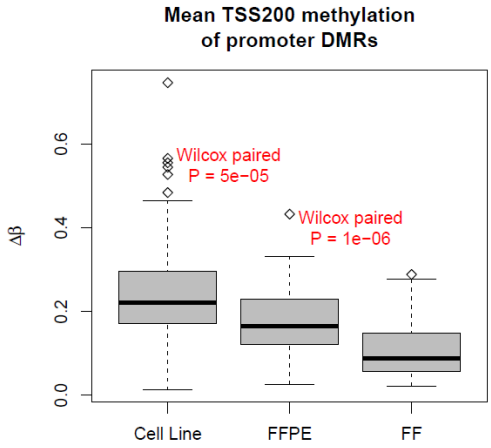


Figure 3.16: Comparison of methylation differences across all datasets. Boxplots of changes in mean beta value (Δβ) of promoter DMRs in HPV+ and HPV– cell lines, FFPE tumours and FF tumours. HPV cell lines showed the largest changes in TSS200 methylation, significantly larger than FFPE (Paired Wilcox *P* = 5e-05) which showed significantly higher methylation relative to FF (Paired Wilcox *P* = 1e-06).

Using the same approach for the 74 consistent hypo-MVPs, 5 hypomethylated TSS200 DMRs were derived. Figure 3.17 shows exemplar profiles of a hyper-DMR for *CDH8* and a hypo-DMR for *MEI1*.

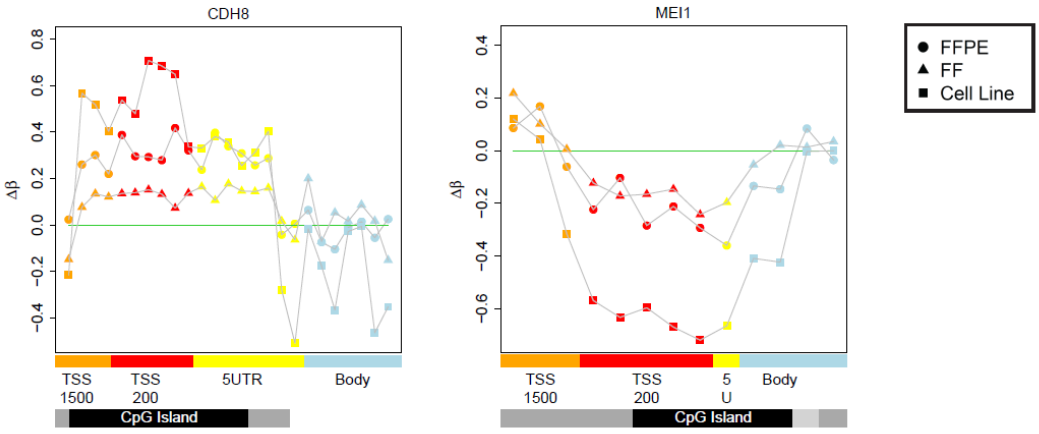


Figure 3.17: Exemplar profiles of a hyper-DMR for CDH8 and a hypo-DMR for MEI1. Comparison of DMR profiles obtained from FFPE HNSCCs, FF HNSCCs and HNSCC cell lines. The profiles clearly show the increasing power to detect MVPs and DMRs dependent on cell type purity (cell line > laser capture microdissected FFPE > FF). Feature annotation is as provided by BeadChip and methylation values are colour-coded accordingly: TSS1500 = orange (1500 bp to 200 bp upstream of the transcription start site), TSS200 = red (200 bp upstream of the transcription start site), 5'UTR = yellow, gene body = blue, CpG islands = black, CpG shores = grey and CpG shelves = light grey.

These profiles highlight the increasing power to detect MVPs and DMRs dependent on cell type purity (cell line > laser capture microdissected FFPE > FF). All DMRs associated with Cadherin genes had sample-permutation estimated $P < 0.05$. Profiles of these are displayed in Figure 3.18. In summary, 43 genes with promoter hypermethylation were found consistently across all datasets (permutation FDR 10%) including multiple Cadherin genes and other PRC2 targets. In addition, five genes were found (*SNTB1*, *CYP7B1, MEI1, ICA1, and FAM163A*) with hypomethylated promoter DMRs.

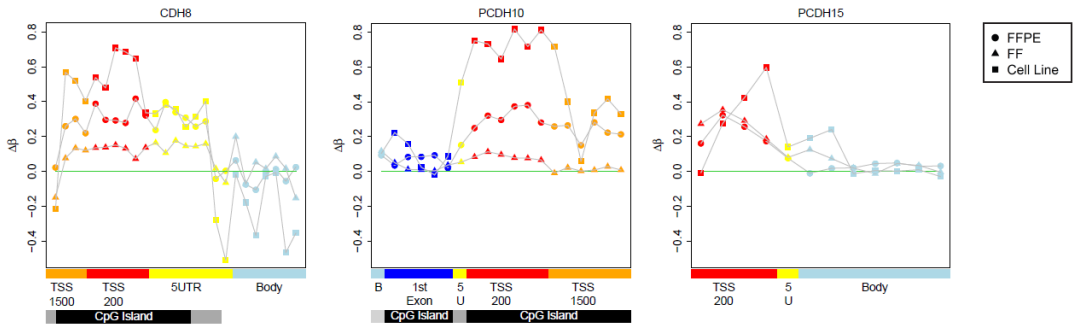


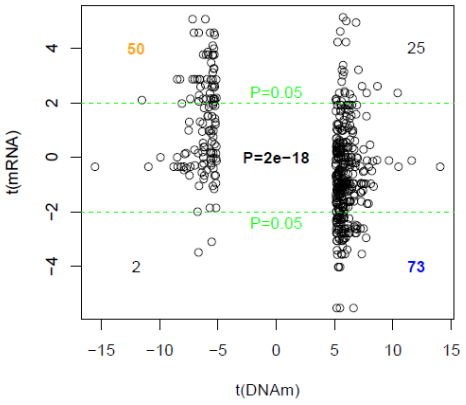
Figure 3.18: Profiles of hypermethylated TSS200 promoter DMRs of Cadherin genes (across FFPE HNSCCs, FF HNSCCs and HNSCC cell lines), namely CDH8, PCDH10 and PCDH15. Feature annotation is as provided by BeadChip and methylation values are colour-coded accordingly: TSS1500 = orange (1500 bp – 200 bp upstream of the transcription start site), TSS200 = red (200 bp upstream of the transcription start site), 5'UTR = yellow, gene body = blue, CpG islands = black, CpG shores = grey and CpG shelves = light grey.

3.1.7 Integration with publicly available gene expression data

For additional functional evidence of the effect of DNA methylation changes on gene expression, the observed methylation differences between HPV+ and HPV- FFPE tumour samples were compared with publicly available gene expression data [20]. The top 500 MVPs mapping to CpG islands were compared with the differential expression *t*-statistics of their associated genes. A significant negative correlation was observed (Wilcox $P = 2e\text{-}18$; Figure 3.19A). A list of genes with consistent TSS200 DMRs across all datasets (FFPE HNSCCs, FF HNSCCs, HNSCC cell lines) and which also exhibited differential gene expression in the independent Pyeon gene

expression data set is shown in Figure 3.19B. Among these were 3 Cadherin genes (*CDH8, PCDH10* and *PCDHB11*). These data are consistent with Cadherin genes being targets for HPV-mediated hypermethylation and transcriptional silencing in HNSCC.

A



B

Gene name	t-DNAm(FFPE)	t-mRNA(Pyeon)
ELMO1	7.48	-2.61
CDH8	**6.93**	**-2.64**
CRMP1	5.89	-3.55
PCDH10	**5.80**	**-2.00**
C14orf162	-5.68	2.35
MSX2	5.62	-3.06
SYN2	5.56	-2.46
PCDHB11	**5.48**	**-4.02**
HTR1E	5.26	-2.24
PITX2	5.22	-2.78

Figure 3.19: Integration of DNA methylation data with public gene expression data.
A) DNA methylation correlates with decreased gene expression: scatterplot of t-statistics between HPV+ and HPV- FFPE cancer samples (top 500 differentially methylated MVPs restricted to CpG islands) shows significant anticorrelation between DNA methylation and gene expression. Gene expression data were from (Pyeon et al., 2007).
B) List of top 10 anticorrelated targets: Differentially methylated genes in promoter region (TSS200) which also exhibit differential gene expression in the independent Pyeon gene expression data set. 3 Cadherin genes are found among the top10 hits (bold).

3.2 Discussion of results obtained in chapter 3.

My findings reported in this chapter represent the most comprehensive epigenetic study of a virus-induced cancer known to date and the first to validate the existence of an HPV-mediated DNA methylation signature in HPV+ HNSCC. Supported by extensive validation using independent samples and different methods, the signature showed a clear skew towards hypermethylation which was most prominent at promoter regions (defined by TSS200). However, there was also significant hypomethylation at gene bodies which, together with promoter hypermethylation, is a clear hallmark of gene silencing [160]. It is well documented for instance that hypermethylation of the promoter region of tumour suppressor genes plays an important role in

cellular transformation [161] and indeed consistent hypermethylation (defined by both hyper-MVPs and hyper-DMRs) was found in the promoter regions of such genes as well a candidate CIMP in the HPV+ samples. CIMPs have been observed in a number of cancers, including colon cancer [162], brain cancer [159] and breast cancer [158]. In all cases reported so far, CIMP has been associated with a favourable clinical outcome. HPV+ HNSCC patients are also associated with a more favourable outcome [15] but the candidate CIMP observed here was only present in a sub-group of four HPV+ patients who all had a shorter survival and recorded death. To my knowledge, this is the first time that a CIMP (albeit a candidate CIMP) is associated with potentially less favourable clinical outcome. Furthermore, I was able to show that my signature (defined by top 1000 MVPs) was independent of gender and predictive for smoking status and length of survival, confirming previous findings [8, 15, 163].

The inclusion of multiple sample types (FFPE, FF and cell lines) in the validation part revealed an important observation with direct implications for projects with an epigenetic biomarker component such as ICGC (http://www.icgc.org/), IHEC (http://www.ihec-epigenomes.org/), OncoTrack (http://www.oncotrack.eu/) and others. While FF samples have emerged as gold-standard for the genomic analysis of cancer, my data demonstrate that archival FFPE samples may be superior for certain epigenomic analyses, particularly when combined with laser-capture microdissection as illustrated in Figure 3.17. The largest differences in DNA methylation levels were consistently observed in cell lines, followed by laser microdissected FFPE, followed by FF. This general trend was expected because DNA methylation is known to be cell-type specific but the evident high level of confounding cellular heterogeneity (resulting in dilution of the respective MVP/DMR signals) in carefully biobanked FF samples is nevertheless noteworthy.

The most interesting finding arising from the GSE analysis is the enrichment of numerous members of the Cadherin superfamily which are targets of PRC2 and are implicated in many cancers and cancer-specific processes [164], including epithelial to mesenchymal transition (EMT), a process by which carcinomas become invasive and acquire the ability to

metastasize [165]. Notable examples include *E-Cadherin* (*CDH1*), *T-Cadherin* (*CDH13*) and *Proto-Cadherin 10* (*PCDH10*) which are recognized tumour suppressor genes and have been found hypermethylated in a number of human cancers [164]. Among the 49 PRC2 targets (defined by consistent hyper-MVPs) identified here were 10 genes of the Cadherin superfamily in HPV+ HNSCC, including *CDH8* and *CDH13* (both also hypermethylated in cervical cancer [166], *CDH18*, *CDH19*, *CDH23*, *PCDH10*, *PCDH15, PCDHB1, PCDHB4 and PCDHB15.* Moreover, the 11 MVPs in 6 Cadherin genes identified amongst the top 1000 MVPs by unsupervised clustering analysis of FFPE HNSCC samples, warrant further investigation as potential biomarkers as they clustered my HPV+ and HPV– samples according to HPV status with high accuracy.

Two lines of evidence were obtained with respect to functional support for the identified hypermethylation signature. First, I was able to partially phenocopy the signature by ectopic expression of the two HPV oncogenes *E6* and *E7* in a HPV- HNSCC cell line. Combinatorial analysis showed that *E6* is one of the main viral effector genes. The underlying mechanism remains unknown and is subject to future work such as analysis of cross-talk between *E6* and DNA methyltransferases (DNMTs), effect of *E6* on *TP53*, number and distribution of viral integration sites into the host genome and the viral methylome itself. Second, integration of publicly available expression data with my DNA methylation data [20] was performed. Among the top 10 anticorrelated (high promoter methylation and low expression) genes were three of the Cadherins, namely *CDH8*, *PCDH10* and *PCDHB11*. The inverse scenario (low promoter methylation and high expression) was also observed and both are likely to contribute to the different clinical behaviour of HPV+ and HPV- HNSCC with regard to survival and response to therapy. Linking these two lines of evidence suggests a possible mechanism whereby HPV could drive tumour progression by promoting epithelial to mesenchymal transition (EMT) [165] through epigenetic silencing of Cadherins in addition to its established role in tumour initiation.

In conclusion, this work significantly advances the understanding of the epigenetic dynamics at genomic loci targeted by oncogenic viruses as

demonstrated here for loci associated with the infection of HPV in HNSCC. Based on the previously established finding (supported by the results of my survival analysis outlined in section 3.1.1) that HPV+ HNSCC patients have a better prognosis than HPV- patients, it is tempting to speculate that this advantage may be partly epigenetically mediated. My results certainly implicate DNA methylation in this process. If confirmed, targeted reprogramming of the identified HPV-mediated hypermethylation signature (or parts of it) in HPV- patients offers an immediate translational application for my finding. Although still at an early experimental stage, targeted reprogramming has recently been demonstrated, including in cancer cells [167, 168]. In the longer-term, these data will contribute to the identification of diagnostic and prognostic markers, as well as putative therapeutic targets.

Chapter 4

Analysis of the Viral Methylome in HPV+ Head and Neck Cancer

This chapter describes the methylation analysis of the HPV genome in my set of HNSCC samples and the validation of the obtained results in these HNSCC samples, in HNSCC cell lines and in an independent sample set of a different type of HPV-induced cancer. It also describes differences in L1 and L2 expression, associated with the methylation status of the identified methylated CpG sites.

4.1 Results.

4.1.1 HPV type 16 was confirmed in all HPV+ samples and methylation within the viral genome was demonstrated:

All FF HPV+ HNSCC were found to contain copies of HPV type 16, as tested by MeDIP-Seq. In brief, the obtained normalized reads of the 3 HPV+ HNSCC samples were successfully aligned against the HPV type 16 genome (NC_001526). The sequenced reads from HPV- HNSCC samples (tested by E6 qPCR initially) that did not align to the HPV type 16 genome (NC_001526) were further aligned to 1776 available viral genomes. None of these reads

aligned to any of the 29 available human papilloma viruses (Table 2.13), obtained from NCBI Genomes (http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239), such as HPV type 18 (NC_001357) and others.

By applying the MeDIP-Seq technology methylation changes within the HPV type 16 genome in HNSCC were demonstrated. After normalization of reads from these 3 HPV+ FF HNSCC samples according to average HPV copy number per cell for all three samples (Sample #39: 96,94 copies per cell; #105: 110.44 copies per cell; #125: 0.66 copies per cell) in reads per million, RPM, the relative methylation levels were calculated within the viral genome (Figure 4.1).

Figure 4.1: HPV type 16 methylome in HNSCC. Sites of methylation within the viral genome were mainly detected at the boundary of the L1/L2 ORF and within the E1 ORF (created in Circos [169])**.**

These methylation changes were particularly observed at the boundary of the L1 and the L2 gene and within the E1 gene (Figure 4.1). Methylation in the boundary region was detected in all samples investigated. The sequence of this part of the viral genome harbours 6 CpG sites at genomic position 5600, 5606, 5609, 5615, 5707, and 5724. The first 4 CpG sites are located within a sequence of only 16 bp and the remaining 2 follow 92 bp and 109 bp downstream, respectively (Figure 2.7).

4.1.2 Validation of methylation at the boundary of the L1/L2 ORF of the viral genome in FF HNSCC samples and in an independent set of HPV+ HNSCC cell lines by BS-Seq:

The results obtained by MeDIP-Seq were validated by applying the BS-Seq technique on a 421 bp PCR product of the viral genome, covering the 6 CpG sites of interest and two additional CpG sites in genomic positions 5925 and 5961, in the 3 FF HPV+ HNSCC samples (Figure 4.2). BS-Seq was also used to test the methylation status of this region in 3 HPV+ HNSCC cell lines (UPCI:SCC90, UM:SCC47, 93VU-147T) and methylation at all 6 CpG sites at genomic positions 5600, 5606, 5609, 5615, 5707, and 5724 was observed (Figure 4.2).

4.1.3 5-AZA treatment of HPV+ HNSCC cell lines and its effect on L1 and L2 expression:

Three HPV+ HNSCC cell lines (UPCI:SCC90, 93VU-147T, UM:SCC47) were incubated with increasing amounts of 5-AZA (1 µM, 2 µM, 5 µM, 10 µM) for 72 hours.

Employing L1 RT qPCR and L2 RT qPCR, L1 and L2 viral gene expression levels were measured. No significant increase in L1 expression was observed in any cell line upon treatment with 5-AZA (Figure 4.3).

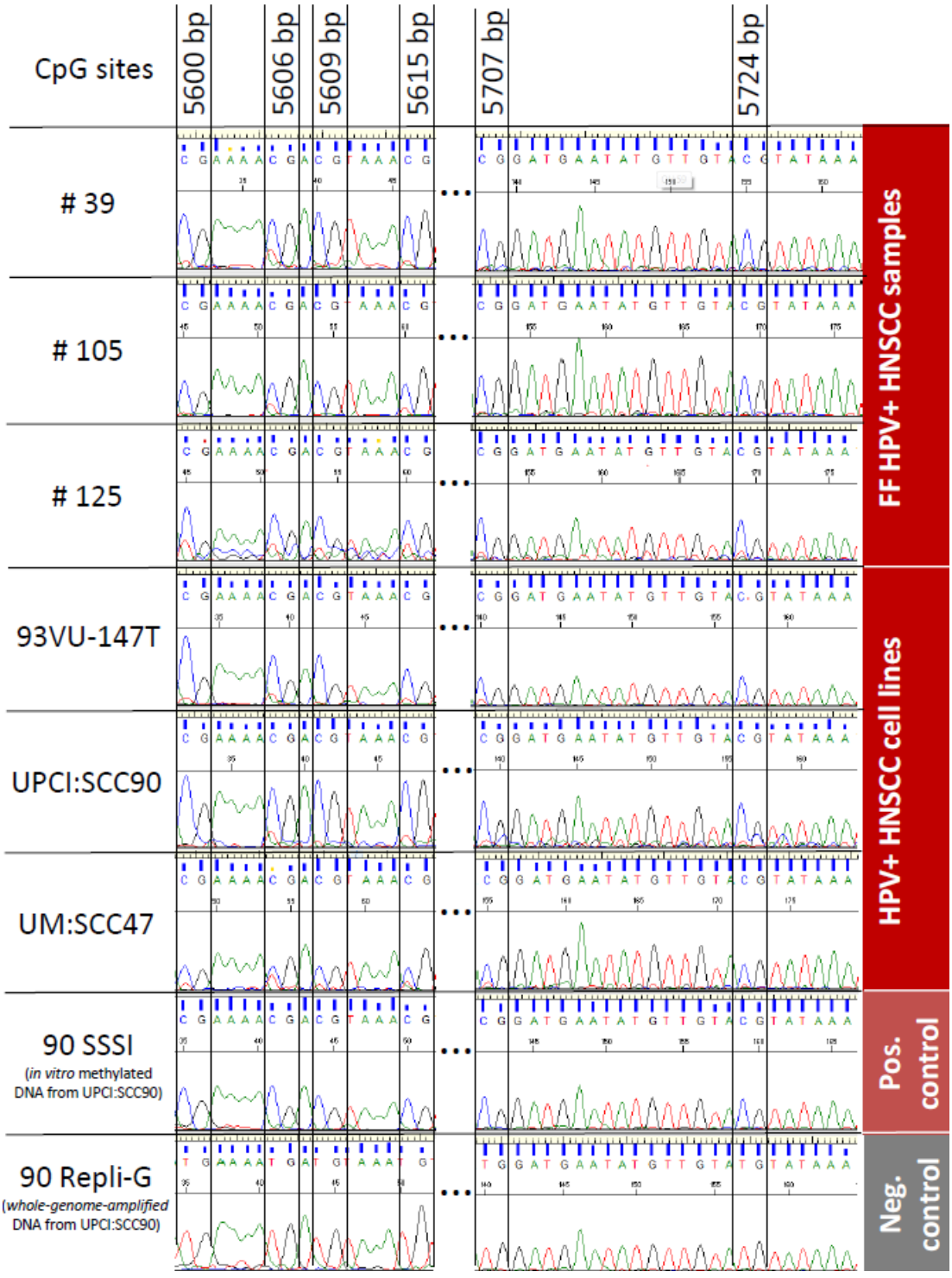


Figure 4.2: Methylation of the L1/2 boundary region is observed in all tested FF HPV type 16+ HNSCC tissue samples and 3 HPV type 16+ HNSCC cell lines (UPCI:SCC90, UM:SCC47, 93VU-147T). Selected BS-Seq traces, covering the 6 CpG sites of interest, are illustrated along with the positive (*in vitro* methylated DNA from UPCI:SCC90) and negative (whole-genome amplified DNA from UPCI:SCC90) control at six CpG sites at genomic position 5600, 5606, 5609, 5615, 5707, and 5724 of the HPV type 16 genome.

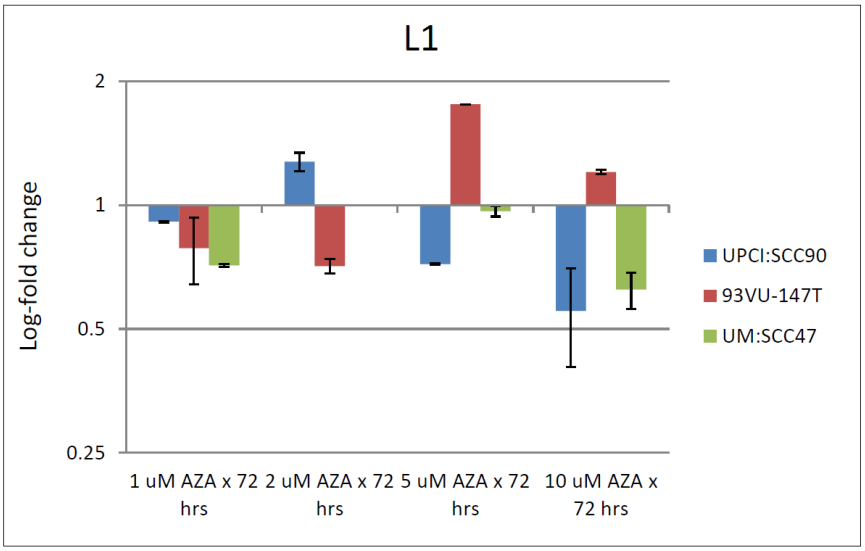


Figure 4.3: Log-fold change of L1 gene expression in UPCI:SCC90, 93VU-147T and UM:SCC47 upon treatment with increasing amounts of 5-AZA (1 µM, 2 µM, 5 µM, 10 µM) for 72 hours.

A significant increase in L2 expression was observed in 93VU-147T upon treatment with 5-AZA at concentrations of 2 µM, 5 µM, and 10 µM for 72 hours. No significant increase of L2 gene expression was observed in UPCI:SCC90 and UM:SCC47 under the same conditions (Figure 4.4).

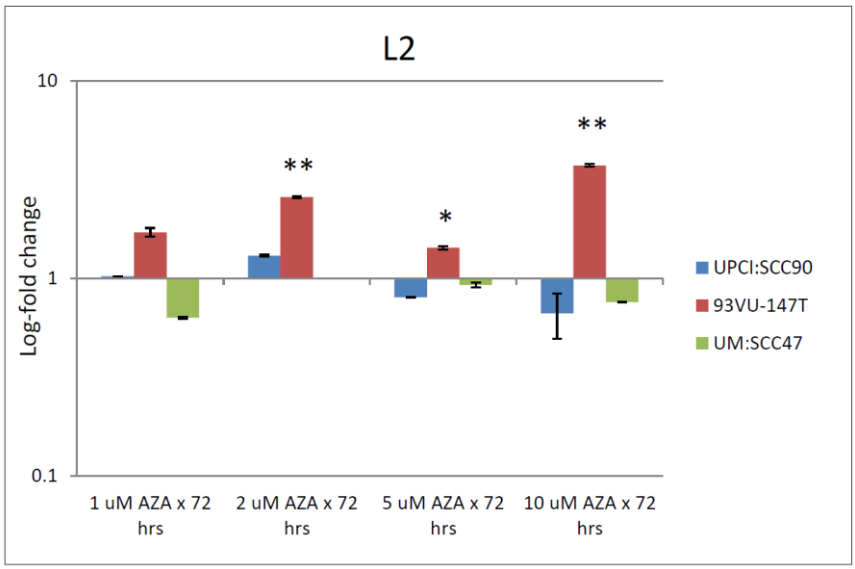


Figure 4.4: Log-fold change of L2 gene expression in UPCI:SCC90, 93VU-147T and UM:SCC47 upon treatment with increasing amounts of 5-AZA (1 µM, 2 µM, 5 µM, 10 µM) for 72 hours. A significant increase in L2 expression was observed in 93VU-147T after incubation with 5-AZA at concentrations of 2 µM ($P < 0.01$; unpaired T-test), 5 µM ($P < 0.05$; unpaired T-test), and 10 µM ($P < 0.01$; unpaired T-test) for 72 hours.

The methylation status of the L1/L2 boundary region upon treatment with 1 µM, 2 µM, 5 µM, and 10 µM of 5-AZA was tested using the BS-Seq technique, as described in method section 2.4.2. No demethylation of this locus was observed in any cell line after incubation of cell lines with the tested concentrations of 5-AZA.

In summary, this preliminary experiment of head and neck cancer cell lines, shows a significant increase in L2 gene expression in a single cell line (93VU-147T) upon treatment with 5-AZA only. No significant gene expression change was found in any other tested cell line. No change of the methylation status at this locus (tested by BS-Seq) was observed upon treatment with 5-AZA.

4.1.4 Integration sites identified within the host genome:

By selecting for paired-end reads of which one aligned with high confidence to HPV type 16 and the other to human DNA, enrichment of these reads ('integration site' reads) within genomic locations was calculated. The probability of an 'integration site' read aligning to a specific genomic region was weighted according to the total number of reads aligning in the chosen window. The resulting empirical *P*-values were log transformed (-log10) and plotted along the HPV type 16 genome. Thereby, viral integration sites within the host genome were identified.

The integration sites across the human genome appear to be random, whereas there is a significant enrichment of potential integration sites within the viral E2 (E4 and E5) region, as illustrated in Figure 4.5 and Figure 4.6. A possible explanation for this is discussed in section 4.2.

