Mechanisms of human papillomavirus and host gene transcriptional deregulation in cervical carcinogenesis

This dissertation is submitted for the degree of Doctor of Philosophy

by

Emma Louise Antoinette Drane

Downing College

November 2017

Declaration of Authorship

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

Signed

Emma Louise Antoinette Drane

Date

Acknowledgements

This thesis would not be possible without the contribution and encouragement of many people and I would like to send my warmest thanks to them all.

I would particularly like to thank my supervisor Professor Nick Coleman for giving me the opportunity to come to Cambridge and take on such an interesting project. I am very grateful for all his support and guidance, and for trusting me to take this project in a new direction. I would also like to thank my postdoc supervisor Dr. Ian Groves for his rigorous lab training, enabling me to transition quickly into the Biological Sciences environment as well as his considerable time and effort in helping me bring this thesis together.

I would like to send my sincere thanks to Marco for making our collaboration such a success and for making my time at the Babraham Institute so enjoyable. I am so very grateful for Marco's wealth of knowledge and patience when teaching me and answering my never-ending questions. I would like to wish him and Jo all the luck for the future.

I am hugely grateful for and humbled by the support and friendship of all members of the Coleman Lab; Cinzia and Dawn, you were never too busy to offer guidance and give me your time — thank you for looking after me and being such wonderful influences. Thank you to other PhD students sharing this experience with me: Justyna, Marta, Shivani, Valtteri, Jenny and Luz who have also shared the intense highs and lows throughout the three years of my research.

I would like to thank my family: all parents, grandparents, brothers and sister for your support and excitement whilst sharing this PhD experience with me (but never asking too many questions). Finally, I would like to say a separate thank you to my husband, Rob. Not only did he move house and career to enable me to live closer to the lab, he took on the extra challenge of having me as his wife! I will be forever grateful for his unwavering support, unconditional love and well-timed doses of perspective; I couldn't have done this PhD without you — thank you.

Summary

Cervical malignancy is the fourth most common cause of cancer-related mortality in women worldwide; infection with high-risk human papillomavirus (HRHPV) is responsible for over 500,000 cases of cervical carcinoma each year, approximately 90% of which are squamous cell carcinomas (SCCs). Over half of all HPV-positive cervical SCCs are caused by the deregulated expression of HPV16 oncogenes E6 and E7 in proliferating basal cells of the cervical squamous epithelium. The major risk factor associated with cervical neoplastic progression is integration of HRHPV into the host genome, which is detected in ~85% of HPV16-positive cervical carcinomas. The work presented in this doctoral thesis sought to provide insights into our understanding of the process of HPV16 integration as well as to elucidate mechanisms that deregulate both virus and host gene expression following integration.

The W12 cell model system used in this project is a polyclonal cervical keratinocyte line generated by explant culture of a low-grade cervical squamous intraepithelial lesion (LSIL) that arose following natural infection with HPV16. Through single cell cloning of a long-term culture W12 series, twenty-four isogenic clones, each containing a different site of HPV16 integration, were developed. The W12 clones were isolated in the absence of selective pressure, and as such represent the range of integration events that occur in a pre-malignant lesion at the early stages of carcinogenesis, prior to integrant selection. Despite identical genetic backgrounds, expression levels of oncogenes E6 and E7 varied up to 16-fold between the W12 clones. Expression of HPV oncogenes is ultimately determined by transcription factor binding to the non-coding long control region (LCR) of the viral genome. The initial result of this study found that genomic mutations affecting transcription factor binding at the LCR of the W12 clones was not a cause of differential viral expression,

concluding that epigenetic control may be at play.

In order to provide a tractable system, cells without full-length HPV16 concatemerisation and with four or less copies of integrated virus DNA per cell were used for further analysis. Higher levels of virus expression per template were associated with increased levels of histone post-translational modification (PTM) hallmarks of transcriptionally active chromatin and reduced levels of repressive hallmarks. There was greater abundance of the active/elongating form of the RNA polymerase-II enzyme (RNAPII-Ser2P), together with CDK9, the component of positive transcription elongation factor-b (P-TEFb) responsible for the Ser2 phosphorylation. The changes observed were functionally significant, as cells with higher HPV16 expression per template showed greater sensitivity to depletion and/or inhibition of histone acetyl transferases and CDK9, as well as reduced sensitivity to histone deacetylase inhibition.

Employing next generation sequencing data available for five representative W12 clones, the sites of HPV16 host integration were identified. This confirmed that the virus preferentially inserts into areas of active and open regions of host chromatin, as indicated by the abundance of active PTMs and DNaseI sites and absence of repressive PTMs. HPV16 integration occurs both within genes and at intergenic regions. Features of the integration sites confirm integration occurs either via direct insertion or by a looping mechanism whereby adjacent regions of the host are amplified resulting in local rearrangements. The genomic sequence of the host at the specific site of virus integration showed increased levels of microhomology with the virus genome, hence a mechanism of microhomology-mediated end-joining-dependent integration is likely. HPV16 integration is also associated with changes in host gene expression at least 2.5 megabases away from the integration locus; in the cases where HPV16 integrated directly into a host gene the introduction of the HPV16 promoter resulted a dramatic increase in the expression of downstream exons.

The three-dimensional (3D) structure of the nucleus and physical interactions between stretches of the genome over long distances (i.e. enhancer and promoters) are known to exert an additional level of gene regulation. Identification of 3D virus-host interactions in the W12 clones employing the newly developed and unique ‘Sequence

Capture of Regions Interacting with Bait Loci Hi-C' (SCRiBL-Hi-C) protocol showed that both short- (~50 kb), and long-range (~1 Mb) interactions occur during the early stages of carcinogenesis. Direct HPV16-host 3D interactions were shown to be associated with host gene expression changes, and, in addition, insertion of the virus can disrupt normal host architecture.

Together, the data in this thesis indicate that transcription and subsequent expression of the HPV16 genome is controlled by multiple layers of epigenetic regulation. As such, therapeutics targeting the viral epigenome could be beneficial in modulating HPV16 expression in cases of cervical carcinoma. Virus-host interactions should be further investigated to determine whether changes to the host genome as a result of virus integration make a significant contribution to early cell selection events during carcinogenesis.