Figure 4.5: Potential integration sites of HPV type 16 in the human genome and illustration of location of 'integration site reads' in the HPV genome, representing the genomic region of the virus disrupted during the integration process. Links coloured dark blue indicate those located in the HPV region most significantly enriched for integration sites. HPV and human genome drawn at different scale (created in Circos [169]).

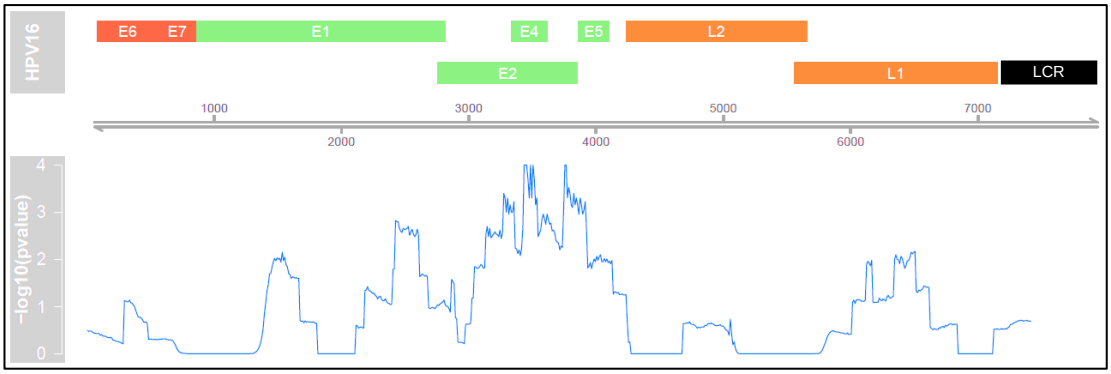


Figure 4.6: Enrichment of potential integration sites in the HPV genome as measured by –log transformed empirical *P*-value. Region containing the E2, E4 and E5 genes display the most significant degree of enrichment.

4.2 Discussion of results obtained in chapter 4.

In line with previous reports, showing that HPV type 16 is found in almost all cases of HPV+ HNSCC [28], all FF HPV+ HNSCC samples tested in this thesis had integrated copies of HPV type 16 only.

By employing the MeDIP-Seq technology on HNSCC samples, allowing the interrogation of 100% of the HPV methylome, methylation within one region of the HPV type 16 genome was detected consistently across all samples. This 124 bp region lies at the boundary of the L1 and the L2 gene and harbours a total of 6 CpG sites.

Methylation in this specific location has previously been shown in cervical cancer [103, 114]. Fernandez et al. showed that methylated CpG sites within this region of both the HPV type 16 and HPV type 18 genome is present in the majority of cervical cancer samples tested, rather than in premalignant cervical lesions and in carriers [103]. As part of their comprehensive HPV type 16 methylome analysis Park et al. assessed the methylation status of this region in advanced stage III/IV HNSCC patients [119]. Although, not specifically mentioned by the authors, they found consistent methylation of CpGs within this region across every sample tested. The main result presented by Park et al. was that large parts of the HPV type 16 genome, in particular the long-control region (LCR), which controls transcription of the E6

and E7 oncogenes [25], was found to be unmethylated in the majority of examined cases. In line with the findings by Park et al. the LCR region is found unmethylated in all FF HPV+ HNSCC samples tested in this thesis, suggesting that methylation in this genomic location may be a rare event in HNSCC. The LCR region harbours multiple binding sites for E2. Methylation within the LCR is known to prevent the binding of E2, thereby increasing the expression of E6 and E7, the two key oncogenes [104]. Thus, one would expect to find methylated CpG sites within this region. One explanation may be that, as the E2 gene is often disrupted or lost upon viral integration, the selective pressure to methylate this genomic segment is lost in the absence of the transcriptional repressor E2 [119].

Based on the findings by Fernandez et al., it is conceivable that methylation of the viral genome plays an important role in HPV-induced carcinogenesis, rather than a bystander role only. Moreover, the boundary of the L1 and L2 gene is reported more frequently methylated in cervical cancer samples than in premalignant and normal tissue [114]. Hence, the question arises whether methylation of this site is directed by the virus or rendered by a potential host defence mechanism. The HPV genome does not encode for any DNA methyltransferase (DNMT) and thus it is believed that the viral genome is methylated by the human host cell DNMTs [104]. This may either be directed by the virus through cross-talk between viral proteins and DNMTs (in order to regulate intrinsic cell-differentiation dependent [113] and temporal viral gene expression) or it may represent a potential host defence mechanism, with the host aiming to alter viral gene expression to avoid viral proteins disrupting normal cellular function. Assuming a cybernetic model of carcinogenesis, it is likely that both processes take place at the same time with abundant paths of interaction.

Having confirmed the absence of methylation within the late gene promoter *P670* (within the E7 viral gene region) by MeDIP-Seq, it was tempting to test whether methylation at the L1/L2 boundary region may have some influence on the expression of late genes. Preliminary L1 and L2 gene expression data after 5-AZA treatment in HPV+ HNSCC cell lines showed no significant change of expression in tested cell lines apart from a significant change in L2

expression in 93VU-147T. No change of methylation of the L1/2 boundary region was observed upon treatment with 5-AZA. This region of interest remains fully methylated at all concentrations of 5-AZA used and suggests that this locus may be resistant to the effect of 5-AZA. This of course raises the question why a change in L2 expression was observed in 93VU-147T, although the methylation status of the tested CpG sites remained unchanged. This may point towards an epigenetic mechanism of L2 regulation which is unrelated to changes in the examined region. However, it has to be acknowledged that these expression data are very preliminary and potential explanations provided are only speculative. Further experiments need to be carefully planned and are subject to future work.

Irrespective of the potential effects of methylation in this specific location, it may serve as a diagnostic biomarker for HPV+ HNSCC. Park et al. employed MSP to test for LCR methylation status in serum and saliva samples of HPV+ HNSCC patients. As no control group was included in that study, it can be regarded as a successful feasibility trial [119]. However, based on the findings discussed in this section and the results by Brandsma et al. [114] and Fernandez et al. [103], testing the methylation status of this region at the boundary of L1/L2 in serum and saliva samples may represent a far more powerful diagnostic biomarker.

Apart from methylation at the L1/2 boundary region of the viral genome, methylation within the E1 gene was detected to varying degrees in the FF HPV+ HNSCC tissue dataset. The significance of these findings remains to be investigated.

One of the main advantages of having employed MeDIP-Seq is that potential integration sites can be detected by selecting for paired-end reads that align both to the viral and host genome (‘integration site reads’). This is a far more reliable method than calculating the ratio of E2 and E6 viral gene expression as a measure of integrated vs. episomal HPV type 16 [119], owing to the fact that E2 is frequently disrupted during the integration of the viral genome into the host genome [42]. My data show that integration sites across the human genome appear to be random. However, there was a significant enrichment

of potential ‘integration site reads’ within the viral E2 (E4 and E5) region, suggesting that this is a common location of disruption of the viral sequence upon integration into the host genome in HNSCC. This is in line with evidence, showing that disruption of the HPV genome at this location upon integration is a common and early event in the history of cervical HPV infection [42]. Limitations to the current findings are that only a small set of samples were used for analysis and that relatively few ‘integration site reads’ were detected Thus, the described integration sites within the host genome will have to be evaluated further in a larger set of samples.

In conclusion, the obtained methylation data, having mapped the entire HPV type 16 methylome in HNSCC and having detected methylation at CpG sites within a region at the boundary of the L1 and L2 gene, along with findings reported in cervical cancer, suggest that this may not only represent a significant event in HPV-induced carcinogenesis, but may also lead the way to the identification of novel diagnostic markers for HPV+ HNSCC.

Chapter 5

Genomic Analysis of HPV+ and HPV- HNSCC

This chapter describes the results of the genomic analysis of HPV+, compared with HPV- HNSCC, and discusses the findings.

5.1 Results.

5.1.1 Patient demographic data:

The median age was slightly higher in the HPV- group (58 vs. 56.5 years) (Table 2.4). The male to female ratio was similar between the groups, and the majority of cases showed moderately or poorly differentiated histology with evidence of lymph nodal involvement at presentation. In my cohort, as predicted, the vast majority of HPV- cases were active smokers and/or heavy alcohol users (Figure 5.1).

5.1.2 Exon-sequencing:

Paired-end sequencing of hybrid-captured DNA, targeting 3,230 exons in 182 genes often mutated in cancer was employed. The sequence analysis revealed that HPV+ and HPV- oropharyngeal carcinomas cluster into two distinct subgroups, with few overlapping genetic alterations (Figure 5.1 and Figure 5.2A). *TP53, PIK3CA, CCND1, CDKN2A together with CDKN2B, PTEN, FBXW7, and SOX2* were among the most frequently genetically altered genes (>10% of HNSCC cases).

TP53 mutations are detected in 100% of HPV- samples (Figure 5.1) and the list of observed *TP53* mutations is illustrated in Table 5.1. HPV+ and HPV-

HNSCC samples clustered, but not into two distinct subgroups after exclusion of the *TP53* mutation data (Figure 5.2B).

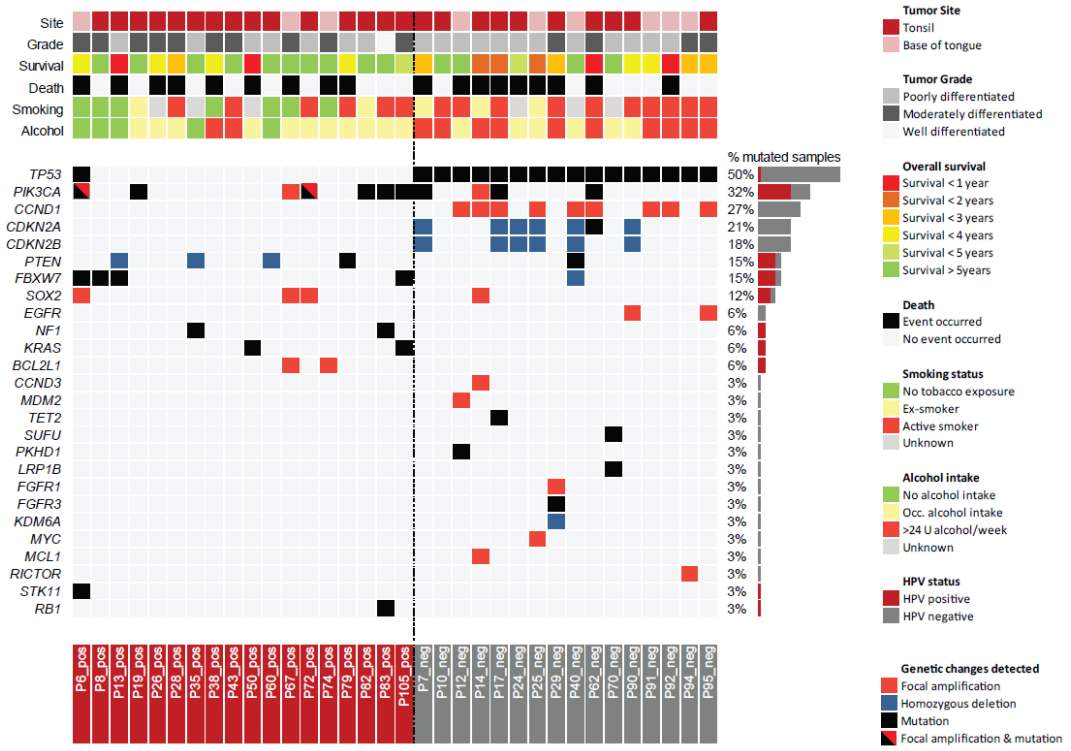


Figure 5.1: Illustration of somatic events in HPV+ and HPV- HNSCC revealed by deep sequencing. Relevant demographic and histological data are described above the heatmap of genomic changes. The colour coding of the observed changes and patient characteristics are explained in the key on the right.

Copy number alterations in *CCND1* amplification and *CDKN2A/B* deletions were exclusively detected in HPV- cases (in around 55% and 40% of cases respectively). Comparing the overall survival of patients with a *CCND1* amplification (n=9), with the ones without a copy number alteration in this genomic location (n=25), a trend towards a worse outcome for patients harbouring the amplification is observed (Figure 5.3; P = 0.0855; Log Rank test). The Cox regression model (proportional hazard model) which was fitted showed a weak trend towards *CCND1* amplification in HNSCC (P ~ 0.0945 with a hazard ratio of ~2.3, corresponding to a 130% increase in the risk of dying at any moment).

Sample names	TP53 mutations
P6_pos	R290C
P8_pos	
P13_pos	
P19_pos	
P26_pos	
P28_pos	
P35_pos	
P38_pos	
P43_pos	
P50_pos	
P60_pos	
P67_pos	
P72_pos	
P74_pos	
P79_pos	
P82_pos	
P83_pos	
P105_pos	
P7_neg	R175H
P10_neg	Y234H
P12_neg	Y220S
P14_neg	R273L
P17_neg	G154fs
P24_neg	L130fs
P25_neg	Q165
P29_neg	Y236
P40_neg	R306
P62_neg	Y220C
P70_neg	Q104
P90_neg	L114fs, L330fs
P91_neg	R337L
P92_neg	R335L, G334V
P94_neg	T155P
P95_neg	920-1C>T splice

Table 5.1: List of *TP53* mutations revealed by deep sequencing in HPV+ and HPV- HNSCC samples.

PIK3CA mutation or amplification, and *PTEN* inactivation by gene copy loss or mutation were seen in over 60% of HPV+ tumours, and in 31% HPV- tumours. *FBXW7* alterations were present in over 15% of all samples and *SOX2* amplification in 12% of cases (Figure 5.1).

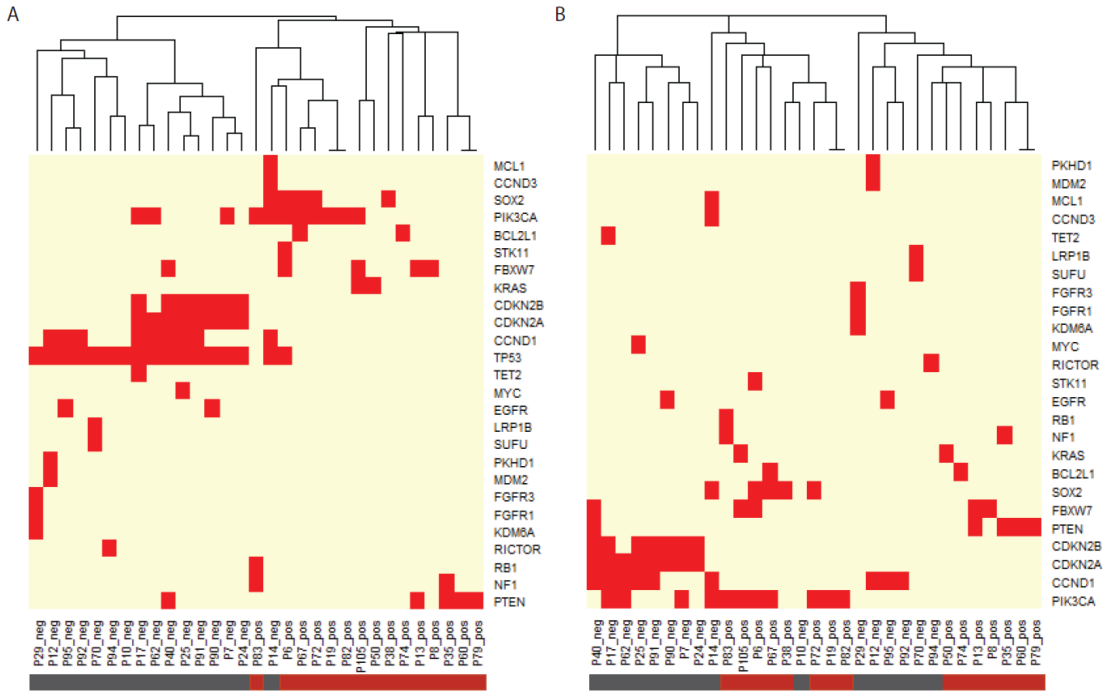


Figure 5.2: (A) Hierarchical clustering of HPV+ and HPV- HNSCC samples using all detected genetic changes. One HPV+ sample (P83_pos; heavy smoker) has an *RB1* mutation and clusters with the HPV- group.
(B) Hierarchical clustering excluding mutations within *TP53* reveals that HPV+ and HPV- tumours do not cluster into two distinct subgroups.

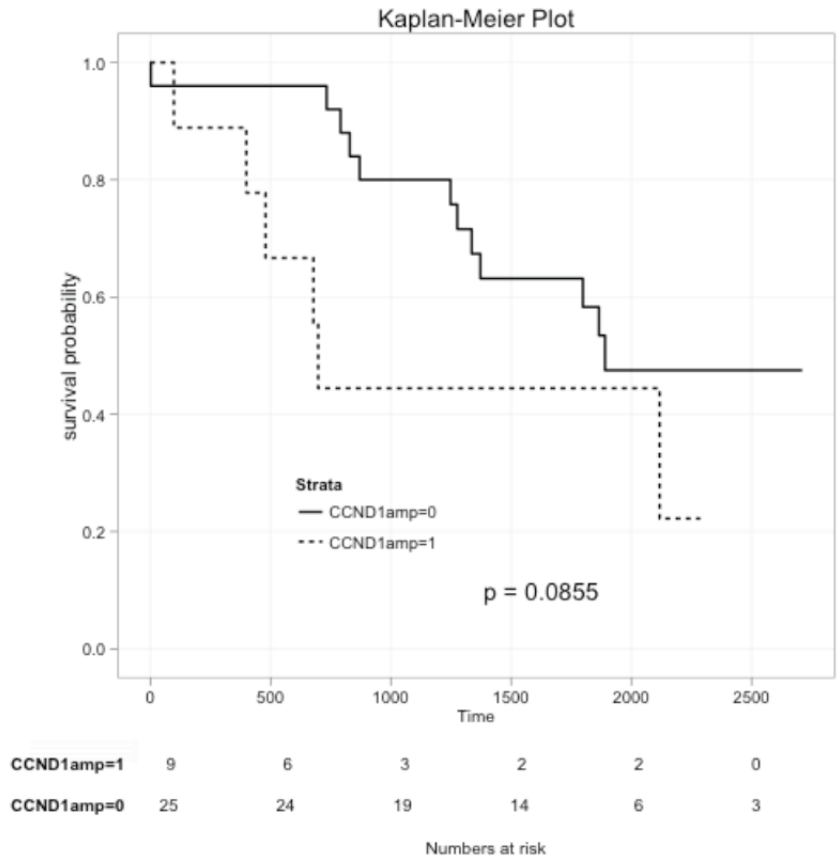


Figure 5.3: Kaplan Meier Plot. Survival analysis, comparing the overall survival of 9 HNSCC patients with a *CCND1* amplification in their tumour tissue with the 25 patients without. A trend was observed for overall survival (log rank test; *P* = 0.0855)

5.1.3 Validation of obtained results:

To validate the results, Infinium CNV profiling, Sequenom OncoCarta panels v1.0 and v3.0 and immunohistochemistry was applied. Copy number gains and losses detected by next-generation sequencing (NGS) were interrogated by Infinium CNV profiling. 42 of 44 (95%) copy number alterations detected by sequencing were confirmed (Figure 5.4).

In order to obtain a global picture, I mapped all copy number alterations detected by Infinium CNV profiling. Comparing the obtained genome-wide copy number alteration profiles between HPV+ and HPV- HNSCCs revealed that overall, similar genomic regions harbour concordant copy number changes in both groups (in particular in chromosomes 3, 7 and 14).

Amplification of 3q seems to be a particular target in both HPV+ and HPV- HNSCC lesions (Figure 5.5).

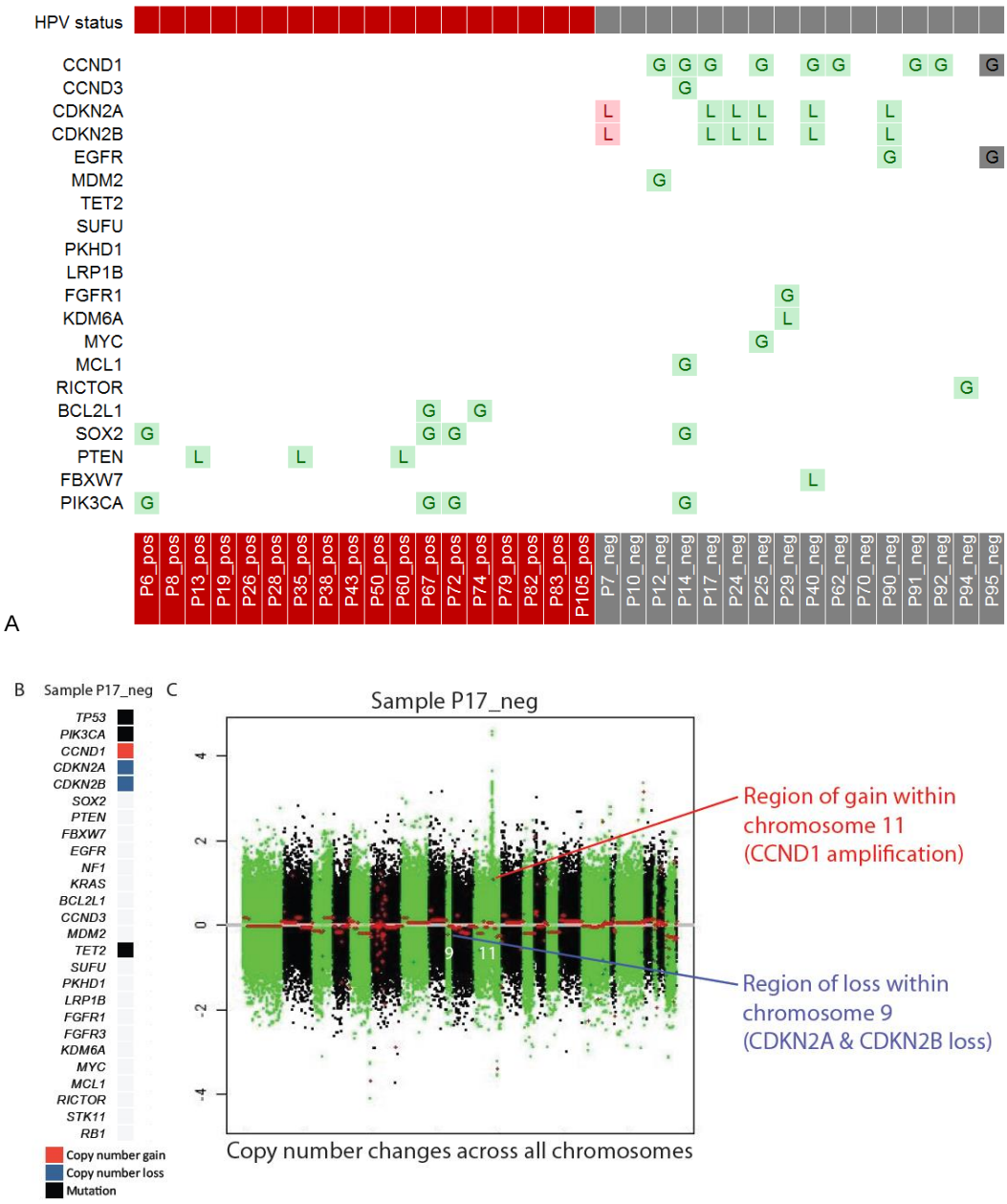


Figure 5.4: Validation of copy number changes by Infinium CNV profiling across all samples. A) 42 of 44 (95%) copy number alterations detected by sequencing were confirmed (green: confirmed, pink: not confirmed, grey: no data);

B) Genetic changes in 'P17_neg' detected by NGS (extracted from Figure 1); C) Illustration of copy number changes (obtained from Infinium CNV profiling) in 'P17_neg'. Both the loss of the *CDKN2A* and *CDKN2B* genes (in a region of loss within chromosome 9) and the gain of the *CCND1* gene (in an amplified region of chromosome 11) are shown. Y-axis: log fold change of copy number, X-axis: copy number changes across all chromosomes.

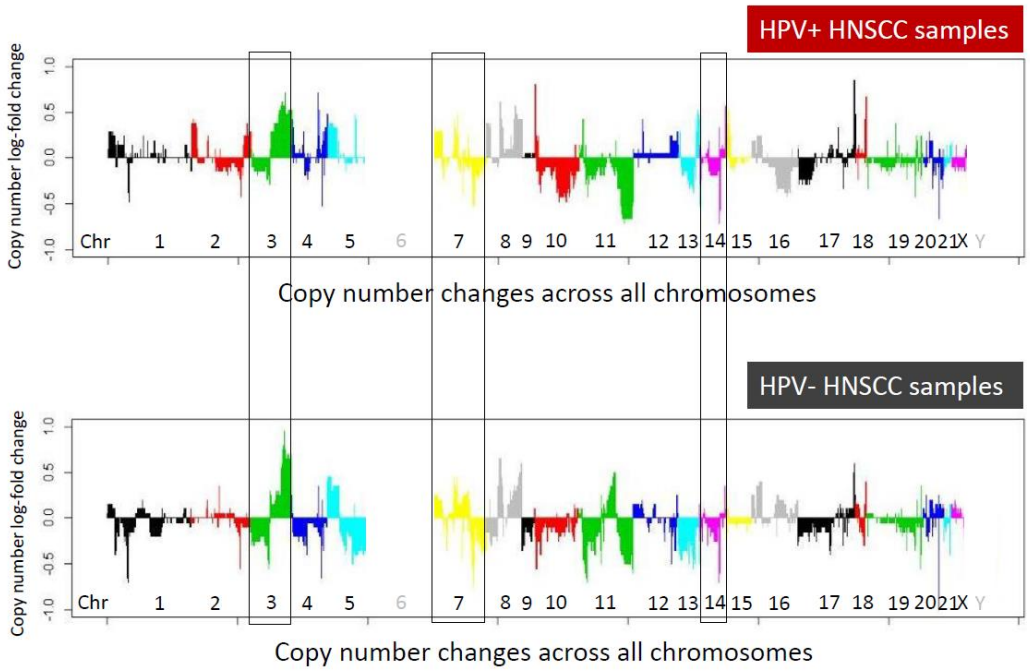


Figure 5.5: Infinium CNV profiling of HPV+ and HPV- HNSCC samples.
Comparing the obtained genome-wide copy number alteration profiles between the two groups, it is evident that the same regions are found to harbour concordant copy number changes in both HPV+ and HPV- HNSCC samples (in particular in chromosomes 3, 7 and 14). This may point towards common molecular pathways in the pathogenesis of both HPV+ and HPV- cancers. Chromosome 6 (MHC regions) and Y chromosome are not shown.

Furthermore, the detected mutations by NGS were validated by Sequenom OncoCarta panels v1.0 and v3.0 (Figure 5.6). Since the applied NGS targets all exons of all genes tested, whilst Sequenom OncoCarta panels target only specific mutational hotspots of certain genes, it follows that the majority of NGS detected mutations were not included in the Sequenom analysis. Eight out of 9 mutations that were detected by NGS were also confirmed by Sequenom. One *PIK3CA* mutation in sample P72_pos was called at 1% allele frequency by NGS, and this mutation was therefore unlikely to be detected by Sequenom analysis.

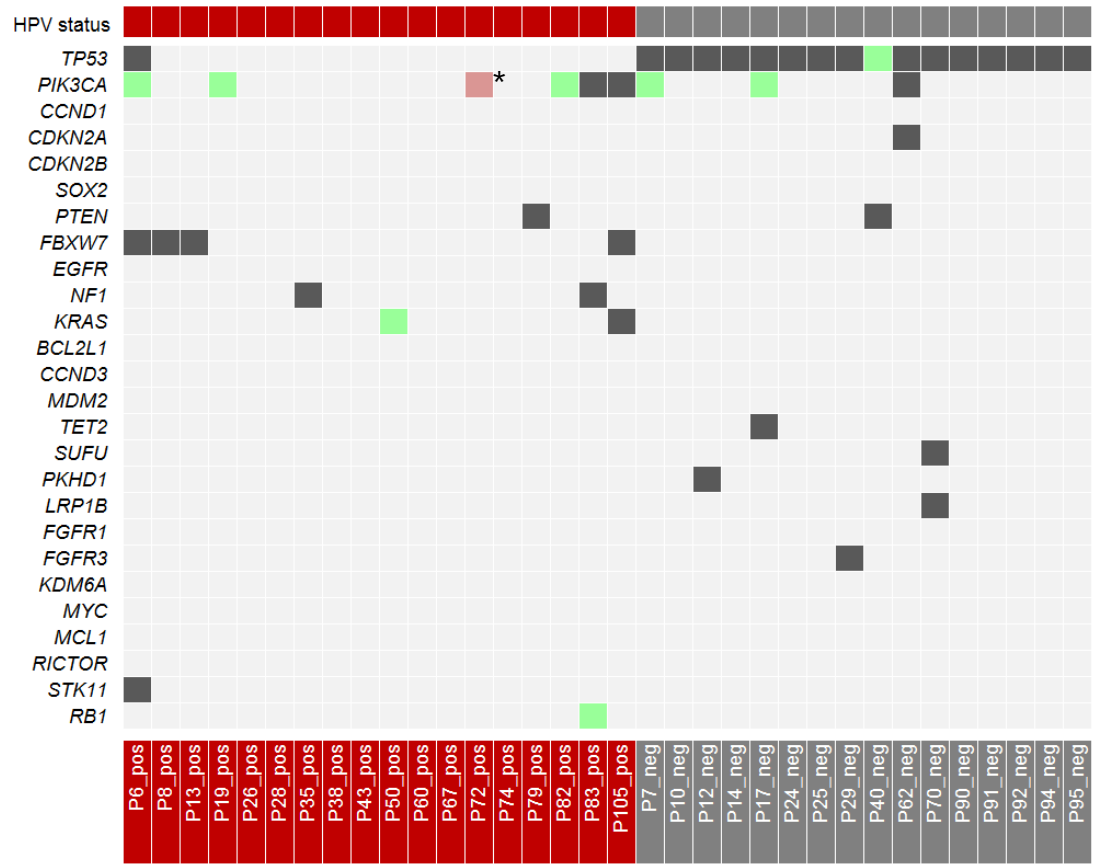


Figure 5.6: Validation of detected mutations by Sequenom OncoCarta panels v1.0 and v3.0. Mutations in HNSCC samples detected by deep sequencing were validated using the OncoCarta panels v1.0 and v3.0. 8 out of 9 mutations that were successfully tested on the Oncocarta panel were confirmed (green: confirmed, pink: not confirmed, grey: n/a). *The *PIK3CA*_E545K mutation in sample P72_pos was called at 1% allele frequency by NGS, and this mutation was therefore unlikely to be detected by Sequenom analysis.

In order to obtain a further line of evidence, findings from the genomic analysis were validated by immunohistochemistry for two frequently genetically altered genes: *CCND1* and *PTEN.* Genomic alterations in *CCND1* were confirmed by Cyclin D1 immunochemistry with strong expression of the Cyclin D1 protein in 8 of 9 *CCND1* amplified cases (and intermediate expression in the remaining case). Representative samples are shown in Figure 5.7.

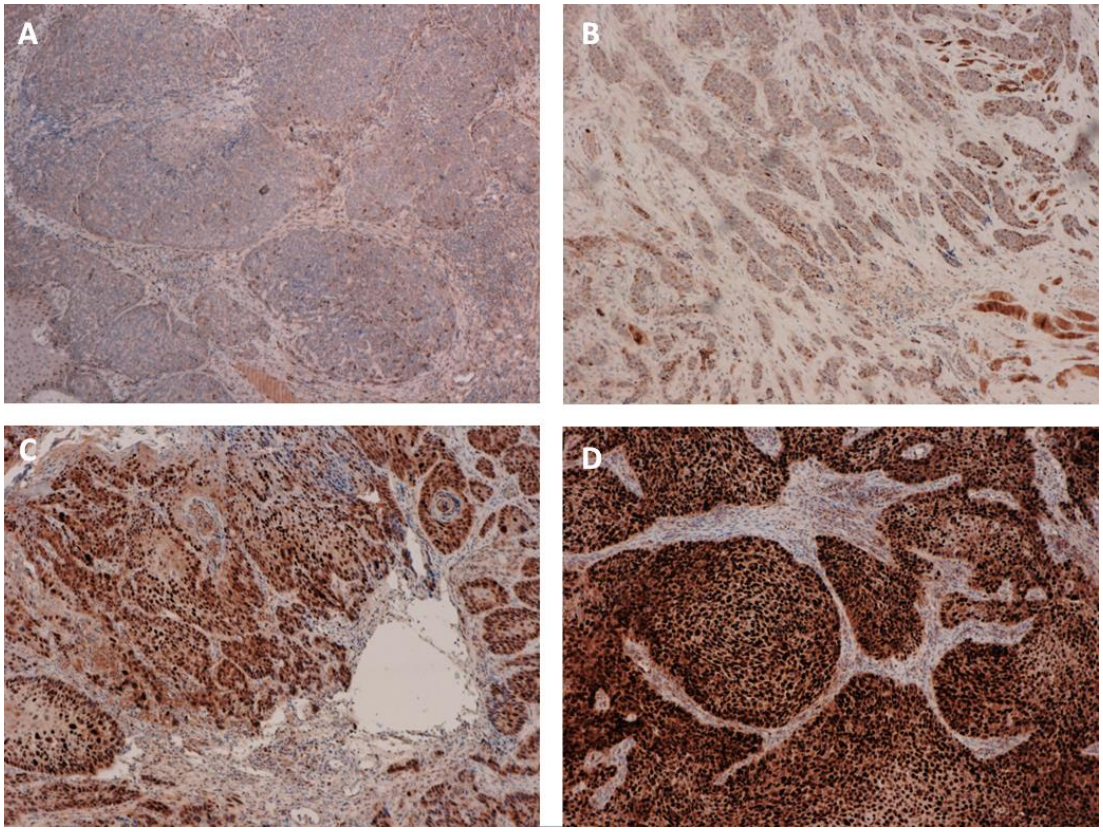


Figure 5.7: Validation of detected copy number alterations of *Cyclin D1* (*CCND1*) by immunohistochemistry.

Staining of HNSCC samples for Cyclin D1 confirmed strong expression in 8 of 9 *CCND1* amplified cases (and intermediate expression in the remaining case) compared with samples harbouring no copy number alteration; Representative samples shown: Low levels of *CCND1* expression in the tumour tissue of sample 'P38_pos' (A) and sample 'P29_neg' (B); Deep-sequencing: No CNA; High levels of Cyclin D1 expression in the tumour tissue of sample 'P12_neg' (C) and sample 'P17_neg' (D); Deep-sequencing: *CCND1* copy number gain.

PTEN loss and mutation was validated by immunohistochemistry (Figure 5.8). *PTEN* staining was negative in all cases in which deep sequencing revealed a homozygous deletion or mutation. Four additional samples displayed low PTEN protein expression. In three of these cases a heterozygous deletion/single copy loss of *PTEN* was present, as detected by NGS. In the remaining sample other mechanisms may explain the loss of expression, such as an epigenetic alteration or changes in the posttranscriptional regulation of *PTEN*.

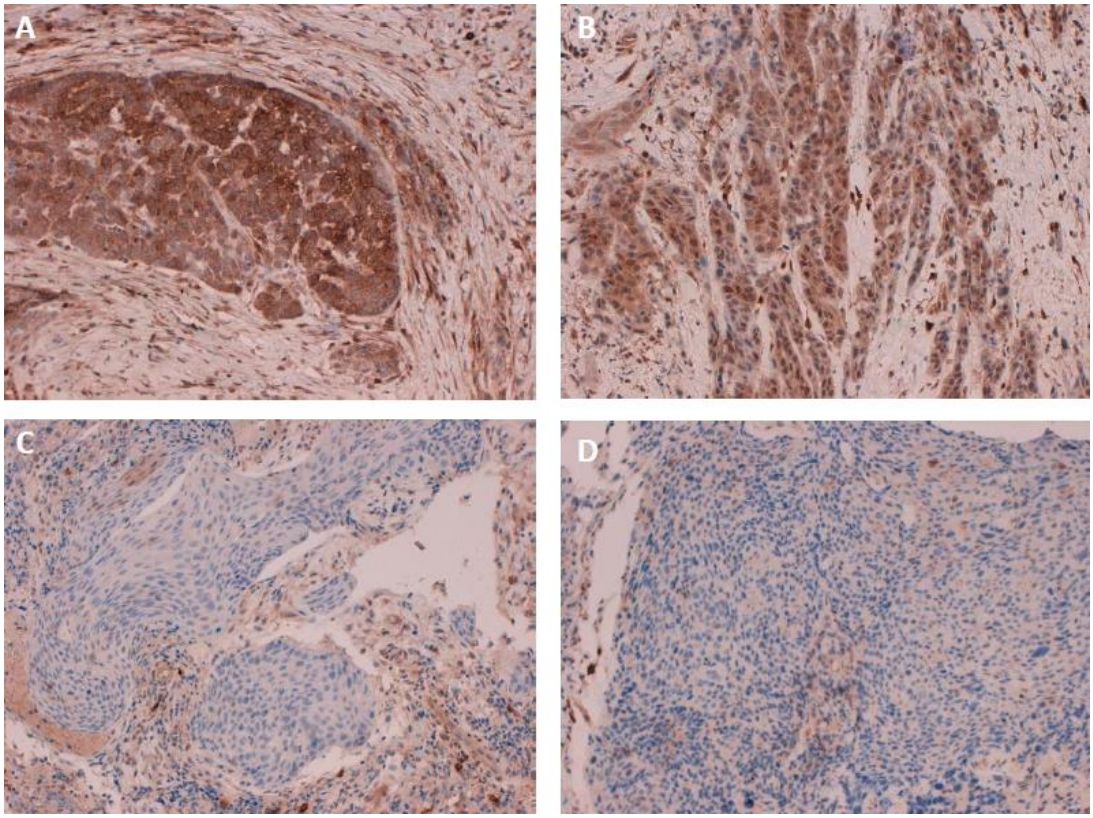


Figure 5.8: Validation of detected *PTEN* copy number loss by immunohistochemistry.
Staining of HNSCC samples for PTEN was negative in all cases in which deep sequencing revealed a homozygous deletion or mutation.
Representative samples shown: Abundant PTEN expression in the tumour tissue of sample 'P26_pos' (A) and sample 'P70_neg' (B); Deep-sequencing: No CNA; Absence of PTEN protein in the tumour tissue of sample 'P60_pos' (C) and sample 'P13_pos' (D); Deep-sequencing: *PTEN* copy number loss.

5.2 Discussion of results obtained in chapter 5.

Overall, sequence analysis revealed that HPV+ and HPV- oropharyngeal carcinomas cluster into two distinct subgroups, with few overlapping genetic alterations. These data concur with epidemiological and clinical data, indicating that HPV+ HNSCC is a distinct disease entity [8, 9].

The fact that targeted deep next-generation sequencing revealed that *TP53* is mutated in 100% of HPV- samples indicates that *TP53* is likely to be a ubiquitous early event in the pathogenesis of oropharyngeal HNSCC. Previous studies suggested that 60-85% of HNSCC cases test positive for

TP53 mutations [51]. In HPV+ disease, E6 leads to *TP53* functional inactivation. Consistent with this, only one *TP53* mutation was identified in an HPV+ tumour. However, this mutation (R290C, Table 5.1) causes only a 40% decrease in *TP53* function and has been detected in sarcomas harbouring *MDM2* amplification [170, 171]. One caveat in my study is that all HPV- samples analysed were also p16 negative, thus it remains possible that in HPV- samples with elevated p16 expression (e.g. through *RB1* mutation), the frequency of *TP53* mutation is less than 100%.

The obtained data for HPV- oropharyngeal cancer indicate that the frequency of *CCND1* amplification (in around 55% of cases) and *CDKN2A/B* deletions (in around 40% of cases) are higher than previously reported [172]. *CCND1* amplification has also been described in 12% of non-small cell lung cancers [173] and in up to 41% oesophageal squamous cell carcinoma [174], suggesting that this could be one of the more common genetic alterations linked to smoking-induced epithelial malignancy. In HPV+ cancer, the oncoprotein E7 leads to cell cycle dysregulation by substituting for cyclin D gain-of-function and cyclin dependent kinase inhibitor loss-of-function activities. Overall, this indicates that direct dysregulation of the cell cycle is a key mechanism for oropharyngeal tumours to evolve.

HPV+ HNSCC samples frequently harbour mutations or CNVs in genes implicated in activation of the PI3K/AKT/mTOR pathway. In particular, *PIK3CA* mutation and *PTEN* inactivation by gene copy loss or mutation were seen in over 60% of HPV+ tumours, and in 31% of HPV- tumours. These findings may help to explain the high frequency of PI3K pathway activation in HPV+ HNSCC samples and the efficacy of mTOR inhibitors in xenograft studies with HPV+ cell lines previously reported [175]. It will be important to audit both the sequence and copy number of the *PIK3CA* and *PTEN* genes if such agents are tested in clinical trials for HPV+ HNSCC.

The obtained results suggest that mutations in *FBXW7* may be enriched in HPV+ disease. FBXW7 is an E3 ubiquitin ligase that targets a number of growth-promoting proteins for proteasomal degradation, including Cyclin E,

MYC, NOTCH and mTOR [176, 177]. Loss of *FBXW7* occurs in combination with *NOTCH* gain-of-function mutations in T-ALL [178], suggesting it may be an important target for FBXW7 ligase activity in these tumours. In contrast, HNSCC frequently display *NOTCH* loss-of-function-mutations [110, 179], thus in HNSCC, other substrates such as Cyclin E, MYC or mTOR may be the relevant targets for FBXW7.

The *SOX2* and *PIK3CA* genes both reside on the long arm of chromosome 3 (3q26) and these genes were amplified in three HPV+ samples and one HPV- tumour. Amplification of 3q seems to be a particular target in both HPV+ and HPV- HNSCC lesions (Figure 5.5). While *PIK3CA* amplifications have previously been reported in HPV+ HNSCC [180, 181], *SOX2* has recently been proposed as the critical target of 3q gains observed at a high frequency in squamous lung cancer [182] and in oesophageal squamous cell carcinoma [183]. *SOX2* is also frequently amplified and overexpressed in oral squamous cell carcinoma [184]. Furthermore, *SOX2* expression is upregulated in a subpopulation of putative HNSCC stem cells that displays characteristics of epithelial to mesenchymal transition (EMT), associated with increased propensity for metastasis [185].

I also demonstrate inactivating mutations in *NF1* and *STK11* in HPV+ HNSCC for the first time. Loss of *STK11* is associated with metastasis in head and neck cancer [186].

Beyond the genes directly involved in signalling and cell cycle, I found amplifications in genes implicated in preventing apoptosis: *BCL2L1* (6% amplification) and *MCL1* (3% amplification), suggesting that suppression of apoptosis may also contribute to HNSCC pathogenesis.

Receptor tyrosine kinase mutations, *FGFR1*, *FGFR3* and *EGFR*, were only observed in HPV- tumours at low frequency.

Overall, my data strongly support a causal role for HPV in oropharyngeal carcinogenesis by overcoming the requirement for genetic lesions in the TP53 and RB1 tumour suppressor pathways evident in the HPV- tumours.

The detection of frequent PI3K/AKT/mTOR pathway alterations in HPV+ tumours is consistent with a recent report demonstrating PI3K pathway activation and sensitivity to mTOR inhibition in both cervical carcinoma and HPV+ HNSCC [175]. Together, these studies provide a rationale for the testing of PI3K pathway inhibitors in HPV+ HNSCC. In HPV- tumours, the frequent alteration of *CDKN2A/B* and/or *CCND1* suggests that, if supported by functional data, trials with CDK inhibitors may be indicated. My data support the observations by gene expression microarrays and by genome-wide methylation studies that HPV+ HNSCC is a distinct entity, with a distinct set of somatic alterations. However, it would appear that a core set of pathways (TP53, RB1/cell cycle and PI3K/AKT/mTOR) is compromised in both HPV+ and HPV- oropharyngeal tumours, thus targeted therapies directed against one or more of these pathways could be efficacious in both contexts.

Chapter 6

Integration of obtained data – Updating the current model of HNSCC

Having discussed the individual results in chapters 3, 4, and 5, this chapter will endeavour to collate the various results in order to integrate them and to present an updated model of head and neck cancer.

6.1. Results

6.1.1 Integration of obtained methylation data from HPV+ and HPV- HNSCC with publicly available methylation data on HPV-driven cancer (cervical cancer) and smoking-induced cancer (lung cancer):

In order to test the effect of HPV on DNA methylation further, methylation data obtained from the 18 HPV+ and 14 HPV- HNSCC samples were integrated with publicly available methylation data on HPV-induced vs. smoking-induced cancer, 48 cervical cancer samples [154] and 59 lung cancer samples [153], respectively. Using a selection of HPV-associated vs. smoking-associated features (see method section), identified by comparing HPV+ with HPV- HNSCC, multidimensional scaling of the datasets using a simple Euclidean distance measure was applied and distances were plotted. An overlap of cervical cancer samples and HPV+ HNSCC samples was observed (Figure 6.1). Significance of this observation was further tested using a Wilcoxon rank sum test on intersample distances. When focusing

upon HPV+/HPV- methylation signatures the methylation pattern of cervical cancer samples was more closely related to the HPV+ signature observed in HNSCC (P < 2.2e-16). This suggests that HPV induces a distinct methylation signature that is independent of tissue-specific DNA methylation.

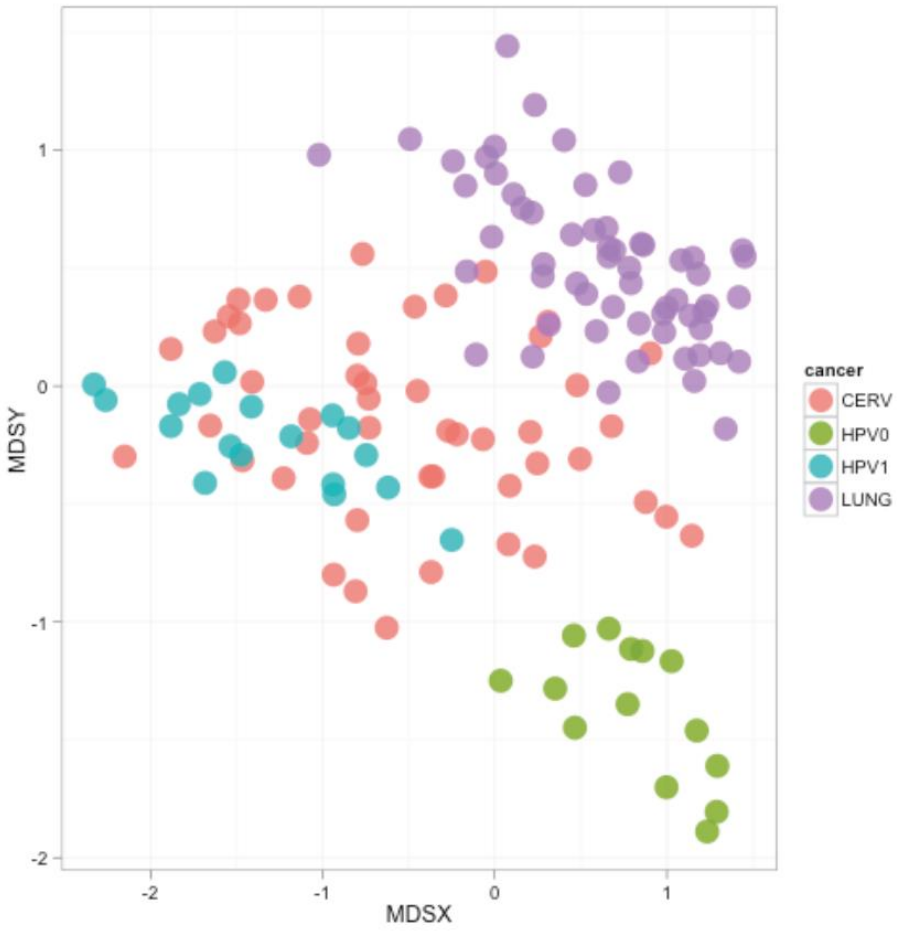


Figure 6.1: Multidimensional scaling using the 4 datasets, comprising of 48 cervical cancer samples (CERV; pink), 59 lung cancer samples (LUNG; purple), 18 HPV+ HNSCC samples (HPV1; light-blue) and 14 HPV- HNSCC samples (HPV0; green), using a selection of HPV-associated vs. smoking associated features identified by comparing HPV+ vs. HPV- HNSCC.

6.1.2 Integration of obtained genomic data from HPV+ and HPV- HNSCC with mutation data from HPV-driven cancer (cervical cancer) and smoking-induced cancer (lung cancer):

The genomic data obtained from 18 HPV+ and 16 HPV- HNSCC tissue samples (chapter 5) were integrated with publicly available mutation data on HPV-induced vs. smoking-induced cancer. Mutation data for cervical cancer and lung cancer samples were obtained from the COSMIC database (http://www.sanger.ac.uk/genetics/CGP/cosmic) with one report cited from the literature [187]. The most common mutations in these cancers are summarized in Table 6.1.

	HPV+ HNSCC	HPV- HNSCC	Cervical Cancer	Lung Cancer
TP53	6%	100%	6%	36%; >90%*
PIK3CA	33%	19%	11%	3%
KRAS	11%	0%	9%	16%
EGFR	0%	13%	0%	27%
PTEN	22%	6%	5%	3%
CDKN2A	0%	44%	9%	13%
STK11	6%	0%	16%	9%

* 36% of cases reported to be mutated in lung adenocarcinoma (COSMIC); >90% of cases reported to be mutated in SCLC [187].

Table 6.1: Percentage of somatic mutations (and genomic alterations in the cases of HNSCC) found within the listed genes, comparing the obtained NGS data on HPV+ HNSCC and HPV- HNSCC with publicly available mutation data on cervical cancer and lung cancer (Data obtained from the COSMIC database; http://www.sanger.ac.uk/genetics/CGP/cosmic).

TP53 mutations are rarely found in the two investigated HPV-induced cancers (outlined in Table 6.1) in which viral protein E6 causes the degradation of TP53. *PIK3CA* amplifications and mutations are found in both HPV-induced cancer and cancers caused by other carcinogens, although the mutation rates for this gene appear to be higher in HPV-induced cancers. *KRAS* mutations (commonly found in tumours of patients with a history of smoking [188] are reported to be present in 16% of lung cancers. *KRAS*

mutations were detected in my set of HPV+ HNSCCs in 2 cases, one patient is a heavy smoker and in the other patient the smoking status is unknown, suggesting that this may be a smoking-induced mutation. *EGFR* mutations seem to be an exclusive feature of smoking-induced cancer (13% and 27% in HPV- HNSCC and lung cancer, respectively, compared with 0% in HPV-associated cancers). Moreover, *PTEN* mutations are present in all types of evaluated sample sets of cancer. In line with previous results, showing that *CDKN2A* loss is a common feature of HPV- HNSCC, *CDKN2A* loss is observed in nearly half of all cases of HPV- HNSCC. *STK11*, which was reported to be a highly significantly mutated gene in lung cancer, is also found mutated in HPV+ HNSCC and cervical cancer.

Moreover, pathways affected and genetic changes observed in lung adenocarcinoma in a previous study [189] were integrated with the obtained genomic data on HPV+ and HPV- HNSCC. A remarkable overlap of genetic alterations between these different cancers is observed (Figure 6.2).

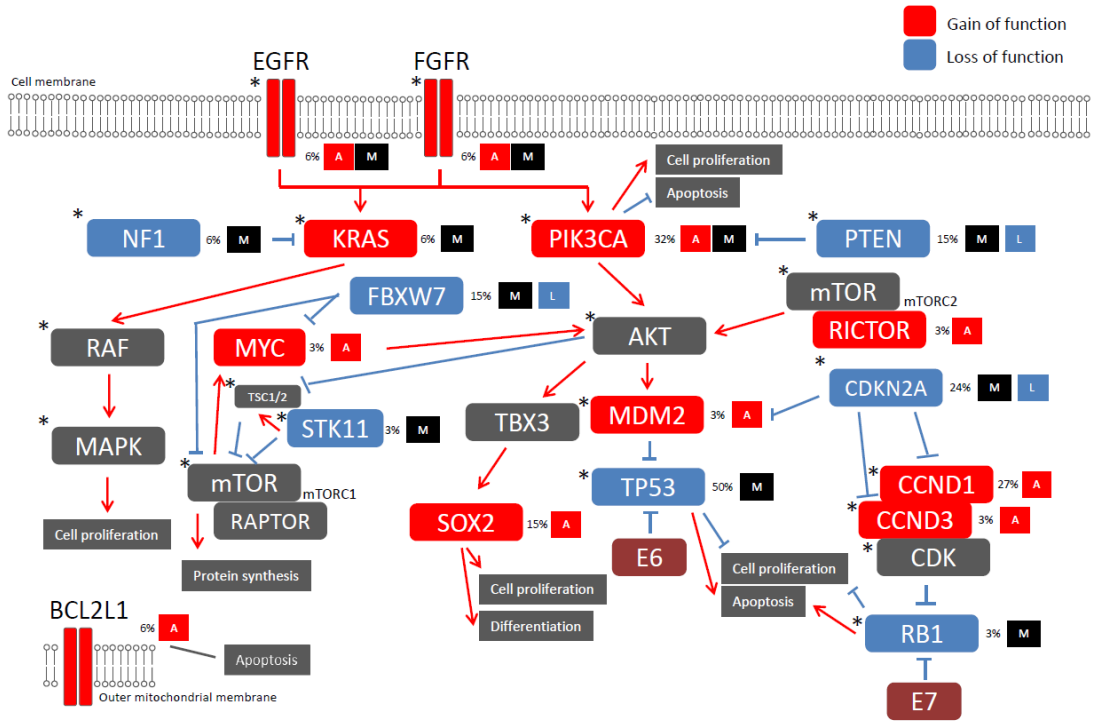


Figure 6.2: Mutated pathways in HNSCC (according to mutations detected by NGS of selected genes as described) and integration of these data with mutation data from lung adenocarcinoma (*=significantly mutated in lung adenocarcinoma cases).