Abbreviations

BAC - bacterial artificial chromosome

BCA - bicinchoninic acid

BET - bromodomain and extra-terminal

BRD4 - bromodomain-containing protein 4

BSA - bovine serum albumin

cDNA - complementary DNA

CDK - cyclin dependent kinase

CFS - common fragile site

ChIP - chromatin immunoprecipitation

CIN - cervical intraepithelial neoplasia

CNV - copy number variation

CO_2 - carbon dioxide

Ct - threshold cycle

CTD - carboxy-terminal domain

DAPI - 4',6-diamidino-2-phenylindole

DSB - double strand break

dH_2O - distilled water

DMEM - Dulbecco's Modified Eagle's Medium

DMSO - dimethyl sulfoxide

DNMT - DNA methyltransferase

dNTP - deoxynucleotide triphosphate

DSIF - DRB-sensitivity-inducing factor

E2BS - E2 binding site

E6-AP - E6-associated protein

EBV - Epstein-Barr virus

ECL - enhanced standard chemiluminescence

EDTA - Ethylenediaminetetraacetic acid

EGF - epidermal growth factor

FBS - Fetal bovine serum

FDA - Food and Drug Administration

FIGO - Federation of Gynecology and Obstetrics

FISH - fluorescence in-situ hybridisation

FoSTeS - fork stalling and template switching

gDNA - genomic DNA

GFP - green fluorescent protein

GMEM - Glasgow Minimum Essential Medium

GTF - general transcription factor

H2AK5ac - histone 2A lysine 5 acetyl

H3ac - hisone 3 acetylation

H3K4me1 - histone H3 lysine 4 monomethyl

H3K4me3 - histone H3 lysine 4 trimethyl

H3K9me2 - histone H3 lysine 9 dimethyl

H3K9me3 - histone H3 lysine 9 trimethyl

H3K27ac - histone H3 lysine 27 acetyl

H3K27me2 - histone H3 lysine 27 dimethyl

H3K27me3 - histone H3 lysine 27 trimethyl

HAT - histone acetyl transferase

HBV - Hepatitis B

HCV - Hepatitis C

HDAC - histone deacetylase

HHV-8 - human herpes virus 8

HIF - hypoxia inducible factor

HK - gene housekeeping gene

HMT - histone methyltranserases

HPV - human papillomavirus

HRHPV - high-risk human papillomavirus

HRP - horseradish peroxidase

HSIL - high-grade squamous intraepithelial lesions

hTERT - human telomerase reverse transcriptase

IP - immunoprecipitation

KD - knock down

KO - knockout

LB - lysogeny broth

LBC - liquid based cytology

LCR - long control region

LRHPV - low-risk human papillomavirus

LSIL - low-grade squamous intraepithelial lesions

MAP - mitogen-activated protein

MEM - Minimum Essential Medium

MMBIR - microhomology-mediated break-induced replication

mRNA - messenger RNA

NCX - normal cervix

NELF - negative elongation factor

NEB - New England Biolabs

NGS - next generation sequencing

NICE - National Institute of Clinical Excellence

NS - not significant

NTC - non-targeting control

ORF - open reading frame

Ori - origin of replication

pA - polyadenylation site

Par - parental

PBS - phosphate-buffered saline

PCR - polymerase chain reaction

PIC - protease inhibitor cocktail

PFA - paraformaldehyde

PMSF - phenylmethylsulfonyl fluoride

PRC - polycomb repressive complex

pRB - retinoblastoma protein

P-TEFb - postive elongation factor b

PTM - post-translational modification

PVDF - polyvinylidene difluoride

QC - quality control

qRT-PCR - quantitative reverse transcriptase polymerase chain reaction

RCF - relative centrifugal force

RFU - relative fluorescence unit

RIPA - radioimmunoprecipitation assay

RNAPII - RNA polymerase II

RNAPII Ser2P - RNA polymerase II serine 2 phosphorylated

RNAPII Ser5P - RNA polymerase II serine 5 phosphorylated

RNA-seq - RNA sequencing

rpm - revolutions per minute

RT - room temperature

SCC - squamous cell carcinoma

SCRiBL - sequence capture of regions interacting with bait loci

SDS-PAGE - sodium dodecyl sulfate polyacrylamide gel electrophoresis

SEC - super elongation complex

SEM - standard error of the mean

siNTC - non-targeting siRNA

siRNA - short interfering RNA

SNP - single nucleotide polymorphism

snRNP - small nuclear ribonucleoprotein particle

SSC - saline-sodium citrate

TIP60 - Tat-interacting protein 60 kDa

TF - transcription factor

TSS - transcription start site

VLP - virus-like particle

Contents

1 Introduction **1**

1.1 Cervical cancer 2

1.2 Human Papillomavirus (HPV) and Cervical Cancer 4

1.3 Clinical management of cervical cancer 6

1.4 The HPV genome 9

1.4.1 Virus structure 9

1.4.2 Genome organisation 9

1.4.3 HPV16 oncogenes 10

1.4.4 HPV16 long control region (LCR) 15

1.4.5 HPV16 life cycle 20

1.5 Integration of HRHPV and cervical carcinogenesis 21

1.6 W12: *in vitro* model of cervical neoplastic progression 24

1.7 Epigenetic regulation of the eukaryotic genome 26

1.7.1 DNA Methylation 26

1.7.2 MicroRNAs 27

1.7.3 Chromatin structure 28

1.7.4 Post-translational histone modifications 30

1.8 Role of HPV16 viral oncogenes in modulating epigenetic mechanisms 32

1.8.1 HPV16 and DNA methylation 32

1.8.2 HPV16 and microRNAs 33

1.8.3 HPV16 and histone-modifying enzymes and chromatin remodelling complexes 33

1.9 Deregulated epigenetic regulation and cancer 34

1.10 RNA polymerase II-dependent transcription 35
1.11 Rationale and aims of the investigation 38
1.11.1 Hypothesis 39
1.11.2 Published work 40

2 Materials and Methods **41**

2.1 Cell culture and cell treatments 42
2.1.1 Cell lines and cell culture maintenance 42
2.1.2 Resuscitation of established cell lines from liquid nitrogen . . . 43
2.1.3 Subculture of cell lines 43
2.1.4 Subculture and X-ray irradiation of G3T3 cell line 44
2.1.5 Cryopreservation of cell cultures 44
2.1.6 Cell treatment with small molecule inhibitors 45
2.1.7 Cell transfection 46
2.2 DNA analysis 47
2.2.1 DNA extraction 47
2.2.2 Polymerase chain reaction of HPV16 long control region (LCR) 48
2.2.3 Polymerase chain reaction of virus-host breakpoints 49
2.2.4 DNA gel extraction 50
2.2.5 DNA sequencing analysis 51
2.2.6 quantitative-PCR (qPCR) analysis of gDNA samples 51
2.2.7 Calculating primer amplification efficiencies 52
2.3 RNA analysis 53
2.3.1 RNA extraction from cell lines 53
2.3.2 cDNA synthesis 54
2.3.3 qPCR analysis of cDNA samples 54
2.3.4 Quantification of transcript level changes 55
2.4 Protein analysis 56
2.4.1 Total protein extraction 56
2.4.2 Protein quantification 56
2.4.3 Protein sample preparation and SDS-PAGE separation 57

2.4.4 Western blotting . . . 57
2.5 Chromatin immunoprecipitation (ChIP) assays . . . 59
2.5.1 Cell fixation . . . 59
2.5.2 Shearing of chromatin . . . 60
2.5.3 Immunoprecipitation . . . 60
2.5.4 Elution of chromatin and cross-link reversal . . . 60
2.5.5 ChIP-qPCR . . . 61
2.6 Statistical analysis . . . 61
2.7 Fluorescent *in situ* hybridisation (FISH) . . . 64
2.7.1 Preparation of FISH slides . . . 64
2.7.2 Growing BAC colonies . . . 64
2.7.3 Extraction of BAC DNA . . . 65
2.7.4 Extraction of HPV16 plasmid DNA . . . 66
2.8 Generating directly labelled DNA probes by nick translation . . . 67
2.8.1 Nick translation . . . 67
2.8.2 Coupling fluorescent dye . . . 67
2.8.3 Probe precipitation . . . 68
2.9 DNA FISH . . . 68
2.10 DNA FISH analysis . . . 70
2.10.1 Microscope analysis . . . 70
2.10.2 Determining probe distance . . . 70

3 HPV16 oncogene expression levels during early cervical carcinogenesis are determined by the balance of epigenetic chromatin modifications at the integrated virus genome 72

3.1 Introduction . . . 73
3.2 Results . . . 76
3.2.1 Genetic mutation of the HPV16 LCR is not responsible for the differential oncogene expression of the W12 integrant clones. . 76

3.2.2 HPV16 oncogene expression from the W12 integrant clones depends upon the level of association of activating or repressive chromatin marks. 79
3.2.3 Level of HPV16 integrant transcription per template correlates with association of histone acetylation modifying enzymes . . . 83
3.2.4 HPV16 transcript levels per template correlate with active RNA polymerase II (RNAPII) level and activating complex P-TEFb 94
3.3 Discussion 101

4 Adapting SCRiBL Hi-C methodology to capture the integrated HPV16 genome 107

4.1 Introduction 108
4.2 Materials and Methods 112
4.2.1 Hi-C protocol (SCRiBL) 112
4.2.2 Part I: Generation of Hi-C libraries 112
4.2.3 Part IIa: Generation of biotinylated RNA for target enrichment (SCRiBL) 120
4.2.4 Part IIb: Generation of biotinylated RNA for target enrichment (Capture-seq) 123
4.2.5 Part III: Solution hybrid capture of Hi-C library 125
4.2.6 Bioinformatic analysis performed by Jack Monahan 131
4.3 Results 134
4.3.1 Modifications to the 'Sequence Capture of Regions Interacting with Bait Loci' (SCRiBL) protocol for production of Hi-C and captured libraries from the W12 clones. 134
4.3.2 Short-range interactions of W12 clone Hi-C libraries are detected using chromosome conformation capture (3C) assays. . 137
4.3.3 Verification of in-nucleus ligation efficiency of W12 clone Hi-C libraries by PCR digest assay. 141

4.3.4 Determination of conditions for final PCR amplification of W12 clone Hi-C libraries. 143
4.3.5 Generation of RNA baits for capture-sequencing to detect virus-host breakpoints in the W12 clones. 144
4.3.6 Design and generation of RNA baits for the production of W12 clone SCRiBL libraries. 148
4.3.7 Enrichment of HPV16 genome from W12 clone Hi-C libraries through capture with biotinylated RNA baits. 152
4.3.8 Quality assessment of SCRiBL library NGS using the HiCUP pipeline. 153
4.4 Discussion 156

5 Integrated HPV16 genomes interact with host chromosomes three-dimensionally (3D) modulating nuclear architecture and host gene expression. 165

5.1 Introduction 166
5.2 Results 168
5.2.1 Integrated HPV16 genomes interact in 3D with host chromosomes 168
5.2.2 W12 integrant clone 5' and 3' virus-host breakpoints identification at nucleotide resolution. 172
5.2.3 HPV16 integrates into regions of open and active host chromatin. 178
5.2.4 Short- and long-range 3D interactions occur between the HPV16 and host genomes regardless of cell selection during early cervical carcinogenesis. 180
5.2.5 HPV16 integration can disrupt local host genome architecture and affects the expression of host genes adjacent to the integration site. 186
5.2.6 HPV16 integration results in virus-host fusion transcripts . . . 196
5.3 Discussion 198

6 Concluding discussion and future work 208

7 Appendix 1 **218**

8 Appendix 2 **225**

Bibliography **240**

Chapter 1

Introduction

1.1 Cervical cancer

Cervical cancer is the fourth most common cause of cancer-related mortality in women worldwide, with an estimated 266,000 women expected to die from the disease each year[1]. The global burden of cervical cancer is largely dependent on geographical location. The mortality rate varies 18-fold between different regions of the world; 86% of all cases arise in less developed regions, including Melanesia, Eastern and Middle Africa, where mortality rates are in excess of 20 per 100,000[1](Figure 1.1). The discrepancy can be explained by the introduction of effective, population-wide screening programs in developed countries, which have served to dramatically reduce the incidence of cervical carcinoma and associated mortality.