6.1.3 Integration of the obtained epigenetic and genetic changes in head and neck cancer models into the Hanahan and Weinberg model:

As outlined in chapter 1, cancer is caused largely by the acquisition of biological capabilities, rendered by changes in the genome and the epigenome in a partly stochastic and partly directed multi-step process, with underlying genomic instability. Using the obtained data from HPV+ and HPV- HNSCC, observed genomic and epigenomic alterations were linked to hallmark capabilities (Figure 6.3), as defined by Hanahan and Weinberg [54].

Figure 6.3: Link of the genomic and epigenomic changes found in HNSCC to the 10 hallmark capabilities of cancer (adapted from [54]).

Genomic and epigenomic alterations underlie the observed phenotypic abnormalities and, apart from infection with a high-risk strain of HPV in HPV+ cancers, are mediated by exposure to environmental or life-style factors in addition to genetic factors.

6.1.4 An updated model of HPV-driven and HPV independent head and neck carcinogenesis:

In the following section, an updated model of HPV-driven and HPV independent carcinogenesis is presented which incorporates findings from my thesis which can be summarized as the UCL/UCLH Head and Neck Cancer Genome and Epigenome Project.

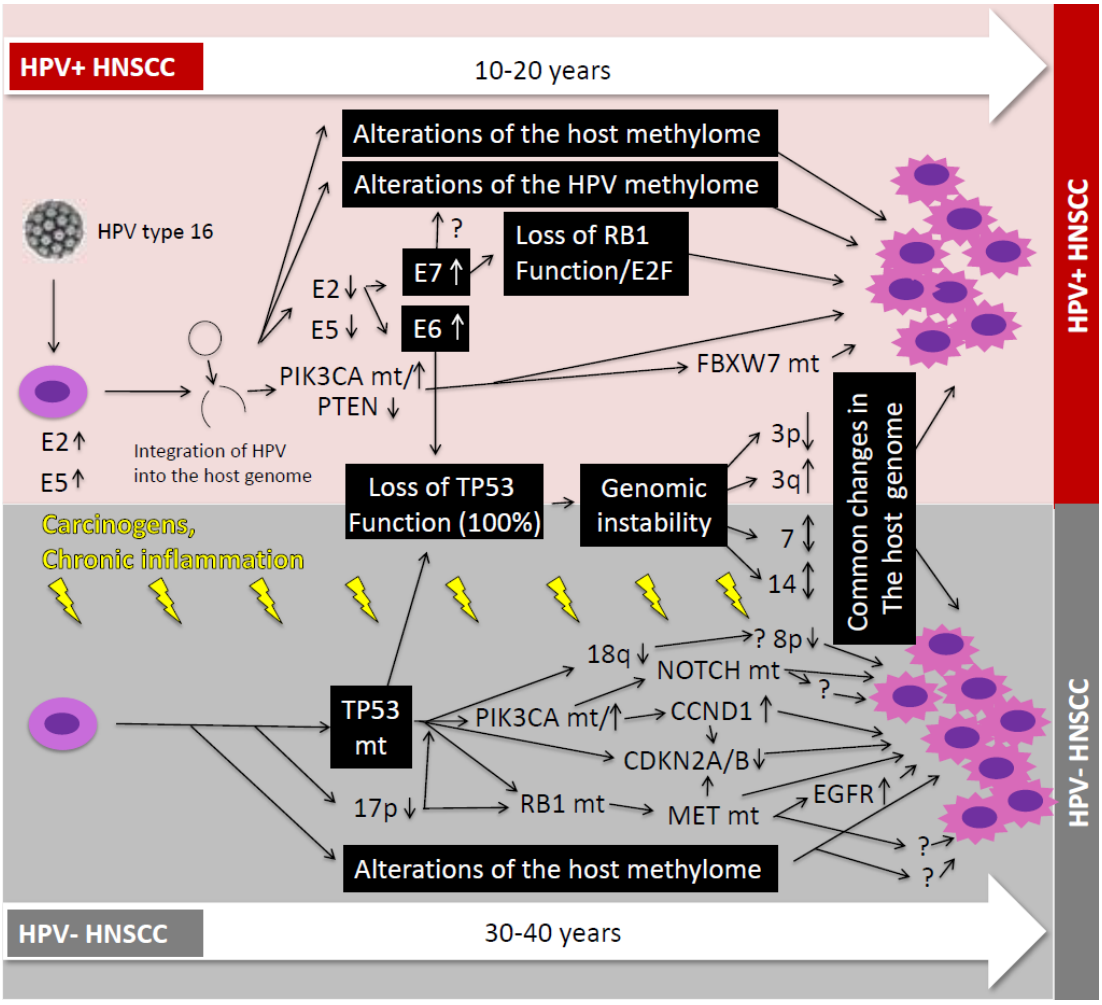


Figure 6.4: Updated model of HPV-driven and HPV independent carcinogenesis with a focus on head and neck cancer. It reveals partially distinct and partially overlapping pathways of carcinogenesis in these two distinct subtypes.

It reveals partially distinct and partially overlapping pathways of carcinogenesis in these two distinct subtypes with regard to clinical behaviour and epidemiological associations. Loss of TP53 function is central in malignant transformation in both cancers (present in 100% of cases in my set

of HNSCC). The resulting genomic instability renders the acquisition of common genomic alterations (such as gains and losses of entire chromosomal regions; see chapter 5). Alterations of the host methylome and the viral methylome, outlined in Figure 6.4, may be equally important and are discussed in chapters 3, 4, and 7.

6.1.5 Utilizing the obtained data – on the way to a diagnostic biomarker:

This section aims to test the suitability of the described genetic and epigenetic changes in HPV+ HNSCC, compared with HPV- HNSCC, for the development of biomarkers.

In chapter 5 the obtained mutational profiles of HPV+ HNSCC and HPV- HNSCC (targeted NGS data) were used to cluster samples into groups. In the case when *TP53* was part of the panel of candidate markers used, samples clustered into HPV+ and HPV- subgroups (apart from one sample). In the case when *TP53* was excluded from the analysis, samples did not cluster into two groups according to HPV status.

In order to increase the discriminatory power of the performed test, a panel of candidate epigenetic markers was included in the hierarchical clustering analysis. 11 MVPs annotated to Cadherin genes (and shown to be a strong classifier of HPV status in my set of HNSCC, as described in chapter 3) were added to the panel of candidate genetic markers. Hierarchical clustering analysis on HPV+ and HPV- HNSCC samples was then repeated, employing this combinatorial panel of genetic and epigenetic candidate markers.

Using this combinatorial panel, HPV+ and HPV- HNSCC samples cluster into two distinct subgroups (100% of cases; Figure 6.5 A). Even when excluding the mutation status of TP53 in samples from the analysis, HPV+ and HPV- tumours still clustered into two distinct subgroups with one outlier (Figure 6.5 B), using the same panel and clustering method.

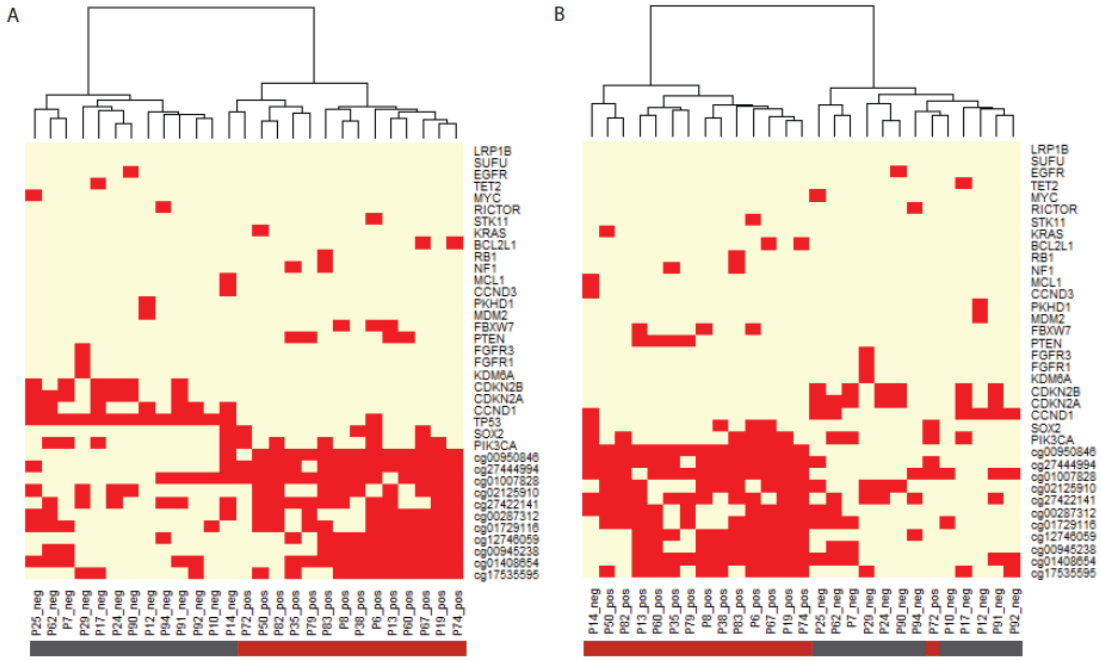


Figure 6.5: Hierarchical Clustering (Euclidean distance measure) of HPV+ and HPV- HNSCC samples according to a panel of genetic (NGS data; chapter 5) and a panel of epigenetic markers (11 Cadherin-annotated probes; chapter 3).

A) Hierarchical clustering reveals that HPV+ and HPV- tumours cluster into two distinct subgroups using the combinatorial panel of markers.

B) After having excluded mutations within *TP53* and using the same clustering method, HPV+ and HPV- tumours still cluster into two distinct subgroups with the exception of one outlier.

In summary, this combinatorial panel of genetic and epigenetic candidate markers is likely to identify HPV+ and HPV- patients irrespective of *TP53* status. *TP53* mutational status has not been examined in cases of mixed p16 staining or in cases, showing a positive p16 and negative E6 qPCR result and vice versa (excluded from analysis).

6.2 Discussion of results obtained in chapter 6:

The results obtained in this chapter clearly show that factors contributing to HPV-induced head and neck carcinogenesis are distinct from HPV- HNSCC, but that common pathways are affected, with a cell gaining the spectrum of hallmark capabilities, as outlined by Hanahan and Weinberg [54], during

malignant transformation. The potential of the detected genetic and epigenetic changes to serve as a diagnostic biomarker is evaluated, but more work is needed to test the specificity and sensitivity of this panel of candidate markers in the clinical setting. Lastly, the obtained data were integrated into an updated model of head and neck carcinogenesis and shall be discussed in more detail in chapter 7.

Chapter 7

General Discussion and Future Directions

Having discussed results in chapters 3, 4, 5 and 6 individually, this chapter summarizes the key findings of my work and evaluates these in the context of other HPV-induced and smoking-induced cancers. Furthermore, an updated model of head and neck carcinogenesis (defined in chapter 6), the value of observed genetic and epigenetic changes for the establishment of novel biomarkers and future directions will be discussed.

In chapter 3, I tested the possible involvement of epigenetic modulation by HPV in HNSCC by conducting a genome-wide DNA methylation analysis. Using laser-capture microdissection of 42 FFPE HNSCCs, DNA methylation profiles of 18 HPV+ and 14 HPV- samples were generated. Unsupervised clustering over the most methylation variable positions (MVPs) showed that samples segregated according to HPV status, but also that HPV+ tumours are heterogeneous. Amongst the MVPs there was significant enrichment of those mapping to transcriptional start sites, leading to the identification of a CIMP in a sub-group of the HPV+ tumours. Supervised analysis revealed a strong preponderance (87%) of the MVPs towards hypermethylation in HPV+ HNSCC. This signature was validated in two sets of independent HPV+/HPV- HNSCC samples (FF and cell lines) using two independent methods (Infinium 450k and MeDIP-seq). Grouping of MVPs into functionally more significant DMRs and assigning them to genes for gene set enrichment analysis identified 43 hypermethylated promoter DMRs, including three Cadherins of the Polycomb group target genes. Integration with independent expression data showed a strong negative correlation, especially for the Cadherin gene family members. Combinatorial ectopic expression of the two HPV oncogenes (*E6* and *E7*) in an HPV- HNSCC cell line partially

phenocopied the hypermethylation signature observed in HPV+ HNSCC tumours and established *E6* as the main viral effector gene. My data establish archival FFPE tissue to be highly suitable for this type of methylome analysis and suggest that HPV modulates the HNSCC epigenome through hypermethylation of Polycomb repressive complex 2 target genes such as Cadherins which are implicated in tumour progression and metastasis.

In addition to the evaluation of methylation changes in the host methylome, the methylation status of the HPV genome in HNSCC was determined. HPV type 16 was confirmed in all HPV+ samples tested and methylation within the viral genome was demonstrated. Methylation which was detected at the boundary of the L1/L2 gene of the viral genome was validated in FF HNSCC samples and in an independent set of HPV+ HNSCC cell lines by BS-Seq. Using the MeDIP-Seq technology, which allowed me to identify integration sites of the virus in the context of HPV+ HNSCC, I showed that integration sites across the human genome appear to be random, whereas there is a significant enrichment of potential integration sites within the viral E2 (E4 and E5) region in line with previous reports in cervical cancer [190]. In summary, the obtained methylation data from the entire HPV type 16 methylome in HNSCC may lead the way to the identification of novel diagnostic markers, detecting this change in extracted DNA from patients' saliva or tissue samples.

For the genomic analysis described in chapter 5, paired-end sequencing of hybrid-captured DNA, targeting 3,230 exons in 182 genes often mutated in cancer was employed. Infinium copy number variation (CNV) profiling, Sequenom MassArray sequencing and immunohistochemistry were used for the validation of the results. The main results were that HPV+ and HPV- HNSCCs clustered into two distinct subgroups. *TP53* mutations were detected in 100% of HPV- cases and abrogation of the G1/S checkpoint by *CDKN2A/B* deletion and/or *CCND1* amplification were found in the majority of HPV- tumours. Various mutations within the PI3 kinase (PI3K) signalling pathway were found in both groups. These findings strongly support a causal role for HPV, acting via TP53 and RB1 pathway inhibition, in the pathogenesis of a subset of oropharyngeal cancers. The high frequency of

PI3K pathway alterations in both HPV+ and HPV- HNSCCs implicates activation of this pathway in oropharyngeal tumourigenesis. Furthermore, studies of CDK inhibitors in HPV- disease are warranted. These findings indicate that therapeutic stratification according to somatic genomic changes, in addition to HPV status, could be the most appropriate management for these cancers in the future.

In chapter 6, methylation data from HPV+ and HPV- HNSCC (chapter 3) were integrated with publicly available methylation data on HPV-driven cancer (cervical cancer) and smoking-induced cancer (lung cancer). A significant overlap of cervical cancer samples and HPV+ HNSCC was observed when applying multidimensional scaling, suggesting that HPV induces a methylation signature independent of the methylation pattern observed between tissues. Integration of obtained genomic data from HPV+ and HPV- HNSCC with mutation data from HPV-driven cancer (cervical cancer) and smoking-induced cancer (lung cancer) revealed that HPV+ HNSCC and cervical cancer, as well as HPV- HNSCC and lung cancer, harbour similar mutations which can be regarded as the main driver mutations across these cancers.

On the basis of these findings and having integrated all applicable data into the Hanahan and Weinberg model, I present an updated model of HPV-driven and HPV independent head and neck carcinogenesis in chapter 6.

It is noteworthy that partially distinct and partially overlapping pathways of carcinogenesis appear to play a key role in HPV+ vs. HPV- HNSCC. Loss of TP53 function is clearly central in the process of malignant transformation in both cancers. Mutations in the TP53 gene are present in 100% of cases in my set of HPV- HNSCC samples and have been reported in over 90% of SCLC cases [187]. In HPV+ cancers E6 leads to the degradation of TP53. Irrespective of the different mechanisms in the two subtypes, both resulting in genomic instability, this renders the acquisition of common genomic alterations (such as gains and losses of entire chromosomal regions which I demonstrated by Infinium CNV profiling of HPV+ and HPV- HNSCC; chapter 5). One explanation why these regions of gains and losses are present in

similar genomic locations in both groups could be that these alterations are acquired under the selective pressure being exerted in the same microenvironment in which both cancers evolve. In contrast, specific effects of HPV on the host genome and epigenome (independent of a loss of TP53 function and distinct from those present in e.g. smoking-induced HNSCC) may be equally important and make HPV+, compared with HPV- HNSCC, still appear as a distinct genetic and epigenetic entity. However, it is difficult to know which proportion of genetic and epigenetic changes are driving the process of malignant transformation and which proportion can be regarded as passenger mutations or simple bystanders which do not play any role in the process of malignant transformation or in the maintenance of the malignant phenotype. It is therefore important to define the specific epigenetic and genetic drivers in the future which underlie the spectrum of hallmark capabilities [54] acquired by a given cell in the development of these subtypes of HNSCC.

Lastly, the potential of the detected genetic and epigenetic changes to serve as a diagnostic biomarker is evaluated. The data clearly show that the obtained NGS data are not yet sufficient to cluster HPV+ and HPV- HNSCC in 100% of cases. Adding a small panel of candidate epigenetic markers, sufficiently clusters these samples into two groups, even if TP53 is excluded. Although this current analysis has various limitations, it shows a promising future approach towards the development of novel biomarkers by combining genetic, epigenetic (and potentially transcriptomic data) which will be analysed in extensive sample collections in the course of ongoing cancer genome and epigenome initiatives (see chapter 1). This will further boost the translation of personalized cancer medicine into clinical practices across the world.

In conclusion, my work has contributed to the understanding of genetic and epigenetic changes at oncogenic loci associated with the progression of HPV+ and HPV- HNSCC. Only a small number of epigenetic alterations in the context of HPV+ and HPV- HNSCC has been described to date. The evaluation of the genetic and epigenetic changes, detected by comprehensive methylome analysis (both of the host and the viral DNA),

genomic analysis and multidimensional analysis of the obtained data with publicly available data, in HPV+ HNSCC, compared with HPV- HNSCC, represents a major step towards advancing the understanding of both HPV+ HNSCC and HPV- HNSCC and may explain the better outcome and survival of patients suffering from this distinct subtype. In the longer-term, these data can be expected to be used for the identification of potential diagnostic and prognostic markers, as well as putative therapeutic targets.

8. References

1. K.D. Hunter, E.K. Parkinson, and P.R. Harrison, *Profiling early head and neck cancer.* Nat Rev Cancer, 2005. **5**(2): p. 127-35.
2. S. Warnakulasuriya, *Global epidemiology of oral and oropharyngeal cancer.* Oral Oncol, 2009. **45**(4-5): p. 309-16.
3. B.A. Conley, *Treatment of advanced head and neck cancer: what lessons have we learned?* J Clin Oncol, 2006. **24**(7): p. 1023-5.
4. T. Rodriguez, A. Altieri, L. Chatenoud, S. Gallus, C. Bosetti, E. Negri, S. Franceschi, F. Levi, R. Talamini, and C. La Vecchia, *Risk factors for oral and pharyngeal cancer in young adults.* Oral Oncol, 2004. **40**(2): p. 207-13.
5. G. D'Souza, A.R. Kreimer, R. Viscidi, M. Pawlita, C. Fakhry, W.M. Koch, W.H. Westra, and M.L. Gillison, *Case-control study of human papillomavirus and oropharyngeal cancer.* N Engl J Med, 2007. **356**(19): p. 1944-56.
6. N. Tran, B.R. Rose, and C.J. O'Brien, *Role of human papillomavirus in the etiology of head and neck cancer.* Head Neck, 2007. **29**(1): p. 64-70.
7. W.M. Mendenhall and H.L. Logan, *Human Papillomavirus and Head and Neck Cancer.* Am J Clin Oncol, 2009: p. Epub ahead of Print.
8. M.L. Gillison, *Human papillomavirus-associated head and neck cancer is a distinct epidemiologic, clinical, and molecular entity.* Semin Oncol, 2004. **31**(6): p. 744-54.
9. M.B. Gillespie, S. Rubinchik, B. Hoel, and N. Sutkowski, *Human papillomavirus and oropharyngeal cancer: what you need to know in 2009.* Curr Treat Options Oncol, 2009. **10**(5-6): p. 296-307.
10. E.M. Sturgis and P.M. Cinciripini, *Trends in head and neck cancer incidence in relation to smoking prevalence: an emerging epidemic of human papillomavirus-associated cancers?* Cancer, 2007. **110**(7): p. 1429-35.
11. L. Hammarstedt, D. Lindquist, H. Dahlstrand, M. Romanitan, L.O. Dahlgren, J. Joneberg, N. Creson, J. Lindholm, W. Ye, T. Dalianis,

and E. Munck-Wikland, *Human papillomavirus as a risk factor for the increase in incidence of tonsillar cancer.* Int J Cancer, 2006. **119**(11): p. 2620-3.

12. A. Nasman, P. Attner, L. Hammarstedt, J. Du, M. Eriksson, G. Giraud, S. Ahrlund-Richter, L. Marklund, M. Romanitan, D. Lindquist, T. Ramqvist, J. Lindholm, P. Sparen, W. Ye, H. Dahlstrand, E. Munck-Wikland, and T. Dalianis, *Incidence of human papillomavirus (HPV) positive tonsillar carcinoma in Stockholm, Sweden: an epidemic of viral-induced carcinoma?* Int J Cancer, 2009. **125**(2): p. 362-6.
13. M.L. Gillison, G. D'Souza, W. Westra, E. Sugar, W. Xiao, S. Begum, and R. Viscidi, *Distinct risk factor profiles for human papillomavirus type 16-positive and human papillomavirus type 16-negative head and neck cancers.* J Natl Cancer Inst, 2008. **100**(6): p. 407-20.
14. K. Settle, M.R. Posner, L.M. Schumaker, M. Tan, M. Suntharalingam, O. Goloubeva, S.E. Strome, R.I. Haddad, S.S. Patel, E.V. Cambell, 3rd, N. Sarlis, J. Lorch, and K.J. Cullen, *Racial survival disparity in head and neck cancer results from low prevalence of human papillomavirus infection in black oropharyngeal cancer patients.* Cancer Prev Res (Phila Pa), 2009. **2**(9): p. 776-81.
15. C. Fakhry, W.H. Westra, S. Li, A. Cmelak, J.A. Ridge, H. Pinto, A. Forastiere, and M.L. Gillison, *Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial.* J Natl Cancer Inst, 2008. **100**(4): p. 261-9.
16. L. Licitra, F. Perrone, P. Bossi, S. Suardi, L. Mariani, R. Artusi, M. Oggionni, C. Rossini, G. Cantu, M. Squadrelli, P. Quattrone, L.D. Locati, C. Bergamini, P. Olmi, M.A. Pierotti, and S. Pilotti, *High-risk human papillomavirus affects prognosis in patients with surgically treated oropharyngeal squamous cell carcinoma.* J Clin Oncol, 2006. **24**(36): p. 5630-6.
17. H.C. Hafkamp, J.J. Manni, A. Haesevoets, A.C. Voogd, M. Schepers, F.J. Bot, A.H. Hopman, F.C. Ramaekers, and E.J. Speel, *Marked differences in survival rate between smokers and nonsmokers with HPV 16-associated tonsillar carcinomas.* Int J Cancer, 2008. **122**(12): p. 2656-64.

18. P. Lohavanichbutr, J. Houck, W. Fan, B. Yueh, E. Mendez, N. Futran, D.R. Doody, M.P. Upton, D.G. Farwell, S.M. Schwartz, L.P. Zhao, and C. Chen, *Genomewide gene expression profiles of HPV-positive and HPV-negative oropharyngeal cancer: potential implications for treatment choices.* Arch Otolaryngol Head Neck Surg, 2009. **135**(2): p. 180-8.
19. I. Martinez, J. Wang, K.F. Hobson, R.L. Ferris, and S.A. Khan, *Identification of differentially expressed genes in HPV-positive and HPV-negative oropharyngeal squamous cell carcinomas.* Eur J Cancer, 2007. **43**(2): p. 415-32.
20. D. Pyeon, M.A. Newton, P.F. Lambert, J.A. den Boon, S. Sengupta, C.J. Marsit, C.D. Woodworth, J.P. Connor, T.H. Haugen, E.M. Smith, K.T. Kelsey, L.P. Turek, and P. Ahlquist, *Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers.* Cancer Res, 2007. **67**(10): p. 4605-19.
21. N.F. Schlecht, R.D. Burk, L. Adrien, A. Dunne, N. Kawachi, C. Sarta, Q. Chen, M. Brandwein-Gensler, M.B. Prystowsky, G. Childs, R.V. Smith, and T.J. Belbin, *Gene expression profiles in HPV-infected head and neck cancer.* J Pathol, 2007. **213**(3): p. 283-93.
22. R.J. Slebos, Y. Yi, K. Ely, J. Carter, A. Evjen, X. Zhang, Y. Shyr, B.M. Murphy, A.J. Cmelak, B.B. Burkey, J.L. Netterville, S. Levy, W.G. Yarbrough, and C.H. Chung, *Gene expression differences associated with human papillomavirus status in head and neck squamous cell carcinoma.* Clin Cancer Res, 2006. **12**(3 Pt 1): p. 701-9.
23. W.H. Westra, J.M. Taube, M.L. Poeta, S. Begum, D. Sidransky, and W.M. Koch, *Inverse relationship between human papillomavirus-16 infection and disruptive p53 gene mutations in squamous cell carcinoma of the head and neck.* Clin Cancer Res, 2008. **14**(2): p. 366-9.
24. W.M. Mendenhall and H.L. Logan, *Human papillomavirus and head and neck cancer.* Am J Clin Oncol, 2009. **32**(5): p. 535-9.
25. J. Doorbar, *Molecular biology of human papillomavirus infection and cervical cancer.* Clin Sci (Lond), 2006. **110**(5): p. 525-41.

26. A.C. Jung, J. Briolat, R. Millon, A. de Reynies, D. Rickman, E. Thomas, J. Abecassis, C. Clavel, and B. Wasylyk, *Biological and clinical relevance of transcriptionally active human papillomavirus (HPV) infection in oropharynx squamous cell carcinoma.* Int J Cancer, 2010. **126**(8): p. 1882-94.
27. R. Herrero, X. Castellsague, M. Pawlita, J. Lissowska, F. Kee, P. Balaram, T. Rajkumar, H. Sridhar, B. Rose, J. Pintos, L. Fernandez, A. Idris, M.J. Sanchez, A. Nieto, R. Talamini, A. Tavani, F.X. Bosch, U. Reidel, P.J. Snijders, C.J. Meijer, R. Viscidi, N. Munoz, and S. Franceschi, *Human papillomavirus and oral cancer: the International Agency for Research on Cancer multicenter study.* J Natl Cancer Inst, 2003. **95**(23): p. 1772-83.
28. A.R. Kreimer, G.M. Clifford, P. Boyle, and S. Franceschi, *Human papillomavirus types in head and neck squamous cell carcinomas worldwide: a systematic review.* Cancer Epidemiol Biomarkers Prev, 2005. **14**(2): p. 467-75.
29. N. Stransky, A.M. Egloff, A.D. Tward, A.D. Kostic, K. Cibulskis, A. Sivachenko, G.V. Kryukov, M.S. Lawrence, C. Sougnez, A. McKenna, E. Shefler, A.H. Ramos, P. Stojanov, S.L. Carter, D. Voet, M.L. Cortes, D. Auclair, M.F. Berger, G. Saksena, C. Guiducci, R.C. Onofrio, M. Parkin, M. Romkes, J.L. Weissfeld, R.R. Seethala, L. Wang, C. Rangel-Escareno, J.C. Fernandez-Lopez, A. Hidalgo-Miranda, J. Melendez-Zajgla, W. Winckler, K. Ardlie, S.B. Gabriel, M. Meyerson, E.S. Lander, G. Getz, T.R. Golub, L.A. Garraway, and J.R. Grandis, *The mutational landscape of head and neck squamous cell carcinoma.* Science, 2011. **333**(6046): p. 1157-60.
30. A.K. Chaturvedi, E.A. Engels, W.F. Anderson, and M.L. Gillison, *Incidence trends for human papillomavirus-related and -unrelated oral squamous cell carcinomas in the United States.* J Clin Oncol, 2008. **26**(4): p. 612-9.
31. A.K. Chaturvedi, E.A. Engels, R.M. Pfeiffer, B.Y. Hernandez, W. Xiao, E. Kim, B. Jiang, M.T. Goodman, M. Sibug-Saber, W. Cozen, L. Liu, C.F. Lynch, N. Wentzensen, R.C. Jordan, S. Altekruse, W.F. Anderson, P.S. Rosenberg, and M.L. Gillison, *Human papillomavirus*

and rising oropharyngeal cancer incidence in the United States. J Clin Oncol, 2011. **29**(32): p. 4294-301.

32. S. Marur, G. D'Souza, W.H. Westra, and A.A. Forastiere, *HPV-associated head and neck cancer: a virus-related cancer epidemic.* Lancet Oncol, 2010. **11**(8): p. 781-9.
33. K.K. Ang, J. Harris, R. Wheeler, R. Weber, D.I. Rosenthal, P.F. Nguyen-Tan, W.H. Westra, C.H. Chung, R.C. Jordan, C. Lu, H. Kim, R. Axelrod, C.C. Silverman, K.P. Redmond, and M.L. Gillison, *Human papillomavirus and survival of patients with oropharyngeal cancer.* N Engl J Med, 2010. **363**(1): p. 24-35.
34. H. zur Hausen, *Papillomaviruses in the causation of human cancers - a brief historical account.* Virology, 2009. **384**(2): p. 260-5.
35. M.E. McLaughlin-Drubin and K. Munger, *Oncogenic activities of human papillomaviruses.* Virus Res, 2009. **143**(2): p. 195-208.
36. M.A. Bedell, J.B. Hudson, T.R. Golub, M.E. Turyk, M. Hosken, G.D. Wilbanks, and L.A. Laimins, *Amplification of human papillomavirus genomes in vitro is dependent on epithelial differentiation.* J Virol, 1991. **65**(5): p. 2254-60.
37. S. Duensing, A. Duensing, E.R. Flores, A. Do, P.F. Lambert, and K. Munger, *Centrosome abnormalities and genomic instability by episomal expression of human papillomavirus type 16 in raft cultures of human keratinocytes.* J Virol, 2001. **75**(16): p. 7712-6.
38. M. Schmitz, C. Driesch, L. Jansen, I.B. Runnebaum, and M. Durst, *Non-random integration of the HPV genome in cervical cancer.* PLoS One, 2012. **7**(6): p. e39632.
39. S.H. Kim, B.S. Koo, S. Kang, K. Park, H. Kim, K.R. Lee, M.J. Lee, J.M. Kim, E.C. Choi, and N.H. Cho, *HPV integration begins in the tonsillar crypt and leads to the alteration of p16, EGFR and c-myc during tumor formation.* Int J Cancer, 2007. **120**(7): p. 1418-25.
40. E. Schwarz, U.K. Freese, L. Gissmann, W. Mayer, B. Roggenbuck, A. Stremlau, and H. zur Hausen, *Structure and transcription of human papillomavirus sequences in cervical carcinoma cells.* Nature, 1985. **314**(6006): p. 111-4.

41. E.S. Hwang, T. Nottoli, and D. Dimaio, *The HPV16 E5 protein: expression, detection, and stable complex formation with transmembrane proteins in COS cells.* Virology, 1995. **211**(1): p. 227-33.
42. S.I. Collins, C. Constandinou-Williams, K. Wen, L.S. Young, S. Roberts, P.G. Murray, and C.B. Woodman, *Disruption of the E2 gene is a common and early event in the natural history of cervical human papillomavirus infection: a longitudinal cohort study.* Cancer Res, 2009. **69**(9): p. 3828-32.
43. M. Scheffner, B.A. Werness, J.M. Huibregtse, A.J. Levine, and P.M. Howley, *The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53.* Cell, 1990. **63**(6): p. 1129-36.
44. J.M. Huibregtse, M. Scheffner, and P.M. Howley, *A cellular protein mediates association of p53 with the E6 oncoprotein of human papillomavirus types 16 or 18.* EMBO J, 1991. **10**(13): p. 4129-35.
45. M. Scheffner, J.M. Huibregtse, R.D. Vierstra, and P.M. Howley, *The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53.* Cell, 1993. **75**(3): p. 495-505.
46. T.D. Kessis, R.J. Slebos, W.G. Nelson, M.B. Kastan, B.S. Plunkett, S.M. Han, A.T. Lorincz, L. Hedrick, and K.R. Cho, *Human papillomavirus 16 E6 expression disrupts the p53-mediated cellular response to DNA damage.* Proc Natl Acad Sci U S A, 1993. **90**(9): p. 3988-92.
47. N. Dyson, P.M. Howley, K. Munger, and E. Harlow, *The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product.* Science, 1989. **243**(4893): p. 934-7.
48. K. Munger, B.A. Werness, N. Dyson, W.C. Phelps, E. Harlow, and P.M. Howley, *Complex formation of human papillomavirus E7 proteins with the retinoblastoma tumor suppressor gene product.* EMBO J, 1989. **8**(13): p. 4099-105.
49. K. Huh, X. Zhou, H. Hayakawa, J.Y. Cho, T.A. Libermann, J. Jin, J.W. Harper, and K. Munger, *Human papillomavirus type 16 E7 oncoprotein associates with the cullin 2 ubiquitin ligase complex, which contributes*

to degradation of the retinoblastoma tumor suppressor. J Virol, 2007. **81**(18): p. 9737-47.

50. R.J. Slebos, M.H. Lee, B.S. Plunkett, T.D. Kessis, B.O. Williams, T. Jacks, L. Hedrick, M.B. Kastan, and K.R. Cho, *p53-dependent G1 arrest involves pRB-related proteins and is disrupted by the human papillomavirus 16 E7 oncoprotein.* Proc Natl Acad Sci U S A, 1994. **91**(12): p. 5320-4.
51. C.R. Leemans, B.J. Braakhuis, and R.H. Brakenhoff, *The molecular biology of head and neck cancer.* Nat Rev Cancer, 2011. **11**(1): p. 9-22.
52. C.H. Chung and M.L. Gillison, *Human papillomavirus in head and neck cancer: its role in pathogenesis and clinical implications.* Clin Cancer Res, 2009. **15**(22): p. 6758-62.
53. L. Ribassin-Majed, R. Lounes, and S. Clemencon, *Efficacy of vaccination against HPV infections to prevent cervical cancer in France: present assessment and pathways to improve vaccination policies.* PLoS One, 2012. **7**(3): p. e32251.
54. D. Hanahan and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
55. M. Lechner, C. Boshoff, and S. Beck, *Cancer Epigenome*, in *Advances in Genetics*. 2010, Elsevier. p. - book chapter in press.
56. Collaborators, *Moving AHEAD with an international human epigenome project.* Nature, 2008. **454**(7205): p. 711-5.
57. K.L. Richards, B. Zhang, K.A. Baggerly, S. Colella, J.C. Lang, D.E. Schuller, and R. Krahe, *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**(3): p. e4941.
58. W.A. Burgers, L. Blanchon, S. Pradhan, Y. de Launoit, T. Kouzarides, and F. Fuks, *Viral oncoproteins target the DNA methyltransferases.* Oncogene, 2007. **26**(11): p. 1650-5.
59. A. Bernat, N. Avvakumov, J.S. Mymryk, and L. Banks, *Interaction between the HPV E7 oncoprotein and the transcriptional coactivator p300.* Oncogene, 2003. **22**(39): p. 7871-81.
60. *Time for the epigenome.* Nature, 2010. **463**(7281): p. 587.

61. S.C. Chuang, A. Agudo, W. Ahrens, D. Anantharaman, S. Benhamou, S. Boccia, C. Chen, D.I. Conway, E. Fabianova, R.B. Hayes, C.M. Healy, I. Holcatova, K. Kjaerheim, P. Lagiou, P. Lazarus, T.V. Macfarlane, M.B. Mahimkar, D. Mates, K. Matsuo, F. Merletti, A. Metspalu, H. Morgenstern, J. Muscat, G. Cadoni, A.F. Olshan, M. Purdue, H. Ramroth, P. Rudnai, S.M. Schwartz, L. Simonato, E.M. Smith, E.M. Sturgis, N. Szeszenia-Dabrowska, R. Talamini, P. Thomson, Q. Wei, D. Zaridze, Z.F. Zhang, A. Znaor, P. Brennan, P. Boffetta, and M. Hashibe, *Sequence Variants and the Risk of Head and Neck Cancer: Pooled Analysis in the INHANCE Consortium.* Front Oncol, 2011. **1**: p. 13.
62. C.C. Cortez and P.A. Jones, *Chromatin, cancer and drug therapies.* Mutat Res, 2008. **647**(1-2): p. 44-51.
63. S.B. Baylin and P.A. Jones, *A decade of exploring the cancer epigenome - biological and translational implications.* Nat Rev Cancer, 2011. **11**(10): p. 726-34.
64. J.C. Chow, Z. Yen, S.M. Ziesche, and C.J. Brown, *Silencing of the mammalian X chromosome.* Annu Rev Genomics Hum Genet, 2005. **6**: p. 69-92.
65. I.M. Morison, J.P. Ramsay, and H.G. Spencer, *A census of mammalian imprinting.* Trends Genet, 2005. **21**(8): p. 457-65.
66. W. Reik, *Stability and flexibility of epigenetic gene regulation in mammalian development.* Nature, 2007. **447**(7143): p. 425-32.
67. R. Lister, M. Pelizzola, R.H. Dowen, R.D. Hawkins, G. Hon, J. Tonti-Filippini, J.R. Nery, L. Lee, Z. Ye, Q.M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A.H. Millar, J.A. Thomson, B. Ren, and J.R. Ecker, *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.
68. P. Schmezer and C. Plass, *[Epigenetic aspects in carcinomas of the head and neck].* HNO, 2008. **56**(6): p. 594-602.
69. J.F. Costello, C. Hong, C. Plass, and D.J. Smiraglia, *Restriction landmark genomic scanning: analysis of CpG islands in genomes by 2D gel electrophoresis.* Methods Mol Biol, 2009. **507**: p. 131-48.

70. J.F. Costello, D.J. Smiraglia, and C. Plass, *Restriction landmark genome scanning.* Methods, 2002. **27**(2): p. 144-9.
71. T.H. Huang, M.R. Perry, and D.E. Laux, *Methylation profiling of CpG islands in human breast cancer cells.* Hum Mol Genet, 1999. **8**(3): p. 459-70.
72. R.A. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J.B. Potash, S. Sabunciyan, and A.P. Feinberg, *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.* Nat Genet, 2009. **41**(2): p. 178-86.
73. J. Sandoval, H. Heyn, S. Moran, J. Serra-Musach, M.A. Pujana, M. Bibikova, and M. Esteller, *Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome.* Epigenetics, 2011. **6**(6): p. 692-702.
74. K.H. Taylor, R.S. Kramer, J.W. Davis, J. Guo, D.J. Duff, D. Xu, C.W. Caldwell, and H. Shi, *Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing.* Cancer Res, 2007. **67**(18): p. 8511-8.
75. S.J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C.D. Haudenschild, S. Pradhan, S.F. Nelson, M. Pellegrini, and S.E. Jacobsen, *Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.* Nature, 2008. **452**(7184): p. 215-9.
76. C.A. Bormann Chung, V.L. Boyd, K.J. McKernan, Y. Fu, C. Monighetti, H.E. Peckham, and M. Barker, *Whole methylome analysis by ultra-deep sequencing using two-base encoding.* PLoS One, 2010. **5**(2): p. e9320.
77. L.M. Butcher and S. Beck, *AutoMeDIP-seq: a high-throughput, whole genome, DNA methylation assay.* Methods, 2010. **52**(3): p. 223-31.
78. N. Li, M. Ye, Y. Li, Z. Yan, L.M. Butcher, J. Sun, X. Han, Q. Chen, X. Zhang, and J. Wang, *Whole genome DNA methylation analysis based on high throughput sequencing technology.* Methods, 2010. **52**(3): p. 203-12.

79. E. Hodges, A.D. Smith, J. Kendall, Z. Xuan, K. Ravi, M. Rooks, M.Q. Zhang, K. Ye, A. Bhattacharjee, L. Brizuela, W.R. McCombie, M. Wigler, G.J. Hannon, and J.B. Hicks, *High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing.* Genome Res, 2009. **19**(9): p. 1593-605.
80. X. Zhang, J. Yazaki, A. Sundaresan, S. Cokus, S.W. Chan, H. Chen, I.R. Henderson, P. Shinn, M. Pellegrini, S.E. Jacobsen, and J.R. Ecker, *Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis.* Cell, 2006. **126**(6): p. 1189-201.
81. D. Zilberman, M. Gehring, R.K. Tran, T. Ballinger, and S. Henikoff, *Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription.* Nat Genet, 2007. **39**(1): p. 61-9.
82. M. Bibikova, J. Le, B. Barnes, S. Saedinia-Melnyk, R. Shen, and K.L. Gunderson, *Genome-wide DNA methylation profiling using Infinium® assay.* Epigenomics, 2009. **1**(1): p. 177-200.
83. M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J.M. Le, D. Delano, L. Zhang, G.P. Schroth, K.L. Gunderson, J.B. Fan, and R. Shen, *High density DNA methylation array with single CpG site resolution.* Genomics, 2011. **98**(4): p. 288-95.
84. S.B. Baylin, *DNA methylation and gene silencing in cancer.* Nat Clin Pract Oncol, 2005. **2 Suppl 1**: p. S4-11.
85. C. Long, B. Yin, Q. Lu, X. Zhou, J. Hu, Y. Yang, F. Yu, and Y. Yuan, *Promoter hypermethylation of the RUNX3 gene in esophageal squamous cell carcinoma.* Cancer Invest, 2007. **25**(8): p. 685-90.
86. P.A. Jones and P.W. Laird, *Cancer epigenetics comes of age.* Nat Genet, 1999. **21**(2): p. 163-7.
87. A.S. Wilson, B.E. Power, and P.L. Molloy, *DNA hypomethylation and human diseases.* Biochim Biophys Acta, 2007. **1775**(1): p. 138-62.
88. M. Lechner, C. Boshoff, and S. Beck, *Cancer epigenome.* Adv Genet, 2010. **70**: p. 247-76.
89. C.J. Marsit, E.A. Houseman, B.C. Christensen, K. Eddy, R. Bueno, D.J. Sugarbaker, H.H. Nelson, M.R. Karagas, and K.T. Kelsey, *Examination of a CpG island methylator phenotype and implications of*

methylation profiles in solid tumors. Cancer Res, 2006. **66**(21): p. 10621-9.

90. S. Sengupta, S. Chakrabarti, A. Roy, C.K. Panda, and S. Roychoudhury, *Inactivation of human mutL homolog 1 and mutS homolog 2 genes in head and neck squamous cell carcinoma tumors and leukoplakia samples by promoter hypermethylation and its relation with microsatellite instability phenotype.* Cancer, 2007. **109**(4): p. 703-12.
91. A.L. Carvalho, C. Jeronimo, M.M. Kim, R. Henrique, Z. Zhang, M.O. Hoque, S. Chang, M. Brait, C.S. Nayak, W.W. Jiang, Q. Claybourne, Y. Tokumaru, J. Lee, D. Goldenberg, E. Garrett-Mayer, S. Goodman, C.S. Moon, W. Koch, W.H. Westra, D. Sidransky, and J.A. Califano, *Evaluation of promoter hypermethylation detection in body fluids as a screening/diagnosis tool for head and neck squamous cell carcinoma.* Clin Cancer Res, 2008. **14**(1): p. 97-107.
92. C.A. Righini, F. de Fraipont, J.F. Timsit, C. Faure, E. Brambilla, E. Reyt, and M.C. Favrot, *Tumor-specific methylation in saliva: a promising biomarker for early detection of head and neck cancer recurrence.* Clin Cancer Res, 2007. **13**(4): p. 1179-85.
93. C.J. Marsit, B.C. Christensen, E.A. Houseman, M.R. Karagas, M.R. Wrensch, R.F. Yeh, H.H. Nelson, J.L. Wiemels, S. Zheng, M.R. Posner, M.D. McClean, J.K. Wiencke, and K.T. Kelsey, *Epigenetic profiling reveals etiologically distinct patterns of DNA methylation in head and neck squamous cell carcinoma.* Carcinogenesis, 2009. **30**(3): p. 416-22.
94. D.T. Hsiung, C.J. Marsit, E.A. Houseman, K. Eddy, C.S. Furniss, M.D. McClean, and K.T. Kelsey, *Global DNA methylation level in whole blood as a biomarker in head and neck squamous cell carcinoma.* Cancer Epidemiol Biomarkers Prev, 2007. **16**(1): p. 108-14.
95. M.A. Sartor, D.C. Dolinoy, T.R. Jones, J.A. Colacino, M.E. Prince, T.E. Carey, and L.S. Rozek, *Genome-wide methylation and expression differences in HPV(+) and HPV(-) squamous cell carcinoma cell lines are consistent with divergent mechanisms of carcinogenesis.* Epigenetics, 2011. **6**(6): p. 777-87.

96. H. Richly, L. Aloia, and L. Di Croce, *Roles of the Polycomb group proteins in stem cells and cancer.* Cell Death Dis, 2011. **2**: p. e204.
97. A. Sparmann and M. van Lohuizen, *Polycomb silencers control cell fate, development and cancer.* Nat Rev Cancer, 2006. **6**(11): p. 846-56.
98. M. Widschwendter, H. Fiegl, D. Egle, E. Mueller-Holzner, G. Spizzo, C. Marth, D.J. Weisenberger, M. Campan, J. Young, I. Jacobs, and P.W. Laird, *Epigenetic stem cell signature in cancer.* Nat Genet, 2007. **39**(2): p. 157-8.
99. J.E. Ohm, K.M. McGarvey, X. Yu, L. Cheng, K.E. Schuebel, L. Cope, H.P. Mohammad, W. Chen, V.C. Daniel, W. Yu, D.M. Berman, T. Jenuwein, K. Pruitt, S.J. Sharkis, D.N. Watkins, J.G. Herman, and S.B. Baylin, *A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing.* Nat Genet, 2007. **39**(2): p. 237-42.
100. Y. Schlesinger, R. Straussman, I. Keshet, S. Farkash, M. Hecht, J. Zimmerman, E. Eden, Z. Yakhini, E. Ben-Shushan, B.E. Reubinoff, Y. Bergman, I. Simon, and H. Cedar, *Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer.* Nat Genet, 2007. **39**(2): p. 232-6.
101. A.P. Feinberg, R. Ohlsson, and S. Henikoff, *The epigenetic progenitor origin of human cancer.* Nat Rev Genet, 2006. **7**(1): p. 21-33.
102. K. Paschos and M.J. Allday, *Epigenetic reprogramming of host genes in viral and microbial pathogenesis.* Trends Microbiol, 2010. **18**(10): p. 439-47.
103. A.F. Fernandez, C. Rosales, P. Lopez-Nieva, O. Grana, E. Ballestar, S. Ropero, J. Espada, S.A. Melo, A. Lujambio, M.F. Fraga, I. Pino, B. Javierre, F.J. Carmona, F. Acquadro, R.D. Steenbergen, P.J. Snijders, C.J. Meijer, P. Pineau, A. Dejean, B. Lloveras, G. Capella, J. Quer, M. Buti, J.I. Esteban, H. Allende, F. Rodriguez-Frias, X. Castellsague, J. Minarovits, J. Ponce, D. Capello, G. Gaidano, J.C. Cigudosa, G. Gomez-Lopez, D.G. Pisano, A. Valencia, M.A. Piris, F.X. Bosch, E. Cahir-McFarland, E. Kieff, and M. Esteller, *The dynamic DNA*

methylomes of double-stranded DNA viruses associated with human cancer. Genome Res, 2009. **19**(3): p. 438-51.

104. A.F. Fernandez and M. Esteller, *Viral epigenomes in human tumorigenesis.* Oncogene, 2010. **29**(10): p. 1405-20.
105. E.A. White, R.E. Kramer, M.J. Tan, S.D. Hayes, J.W. Harper, and P.M. Howley, *Comprehensive analysis of host cellular interactions with human papillomavirus e6 proteins identifies new e6 binding partners and reflects viral diversity.* J Virol, 2012. **86**(24): p. 13174-86.
106. J. Califano, P. van der Riet, W. Westra, H. Nawroz, G. Clayman, S. Piantadosi, R. Corio, D. Lee, B. Greenberg, W. Koch, and D. Sidransky, *Genetic progression model for head and neck cancer: implications for field cancerization.* Cancer Res, 1996. **56**(11): p. 2488-92.
107. P.M. Weinberger, Z. Yu, B.G. Haffty, D. Kowalski, M. Harigopal, J. Brandsma, C. Sasaki, J. Joe, R.L. Camp, D.L. Rimm, and A. Psyrri, *Molecular classification identifies a subset of human papillomavirus--associated oropharyngeal cancers with favorable prognosis.* J Clin Oncol, 2006. **24**(5): p. 736-47.
108. S.J. Smeets, B.J. Braakhuis, S. Abbas, P.J. Snijders, B. Ylstra, M.A. van de Wiel, G.A. Meijer, C.R. Leemans, and R.H. Brakenhoff, *Genome-wide DNA copy number alterations in head and neck squamous cell carcinomas with or without oncogene-expressing human papillomavirus.* Oncogene, 2006. **25**(17): p. 2558-64.
109. N. Reimers, H.U. Kasper, S.J. Weissenborn, H. Stutzer, S.F. Preuss, T.K. Hoffmann, E.J. Speel, H.P. Dienes, H.J. Pfister, O. Guntinas-Lichius, and J.P. Klussmann, *Combined analysis of HPV-DNA, p16 and EGFR expression to predict prognosis in oropharyngeal cancer.* Int J Cancer, 2007. **120**(8): p. 1731-8.
110. N. Stransky, A.M. Egloff, A.D. Tward, A.D. Kostic, K. Cibulskis, A. Sivachenko, G.V. Kryukov, M. Lawrence, C. Sougnez, A. McKenna, E. Shefler, A.H. Ramos, P. Stojanov, S.L. Carter, D. Voet, M.L. Cortes, D. Auclair, M.F. Berger, G. Saksena, C. Guiducci, R. Onofrio, M. Parkin, M. Romkes, J.L. Weissfeld, R.R. Seethala, L. Wang, C. Rangel-Escareno, J.C. Fernandez-Lopez, A. Hidalgo-Miranda, J.

Melendez-Zajgla, W. Winckler, K. Ardlie, S.B. Gabriel, M. Meyerson, E.S. Lander, G. Getz, T.R. Golub, L.A. Garraway, and J.R. Grandis, *The Mutational Landscape of Head and Neck Squamous Cell Carcinoma.* Science, 2011.

111. N. Agrawal, M.J. Frederick, C.R. Pickering, C. Bettegowda, K. Chang, R.J. Li, C. Fakhry, T.X. Xie, J. Zhang, J. Wang, N. Zhang, A.K. El-Naggar, S.A. Jasser, J.N. Weinstein, L. Trevino, J.A. Drummond, D.M. Muzny, Y. Wu, L.D. Wood, R.H. Hruban, W.H. Westra, W.M. Koch, J.A. Califano, R.A. Gibbs, D. Sidransky, B. Vogelstein, V.E. Velculescu, N. Papadopoulos, D.A. Wheeler, K.W. Kinzler, and J.N. Myers, *Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1.* Science, 2011. **333**(6046): p. 1154-7.
112. C. Liang, C.J. Marsit, M.D. McClean, H.H. Nelson, B.C. Christensen, R.I. Haddad, J.R. Clark, R.O. Wein, G.A. Grillone, E.A. Houseman, G. Halec, T. Waterboer, M. Pawlita, J.F. Krane, and K.T. Kelsey, *Biomarkers of HPV in Head and Neck Squamous Cell Carcinoma.* Cancer Res, 2012. **72**(19): p. 5004-5013.
113. S. Vinokurova and M. von Knebel Doeberitz, *Differential methylation of the HPV 16 upstream regulatory region during epithelial differentiation and neoplastic transformation.* PLoS One, 2011. **6**(9): p. e24451.
114. J.L. Brandsma, Y. Sun, P.M. Lizardi, D.P. Tuck, D. Zelterman, G.K. Haines, 3rd, M. Martel, M. Harigopal, K. Schofield, and M. Neapolitano, *Distinct human papillomavirus type 16 methylomes in cervical cells at different stages of premalignancy.* Virology, 2009. **389**(1-2): p. 100-7.
115. M. Kalantari, L.L. Villa, I.E. Calleja-Macias, and H.U. Bernard, *Human papillomavirus-16 and -18 in penile carcinomas: DNA methylation, chromosomal recombination and genomic variation.* Int J Cancer, 2008. **123**(8): p. 1832-40.
116. T. Turan, M. Kalantari, I.E. Calleja-Macias, H.A. Cubie, K. Cuschieri, L.L. Villa, H. Skomedal, H.A. Barrera-Saldana, and H.U. Bernard, *Methylation of the human papillomavirus-18 L1 gene: a biomarker of neoplastic progression?* Virology, 2006. **349**(1): p. 175-83.