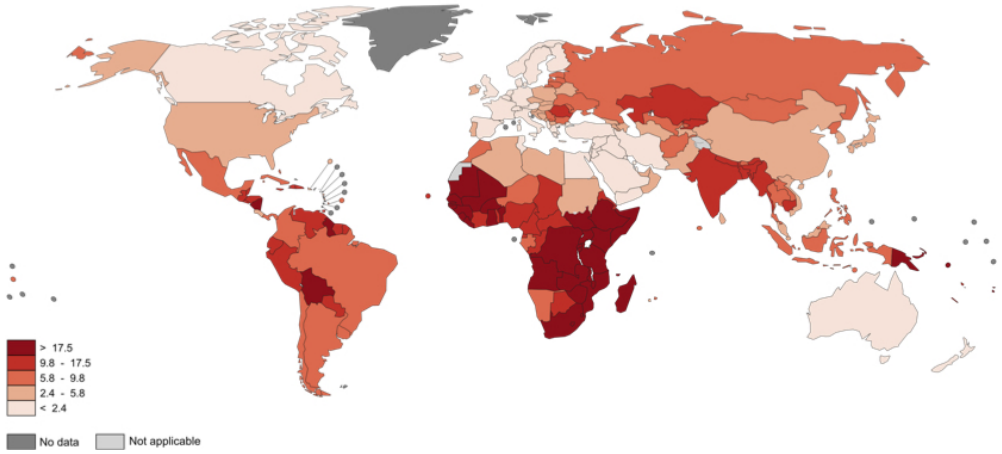

Figure 1.1: Estimated age-standardised rates (per 100,000) of cervical cancer mortality worldwide in 2012. Reproduced from GLOBOCAN, 2012[1].

The classification of cervical cancers is dependent on the type of epithelium it develops from; whilst the vast majority (~85%) of cervical cancers represent squamous cell carcinomas (SCCs), adenocarcinomas and adenosquamous carcinomas are also seen[2]. SCCs arise from squamous cells that cover the outer surface of the cervix (ectocervix); in comparison, adenocarcinomas derive from glandular epithelial cells scattered along the endocervical canal[3].

Cervical SCC is the most prevalent cervical cancer type and develops from a pre-malignant lesion through a spectrum of well-defined transformations. Progression is associated with the expansion of the proliferative compartment within the squamous epithelium and increasing cell atypia. Cervical carcinogenesis is a model for early

detection due to the long and well-known natural history of the disease. Prior to invasive disease, cervical abnormalities can be assessed by cervical cytology employing a test known as the Papanicolaou (Pap) smear, first described by Dr. George Papanicolaou in 1941[4]. The test requires obtaining a sample of cervical epithelial cells, which are then graded on the level of dysplasia and resultant disease severity is inferred. Over the years a number of grading systems have been developed to in an effort to improve and standardise cervical cytology reporting[5]. The first grading system employed was introduced in 1973 and was structured upon a three-tiered classification of cervical intraepithelial neoplasia (CIN), based purely on the histology of the cervical lesion. HPV-induced squamous intraepithelial lesions are regarded as precursor lesions and are graded into three different risk categories depending on proportion of the cervical epithelium occupied by dysplastic cells: CIN1, CIN2 or CIN3 corresponding to one-third, two-thirds or full thickness, respectively[6]. Importantly lesions graded CIN1–3 do not show invasion of the basement membrane, which is characteristic of malignant disease[7]. However, the CIN grading system was not fully clinically relevant; discrepancies between the classification of cervical lesions and biological behaviour led to difficulties in determining effective treatment strategies. As such, a two-tiered Bethesda grading system was devised in 1988[8]. Cervical premalignant lesions are graded into low-grade or high-grade squamous intraepithelial lesions (LSIL or HSIL, respectively), both corresponding to the risk of progression to invasive cancer[9]. To correlate both grading systems, LSILs usually correspond to CIN1, whilst HSILs encompass CIN2 and CIN3 lesions (Figure 1.2).

However, only a subset of cervical precursor lesions progress to cervical cancer; currently because of the lack of predictive markers, clinicians and pathologists are not able to distinguish the lesions that will progress from those that will not[10]. In women who are immunocompetent the rates of spontaneous regression (i.e. without intervention) for a CIN1 or CIN2 lesion range from 40–60% and only a very small percentage progress to develop a higher-grade lesion[11, 10]. Moreover, of those that do develop CIN3, less than half will progress to invasive cancer[6]. Despite this, when a CIN3 lesion is identified healthcare professionals are duty bound to offer surgical intervention as the lesion holds the risk of developing into invasive cancer; as such

there are fewer data available on the spontaneous regression of CIN3 lesions[6].

Following the diagnosis of invasive disease, a patient is staged based on the size and spread of the tumour according to the International Federation of Gynecology and Obstetrics (FIGO) guidelines. Carcinoma of the uterine cervix grows locally, and primarily extends to the uterus and paracervical tissues and the pelvis. As such, tumours that are strictly confined to the cervix are classified as stage I whereas stage II-IV tumours extend beyond the cervix with increasing distance from the initial cervical lesion[12].

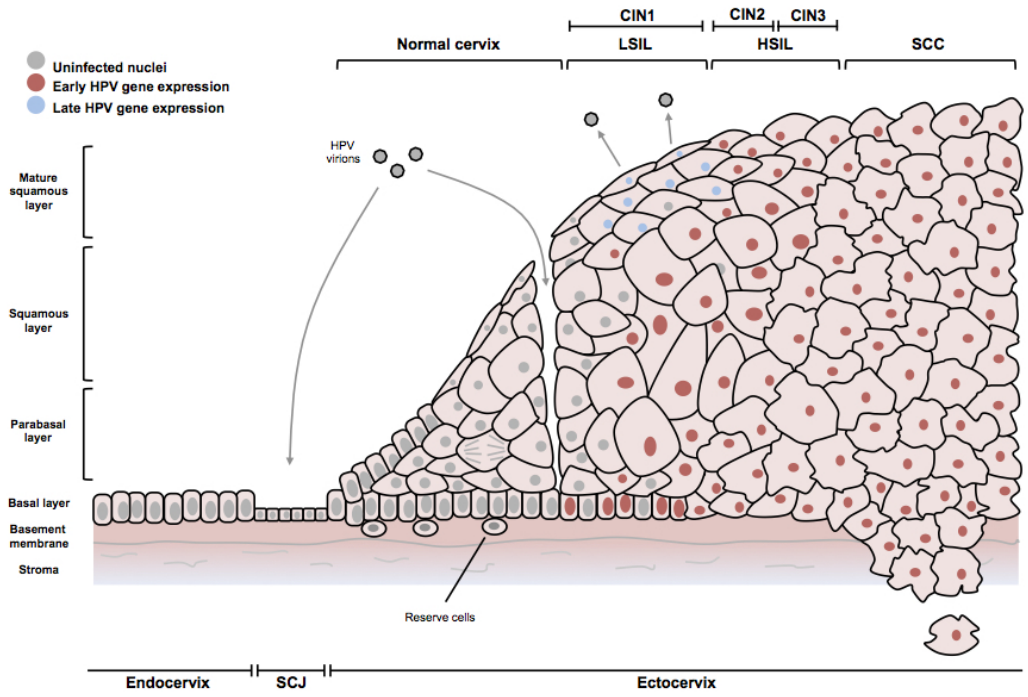

Figure 1.2: Diagram of cervical neoplastic progression. This diagram represents increasing cell atypia and progressive loss of differentiation of cells at the surface of the epithelium, indicative of cervical neoplastic progression. Approximate correspondence between the two-tier Bethesda and the three-tier CIN grading systems are demonstrated (Adapted from Groves & Coleman, 2015[2]).

1.2 Human Papillomavirus (HPV) and Cervical Cancer

Virus infections are responsible for approximately 15% of cancer cases worldwide[13]. Malignancies include Burkitts lymphoma, Kaposi sarcoma and hepatocellular carci-

nomas to name but a few, each caused by infection with Epstein-Barr virus (EBV), human herpes virus 8 (HHV-8/KSHV) and Hepatitis B and C viruses (HBV & HCV), respectively[14]. Cervical cancer is another example of virus-driven disease, with almost all cases (99.7%) caused as a result of persistent infection and ineffective clearance of human papillomaviruses (HPV); as such, infection with HPV is accepted to be a necessary cause of cervical cancer[15, 16]. HPV infection has also been attributable to the pathogenesis of other cancers including those of the vagina (78%), penis (51%), anus (88%), vulva (<48%) and oropharynx (<51%)[17]. These carcinomas are typically caused by infection with the high-risk group of HPVs, which have a higher oncogenic potential than the low-risk types.

Over 200 HPVs have been recognised and are classified by genotype into five evolutionary groups based upon whether they infect cutaneous or mucosal epithelia and their disease associations[18, 19]. The most well studied HPV type is the mucosal alpha papillomaviruses, over forty of which have been shown to infect the anogenital mucosa[20]. As previously mentioned, this group is further subdivided into high-risk (HRHPV) and low-risk groups (LRHPV) based on their oncogenic potential. Whilst LRHPVs, typically HPV6 and 11, can cause benign epithelial hyperplasias (genital warts), they are not associated with intraepithelial neoplasia. In contrast, HRHPVs including HPV16, 18, 31 and 45 are associated with over 90% of cervical malignancies, with HPV16 alone accounting for over half of all cases worldwide[21, 22].

Despite this, infection with HRHPV does not inevitably mean that cervical abnormalities will develop; only 0.3% to 1.2% of initial infections will eventually progress to invasive cervical cancer[23]. Although HPV infection is a common sexually transmitted disease — the overall prevalence of HRHPV infection is 23%[24] — approximately 90% of infections are spontaneously cleared by the host immune system within two years[25, 26]. Persistent infection is seen in only 10-15% of women who are unable to clear the infection[2]; HPV persistence is the main risk factor associated with progression and, as a result, these women are at greater risk of developing cervical cancer[27].

1.3 Clinical management of cervical cancer

As with all cancer types, prevention and early detection are far better than cure. The recognition and understanding that infection with HPV is a necessary factor of cervical cancer facilitated two major developments for cervical cancer prevention: HPV vaccination and HPV DNA testing.

As previously mentioned, the first population-wide cervical screening program the Pap smear test involves taking a sample of epithelial cells from the surface of the cervix, which are then smeared onto a glass slide, stained, and the morphology of the cells assessed. The identification of pre-cancerous or cancerous cells within the sample informs the pathologist of the degree of cervical dysplasia[4, 18]. In parts of the world where routine screening has been implemented, there has been a significant decrease in the incidence and mortality as a result of cervical cancer[28]. Despite this, there are a number of concerns with the Pap smear test due to its subjective nature and lack of sensitivity. The test has a high false negative rate that varies from 30% and 86% and it is inadequately sensitive (as low as 86%)[29]. Furthermore, approximately 8% of Pap smears are inadequate for interpretation due to problems with sample collection and the crude nature of slide preparation. As such, alternative technologies have been investigated. Liquid-based cytology (LBC) was introduced to address the aforementioned issues by making changes to sample processing methodology. Instead of smearing the collected cervical epithelial cells onto a glass slide which is subsequently fixed, the head of the collecting brush is broken off and placed into a preservative fluid, this ensures that most, or all, of the cervical cells are retained. The resulting fluid is then centrifuged to remove cellular debris, such as blood or mucus, before the slide is prepared. Comparisons between Pap and LBC screening methods showed that not only was there a significant decrease in the number of inadequate samples (from 9.1% to 1.5%), test sensitivity was increased by 12% with the introduction of LBC technologies. As such, in 2003 the National Institute of Clinical Excellence (NICE) recommended LBC as the main method for cervical screening[30].

More recently molecular HPV-testing has been suggested as a new approach for screening. Several randomised trials have indicated that molecular techniques

lead to greater diagnostic sensitivity and reproducibility, particularly when detecting CIN2 or CIN3 lesions, when compared with cervical cytology[31]. In addition, HPV testing gives women the choice to self-sample. Low attendance of screening programs remains a major obstacle in the prevention of cervical cancer in both developed and developing countries. As such, self-sampling offers an alternative to attending healthcare clinics and it is likely that more women would engage and be effectively screened. Cervical cancer mainly occurs in unscreened or under-screened women; it is estimated that approximately 50% of women diagnosed with cervical cancer have never had a cervical cytology test and 10% have not had one in the last five years prior to diagnosis[32]. To ensure the most robust screening, there is an argument for co-testing, i.e. cytology and HPV-testing. In this way, the small subset of women with cervical cancer that test negative with common HPV assays but positive for cytology would also be identified[33].

Screening programs facilitate early disease detection, however, only a primary prevention strategy, such as vaccination against HPV infection, can address the fundamental cause of the disease. As previously stated, most HPV infections resolve with time and the virus is cleared as a result of a successfully mediated immune response. Upon natural infection, neutralizing antibodies against the major coat protein L1, which self assemble into virus-like-particles (VLPs), are produced[26]. Based upon the natural clearance of an HPV infection, numerous prophylactic HPV vaccines have been developed and are widely used across the world. In 2006 the Food and Drug Administration (FDA) approved Gardasil (Merck), a quadrivalent vaccine developed for the prevention of disease associated with infection with different HPV types: cervical cancer (HPV 16 & 18), genital intraepithelial neoplasia (HPV 6, 11, 16 & 18), and genital warts (HPV 6 & 11). Subsequently two additional vaccines were approved in 2009 and 2014: Ceravix (GlaxoSmithKline), a bivalent vaccine against HRHPV types 16 and 18, and Gardasil 9 (Merck), the most advanced prophylactic vaccine to date, which protects against nine HPV types: 6, 11, 16, 18, 31, 33, 45, 52 and 58[34, 35].

The immune response to HPV infection and/or vaccination is type-specific. Ideally vaccines would protect against all HRHPV types or should, at least, be tailored

to specific regions of the world as the prevalence and distribution of HRHPV types differ between geographical locations. HPV16 and HPV18 are predominant in Europe, North America and Asia; however, in other continents different strains are more prevalent such as HPV31 in South America and HPV52 in Africa[36]. As such, the current prophylactic vaccines are tailored towards the population of developed countries and are unable to prevent all cases of cervical cancer worldwide, particularly in developing regions where the need is greatest. Furthermore, at present, both vaccines are prohibitively expensive for wide scale use in developing countries with a cost of ~$325–$403 for the recommended three doses[36].

While prophylactic HPV vaccines are able to block an initial infection and effectively prime the immune response against future infections, they are unable to eradicate prior infection. Another area of active investigation is the development of therapeutic vaccines, used to treat established infections. Unlike prophylactic HPV vaccines, which are used to generate neutralizing antibodies against virus particles, therapeutic HPV vaccines are used to stimulate cell-mediated immune responses to specifically target and kill infected cells[37]. The HPV-encoded early proteins, specifically E6 and E7, are the main targets for therapeutic vaccines since they are consistently expressed in HPV-associated malignancies and pre-cancerous lesions and play a crucial role in the generation and maintenance of HPV-associated disease[38]. A number of therapeutic vaccine approaches have been developed and include: live vector-, protein-, nucleic acid-, and cell-based immunisation strategies. These are currently being tested in clinical trials and present as a viable therapeutic option in cervical disease[38, 39].

The development of cervical cancer treatments is much less advanced than primary (HPV vaccination) and secondary (screening via cytology and HPV testing) prevention strategies. Current treatment options are based upon the pre-cancerous or cancer stage and tumour size. A range of surgical options including conisation of cervix, hysterectomy and pelvic trachelectomy are offered to women with early stage disease (CIN3 and stage I cancer) depending on their desire to remain fertile; the 5-year survival of patients presenting with localised disease exceeds 90%. As has been previously mentioned however, the majority of women with CIN3 lesions will not

develop cancer, raising the concern that over 50% of all patients are currently over-treated at this stage of disease progression[6]. However, as the disease progresses the estimated 5-year survival dramatically decreases; for women with cancers that have spread within the region it is 57% and for cancers which have metastasised to distant organs survival is estimated at only 17%[40]. For more advanced stages of the disease, obtaining negative surgical margins becomes increasingly difficult and surgical resection becomes technically more challenging; in these instances if surgery remains a viable option, chemotherapy and/or radiation is given in combination in an attempt to improve treatment results. However, for patients with stage IV disease and distant metastases, systemic chemotherapy with cisplatin remains the best option[41]. Novel treatment options for patients with metastatic cancers and recurrent disease after chemo-radiation therapy are being developed and include anti-angiogenic drugs and antibody-based therapies, the first of which was approved by the FDA in 2014[42].

1.4 The HPV genome

1.4.1 Virus structure

HPVs are relatively small, non-enveloped viruses approximately 55 nm in diameter. They consist of an icosahedral capsid composed of 72 capsomeres containing the viral genome. Capsomeres are composed of two structural late proteins: L1, which accounts for 80% of the virus particle, and the minor capsid protein L2. The HPV16 genome is a double-stranded, circular (episomal) DNA molecule, 7904 base pairs (bp) in length, which is transcribed unidirectionally and all resultant transcripts are polycistronic.