117. T. Turan, M. Kalantari, K. Cuschieri, H.A. Cubie, H. Skomedal, and H.U. Bernard, *High-throughput detection of human papillomavirus-18 L1 gene methylation, a candidate biomarker for the progression of cervical neoplasia.* Virology, 2007. **361**(1): p. 185-93.

118. A. Balderas-Loaeza, G. Anaya-Saavedra, V.A. Ramirez-Amador, M.C. Guido-Jimenez, M. Kalantari, I.E. Calleja-Macias, H.U. Bernard, and A. Garcia-Carranca, *Human papillomavirus-16 DNA methylation patterns support a causal association of the virus with oral squamous cell carcinomas.* Int J Cancer, 2007. **120**(10): p. 2165-9.

119. I.S. Park, X. Chang, M. Loyo, G. Wu, A. Chuang, M.S. Kim, Y.K. Chae, S. Lyford-Pike, W.H. Westra, J.R. Saunders, D. Sidransky, and S.I. Pai, *Characterization of the methylation patterns in human papillomavirus type 16 viral DNA in head and neck cancers.* Cancer Prev Res (Phila), 2011. **4**(2): p. 207-17.

120. S.J. Smeets, A.T. Hesselink, E.J. Speel, A. Haesevoets, P.J. Snijders, M. Pawlita, C.J. Meijer, B.J. Braakhuis, C.R. Leemans, and R.H. Brakenhoff, *A novel algorithm for reliable detection of human papillomavirus in paraffin embedded head and neck cancer specimen.* Int J Cancer, 2007. **121**(11): p. 2465-72.

121. A.G. Schache, T. Liloglou, J.M. Risk, A. Filia, T.M. Jones, J. Sheard, J.A. Woolgar, T.R. Helliwell, A. Triantafyllou, M. Robinson, P. Sloan, C. Harvey-Woodworth, D. Sisson, and R.J. Shaw, *Evaluation of human papilloma virus diagnostic testing in oropharyngeal squamous cell carcinoma: sensitivity, specificity, and prognostic discrimination.* Clin Cancer Res, 2011. **17**(19): p. 6262-71.

122. S. Thavaraj, A. Stokes, E. Guerra, J. Bible, E. Halligan, A. Long, A. Okpokam, P. Sloan, E. Odell, and M. Robinson, *Evaluation of human papillomavirus testing for squamous cell carcinoma of the tonsil in clinical practice.* J Clin Pathol, 2011. **64**(4): p. 308-12.

123. D.G. Ginzinger, *Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream.* Exp Hematol, 2002. **30**(6): p. 503-12.

124. A.A. Dussault and M. Pouliot, *Rapid and simple comparison of messenger RNA levels using real-time PCR.* Biol Proced Online, 2006. **8**: p. 1-10.
125. M. Weber, J.J. Davies, D. Wittig, E.J. Oakeley, M. Haase, W.L. Lam, and D. Schubeler, *Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells.* Nat Genet, 2005. **37**(8): p. 853-62.
126. F. Mohn, M. Weber, D. Schubeler, and T.C. Roloff, *Methylated DNA immunoprecipitation (MeDIP).* Methods Mol Biol, 2009. **507**: p. 55-64.
127. L.M. Butcher and S. Beck, *AutoMeDIP-seq: A high-throughput, whole genome, DNA methylation assay.* Methods, 2010. **Article in press**.
128. G.A. Wilson, P. Dhami, A. Feber, D. Cortázar, Y. Suzuki, R. Schulz, P. Schär, and S. Beck, *Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers.* GigaScience, 2012.
129. H. Li and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
130. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
131. L. Chavez, J. Jozefczuk, C. Grimm, J. Dietrich, B. Timmermann, H. Lehrach, R. Herwig, and J. Adjaye, *Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage.* Genome Res, 2010. **20**(10): p. 1441-50.
132. C. Thirlwell, M. Eymard, A. Feber, A. Teschendorff, K. Pearce, M. Lechner, M. Widschwendter, and S. Beck, *Genome-wide DNA methylation analysis of archival formalin-fixed paraffin-embedded tissue using the Illumina Infinium HumanMethylation27 BeadChip.* Methods, 2010. **52**(3): p. 248-54.
133. P. Du, W.A. Kibbe, and S.M. Lin, *lumi: a pipeline for processing Illumina microarray.* Bioinformatics, 2008. **24**(13): p. 1547-8.
134. A.E. Teschendorff, U. Menon, A. Gentry-Maharaj, S.J. Ramus, S.A. Gayther, S. Apostolidou, A. Jones, M. Lechner, S. Beck, I.J. Jacobs,

and M. Widschwendter, *An epigenetic signature in peripheral blood predicts active ovarian cancer.* PLoS One, 2009. **4**(12): p. e8274.

135. A.E. Teschendorff, J. Zhuang, and M. Widschwendter, *Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies.* Bioinformatics, 2011. **27**(11): p. 1496-505.
136. S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks, *Evaluation of the Infinium Methylation 450K technology.* Epigenomics, 2011. **3**(6): p. 771-84.
137. J.D. Storey and R. Tibshirani, *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.
138. M. Zhao, E. Rosenbaum, A.L. Carvalho, W. Koch, W. Jiang, D. Sidransky, and J. Califano, *Feasibility of quantitative PCR-based saliva rinse screening of HPV for head and neck cancer.* Int J Cancer, 2005. **117**(4): p. 605-10.
139. F. Wang-Johanning, D.W. Lu, Y. Wang, M.R. Johnson, and G.L. Johanning, *Quantitation of human papillomavirus 16 E6 and E7 DNA and RNA in residual material from ThinPrep Papanicolaou tests using real-time polymerase chain reaction analysis.* Cancer, 2002. **94**(8): p. 2199-210.
140. M. Frommer, L.E. McDonald, D.S. Millar, C.M. Collis, F. Watt, G.W. Grigg, P.L. Molloy, and C.L. Paul, *A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.* Proc Natl Acad Sci U S A, 1992. **89**(5): p. 1827-31.
141. H. Caren, A. Djos, M. Nethander, R.M. Sjoberg, P. Kogner, C. Enstrom, S. Nilsson, and T. Martinsson, *Identification of epigenetically regulated genes that predict patient outcome in neuroblastoma.* BMC Cancer, 2011. **11**: p. 66.
142. J. Chaisaingmongkol, O. Popanda, R. Warta, G. Dyckhoff, E. Herpel, L. Geiselhart, R. Claus, F. Lasitschka, B. Campos, C.C. Oakes, J.L. Bermejo, C. Herold-Mende, C. Plass, and P. Schmezer, *Epigenetic screen of human DNA repair genes identifies aberrant promoter methylation of NEIL1 in head and neck squamous cell carcinoma.* Oncogene, 2012.

143. K.L. Bennett, M. Karpenko, M.T. Lin, R. Claus, K. Arab, G. Dyckhoff, P. Plinkert, E. Herpel, D. Smiraglia, and C. Plass, *Frequently methylated tumor suppressor genes in head and neck squamous cell carcinoma.* Cancer Res, 2008. **68**(12): p. 4494-9.
144. Y. Wu, L. Meng, H. Wang, Q. Xu, S. Wang, S. Wu, L. Xi, Y. Zhao, J. Zhou, G. Xu, Y. Lu, and D. Ma, *Regulation of DNA methylation on the expression of the FHIT gene contributes to cervical carcinoma cell tumorigenesis.* Oncol Rep, 2006. **16**(3): p. 625-9.
145. T. Bai, T. Tanaka, K. Yukawa, and N. Umesaki, *A novel mechanism for acquired cisplatin-resistance: suppressed translation of death-associated protein kinase mRNA is insensitive to 5-aza-2'-deoxycitidine and trichostatin in cisplatin-resistant cervical squamous cancer cells.* Int J Oncol, 2006. **28**(2): p. 497-508.
146. D. Mossman, K.T. Kim, and R.J. Scott, *Demethylation by 5-aza-2'-deoxycytidine in colorectal cancer cells targets genomic DNA whilst promoter CpG island methylation persists.* BMC Cancer, 2010. **10**: p. 366.
147. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo, *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
148. S.A. Forbes, N. Bindal, S. Bamford, C. Cole, C.Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J.W. Teague, P.J. Campbell, M.R. Stratton, and P.A. Futreal, *COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.* Nucleic Acids Res, 2011. **39**(Database issue): p. D945-50.
149. D. Fumagalli, P.G. Gavin, Y. Taniyama, S.I. Kim, H.J. Choi, S. Paik, and K.L. Pogue-Geile, *A rapid, sensitive, reproducible and cost-effective method for mutation profiling of colon cancer and metastatic lymph nodes.* BMC Cancer, 2010. **10**: p. 101.
150. R.K. Thomas, A.C. Baker, R.M. Debiasi, W. Winckler, T. Laframboise, W.M. Lin, M. Wang, W. Feng, T. Zander, L. MacConaill, J.C. Lee, R. Nicoletti, C. Hatton, M. Goyette, L. Girard, K. Majmudar, L. Ziaugra,

K.K. Wong, S. Gabriel, R. Beroukhim, M. Peyton, J. Barretina, A. Dutt, C. Emery, H. Greulich, K. Shah, H. Sasaki, A. Gazdar, J. Minna, S.A. Armstrong, I.K. Mellinghoff, F.S. Hodi, G. Dranoff, P.S. Mischel, T.F. Cloughesy, S.F. Nelson, L.M. Liau, K. Mertz, M.A. Rubin, H. Moch, M. Loda, W. Catalona, J. Fletcher, S. Signoretti, F. Kaye, K.C. Anderson, G.D. Demetri, R. Dummer, S. Wagner, M. Herlyn, W.R. Sellers, M. Meyerson, and L.A. Garraway, *High-throughput oncogene mutation profiling in human cancer.* Nat Genet, 2007. **39**(3): p. 347-51.

151. B. Djordjevic, B.T. Hennessy, J. Li, B.A. Barkoh, R. Luthra, G.B. Mills, and R.R. Broaddus, *Clinical assessment of PTEN loss in endometrial carcinoma: immunohistochemistry outperforms gene sequencing.* Mod Pathol, 2012. **25**(5): p. 699-708.
152. J.S. Reis-Filho, K. Savage, M.B. Lambros, M. James, D. Steele, R.L. Jones, and M. Dowsett, *Cyclin D1 protein overexpression and CCND1 amplification in breast carcinomas: an immunohistochemical and chromogenic in situ hybridisation analysis.* Mod Pathol, 2006. **19**(7): p. 999-1009.
153. S.A. Selamat, B.S. Chung, L. Girard, W. Zhang, Y. Zhang, M. Campan, K.D. Siegmund, M.N. Koss, J.A. Hagen, W.L. Lam, S. Lam, A.F. Gazdar, and I.A. Laird-Offringa, *Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression.* Genome Res, 2012. **22**(7): p. 1197-211.
154. J. Zhuang, A. Jones, S.H. Lee, E. Ng, H. Fiegl, M. Zikan, D. Cibula, A. Sargent, H.B. Salvesen, I.J. Jacobs, H.C. Kitchener, A.E. Teschendorff, and M. Widschwendter, *The dynamics and prognostic potential of DNA methylation changes at stem cell gene loci in women's cancer.* PLoS Genet, 2012. **8**(2): p. e1002517.
155. V.K. Rakyan, H. Beyan, T.A. Down, M.I. Hawa, S. Maslau, D. Aden, A. Daunay, F. Busato, C.A. Mein, B. Manfras, K.R. Dias, C.G. Bell, J. Tost, B.O. Boehm, S. Beck, and R.D. Leslie, *Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis.* PLoS Genet, 2011. **7**(9): p. e1002300.

156. G.K. Smyth, *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
157. C.G. Bell, A.E. Teschendorff, V.K. Rakyan, A.P. Maxwell, S. Beck, and D.A. Savage, *Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus.* BMC Med Genomics, 2010. **3**: p. 33.
158. F. Fang, S. Turcan, A. Rimner, A. Kaufman, D. Giri, L.G. Morris, R. Shen, V. Seshan, Q. Mo, A. Heguy, S.B. Baylin, N. Ahuja, A. Viale, J. Massague, L. Norton, L.T. Vahdat, M.E. Moynahan, and T.A. Chan, *Breast cancer methylomes establish an epigenomic foundation for metastasis.* Sci Transl Med, 2011. **3**(75): p. 75ra25.
159. H. Noushmehr, D.J. Weisenberger, K. Diefes, H.S. Phillips, K. Pujara, B.P. Berman, F. Pan, C.E. Pelloski, E.P. Sulman, K.P. Bhat, R.G. Verhaak, K.A. Hoadley, D.N. Hayes, C.M. Perou, H.K. Schmidt, L. Ding, R.K. Wilson, D. Van Den Berg, H. Shen, H. Bengtsson, P. Neuvial, L.M. Cope, J. Buckley, J.G. Herman, S.B. Baylin, P.W. Laird, and K. Aldape, *Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma.* Cancer Cell, 2010. **17**(5): p. 510-22.
160. M.M. Suzuki and A. Bird, *DNA methylation landscapes: provocative insights from epigenomics.* Nat Rev Genet, 2008. **9**(6): p. 465-76.
161. M. Kulis and M. Esteller, *DNA methylation and cancer.* Adv Genet, 2010. **70**: p. 27-56.
162. D.J. Weisenberger, K.D. Siegmund, M. Campan, J. Young, T.I. Long, M.A. Faasse, G.H. Kang, M. Widschwendter, D. Weener, D. Buchanan, H. Koh, L. Simms, M. Barker, B. Leggett, J. Levine, M. Kim, A.J. French, S.N. Thibodeau, J. Jass, R. Haile, and P.W. Laird, *CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer.* Nat Genet, 2006. **38**(7): p. 787-93.
163. G.M. Poage, R.A. Butler, E.A. Houseman, M.D. McClean, H.H. Nelson, B.C. Christensen, C.J. Marsit, and K.T. Kelsey, *Identification*

of an Epigenetic Profile Classifier That Is Associated with Survival in Head and Neck Cancer. Cancer Res, 2012.

164. G. Berx and F. van Roy, *Involvement of members of the cadherin superfamily in cancer.* Cold Spring Harb Perspect Biol, 2009. **1**(6): p. a003129.
165. J.P. Thiery, *Epithelial-mesenchymal transitions in tumour progression.* Nat Rev Cancer, 2002. **2**(6): p. 442-54.
166. F.E. Henken, S.M. Wilting, R.M. Overmeer, J.G. van Rietschoten, A.O. Nygren, A. Errami, J.P. Schouten, C.J. Meijer, P.J. Snijders, and R.D. Steenbergen, *Sequential gene promoter methylation during HPV-induced cervical carcinogenesis.* Br J Cancer, 2007. **97**(10): p. 1457-64.
167. A.G. Rivenbark, S. Stolzenburg, A.S. Beltran, X. Yuan, M.G. Rots, B.D. Strahl, and P. Blancafort, *Epigenetic reprogramming of cancer cells via targeted DNA methylation.* Epigenetics, 2012. **7**(4).
168. D.J. Gregory, L. Mikhaylova, and A.V. Fedulov, *Selective DNA demethylation by fusion of TDG with a sequence-specific DNA-binding domain.* Epigenetics, 2012. **7**(4).
169. M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S.J. Jones, and M.A. Marra, *Circos: an information aesthetic for comparative genomics.* Genome Res, 2009. **19**(9): p. 1639-45.
170. M. Ito, L. Barys, T. O'Reilly, S. Young, B. Gorbatcheva, J. Monahan, S. Zumstein-Mecker, P.F. Choong, I. Dickinson, P. Crowe, C. Hemmings, J. Desai, D.M. Thomas, and J. Lisztwan, *Comprehensive Mapping of p53 Pathway Alterations Reveals an Apparent Role for Both SNP309 and MDM2 Amplification in Sarcomagenesis.* Clinical Cancer Research, 2011. **17**(3): p. 416-426.
171. A. Petitjean, E. Mathe, S. Kato, C. Ishioka, S.V. Tavtigian, P. Hainaut, and M. Olivier, *Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database.* Human Mutation, 2007. **28**(6): p. 622-629.

172. S.M. Gollin, *Chromosomal alterations in squamous cell carcinomas of the head and neck: window to the biology of disease.* Head Neck, 2001. **23**(3): p. 238-53.
173. E. Cerami, N. Schultz, B. Gross, C. Byrne, A. Jacobsen, J. Gao, E. Larsson, B.A. Aksoy, S.O. Sumer, Y. Sun, S. Sonlu, U. Dogrusoz, and C. Sander, *cBio Cancer Genomics Portal*, C.B.C.a.M.S.-K.C.C.a.t.i.-V.I.V.R.G.o.t.C.E.D.a.B. University, Editor. 2012, Computational Biology Center at Memorial Sloan-Kettering Cancer Center and the i-Vis (Information Visualization) Research Group of the Computer Engineering Department at Bilkent University: New York.
174. M.T. Wang, G. Chen, S.J. An, Z.H. Chen, Z.M. Huang, P. Xiao, X.S. Ben, Z. Xie, S.L. Chen, D.L. Luo, J.M. Tang, J.Y. Lin, X.C. Zhang, and Y.L. Wu, *Prognostic significance of cyclinD1 amplification and the co-alteration of cyclinD1/pRb/ppRb in patients with esophageal squamous cell carcinoma.* Dis Esophagus, 2011.
175. A.A. Molinolo, C. Marsh, M. El Dinali, N. Gangane, K. Jennison, S. Hewitt, V. Patel, T.Y. Seiwert, and J.S. Gutkind, *mTOR as a Molecular Target in HPV-Associated Oral and Cervical Squamous Carcinomas.* Clin Cancer Res, 2012. **18**(9): p. 2558-68.
176. C. Schulein, M. Eilers, and N. Popov, *PI3K-dependent phosphorylation of Fbw7 modulates substrate degradation and activity.* FEBS Lett, 2011. **585**(14): p. 2151-7.
177. K.T. Siu, M.R. Rosner, and A.C. Minella, *An integrated view of cyclin E function and regulation.* Cell Cycle, 2012. **11**(1): p. 57-64.
178. C. Kox, M. Zimmermann, M. Stanulla, S. Leible, M. Schrappe, W.D. Ludwig, R. Koehler, G. Tolle, O.R. Bandapalli, S. Breit, M.U. Muckenthaler, and A.E. Kulozik, *The favorable effect of activating NOTCH1 receptor mutations on long-term outcome in T-ALL patients treated on the ALL-BFM 2000 protocol can be separated from FBXW7 loss of function.* Leukemia, 2010. **24**(12): p. 2005-13.
179. N. Agrawal, M.J. Frederick, C.R. Pickering, C. Bettegowda, K. Chang, R.J. Li, C. Fakhry, T.X. Xie, J. Zhang, J. Wang, N. Zhang, A.K. El-Naggar, S.A. Jasser, J.N. Weinstein, L. Trevino, J.A. Drummond, D.M. Muzny, Y. Wu, L.D. Wood, R.H. Hruban, W.H. Westra, W.M. Koch,

J.A. Califano, R.A. Gibbs, D. Sidransky, B. Vogelstein, V.E. Velculescu, N. Papadopoulos, D.A. Wheeler, K.W. Kinzler, and J.N. Myers, *Exome Sequencing of Head and Neck Squamous Cell Carcinoma Reveals Inactivating Mutations in NOTCH1.* Science, 2011.

180. R. Redon, D. Muller, K. Caulee, K. Wanherdrick, J. Abecassis, and S. du Manoir, *A simple specific pattern of chromosomal aberrations at early stages of head and neck squamous cell carcinomas: PIK3CA but not p63 gene as a likely target of 3q26-qter gains.* Cancer Res, 2001. **61**(10): p. 4122-9.
181. W.G. Yarbrough, A. Whigham, B. Brown, M. Roach, and R. Slebos, *Phosphoinositide kinase-3 status associated with presence or absence of human papillomavirus in head and neck squamous cell carcinomas.* Int J Radiat Oncol Biol Phys, 2007. **69**(2 Suppl): p. S98-101.
182. F. McCaughan, J.C. Pole, A.T. Bankier, B.A. Konfortov, B. Carroll, M. Falzon, T.H. Rabbitts, P.J. George, P.H. Dear, and P.H. Rabbitts, *Progressive 3q amplification consistently targets SOX2 in preinvasive squamous lung cancer.* Am J Respir Crit Care Med, 2010. **182**(1): p. 83-91.
183. Y. Gen, K. Yasui, Y. Zen, K. Zen, O. Dohi, M. Endo, K. Tsuji, N. Wakabayashi, Y. Itoh, Y. Naito, M. Taniwaki, Y. Nakanuma, T. Okanoue, and T. Yoshikawa, *SOX2 identified as a target gene for the amplification at 3q26 that is frequently detected in esophageal squamous cell carcinoma.* Cancer Genet Cytogenet, 2010. **202**(2): p. 82-93.
184. K. Freier, K. Knoepfle, C. Flechtenmacher, S. Pungs, F. Devens, G. Toedt, C. Hofele, S. Joos, P. Lichter, and B. Radlwimmer, *Recurrent copy number gain of transcription factor SOX2 and corresponding high protein expression in oral squamous cell carcinoma.* Genes Chromosomes Cancer, 2010. **49**(1): p. 9-16.
185. C. Chen, Y. Wei, M. Hummel, T.K. Hoffmann, M. Gross, A.M. Kaufmann, and A.E. Albers, *Evidence for epithelial-mesenchymal*

transition in cancer stem cells of head and neck squamous cell carcinoma. PLoS One, 2011. **6**(1): p. e16466.

186. E.R. Kline, S. Muller, L. Pan, M. Tighiouart, Z.G. Chen, and A.I. Marcus, *Localization-specific LKB1 loss in head and neck squamous cell carcinoma metastasis.* Head Neck, 2011. **33**(10): p. 1501-12.

187. Wistuba, II, A.F. Gazdar, and J.D. Minna, *Molecular genetics of small cell lung carcinoma.* Semin Oncol, 2001. **28**(2 Suppl 4): p. 3-13.

188. F. Le Calvez, A. Mukeria, J.D. Hunt, O. Kelm, R.J. Hung, P. Taniere, P. Brennan, P. Boffetta, D.G. Zaridze, and P. Hainaut, *TP53 and KRAS mutation load and types in lung cancers in relation to tobacco smoke: distinct patterns in never, former, and current smokers.* Cancer Res, 2005. **65**(12): p. 5076-83.

189. L. Ding, G. Getz, D.A. Wheeler, E.R. Mardis, M.D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D.M. Muzny, M.B. Morgan, L. Fulton, R.S. Fulton, Q. Zhang, M.C. Wendl, M.S. Lawrence, D.E. Larson, K. Chen, D.J. Dooling, A. Sabo, A.C. Hawes, H. Shen, S.N. Jhangiani, L.R. Lewis, O. Hall, Y. Zhu, T. Mathew, Y. Ren, J. Yao, S.E. Scherer, K. Clerc, G.A. Metcalf, B. Ng, A. Milosavljevic, M.L. Gonzalez-Garay, J.R. Osborne, R. Meyer, X. Shi, Y. Tang, D.C. Koboldt, L. Lin, R. Abbott, T.L. Miner, C. Pohl, G. Fewell, C. Haipek, H. Schmidt, B.H. Dunford-Shore, A. Kraja, S.D. Crosby, C.S. Sawyer, T. Vickery, S. Sander, J. Robinson, W. Winckler, J. Baldwin, L.R. Chirieac, A. Dutt, T. Fennell, M. Hanna, B.E. Johnson, R.C. Onofrio, R.K. Thomas, G. Tonon, B.A. Weir, X. Zhao, L. Ziaugra, M.C. Zody, T. Giordano, M.B. Orringer, J.A. Roth, M.R. Spitz, Wistuba, II, B. Ozenberger, P.J. Good, A.C. Chang, D.G. Beer, M.A. Watson, M. Ladanyi, S. Broderick, A. Yoshizawa, W.D. Travis, W. Pao, M.A. Province, G.M. Weinstock, H.E. Varmus, S.B. Gabriel, E.S. Lander, R.A. Gibbs, M. Meyerson, and R.K. Wilson, *Somatic mutations affect key pathways in lung adenocarcinoma.* Nature, 2008. **455**(7216): p. 1069-75.

190. S.A. Tonon, M.A. Picconi, P.D. Bos, J.B. Zinovich, J. Galuppo, L.V. Alonio, and A.R. Teyssie, *Physical status of the E2 human papilloma*

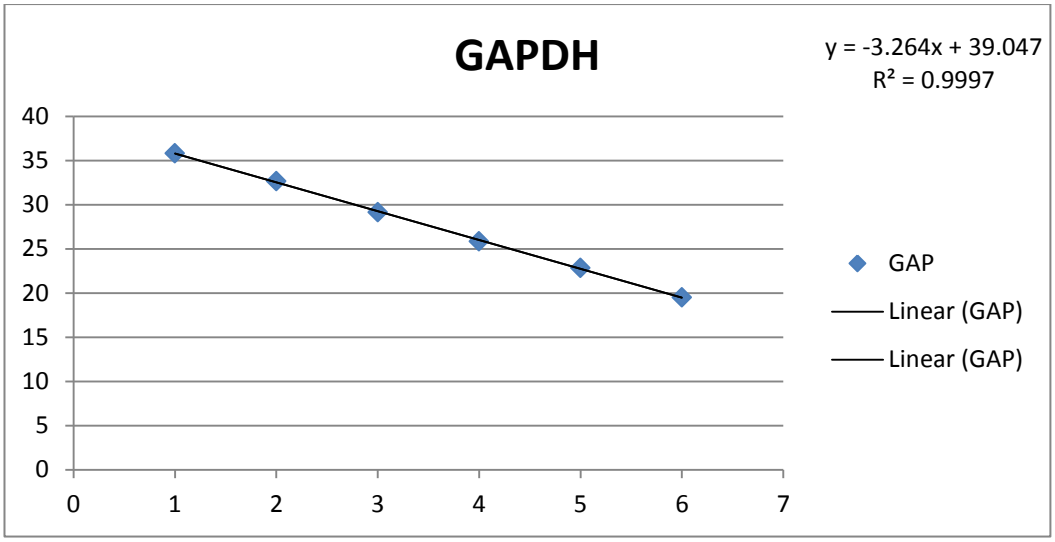
virus 16 viral gene in cervical preneoplastic and neoplastic lesions. J Clin Virol, 2001. **21**(2): p. 129-34.

191. H.U. Bernard, R.D. Burk, Z. Chen, K. van Doorslaer, H. Hausen, and E.M. de Villiers, *Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments.* Virology, 2010. **401**(1): p. 70-9.