1.4.2 Genome organisation

The HPV16 genome is functionally divided into three regions: early, late and a long control region (LCR), the domains of which are separated by two polyadenylation (pA) sites. The early (E) region of the genome encodes seven common open reading frames (ORFs): E6, E7, E1, E2, E4, E5 and E8 which are required for the regulation

of viral DNA replication and viral gene expression[43]. Only E6 and E7, and possibly E1 and E2 proteins are true early proteins in so much as they can be detected in basal epithelial cells[44]. E1, E2, E4 and E5 are expressed in the supra-basal layers and are considered intermediate proteins — actually the expression of E1 and E2 viral replication proteins and transcription factors is greatest in the mid- to upper layers of the epithelium (Figure 1.3 A)[45, 43]; currently, there is insufficient data to be sure of the sites of E8 expression. During the replicative stage if HPV infection, E4 protein is the first, and most abundant, late protein to be expressed in the mid to upper layers of the epithelium which is also the location of maximum E5 expression[43]. The late (L) region encodes the structural proteins L1 and L2 that form a virus capsid; these genes are expressed only in the final stages of cellular differentiation in the upper most, granular layer of the epithelium where viral DNA is packaged in the capsid to be released to infection other cells (Figure 1.3 A)[46]. The LCR region is an 850 bp non-coding, regulatory region that contains promoter sequences that direct transcription of both the early and the late genes[47], the origin of replication (*Ori*), as well as multiple cis-acting sequences that regulate polyadenylation and viral late mRNA stability[48]. The HPV16 genome contains two major promoters; the p97 promoter lies upstream of the E6 ORF and is controlled primarily by upstream cis-elements in the LCR, which are responsible for early gene expression. The second 'late' promoter, p670, lies within the E7 ORF and is responsible for late gene expression, only being induced in differentiated keratinocytes (Figure 1.3 B).

1.4.3 HPV16 oncogenes

When viruses function as carcinogenic agents, they are able to employ a variety of mechanisms that result in cellular immortalisation and the transformation of human cells. Immortalisation and direct transformation of infected cells occurs through the expression of viral oncogenes, which are able to inactivate regulators of genome stability, cell viability and cell cycle; the tumour suppressor proteins p53 and retinoblastoma protein (pRB), two key cell regulators, have been shown to be targeted for degradation by a number of different virus oncogenes[49]. HPV16 encodes two onco-

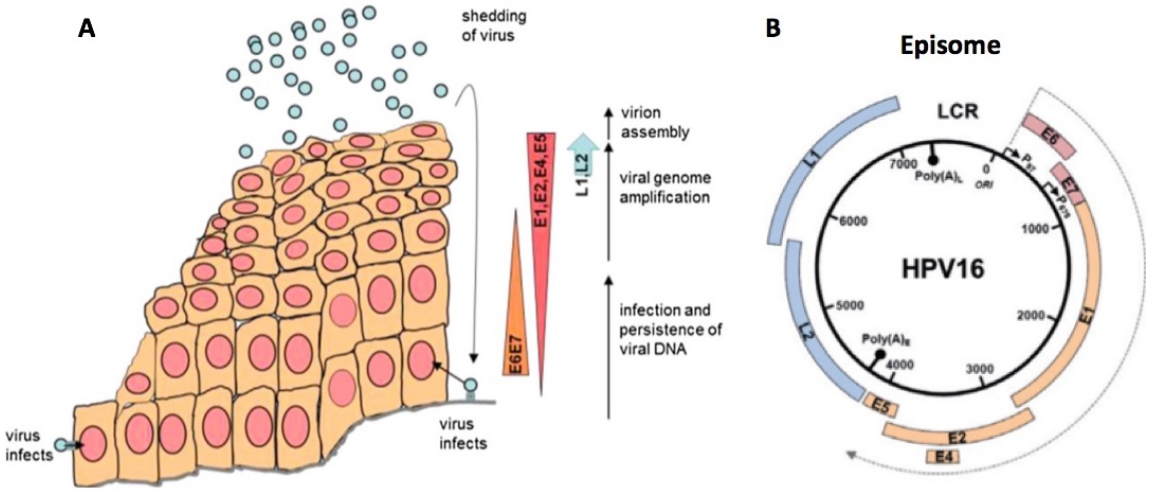

Figure 1.3: Genome organisation and the physical states of the HPV genome. A) Schematic diagram of the HPV16 life cycle in a differentiating epithelium. Viruses are show as light blue circles. Keratinocytes are in light orange color. Nuclei are colored pink. The basement membrane is drawn with a grey line. The key events in the virus replication cycle are indicated to the right hand side of the diagram of the epithelium together with a schematic diagram of the gene expression program of the virus within the infected epithelium. Shading on the arrows represents the quantity of expression of each protein subset during the virus replication cycle. (Adapted from Graham & Faizo, 2017)[43]. B) The genomic organisation of HPV16, highlighting the early (E) region, the late (L) region and the long control region (LCR). The early (p97) and late (p670) promoters and early (A_E) and late (A_L) polyadenylation sites are also indicated. (Adapted from Groves & Coleman, 2015[2]).

genic proteins E6 and E7, which function synergistically to infer limitless replicative potential, evasion of apoptosis and genome instability, all of which are hallmarks of cancer[50]. The many functions of the HPV16 oncogenes are summarised in Figure 1.4.

The HPV16 E6 gene encodes a basic protein of approximately 150 amino acids (19 kDa) that contains two-zinc-binding regions. These binding domains are able to associate with and degrade numerous cellular proteins including the major cell-cycle checkpoint tumour suppressor protein p53[51]. The primary method of E6-directed degradation of p53 is facilitated by the formation of a complex comprised of E6 and the E3-ubiquitin ligase called E6-associated protein (E6-AP), which is able to bind the p53 protein. When the E6/E6-AP complex binds p53, p53 becomes rapidly ubiquitinated resulting in subsequent proteasome-mediated degradation[52]. E6 is also able to inhibit the expression of p53-regulated genes in an E6-AP-independent manner by preventing the co-activating capacity of the histone acetyl transferase p300[53, 54]. Under normal conditions, activated p53 functions include the initiation of DNA repair pathways, cell cycle arrest, cell metabolism and/or apoptosis; however in the presence of E6, p53 cannot accumulate, and the ability of the cell to arrest mitosis in response to DNA damage is removed.

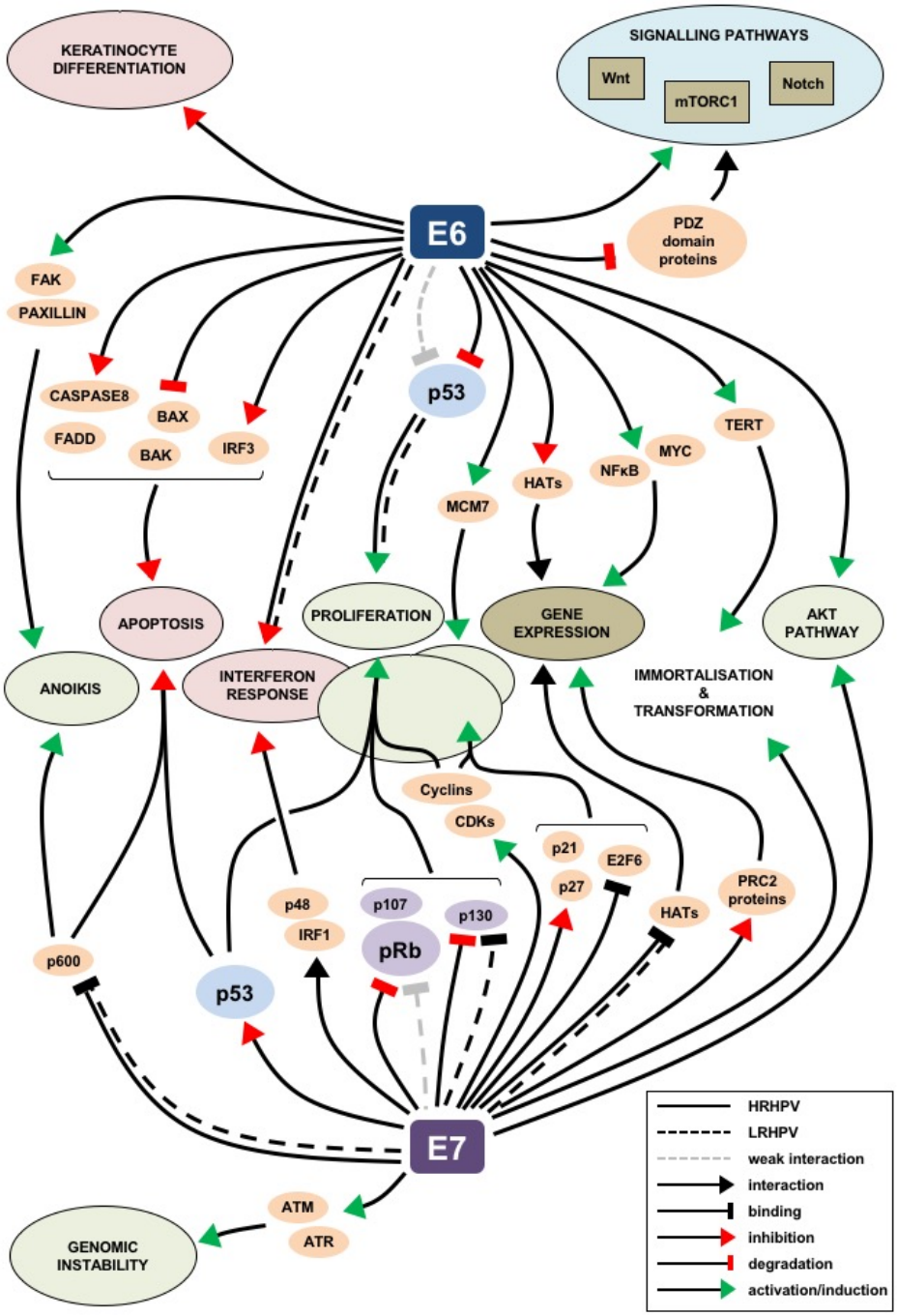

Figure 1.4: Important functions of high-risk HPV E6 and E7 proteins. The figure gives an overview of important direct and indirect effects of the α-genus HPV E6 and E7 proteins on cellular pathways and processes. Important roles of E6 and E7 include degradation of cellular p53 and pRb, respectively. Red oval, general down-regulation of cellular process or pathway; green oval, general up-regulation of cellular process or pathway; brown oval, modulation of cellular process or pathway (Adapted from Groves & Coleman, 2015[2]).

In addition to p53 degradation, HRHPV E6 oncoproteins have developed additional mechanisms to inhibit the apoptosis response by enhancing the degradation or the induction of proteolytic inactivation of the pro-apoptotic proteins BAK and FADD, respectively[55]. Preventing natural cellular apoptosis facilitates carcinogenesis by allowing mutations to accumulate as cells with damaged DNA continue to replicate.

HRHPV E6 is also able to stimulate telomerase expression and activity, thereby enabling replicative immortality[55]. In normal cells telomeres serve to protect the ends of chromosomes from DNA damage including illegitimate fusions, and shorten at a constant rate with progressive cell divisions as a result of tightly repressed telomerase. However, in most immortalised cells — including 85–90% of cell derived from human cancers — the expression of telomerase is increased resulting in the maintenance of telomere length and the absence of cellular senescence[56]. In HRHPV infected cells telomerase activity is stimulated via E6/E6-AP mediated ubiquitination and the subsequent degradation of the transcriptional repressor NFX1-91. Degradation of NFX1-91 results in transcriptional activation of the hTERT (human telomerase reverse transcriptase) gene and additionally has a role in HPV16 E6 activation of the oncogenic transcription factor NF-$\kappa\beta$[57, 55]. Furthermore, E6 is able to manipulate normal cell properties including changes to cell-cell adhesion and cell polarity properties by targeting PDZ domain-containing proteins for degradation, resulting in cell transformation[58].

E7 is a phosphoprotein of approximately 100 amino acids (13 kDa) that contains a short motif (CR2) that mediates the interaction with the retinoblastoma tumour suppressor protein (pRB) and its related proteins p130 and p107[59]; these proteins are linked to cell cycle control and are degraded by the HRHPV E7 oncoprotein. Normally, pRB binds and inactivates the transcription factors E2F 1-3, which maintain cells in a quiescent state in the G_0/G_1 phase of the cell cycle. However, E7 proteins target the active, hypo-phosphorylated form of pRb for proteasomal degradation resulting in E2F-regulated transcription. Resultant transcription of cyclin A and cyclin E positive regulators of cyclin dependent kinases (CDKs) induces cell cycle progression into S-phase and sustained proliferative signaling[55]. E7 also

interacts with the 600 kDa pRb-associated factor p600; p600 is known to play a role in integrin-mediated signaling, and knockdown of p600 sensitises cells to apoptosis. The E7-p600 interaction has been suggested to be an essential mechanism, independent of pRB inactivation, for oncogene-induced cellular transformation[58]. Finally, expression of HRHPV E7 can also cause genomic instability by inducing centrosome amplification; this can lead to aneuploidy and structural chromosomal instability[60]. The overall result of E7 activity is to allow cell growth without differentiation, which can lead to immortalization.

The loss of function of both the p53 and pRb pathways through degradation and inhibition plays a significant role in the development of most human cancers, and confers a growth advantage to affected cells. Loss of pRb results in hyperproliferation triggering apoptosis, which is blocked by the desensitisation of cells to checkpoint signals as a result of the loss of p53. The cooperative action of E6 and E7 leads to the emergence of a clonal population of cells with a growth advantage with a predisposition for transformation and malignant progression[61, 62]. This dogma has been illustrated using the W12 cell system (see section 1.5), and this thesis aims to determine epigenetic mechanisms that regulate the expression of HPV16 E6 and E7 prior to malignancy.

1.4.4 HPV16 long control region (LCR)

The HPV16 LCR consists of non-coding DNA that contains a large number of cis-responsive elements that control the replication and transcription of the virus. They are extremely important for the viral biology, as they couple the expression of the early and late genes to the differentiation of the squamous epithelia and affect the amount of viral gene expression by feedback control. The regulatory elements are short DNA sequence motifs that are recognised and bound by regulatory proteins such as: transcriptional activators, repressor, terminators and initiators of replication. Once bound to the HPV16 genome, regulatory proteins affect the association of the pre-initiation complex of the basal transcription machinery and thereby modulate the rate of transcription initiation and elongation as well as DNA replication[63].

The HPV16 LCR can be divided into three sections, each with different functions; the 5' segment contains transcription termination signals and a nuclear matrix attachment region, which mediates structural organisation of chromatin; the central segment that contains epithelial cell-specific enhancers; and the 3' segment that contains the replication origin and the E6 promoter (p97)[64] (Figure 1.5).

The 3' segment containing the early promoter (p97) has a characteristic arrangement of four cis-responsive elements including a TATA box which serves to bind the transcription pre-initiation complex, a G-rich hexamer (GGGCGT) binding site for the activating transcription factor Sp1, and two E2 binding sites; E2BS1 and E2BS2 positioned between the Sp1 binding site and the TATA box. The central LCR contains a large transcriptional enhancer region composed of numerous cis-responsive elements that are affected individually, but which modulate synergistically or antagonistically p97 promoter activity over a range of 2–3 orders of magnitude[63]. The most noted transcriptional activators known to bind to the HPV16 enhancer region of the LCR are NF1, AP1 (the principal activator of the epithelial enchancer), Oct1 and TEF1. However, transcriptional activation can be antagonised through the binding of repressive transcription factors, for example, YY1 and CDP, which act as negative regulatory elements or silencers[63, 64]. The 5' LCR contains the termination site for the translation of the late genes L2 and L1 (Poly(A)_L - AATAA) and also contains a 79 nt RNA element termed the negative regulatory element (NRE) or late regulatory element (LRE); the LRE is located at the end of the L1 open reading frame and spans the start of the late 3' untranslated region; the element is a conserved feature of papillomaviruses and inhibits late gene expression in undifferentiated epithelial cells[43]. The chromatin structure of the HPV16 LCR region also contributes to the level of viral transcription. The HPV16 LCR is organised in the form of two specifically position nucleosomes, one overlapping with the viral enhancer and one with origin of replication and early promoter. Nucleosome positional transcription experiments have revealed that the function of the early promoter is diminished due to the reduced accessibility of the promoter sequences to transcriptional activators, i.e. AP1 and the basic transcriptional machinery. However, post-translational modification of the nucleosomes (histone tails) and an excess of transcription factors AP1

and Sp1 can overcome this repression to enhance transcriptional activity[65].

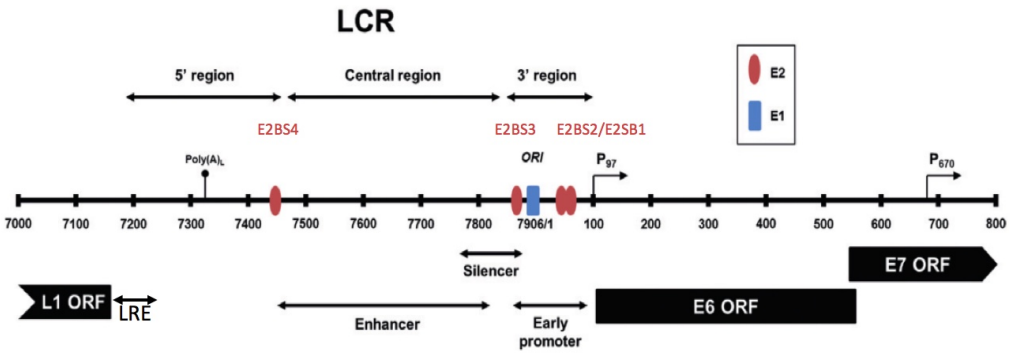

Figure 1.5: The long control region (LCR) of the HPV16 genome. The HPV16 LCR is positioned between the late and early virus gene regions. The virus replication protein E1 binds as a dimer of hexamers at the origin of replication (*Ori*), while the virus transcription factor E2 associates as a dimer at four E2 binding sites (E2BSs). Transcription of early genes occurs from the early promoter (p97) and is dictated by the binding of numerous host transcription factors and the virus E2 dimer across the enhancer, silencer and promoter regions. Activation of transcription from the late promoter (p670) is dependent upon cell differentiation and binding of differentiation-associated transcription factors. Poly(A)$_L$, late polyadenylation site; ORF, open reading frame associations (Adapted from Groves & Coleman, 2015[2]).