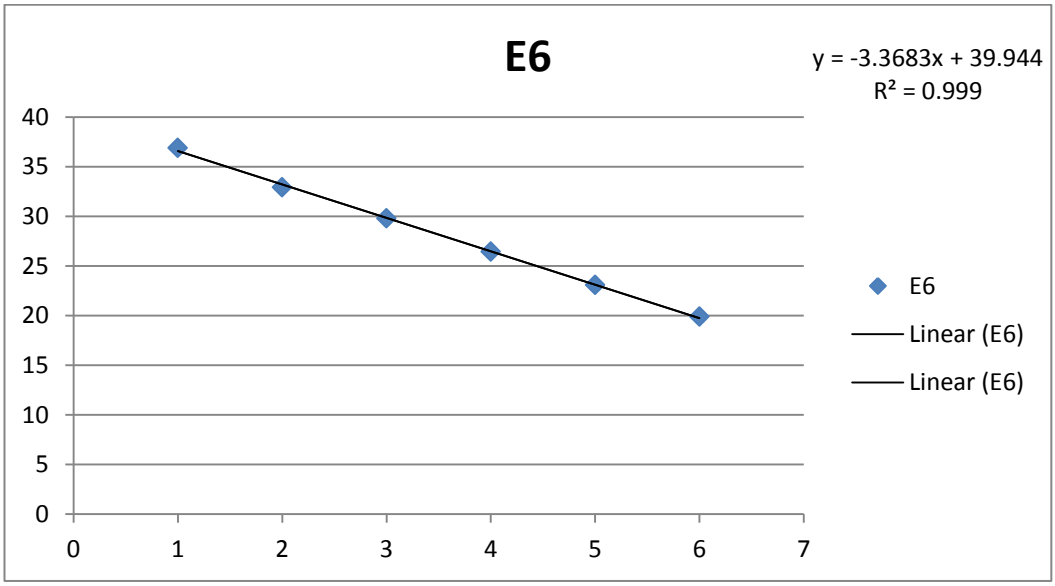
192. M. Gottschling, M. Goker, A. Stamatakis, O.R. Bininda-Emonds, I. Nindl, and I.G. Bravo, *Quantifying the phylodynamic forces driving papillomavirus evolution.* Mol Biol Evol, 2011. **28**(7): p. 2101-13.

9. Appendix

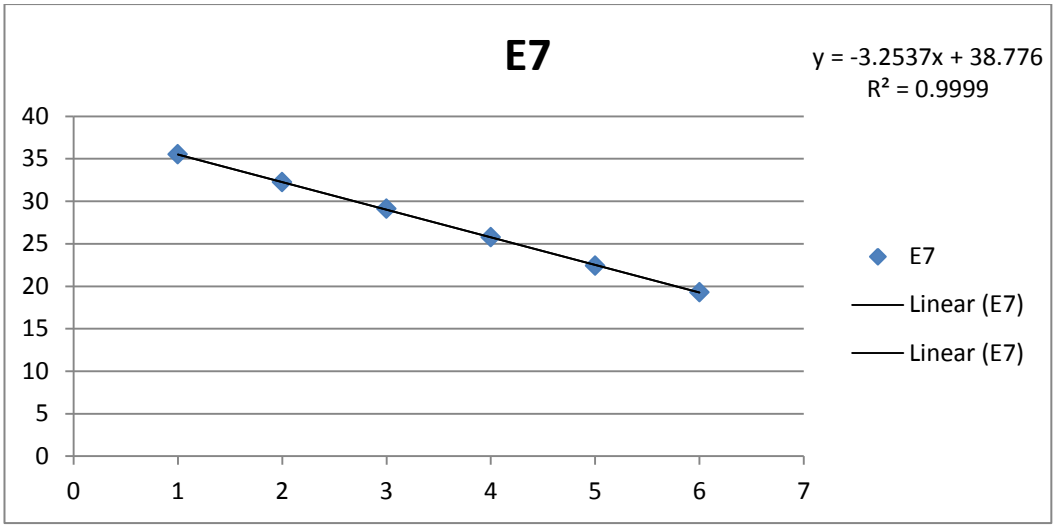
Appendix Figure A.1: Maximum Likelihood based phylogenetic tree of 207 PVs as inferred from a combined E1-E2-L1 amino acid sequence analysis. PV taxonomic units are indicated in Greek letters (‘genera’) and numbers (‘species’: [191]); obtained from [192].



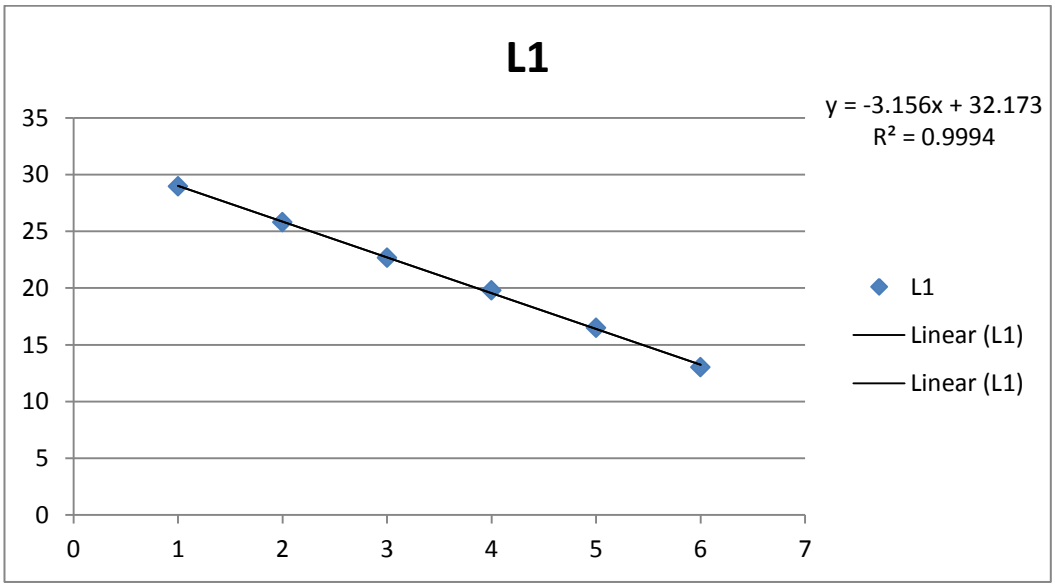
Appendix Figure A.2: Standard curve for GAPDH (slope= -3.264, primer efficiency = 102.48 %)



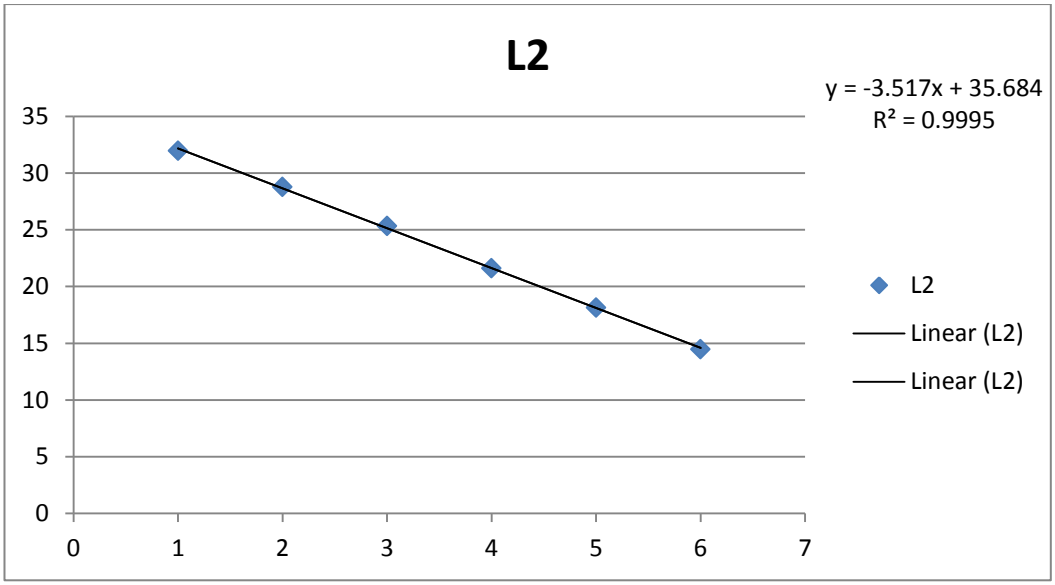
Appendix Figure A.3: Standard curve for E6 (slope= -3.3683, primer efficiency = 98.1 %).



Appendix Figure A.4: Standard curve for E7 (slope= -3.2537, primer efficiency = 102.93 %).



Appendix Figure A.5: Standard curve for L1 (slope= -3.156, primer efficiency = 107.42 %).



Appendix Figure A.6: Standard curve for L2 (slope= -3.517, primer efficiency = 92.46 %).

A

Gene name				
ABL1	*CDK6*	*FLT4*	*MEN1*	*PTPN11*
ABL2	*CDK8*	*FOXP4*	*MET*	*PTPRD*
AKT1	*CDKN2A*	*GATA1*	*MITF*	*RAF1*
AKT2	*CDKN2B*	*GNA11*	*MLH1*	*RARA*
AKT3	*CDKN2C*	*GNAQ*	*MLL*	*RB1*
ALK	*CEBPA*	*GNAS*	*MPL*	*RET*
APC	*CHEK1*	*GPR124*	*MRE11A*	*RICTOR*
AR	*CHEK2*	*GUCY1A2*	*MSH2*	*RPTOR*
ARAF	*CRKL*	*HOXA3*	*MSH6*	*RUNX1*
ARFRP1	*CRLF2*	*HRAS*	*MTOR*	*SMAD2*
ARID1A	*CTNNB1*	*HSP90AA1*	*MUTYH*	*SMAD3*
ATM	*DDR2*	*IDH1*	*MYC*	*SMAD4*
ATR	*DNMT3A*	*IDH2*	*MYCL1*	*SMARCA4*
AURKA	*DOT1L*	*IGF1R*	*MYCN*	*SMARCB1*
AURKB	*EGFR*	*IGF2R*	*NF1*	*SMO*
BAP1	*EPHA3*	*IKBKE*	*NF2*	*SOX10*
BCL2	*EPHA5*	*IKZF1*	*NKX2-1*	*SOX2*
BCL2A1	*EPHA6*	*INHBA*	*NOTCH1*	*SRC*
BCL2L1	*EPHA7*	*INSR*	*NPM1*	*STAT3*
BCL2L2	*EPHB1*	*IRS2*	*NRAS*	*STK11*
BCL6	*EPHB4*	*JAK1*	*NTRK1*	*SUFU*
BRAF	*EPHB6*	*JAK2*	*NTRK2*	*TBX22*
BRCA1	*ERBB2*	*JAK3*	*NTRK3*	*TET2*
BRCA2	*ERBB3*	*JUN*	*PAK3*	*TGFBR2*
CARD11	*ERBB4*	*KDM6A*	*PAX5*	*TNFAIP3*
CBL	*ERCC2*	*KDR*	*PDGFRA*	*TNKS*
CCND1	*ERG*	*KIT*	*PDGFRB*	*TNKS2*
CCND2	*ESR1*	*KRAS*	*PHLPP2*	*TOP1*
CCND3	*EZH2*	*LRP1B*	*PIK3CA*	*TP53*
CCNE1	*FANCA*	*LRP6*	*PIK3CG*	*TSC1*
CD79A	*FBXW7*	*LTK*	*PIK3R1*	*TSC2*
CD79B	*FGFR1*	*MAP2K1*	*PKHD1*	*USP9X*
CDH1	*FGFR2*	*MAP2K2*	*PLCG1*	*VHL*
CDH2	*FGFR3*	*MAP2K4*	*PRKDC*	*WT1*
CDH20	*FGFR4*	*MCL1*	*PTCH1*	
CDH5	*FLT1*	*MDM2*	*PTCH2*	
CDK4	*FLT3*	*MDM4*	*PTEN*	

B

Gene name
ALK
BCR
BRAF
EGFR
ETV1
ETV4
ETV5
ETV6
EWSR1
MLL
RAF1
RARA
RET
TMPRSS2

Appendix Table A.1: 182 genes sequenced across entire coding sequence (A) and 14 genes sequenced across selected introns (B).

OncoCarta v1.0			OncoCarta v3.0	
Gene_Assay ID	**Mutation**		**Gene_Assay ID**	**Mutation**
ABL1_1	G250E		ABL1_301	M244V
ABL1_10	F317L		ABL1_302	L248V
ABL1_11	M351T		AKT1_303	E17K
ABL1_12	E355G		APC_304	E1306*
ABL1_13	F359V		APC_305	E1309fs*4
ABL1_14	H396R		APC_306	APC_E1379*
ABL1_2	Q252H		APC_307	APC_Q1338*
ABL1_3	Y253H		APC_308	Q1367*
ABL1_4	Y253F		APC_309	Q1378*
ABL1_5	E255K		APC_310	Q1429*
ABL1_6	E255V		APC_311	R1114*
ABL1_7	D276G		APC_312	R1450*
ABL1_8	F311L		APC_313	R876*
ABL1_9	T315I		APC_314	S1465fs*3
AKT1_1	rs11555435		APC_315	T1661fs*9
AKT1_2	rs11555431		BRAF_316	D587A
AKT1_3	rs11555432		BRAF_317	D587E
AKT1_4	rs12881616		BRAF_318	D594E
AKT1_5	rs11555433		BRAF_319	E586K
AKT1_6	rs11555436		BRAF_320	F595S
AKT1_7	rs34409589		BRAF_321	G469A
AKT2_1	S302G		BRAF_322	G466V
AKT2_2	R371H		BRAF_323	I592M
BRAF_1	G464R		BRAF_324	I592V
BRAF_10	F595L		BRAF_325	K601del
BRAF_11	G596R		BRAF_326	N581S
BRAF_12-13 (CA)	L597S		BRAF_327	R444W
BRAF_12-13 (CA)	L597R		BRAF_328	S605F
BRAF_12-13 (CA)	L597Q		BRAF_329	S605N
BRAF_12-13 (CA)	L597V		BRAF_330	T599_V600insTT
BRAF_14	T599I		BRAF_331	V471F
BRAF_15-16 (CA)	V600E		BRAF_332	V600A
BRAF_15-16 (CA)	V600K		BRAF_333	V600D
BRAF_15-16 (CA)	V600R		BRAF_334	V600M
BRAF_15-16 (CA)	V600L		CDKN2A_335	E61*
BRAF_17	K601N		CDKN2A_336	E69*
BRAF_18	K601E		CDKN2A_337	E88*
BRAF_2	G464V		CDKN2A_338	R58*
BRAF_2	G464E		CDKN2A_339	D84Y
BRAF_4	G466R		CDKN2A_340	R80*
BRAF_5	F468C		CDKN2A_341	H83Y
BRAF_6-8 (CA)	G469S		CSF1R_342	L301S

BRAF_6-8 (CA)	G469E		CSF1R_342	L301*
BRAF_6-8 (CA)	G469A		CSF1R_343	Y969C
BRAF_6-8 (CA)	G469V		CSF1R_343	Y969F
BRAF_6-8 (CA)	G469R		CSF1R_344	Y969H
BRAF_6-8 (CA)	G469R		CSF1R_345	Y969*
BRAF_9	D594V|G		CTNNB1_346	A13T
CDK_1	R24C		CTNNB1_347	A21T
CDK_2	R24H		CTNNB1_348	D32G
EGFR_1	R108K		CTNNB1_348	D32A
EGFR_10	S768I		CTNNB1_348	D32V
EGFR_11	V769_D770insASV		CTNNB1_349	G34R
EGFR_12	V769_D770insCV		CTNNB1_350	G34V
EGFR_13	D770_N771>AGG		CTNNB1_350	G34E
EGFR_13	V769_D770insASV		CTNNB1_351	S33C
EGFR_13	V769_D770insASV		CTNNB1_351	S33F
EGFR_14	D770_N771insG		CTNNB1_351	S33Y
EGFR_15	N771_P772>SVDNR		CTNNB1_352	S37A
EGFR_16	P772_H773insV		CTNNB1_352	S37P
EGFR_17	H773>NPY		CTNNB1_353	S37F
EGFR_18	H773_V774insNPH		CTNNB1_353	S37C
EGFR_18	H773_V774insPH		CTNNB1_353	S37Y
EGFR_18	H773_V774insH		CTNNB1_354	S45F
EGFR_19	V774_C775insHV		CTNNB1_354	S45Y
EGFR_2	T263P		CTNNB1_354	S45C
EGFR_20	T790M		CTNNB1_355	S45P
EGFR_21	L858R		CTNNB1_355	S45A
EGFR_22	L861Q		CTNNB1_356	T41A
EGFR_3	A289V		CTNNB1_356	T41S
EGFR_4	G598V		CTNNB1_356	T41P
EGFR_5	E709K		CTNNB1_357	T41I
EGFR_5	E709H		CTNNB1_358	T41I
EGFR_6	E709A		CTNNB1_359	V22_G38del
EGFR_6	E709G		CTNNB1_360	V22A
EGFR_6	E709V		CTNNB1_361	W25_D32del
EGFR_7	G719S		EGFR_362	A750P
EGFR_7	G719C		EGFR_363	D761N
EGFR_8	G719A		EGFR_364	D761Y
EGFR_9	M766_A767insAI		EGFR_365	E734K
EGFR_M (CA)	E746_T751del		EGFR_366	E746_S752>A
EGFR_M (CA)	E746_A750del		EGFR_367	E746_S752>D
EGFR_M (CA)	E746_A750del		EGFR_368	E746_T751>A
EGFR_M (CA)	E746_T751del		EGFR_369	E746_T751del
EGFR_M (CA)	E746_A750del		EGFR_370	E746K
EGFR_M (CA)	E746_T751del, S752D		EGFR_371	G719D

EGFR_M (CA)	L747_E749del		EGFR_372	G735S
EGFR_M (CA)	L747_T750del		EGFR_373	G810D
EGFR_M (CA)	L747_S752del		EGFR_374	G810S
EGFR_M (CA)	L747_T751del		EGFR_375	H773R
EGFR_M (CA)	L747_S752del, P753S		EGFR_376	K745R
EGFR_M (CA)	L747_T751del		EGFR_377	L730F
EGFR_M (CA)	A750P		EGFR_378	L747_P753>Q
EGFR_M (CA)	T751A		EGFR_379	L747_P753>S
EGFR_M (CA)	T751P		EGFR_380	L747_R748>FP
EGFR_M (CA)	T751I		EGFR_381	L747_T751>P
EGFR_M (CA)	S752I/F		EGFR_382	L747_T751>S
EGFR_M (CA)	L747_Q ins		EGFR_381	L747_T751>P
EGFR_M (CA)	E746_T751del, I ins		EGFR_383	L858M
EGFR_M (CA)	E746_A750del, T751A		EGFR_384	L858R
EGFR_M (CA)	E746_T751del, V ins		EGFR_385	N771_P772>SVDNR
EGFR_M (CA)	E746_A750del, V ins		EGFR_386	P733L
EGFR_M (CA)	L747_E749del, A750P		EGFR_387	P753S
EGFR_M (CA)	L747_T750del, P ins		EGFR_388	S752_I759del
EGFR_M (CA)	L747_S752del, Q ins		EGFR_389	S752Y
EGFR_M (CA)	T751 Undefined		EGFR_390	T790M
EGFR_M13R	S752_I759del		EGFR_391	V742A
ERBB2_1	L755P		EGFR_392	W731*
ERBB2_2	G776S		ERBB2_393	D769H
ERBB2_2	G776LC		ERBB2_394	V777L
ERBB2_3	G776VC		FLT3_395	D835del
ERBB2_6	A775_G776insYVMA		FLT3_396	D835E
ERBB2_7	P780_Y781insGSP		FLT3_396	D835E
ERBB2_8	P780_Y781insGSP		FLT3_397	I836M
ERBB2_9	S779_P780insVGS		HRAS_398	G12R
FGFR1_1	S125L		HRAS_398	G12C
FGFR1_2	P252T		JAK3_399	A572V
FGFR3_3	G370C		JAK3_400	P132T
FGFR3_4	Y373C		JAK3_401	V722I
FGFR3_5	A391E		KIT_402	N822K
FGFR3_6	K650Q		KIT_403	T670I
FGFR3_6	K650E		KIT_404	V654A
FGFR3_7	K650T		KRAS_405	A146T
FGFR3_7	K650M		KRAS_406	G13A
FLT3_1	I836del		KRAS_407	G13R
FLT3_2	Control1		KRAS_408	L19F
FLT3_2	Control2		KRAS_409	Q22K
FLT3_2	Control3		MET_410	H1112R
FLT3_3	Control1		MET_411	H1112Y

FLT3_3	Control2		MET_412	M1268T
FLT3_3	Control3		MET_413	T1010I
FLT3_4	D835H		MET_414	Y1248C
FLT3_4	D835Y		MET_415	Y1248H
HRAS_2	G12V		MLH1_416	V384D
HRAS_2	G12D		MYC_417	A59V
HRAS_3	G13C		MYC_418	N101T
HRAS_3	G13R		MYC_419	P260A
HRAS_3	G13S		MYC_420	P57S
HRAS_5	Q61H		MYC_421	S77F
HRAS_5	Q61H		MYC_422	T73I
HRAS_6	Q61L		PDGFRA_423	D842_D846>E
HRAS_6	Q61R		PDGFRA_424	D842_D846>G
HRAS_6	Q61P		PDGFRA_425	D842_D846>N
HRAS_7	Q61K		PDGFRA_426	D842_M844del
JAK2_1	V617F		PDGFRA_427	D842_S847>EA
KIT_1	D52N		PDGFRA_428	D842I
KIT_10	V559del		PDGFRA_429	D842V
KIT_11	V559_V560del		PDGFRA_430	D842Y
KIT_12	V560del		PDGFRA_431	H845_N848>P
KIT_13.1-13.2 (CA)	P551_V555del		PDGFRA_432	R841_D842del
KIT_14.2	Y553_Q556del		PDGFRA_433	S566_E571>R
KIT_15	Y570_L576del		PDGFRA_434	S566_E571>R
KIT_16	E561K		PIK3CA_435	E545A
KIT_17	L576P		PIK3CA_435	E545G
KIT_18	P585P		PIK3CA_436	N1068fs*4
KIT_19	D579del		PIK3CA_437	Y1021C
KIT_2	Y503_F504insAY		PTEN_438	K267fs*9
KIT_20	K642E		PTEN_439	K6fs*4
KIT_21	D816V		PTEN_440	N323fs*2
KIT_22	D816H		PTEN_441	N323fs*21
KIT_22	D816Y		PTEN_442	P248fs*5
KIT_23	V825A		PTEN_443	R130Q
KIT_24	E839K		PTEN_443	R130fs*4
KIT_25	M552L		PTEN_444	R130G
KIT_26	Y568D		PTEN_444	R130*
KIT_27	F584S		PTEN_445	R173C
KIT_3	W557R		PTEN_446	R173H
KIT_3	W557R		PTEN_447	R233*
KIT_3	W557G		PTEN_448	R335*
KIT_4	V559D		PTEN_449	V317fs*3
KIT_4	V559A		RB1_450	C706F
KIT_4	V559G		RB1_451	E137*
KIT_5	V559I		RB1_452	E748*

KIT_6	V560D		RB1_453	L199*
KIT_6	V560G		RB1_454	L660fs*2
KIT_7	K550_K558del		RB1_455	R320*
KIT_8	K558_V560del		RB1_456	R358*
KIT_9	K558_E562del		RB1_457	R455*
KRAS_1-2 (CA)	G12V		RB1_458	R552*
KRAS_1-2 (CA)	G12A		RB1_459	R556*
KRAS_1-2 (CA)	G12D		RB1_460	R579*
KRAS_1-2 (CA)	G12C		RET_461	A883F
KRAS_1-2 (CA)	G12S		RET_462	C634R
KRAS_1-2 (CA)	G12R		RET_463	C634W
KRAS_1-2 (CA)	G12F		RET_464	C634Y
KRAS_4	G13V		RET_465	D631_L633>E
KRAS_4	G13D		RET_466	D631G
KRAS_5	A59T		RET_467	D898_E901del
KRAS_6	Q61E		RET_468	E632_A640>VRP
KRAS_6	Q61K		RET_469	E632_L633del
KRAS_7	Q61L		RET_470	E632_L633>V
KRAS_7	Q61R		RET_471	E768D
KRAS_7	Q61P		RET_472	F612_C620del
KRAS_8	Q61H		RET_473	F612_C620del
KRAS_8	Q61H		RET_474	M918T
MET_1	R970C		SRC_475	Q531*
MET_2	T992I		STK11_476	D194N
MET_3	Y1230C		STK11_477	D194V
MET_4	Y1235D		STK11_478	E199K
MET_5	M1250T		STK11_478	E199*
NRAS_1	G12V		STK11_479	E57fs*7
NRAS_1	G12A		STK11_480	F264fs*22
NRAS_1	G12D		STK11_481	G196V
NRAS_2	G12C		STK11_482	P281fs*6
NRAS_2	G12R		STK11_482	P281L
NRAS_2	G12S		STK11_483	P281fs*6
NRAS_3	G13V		STK11_483	P281L
NRAS_3	G13A		STK11_484	Q170*
NRAS_3	G13D		STK11_485	Q37*
NRAS_4	G13C		STK11_486	W332*
NRAS_4	G13R		TP53_487	G245S
NRAS_4	G13S		TP53_488	R175H
NRAS_5	A18T		TP53_489	R248Q
NRAS_6	Q61L		TP53_490	R248W
NRAS_6	Q61R		TP53_491	R273C
NRAS_6	Q61P		TP53_492	R273H
NRAS_7	Q61H		TP53_493	R306*

NRAS_8	Q61E		VHL_494	F148fs*11
NRAS_8	Q61K		VHL_495	L158Q
PDGFRA_1	V561D		VHL_496	L85P
PDGFRA_10	I843_S847>T		VHL_497	L89H
PDGFRA_11	D842V		VHL_498	P81S
PDGFRA_2	T674I		VHL_499	R161*
PDGFRA_3	F808L		VHL_500	R167W
PDGFRA_4	D846Y			
PDGFRA_5	N870S			
PDGFRA_6	D1071N			
PDGFRA_7	D842_H845del			
PDGFRA_8	I843_D846del			
PDGFRA_9	S566_E571>K			
PIK3CA_1	R88Q			
PIK3CA_10	H1047Y			
PIK3CA_12	R38H			
PIK3CA_13	C901F			
PIK3CA_14	M1043I			
PIK3CA_14	M1043I			
PIK3CA_2	N345K			
PIK3CA_3	C420R			
PIK3CA_4	P539R			
PIK3CA_5	E542K			
PIK3CA_6	E545K			
PIK3CA_7	Q546K			
PIK3CA_8	H701P			
PIK3CA_9	H1047R			
PIK3CA_9	H1047L			
RET_1	C634R			
RET_2	C634W			
RET_3	C634Y			
RET_4	E632_L633del			
RET_5	M918T			
RET_6	A664D			

Appendix Table A.2: List of mutations of the displayed genes covered on the Oncocarta v1 and v3 panels (assay TP53_488/R175H did not work in my set of samples). Table kindly provided by Sequenom Inc. (Sequenom, US)