The role of E2 in transcriptional activation

Notably, papillomavirus genomes themselves encode a transcription factor, protein E2, which regulates viral gene expression through binding to the four E2 binding sites in the LCR (Figure 1.5). The E2 protein consists of a DNA-binding domain and a transactivation domain that are linked by a serine-arginine-rich hinge region and forms a highly stable dimer that binds to the E2 bindings sites located across the LCR; two E2BSs are proximal to the viral early promoter, the third is located at the origin of DNA replication and the fourth is located upstream of the keratinocyte enhancer[66]. The E2 proteins function primarily by recruiting cellular factors to the viral genome, which activate or repress transcriptional processes depending on the context of the binding sites and nature of the associated cellular factors. Binding of E2 to the E2BS1 induces transcriptional activation of the early promoter (p97), resulting in enhanced productions of the E6, E7 and E2 proteins. When E2 concentration rises to a high level, it also binds to the low affinity E2-binding sites E2BS2, 3, 4, respectively and represses transcription from the early promoter by competeing with cellular transcription factors for binding sites[67]. Association of HPV16 E2 to E2BS1 and E2BS2 in close proximity to the HPV promoter sterically hinders the binding of cellular factors such as Sp1 and TBP to proximal promoter elements in the viral genome. Additionally, the binding of E2 protein to the E2BSs proximal to p97 promoter represses transcription through steric hindrance of the interaction with the transcriptional initiation factor TFIID at the proximal TATA box and subsequent Poll II pre-initiation complex assemble[68, 69].

Activation of E2 proteins requires the cooperation of at least two E2 dimers bound to their cognate sequences; these two E2 dimers could be adjacent (e.g. E2BS1 and E2BS2) or distant from one another (e.g. E2BS1 and E2BS3). Dimerisation of the HPV16 E2 amino-terminal transactivation domain over large distances has been shown to induce looping of the intervening sequences, bring the proximal enhancer and the proximal promoter together in 3D space, and is another potential mechanism for E2 transcriptional regulation[70, 71]. Experimental evidence shows that enhancer-binding proteins — such as Sp1 — can be targeted to promoter regions via direct

interaction with E2 proteins that bind proximal to transcription start site through the formation of stable DNA loops[72]; in this context, the activity of the p97 promoter is increased as a result of E2-dependent enhancer looping.

One of the best-characterised interactors of E2 is the bromodomain-containing protein (Brd4), which a member of the bromodomain and extra-terinal domain (BET) protein family and is another cellular chromatin-binding factor that has a role in the regulatory function of E2[69]. In association with the major viral regulatory protein E2, Brd4 is involved with multiple processes of the papillomavirus life cycle, including the initiation of viral replication as well as viral genome segregation and maintenance[73, 69]; Brd4 serves to tether E2 and the viral genomes to mitotic chromosomes in dividing cells, thus ensuring viral genome maintenance[74]. Moreover, Brd4 also plays a role in gene transcription from the viral early promoter[73, 74]; notably Brd4 is required for transcriptional activation function of E2[75, 73, 74]. Brd4 recruits a variety of transcription factors and chromatin regulatory to control transcription; these include the positive transcription elongation factor b (P-TEFb) and general cofactor Mediator. It has been shown that the recruitment of P-TEFb is important for E2s transcription activation activity[73] and additionally the intrinsic histone acetyltransferase (HAT) activity of Brd4 results in strong transcriptional activation of target genes[69]. However, Brd4 has been shown to have a dual role in relation to the regulation of viral transcription and has been identified in a transcriptional silencing complex assembled by HPV E2. In this context, Brd4 acts as a cellular co-repressor that reduces the activity of the early promoter resulting in decreased expression of the HPV16 genome; recombinant Brd4 and E2 are both necessary and sufficient to replace the purified E2 repressor complex in inhibiting AP-1-dependent HPV transcription in an E2-binding site-specific manner[76].

1.4.5 HPV16 life cycle

HPVs are intracellular parasites and must deliver their genome into host cells, and subsequently use host cellular machinery for viral replication. The viral capsid — composed of structural viral proteins L1 and L2 — plays a key role in the establishment of a viral infection providing the initial site of interaction between the virus particle and the host cell[77]. Throughout the initiation of an HPV16 infection the L1 protein plays an essential role in maintaining the structural integrity of the capsid as well as binding to host cell surface receptors, whereas the L2 protein ensures that the viral genome is trafficked correctly to the host cell nucleus, where viral gene expression can initiate[78]. Infection of the target keratinocyte by HPV16 is a highly complex process; the initial virus-host interaction and virus entry mechanisms as well as the molecules involved are still a subject of scientific debate[77, 78]. Initially, infectious HPV16 particles bind to the basement membrane of the disrupted mucosal epithelium and virus entry occurs in the basal keratinocytes (Figure 1.2). Attachment is believed to occur through association between L1 components of virus capsid and herparan sulphate proteoglycans (HSPG), which are frequently found in the extracellular matrix and on the surface of most cells. This initial attachment results in a conformational change of the viral capsid, which facilitates L2 proteolytic cleavage — the minor capsid protein L2 is cleaved by furin on the cell surface — and the virus binds to an as yet unidentified receptor on the target cell[77, 78]. HPV16 is then internalised into the cell and following endocytic transport and acidification, progressive capsid disassembly occurs[78, 79]. Host cell cyclophilins release the majority of L1 protein from L2 protein, which remains in complex with the viral genome. The L2 protein then translocates across the endocytic membrane to engage factors that mediate transport to the trans-Golgi network (TGN)[79]. After the initiation of mitosis, the HPV16 genome egresses from the TGN and associates with microtubules. During mitosis there is membrane dissolution and nuclear envelope breakdown, which allows the L2:HPV16 DNA complex to migrate along microtubules into the nucleus and to the condensed chromosomes of the host[78, 80].

Once entered into the host cell nucleus, HPVs are reliant on the full complement of

host replication proteins to mediate viral synthesis as the viral replication proteins E1 and E2 are insufficient to complete replication of the virus genome[81]. Once the virus enters the cells of the basal membrane it hijacks the cellular resources in order to replicate its own genetic material and express HPV proteins in a temporal and spatial pattern. Following entry into the cell, HPV genomes are established as extrachromosomal elements (episomes) in the cell nucleus; after initial infection basal keratinocytes undergo proliferation, increasing the number of cells harbouring viral episomes. At this point viral genomes are maintained at low copy number (~10–200) through co-ordinated replication with the host DNA, and early viral gene expression, particularly E1 and E2, is dominant[44]. As infected HPV-containing daughter cells migrate from the basal layer and differentiate, the cell enters the proliferating compartment of the epithelium, inducing the productive phase of the viral life cycle. The resultant up-regulation of the early promoter leads to increased levels of E6 and E7 proteins, which deregulate cell cycle control. The cell is now permissive to viral replication and episomal copy number is dramatically increased to thousands of copies within a single cell[62, 82]. An increase in activity of the late viral promoter (p670) in the mid and upper layers of the epithelium results in the expression of the structural L1 and L2 capsid proteins. Once associated, the late proteins encapsidate newly synthesised viral genomes and produce infectious particles (virions), which are shed from most superficial layers of the epithelium, and the virus spreads[44]. The papillomavirus lifecycle takes 2–3 weeks, the time necessary for a cervical keratinocyte to undergo complete differentiation, migrate from the basal to the upper most layers of the epithelium, and desquamate[83, 82]

1.5 Integration of HRHPV and cervical carcinogenesis

Mechanisms of HPV invasion and replication result in a very poor host humoral and cell-mediated immune response. The replication of virus DNA and virus assembly occurs in cells already destined to die at the skin surface as part of the normal process

of skin shedding; as such, the productive cycle of HPV does not cause virus-induced cytolysis or necrosis, and thus does not cause inflammation or the production of pro-inflammatory cytokines, which would normally lead to a cascade of immunological responses[84, 83]. In addition, by only infecting cells of the basal epithelium, the virus evades immunologically competent cells in the upper layers of the skin. As a result, in some cases, the virus is able to establish a persistent chronic infection, the single most important risk factor for the development of cervical SCC and its precursors[85].