Gene List	nList	nRep	fRep	nOVERLAP	ODDS RATIO	*P*-value	Adj.*P*-value	Genes
ACEVEDO_METHYLATED_IN_LIVER_CANCER_DN	927	778	0.8392665	78	2.3886656	1.60E-10	1.03E-06	SDK1; TNFRSF8; C20orf151; CNTNAP4; CSRP3; KCNAB2; ITGA8; TREML2; DPP6; KIRREL; OR1C1; PTP4A3; OR6V1; CACNB2; RSU1; RNF17; CA1; TAC1; FAM105A; NFASC; SV2B; FANK1; NAV2; MYO10; CLNK; VPS8; ZAP70; ZNF660; COL24A1; C14orf53; SLC26A4; SORCS3; LIMCH1; CHRM3; DEFB108B; RPTOR; OR2B11; POTEE; TMEM132D; OR10K1; PTPRN2; SPATA18; CASQ2; PXDNL; ST6GALNAC3; MAGEH1; STK32B; RUNX1T1; ESD; A2BP1; JAKMIP1; DOCK3; RGS6; ZNF239; SLCO5A1; IL2RA; GPR139; OLFML2B; CLSTN2; MAL2; GSG1; KIAA1026; DENND3; MYT1L; SPAG11B; ANK2; ADARB2; CNR1; NLRP3; VIPR2; KIF26B; RADIL; FBXW7; ODZ2; SLC6A8; CDH13; ZNF804A; CCDC141;
AACTTT_UNKNOWN	1475	1429	0.9688136	114	1.875692	1.07E-08	3.43E-05	IGF1; BTBD3; RORA; SSBP3; UST; KIFAP3 ESRRG; SLC4A10; ARID1B; CREB5; SH3BP4; TOM1; NOL4; CACNA1H; HCN1; IQGAP2; GPR65; LSAMP; OTP; NEUROD4; RASSF2; NRXN1; HOXC11; CACNB2; CACNA1D; GPR133; CA3; CA10; TAC1; ROR1; KCNQ4; HCRTR2; NRG1; MSX2; MRVI1; MYBPC1; BCOR; LDB2; SEMA3A; SYNE1; NOX3; LUM; GRIN3A; PCDH10; ZNF219; SNTG1; PRKAG2; TREM1; SORCS3; PDE4D; CHRDL1; KIRREL3; RBMS3; REPS1; PTPRG; SHANK2; TEAD1; CHST8; SFRS8; TECTA; COL19A1; STC1; TRPS1; CAMK2D; CASQ2; RPS6KA2; CNTNAP2; STK32B; HPSE2; CBFA2T3; CTNNA3; ESR1; A2BP1; DOCK3; SESN2; RBPMS; CTNND2; CTNND1; CDYL; LRRTM4; SEMA5A; TRERF1; HOXA2; KCNMA1; DLG2; GFPT2; JAZF1; KCNJ3; SMOC1; NXF1; ELAVL4; CLSTN2; NRK; C10orf11; MRPS31; MAPK10; MAP2K6; NELL1; GABRB1; KCNIP4; MOS; MRGPRF; COL4A6; C5orf4; AKT3; CPNE5; ADCY2; VIPR2; LRRN1; EBF2; FBXW7; SERPINI1; CDH13; CDK5

BENPORATH_SUZ12_TARGETS (M9898) PMID: 18443585	**1037**	**998**	**0.9623915**	**82**	**1.9049574**	**4.53E-07**	**0.00096895**	PLEC1; ST3GAL1; DPP6; NOL4; KY; CRMP1; OTP; OCA2; PKNOX2; HOXC11; CACNB4; CACNA1D; CA3; CA10; LRRC3B; SIDT1; NRG1; NRCAM; NPTX1; SV2B; EIF4E3; APCDD1L; BATF3; NAV2; NCAM1; GRIN3A; COL24A1; SLC26A4; SORCS3; DKK2; XYLT1; CYFIP1; KIRREL3; PHLDB1; NT5C1A; SLITRK2; LCORL; CTPS2; PRKG1; CHST8; ABCC6; ANKRD11; CSMD1; SPATA18; FGF3; IGSF21; GPC5; RASGRF2; STK32B; HPSE2; ASTN1; MGAT5B; CTNND2; DLC1; KCNMA1; TLX2; USP3; SLCO5A1; KCNJ3; PDE4DIP; EPB41L4A; CFTR; KCNAB1; MT1X CDH23; OLFML2B; CLSTN2; NRK; NELL1; KCNIP4; ZNF365; ST3GAL6; GPR124; COL4A6; DOK6; ADARB2; SLIT1; VIPR2; GPR137B; FAM163A; CDH13; SLC6A3
BENPORATH_ES_WITH_H3K27ME3	**1117**	**1077**	**0.9641898**	**86**	**1.8478054**	**7.77E-07**	**0.00124641**	ADAMTS17; LHX3; PLEC1; DTNB; DPP6; NOL4; KY; ENTPD2; CRMP1; OTP; OCA2; PKNOX2; NRXN1; HOXC11; TFAP2D; CACNB4; CACNA1D; CA10; SIDT1; TAC1; NRG1; MSX2; NPTX1; MYB; SV2B; EIF4E3; BATF3; NAV2; NCAM1; MICB; SYNE1; GRIN3A; COL24A1; PRKAG2; SLC26A4; SORCS3; DKK2; XYLT1; CHRDL1; KIRREL3; PHLDB1; NT5C1A; NRIP3; PSD2; PRKG1; CHST8; PTPRN2; SNTB1; CSMD1; MAST4; CAMK2B; FGF3; IGSF21; GPC5; RPS6KA2; PARD3B; STK32B; HPSE2; PDE10A; ASTN1; OSBP2; ESR1; CTNND2; HOXA2; KCNMA1; TLX2; KCNQ1; SLCO5A1; KCNJ3; PDE4DIP; SLC26A5; EPB41L4A; KCNAB1; CDH23; OLFML2B; CLSTN2; MXRA7; NELL1; KCNIP4; GPR124; COL4A6; DOK6; ADARB2; SLIT1 FAM163A; SLC6A3
ACEVEDO_LIVER_CANCER_WITH_H3K27ME3_UP	**295**	**275**	**0.9322034**	**30**	**2.5446155**	**1.53E-05**	**0.01967642**	C7orf31; SDK1; KIAA1217; COL19A1; MAGEB1; C9; CRISP2; MAST4; OR2H1; TSGA13; MAGEH1; CREB5; CTNNA3; CTNNA3; PRTG; TMEM71; SCAND3; SELP; AKT3; CSRNP3; LIMCH1; TMEM144; RBPMS; CHRM3; OCA2; CACNB2; FMN2; DLC1 ZBTB46; HSPB3; CCDC141
module_100	**529**	**513**	**0.9697543**	**46**	**2.0597127**	**1.95E-05**	**0.01970417**	BTBD3; DTNB; TRPM2; PLEKHB1; PLEKHB1; CNTNAP2; RUNX1T1; GALR3; DPP6; CRMP1; SORBS2; PTP4A3; C16orf45; SCHIP1; RASSF2; NRXN1; DVL1; SLC23A2; GFPT2; SYT1; ROR1; MAGI2; NPTX1; ABCA3; ELAVL4; SV2B; LDB2; RSBN1; MBP; MAPK10; MYO10; TACC2; NCAM1; SYNE1; POLR2J; GNG7; FBXL7; SLIT1; ATP2B2; RBMS3; ZBTB20; SERPINI1; LAMA2; FZR1; PTPRN2; CDH18; CDH18; SLC6A3

module_137	**531**	**515**	**0.9698682**	**46**	**2.0507021**	**2.15E-05**	**0.01970417**	BTBD3; DTNB; TRPM2; PLEKHB1; CNTNAP2; RUNX1T1; GALR3; DPP6; CRMP1; SORBS2 PTP4A3; C16orf45; C16orf45; SCHIP1; RASSF2; NRXN1; DVL1; SLC23A2; GFPT2; SYT1; ROR1; MAGI2; NPTX1; ABCA3; ELAVL4; SV2B; LDB2; RSBN1; MBP; MAPK10; MYO10; TACC2; NCAM1; SYNE1; POLR2J; GNG7; TTLL1; FBXL7; SLIT1; ATP2B2; RBMS3; ZBTB20; SERPINI1; LAMA2; FZR1; PTPRN2; SLC6A3
module_66	**536**	**520**	**0.9701493**	**46**	**2.0285086**	**2.74E-05**	**0.02070088**	BTBD3; DTNB; TRPM2; PLEKHB1; CNTNAP2; RUNX1T1; GALR3; DPP6; CRMP1; SORBS2; PTP4A3; C16orf45; SCHIP1; RASSF2; NRXN1; DVL1; SLC23A2; GFPT2 SYT1; ROR1; MAGI2; NPTX1; ABCA3; ELAVL4; SV2B; LDB2; RSBN1; MBP; MAPK10; MYO10; TACC2; NCAM1; SYNE1; POLR2J; GNG7 FBXL7; SLIT1; ATP2B2; RBMS3; ZBTB20; SERPINI1; LAMA2; FZR1; PTPRN2; CDH18; SLC6A3
TTANTCA_UNKNOWN	**718**	**680**	**0.9470752**	**56**	**1.882154**	**2.98E-05**	**0.02070088**	ATP10D; IGF1; EMP1; TRPS1; CAMK2D; CASQ2; ESRRG; ATRNL1; CREB5; C7orf33; HPSE2; WHSC1L1; OTP; A2BP1; CNGB3; SESN2; CTNND2; HOXC11; TRERF1; DLC1; HOXA2; NR1H4; USP3; CORO1C; DLG2; JAZF1; TAC1; ROR1; HGF; KRT14; FLT1; ABL2; SLCO1B1; MAL2; GSG1; C10orf11; MYO10; SEMA3A; GABRB1; KCNIP4; COL4A6; MMP3; PDE4D; CPNE5; KIRREL3; RBMS3; BCKDHA; LRRN1; PTPRG; SHANK2; EBF2; FBXW7; SLC9A9; HRASLS; TECTA
BENPORATH_EED_TARGETS (M7617) PMID: 1844358	**1062**	**1015**	**0.9557439**	**76**	**1.7076143**	**3.34E-05**	**0.02070088**	PLEC1; NOL4; KY; OTP; OCA2; PKNOX2; HOXC11; TFAP2D; CACNA1D; CA3; CA10; SIDT1; TAC1; NRG1; NRCAM; MAGI2; MSX2; NPTX1; MYB; SV2B; BATF3; NAV2; NCAM1; MICB; GRIN3A; PCDH10; COL24A1; SLC26A4; SORCS3; DKK2; KIRREL3; NT5C1A; NRIP3; SORCS2; LCORL; PSD2; PRKG1; CHST8; COL19A1; CSMD1; CAMK2B; FGF3; IGSF21; GPC5; STK32B; HPSE2 PDE10A; ASTN1; HCCA2; HTR1E; RGS6; CTNND2; BACH2; HOXA2; KCNMA1; TLX2; KCNQ1; SLCO5A1; PDE4DIP; EPB41L4A; CFTR; KCNAB1; MT1X; CDH23; OLFML2B; CLSTN2; NELL1; ZNF365; DENND3; COL4A6; DOK6 ADARB2 SLIT1 VIPR2 FAM163A SLC6A3

module_11	**526**	**510**	**0.9695817**	**45**	**2.0214817**	**3.55E-05**	**0.02070088**	BTBD3; DTNB; TRPM2; PLEKHB1; CNTNAP2; RUNX1T1; GALR3; DPP6; CRMP1; SORBS2; PTP4A3; C16orf45; SCHIP1; RASSF2; NRXN1; DVL1 SLC23A2; GFPT2; SYT1; MAGI2; NPTX1; ABCA3; ELAVL4; SV2B; LDB2; RSBN1; MBP; MAPK10; MYO10; TACC2; NCAM1; SYNE1; POLR2J; GNG7; LPHN1; FBXL7; ADARB1; SLIT1; ATP2B2; RBMS3; ZBTB20; BCKDHA; SERPINI1; FZR1; PTPRN2
ATGTACA,MIR-493	**266**	**262**	**0.9849624**	**28**	**2.4825011**	**4.23E-05**	**0.02106791**	BCOR; TRPS1; HIC2; CTDSPL; SV2B; PREX1; RSBN1; ESRRG; NAV2; ARID1B; TMTC1; KCNIP4; CNOT6; SYNE1; DIP2C; SORCS3; DKK2; NRXN3; PDE4D; C16orf45; CTNND2; CTNND1; ADAM10; PSD2; KCNMA1; SPIRE2; DLG2; CDH13
CAGGTG_V$E12_Q6	**1850**	**1801**	**0.9735135**	**120**	**1.5230387**	**4.27E-05**	**0.02106791**	LY6G5C; ADAMTS17; BTBD3; SSBP3; UST; DTNB; ESRRG; SLC4A10; KCNAB2; TREML2; ENTPD2; CRMP1; CACNA1H; PTP4A3; CNTN5; NRXN3; RASSF2; NRXN1; SYT11; DVL1; CORO1C; SYT1; TAC1; MFAP2; NAV2; TACC2; NCAM1; ITPR2; PRKAG2; XYLT1; NRIP3; PCDH15; SHANK2; COX6A2; PTPRB; TECTA; COL19A1; STC1; CAMK2B; PGM3; PLEKHB1; VWF; CASQ1; GRIA1; CBFA2T3; PFKFB2; MGAT5B; GCNT3; DOCK3; MICAL2; RGS6; TRERF1; KCNMA1; N4BP1; KCNQ1; DLG2; JAZF1; SMOC1; EPB41L4A; ABL2; ABR; MAPK10; KCNIP4; MRGPRF; ATM; LPHN1; ADAM10; RAPGEF3; SERPINI1; PIGR; WWOX; SDK1; GALR3; NOL4; IQGAP2; YPEL1; BBX; CACNB2; CACNA1D; HIVEP3; CA10; ROR1; KCNQ4; HCRTR2 NRG1; HGF; NRCAM; NPTX1; SV2B; SEMA3A; SYNE1; LUM; PDE1C; KIRREL3; CTPS2; CHRNA7 CHST8; TRPS1; ALDH3B1; RASGRF2; MAPK8IP1; ESR1; KIF13B; FA2H; DHRS9; CTNND1; TLX2; DOCK5; NR1H4; ELAVL4; GALNT2; NRK; GALR1; GABRB1; GPR124; ANK2; EPB41L5; UACA; LRRN1; CDH13
TGTTTGY_V$HNF3_Q6	**576**	**552**	**0.9583333**	**47**	**1.9443051**	**5.79E-05**	**0.02652603**	CD19; TRPS1; CNTNAP4; GPLD1; ARID1B; CREB5; CNTNAP2; HPSE2; SCG3; CRY2; KIRREL; WHSC1L1; IQGAP2; A2BP1; ESR1; ZDHHC14; BBX; NRXN3; HOXC11; UBASH3A; DLG2; CA3; BANP; HCRTR2; JAM3; SMOC1; NRG1; ELAVL4; MAPK10; SEMA3A; GABRB1; NCAM1; KCNIP4; RBP3; SORCS3; AKT3; PDE3B; ROBO4; PDE4D; ADARB2; CHRDL1; KIRREL3; PTPRG; EBF2; TEAD1; TERT; TNMD

RNTCANNRNNYNATTW_UNKNOWN	**48**	**48**	**1**	**10**	**5.40694**	**6.20E-05**	**0.02652603**	LRRN1; CASQ2; ELAVL4; NOL4; NRXN1; CHRDL1; MYB; CDK5; TAC1; USP3
VERHAAK_AML_WITH_NPM1_MUTATED_DN	**246**	**241**	**0.9796748**	**26**	**2.5057869**	**6.70E-05**	**0.02685118**	HGF; FAM105A; EMP1; ROBO1; CD19; CD7; MXRA7 MECOM; RPS6KA2; SH3BP4; STK32B; NCAM1; RUNX1T1; LTF; BAALC; ZAP70; BANK1; YPEL1; PDE3B; XYLT1; GNG7; RBPMS; LHFPL2; THSD7A; NT5E; SIDT1
TTGTTT_V$FOXO4_01	**1549**	**1489**	**0.9612653**	**101**	**1.542327**	**9.36E-05**	**0.03532313**	BTBD3; RORA; ROBO1; SSBP3; RASGRP3; ESRRG; CRY2; ENTPD1; CRMP1; PKNOX2; RASSF2; CORO1C; NCAM1; PRKAG2; TEAD1; TNMD; STC1; ESM1; GPLD1; VWF; CASQ2; GRIA1; ADORA1; HPSE2; SCG3; CTNNA3; A2BP1; CNGB3; KRTAP9-2; TRERF1; DLC1; KCNJ16; DLG2; SMOC1; KRT14; MAGEB1; FLT1; CLSTN2; MAPK10; MAP2K6; KCNIP4; MRGPRF; SLC12A8; SLC4A4; COL4A6; AKT3; LPHN1; ADARB2; CNR1; TNXB; SCML4; SDK1; IGF1; C20orf151; CNTNAP4; ARID1B; CREB5; KIRREL; OTP; ZDHHC14; YPEL1; SORT1; HCRTR2; NRG1; MSX2; NPTX1; BCOR; LDB2; SEMA3A; MRPL11; NOX3; ZNF219; RBP3; PDE3B; PDE4D; CHRDL1; SLITRK2; PRKG2; CHST8; ATP10D; EMP1; TRPS1; ZNRF2; WHSC1L1; GABARAPL1; SESN2; TLN2; RCL1; CTNND1; LRRTM4; HOXA2; C10orf81; USP3; MGAT4A; CFTR; CDH19; ELAVL4; NRK; TSGA13; ROBO4; EBF2
V$CEBPDELTA_Q6	**185**	**179**	**0.9675676**	**21**	**2.7470556**	**0.0001005**	**0.03582371**	ANKRD11; OTP; SDK1; A2BP1; ITGA5; RORA; STC1; DOCK3; RAB3IP; VIPR2; GALK2; SYT11; TLX2; SLC4A10; ARID1B; SEMA3A; FBXW7; CREB5; CNTNAP2; C7orf33; TAC1
V$OCT1_02	**180**	**167**	**0.9277778**	**20**	**2.8105184**	**0.0001097**	**0.0370397**	JAM3; OTP; STC1; NRXN3; TRPS1; PDE4D; ELAVL4; HIC2; PKNOX2; HOXC11; CHRDL1; KIRREL3; NRK ESRRG; DNAH9; CACNA1D; SEMA3A; DLG2; KCNIP4; ENTPD1
V$MEIS1_01	**175**	**168**	**0.96**	**20**	**2.7913797**	**0.0001192**	**0.03822413**	SMOC1; NRG1; A2BP1; ANK2; NEUROD4; ABL2; UST; RGS6; ELAVL4; COL11A1; KIRREL3; PHLDB1; ESRRG; FBXW7; CREB5; SLC23A2; NCAM1; NOL4; MRGPRF; HCRTR2
ACEVEDO_LIVER_CANCER_WITH_H3K9ME3_UP	**141**	**122**	**0.8652482**	**16**	**3.1109729**	**0.0001824**	**0.05570731**	FAM105A; ANK2; NRXN3; LIMCH1; KRT20; ENOX1 SULT1C2; TRPM2; CACNB2; LRRN1; FAT3; MAGEH1; MYO16; CLNK; TMEM108; CA1

SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP	**346**	**331**	**0.9566474**	**31**	**2.1434204**	**0.0002001**	**0.05831783**	JAM3; ATP10D; EMP1; ROBO1; ESM1; OLFML2B; CRISPLD2; MFAP2; COL11A1; LDB2; CLEC4A; MXRA7; HEG1; PDE10A; ENTPD1; LUM; GPR65 COL6A3; COL8A2; ST3GAL6; MMP3; ANKRD6; AKT3; MICAL2; XYLT1; HLA-DRB6; RASSF2; FAP; THSD7A; GFPT2; ROR1
TGATTTRY_V$GFI1_01	**243**	**230**	**0.9465021**	**24**	**2.4098435**	**0.0002096**	**0.05844938**	STC1; MYBPC1; ELAVL4; CLSTN2; LDB2; CASQ2; C10orf11; ESRRG; MAPK10; RASGRF2; CREB5; CTNNA2; LSAMP; ESR1; NRXN3; CTNND1; NRXN1; CHRDL1; TRERF1; HOXA2; TEAD1; JAZF1; HIVEP3; CDH13
SMID_BREAST_CANCER_BASAL_DN	**692**	**667**	**0.9638728**	**52**	**1.7658389**	**0.0002251**	**0.06014118**	IGF1; EFCAB2; MAST4; STS; CAMK2B; ICA1; PCLO; ESRRG; ATRNL1; ZNF385D; CNTNAP2; STK32B CBFA2T3; FAM174B; CRY2; CACNA1H; IQGAP2; ESR1; KIF13B; NME5; C16orf45; MTL5; KCNMA1; CACNA1D; KCNJ3; SYT1; PSCA; SIDT1; DARC; HGF; CYBRD1; MSX2; MYB; MCF2L; KCNAB1; ABCA3; CLSTN2; KDM4B; GGT1; ZNF91; ZSCAN18; GOLSYN; COL4A6; C5orf4; DNALI1; FBXL7; CNR1; CNR1; RAPGEF3; LMF1 SERPINI1; PTPRN2; WWOX
DELYS_THYROID_CANCER_DN	**214**	**205**	**0.9579439**	**22**	**2.4841333**	**0.0002552**	**0.06546773**	CYBRD1; KCNAB1; MT1X; LDB2; CASQ2; PCLO; ESRRG; NCAM1; DPP6; LTF; PDE10A; SYNE1; IQGAP2 SLC4A4; C5orf4; SLC26A4 CHRDL1; DLG2; TNXB; LAMA2; WWOX; FXYD6
BENPORATH_PRC2_TARGETS (M8448) PMID: 18443585	**652**	**625**	**0.958589**	**49**	**1.7742791**	**0.0002977**	**0.07344045**	CSMD1; PLEC1; FGF3; IGSF21; GPC5; STK32B; HPSE2; NOL4; KY; ASTN1; OTP; CTNND2; OCA2; PKNOX2; HOXC11; KCNMA1; TLX2; CACNA1D; SLCO5A1; CA10; PDE4DIP; SIDT1; NRG1; EPB41L4A; NPTX1; KCNAB1; CDH23; OLFML2B; CLSTN2; SV2B; BATF3; NAV2; NELL1; NCAM1; GRIN3A; COL24A1; COL4A6; SLC26A4; SORCS3; DKK2; DOK6; ADARB2; SLIT1; KIRREL3; NT5C1A; PRKG1; CHST8; FAM163A; SLC6A3
MORF_CAMK4	**283**	**265**	**0.9363958**	**26**	**2.2512527**	**0.000311**	**0.07386748**	COL19A1; PGM3; PLEKHB1; SOAT2; GPLD1; ZNF157; SUPT3H; TACC2; NOL4; CDH8; SLC26A4; HTR1E; PDE4D; HOXC11; ATP4B; ATP2B2; PHLDB1; RBMS3; PAPPA2; CDR1; COX6A2; CYP2C19; PTPRB; IPO9; FBXL4; HCRTR2
chr6q14	**44**	**30**	**0.6818182**	**7**	**6.236973**	**0.0003799**	**0.08701329**	PHIP; NT5E; CNR1; ANKRD6; HTR1E; BACH2; PGM3

MORF_MDM2	**271**	**255**	**0.9409594**	**25**	**2.2479278**	**0.0004069**	**0.08998674**	COL19A1; FLT1; MFN1; CSRP3; PGM3 PLEKHB1; SOAT2; GPLD1; ZNF157; SUPT3H; MAP2K6 TACC2; PDE10A; HTR1E; HOXC11; ATP4B; ATP2B2; PHLDB1; RBMS3; CDR1; COX6A2; PTPRB; IPO9; FBXL4; HCRTR2
module_117	**706**	**673**	**0.9532578**	**51**	**1.7097179**	**0.0004991**	**0.10419259**	APOB; TNFRSF8; IGF1; TACR3; MPPED2; VNN1 RASGRP3; CD247; AFM; ICA1; ESRRG; CASP5; GPC5 CREB5; ADORA1; KCNAB2; GYG2; NOL4; GPR65; PHACTR2; CNTN5; KIAA0087; IGJ; ZNF239; SULT1C2; CACNB2; CACNB4; HOXA2; NR1H4; IL2RA; CA1; MAGEB1; SV2B; KIR3DL1; NELL1; LGALS2; LTF P2RY6; ABCB11; CNR1; NLRP3; RAPGEF3; MSH4; VIPR2; ALOX5AP; CHRNA7; CDR1; PRKG1; TERT; APOH; SLC6A3

Appendix Table A.3: Gene set enrichment analysis on consistent hyper-MVPs. Overlaps ('nOVERLAP') between the detected gene set and gene sets in the Molecular Signatures Database v3.0 (MSigDB; Broad Institute of MIT and Harvard, US) were computed as described previously (Subramanian, Tamayo, et al.; PNAS; 2005; 102; 15545-15550). The top 30 hits are illustrated.

Gene List	nList	nRep	fRep	nOVERLAP	ODDS RATIO	*P*-value	Adj.*P*-value	Genes
SLEBOS_HEAD_AND_NECK_CANCER_WITH_HPV_UP (M14132) PMID: 16467079	79	77	0.9746835	6	47.736128	1.12E-08	7.18E-05	HLTF; STMN1; MEIS1; RPA2; MEI1; MCM2
RORIE_TARGETS_OF_EWSR1_FLI1_FUSION_UP	25	25	1	3	70.874864	1.82E-05	0.0584143	POU4F1; HLA-E; EMP3
module_252	227	222	0.9779736	5	12.567451	9.00E-05	0.16366696	STMN1; MEIS1; HLTF; RPA2; POU4F1
GNF2_PTPN6	45	44	0.9777778	3	38.059007	0.0001021	0.16366696	HLA-E; EMP3; LIMD2
KALMA_E2F1_TARGETS	11	11	1	2	112.33382	0.0002263	0.2902739	RPA2; MCM2
module_198	292	287	0.982876	5	9.637248	0.000297	0.318176	STMN1; MEIS1; HLTF; RPA2; POU4F1
TMTCGCGANR_UNKNOWN	102	96	0.9411765	3	16.75571	0.0010224	0.718288	CCNJ; ADNP; POU4F1
DNA_BINDING	588	576	0.9795918	6	5.8035072	0.0010974	0.718288	HLTF; KLF11; MCM2; RPA2; POU4F1; TERT
HOMEODOMAIN	293	227	0.774744	4	9.5102122	0.0012088	0.718288	ADNP; MEIS1; NKX2-4; POU4F1
LOPES_METHYLATED_IN_COLON_CANCER_DN	26	25	0.9615385	2	44.047005	0.0012121	0.718288	RPA2; TERT

Appendix Table A.4: Gene set enrichment analysis on consistent hypo-MVPs. Overlaps ('nOVERLAP') between the detected gene set and gene sets in the Molecular Signatures Database v3.0 (MSigDB; Broad Institute of MIT and Harvard, US) were computed as described previously (Subramanian, Tamayo, et al.;PNAS; 2005; 102; 15545-15550). The top 10 hits are illustrated.