Throughout the normal HPV life cycle the viral genome is maintained in an episomal state. Although in pre-invasive SILs the HPV genome is predominantly found in its episomal form, integration of the virus genome has been shown to correlate with the progression of precancerous lesions (CIN2/3) to invasive cancer[86, 27]; indeed, integration of HPV DNA into the host genome has been identified in 86.5% of all cervical SCCs[87]. During HPV DNA integration, the virus genome usually breaks in the E1 and/or E2 ORFs, resulting in a loss of these genes[88]. In contrast, the integrated HPV genomes of selected cells faithfully retain the LCR and full length E6 and E7 ORFs meaning that all genomic elements required for the transcription of the viral oncogenes are maintained. Truncation of E2, which regulates transcription from the viral early promoter, results in the loss of E2-dependent negative feedback and, as a consequence, the expression of the viral oncogenes E6 and E7 is dramatically increased[27]. Additionally, fusion of HPV E6 and E7 encoding oncogene transcripts to transcribed cellular sequences increases the stability of E6 and E7 transcripts produced from the integrated virus. This results in higher steady-state levels of the viral oncoproteins and thus enhanced oncogenic activity[89]. As previously mentioned, not only does increased expression of E6 and E7 result in deregulated cellular proliferation and cell immortalisation, it also induces genomic instability facilitating the malignant transformation of host cells and tumour formation[90]. Additionally, HPV genome integration is associated with progression from polyclonal to monoclonal status in CIN; this indicates that certain integration events confer a selective advantage in a mixed population of cells[91]. Interestingly, where multiple HPV copies arranged as concatemers have integrated, it is often only the most downstream genome that remains transcriptionally active. The additional copies are silenced by DNA methyla-

tion, which again may reflect clonal selection due to an optimal level of viral oncogene expression[92]. Moreover, HPV integration sites are usually found only at a single or few chromosomal loci in clonal cell populations of cervical cancers[93].

The identification of HPV integration sites in various associated malignancies is an active area of investigation. It has been shown that HPV integration frequently occurs at, or in close proximity to, common fragile sites (CFSs) of the human genome[86, 94, 95]; these are specific chromosomal loci particularly prone to breakage and include the genomic loci 8q24.21. The proto-oncogene *C-MYC* is encoded for within this region, the expression of which has been shown to dramatically increase as a result of HPV16 integration[96, 97, 87]. The expression of host genes at or near the integration site is changed as a result of HPV integration, which has been shown to cause a wide-range of somatic mutations, copy number variations and structural rearrangements of the host genome[98]. As well as increased expression of the viral oncogenes, it is likely that alterations to host gene expression may also promote malignant progression.

Determining the mechanism of HPV integration is also a current research area. Integration is not part of the normal life cycle of HPV, and HPVs encode no integrases or polymerases. HPV integration presumably occurs following double strand breaks (DSBs) in host and viral DNA, hence the frequency of integrations occurring in CFSs[99]. In addition to the body of evidence that suggests that HPV integration is non-random there is increasing evidence that HPV drives integration through microhomology-mediated DNA repair pathways. Identical nucleotide sequences between the host and virus is commonly found across 1–10bp either site of the breakpoint[100, 87, 88, 101]. The enrichment of microhomology at HPV integration sites has led researchers to indicate that fork stalling and template switching (FoSTeS) and/or microhomology-mediated break-induced replication (MMBIR) are likely mechanisms for HPV integration[87].

1.6 W12: *in vitro* model of cervical neoplastic progression

Given the nature of the disease, longitudinal investigations of cervical neoplastic progression in vivo are difficult to perform as, once detected, the disease is treated immediately. However, *in vitro* models are not subject to the same constraints and allow unique insight into the development of the invasive phenotype — the W12 model is a unique example of such a system. The 'parental' W12 cell line is a polyclonal population of cervical squamous cells that were generated by explant culture of a naturally occurring HPV16-positive cervical LSIL[102]. Growth in monolayer culture restricts W12 cell differentiation and maintains the phenotype of the basal epithelium, the key site of HRHPV transcriptional deregulation in cervical carcinogenesis (Figure 1.2). At early passages, the HPV16 genome is maintained at ~100–200 episomal copies per cell and when grown in organotypic culture recapitulates an LSIL phenotype[90] (Figure 1.6 A). Individual culture series have been established by independent long-term *in vitro* cultivation (Figure 1.6 C); upon continuous passage over 9–12 months, W12 mirrors the virus and host events seen in neoplastic progression in vivo. The most frequent outcome is the breakdown of episomal persistence, with emergence of cells containing ~1–10 copies of integrated HPV16. Integration of HPV16 causes transcriptional deregulation of the virus resulting in an increased level of oncoproteins E6 and E7 as well as genomic instability. In addition, when grown in organotypic culture, the cells progress to HSIL and, eventually, SCC[103, 90] (Figure 1.6 B). Hence, the W12 cell system remains the best available model for HPV16 driven cervical carcinogenesis.

In a previous study, limiting dilution cloning from an early pass of polyclonal parental W12 cells within series 2 was performed under non-competitive conditions. This generated a panel of twenty-four clones that all arose from a common genetic background and differ only by the site of HPV16 genome integration into the host chromosomes. As such, the range of integration events that exist prior to episome clearance and integrant emergence were identified, regardless of whether they had a selective advantage in mixed cell populations[104]. As such, the W12 clones represent

a unique system to examine the host and virus factors that determine selection of a particular HPV16 integrant from the range that exists in a typical polyclonal population of pre-malignant cervical keratinocytes.

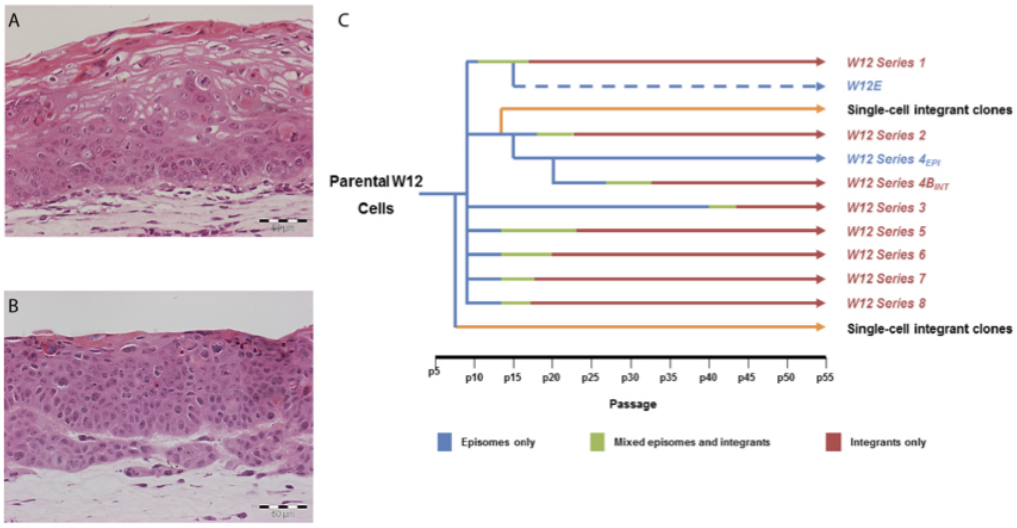

Figure 1.6: The W12 cell system accurately models early cervical carcinogenesis. (A) At early passages, polyclonal W12 cells recapitulate the LSIL from which they were derived. (B) Following long-term passage, they progress through HSIL to SCC. (C) Independent long-term culture series of W12 are characterised by spontaneous emergence of integrant-containing cells, although episome-driven progression may also occur (e.g. Series 4_{EPI}). The 24 W12 clones were generated from W12 Series 2 cells (third row).

1.7 Epigenetic regulation of the eukaryotic genome

For many years cancer research has focussed on the identification of genetic defects leading to carcinogenesis. Following the initial discovery in 1983 of a point mutation in the c-Ha-ras oncogene resulting in a human transforming gene, altered p21,[105] a large number of studies have focussed on the identification of novel DNA mutations associated with tumour cell growth. It is, however, recognised that tumorigenesis is a multistep process characterised by the accumulation of multiple interconnected alterations including genetic, cytogenetic and epigenetic changes[106]. Epigenetics was first described as 'changes in phenotype without changes in genotype'[107], whereby epigenetic mechanisms transduce the inheritance of gene expression patterns without altering the underlying DNA sequence; these mechanisms explain how two identical genotypes can give rise to different phenotypes in response to the same environmental stimulus. Epigenetic regulation of gene expression is a well-established phenomenon that plays a role in a diverse range of biological processes, and its deregulation has been increasingly recognised as a hallmark of cancer[108, 109, 110]. Four distinct mechanisms contribute to the field of epigenetics; DNA methylation, post-translational modification (PTM) of histone proteins, chromatin remodelling and noncoding RNAs[110]. Technological advances such as next-generation sequencing in combination with the development of modification or site-specific antibodies have enabled ChIP-seq analysis of the epigenome at, or near base-pair resolution. Comparisons between control and abnormal cells and tissues have revealed aberrant patterns of epigenetic marks associated with cancer cells[111, 112], as such, they represent new biological targets in the development of cancer therapies[113, 114].

1.7.1 DNA Methylation

DNA methylation is an epigenetic mark involving the covalent addition of a methyl group (-CH_3) to the C-5 position of the cytosine ring of DNA[115]. Methylation of cytosine occurs almost exclusively in the context of CpG dinucleotides and it catalysed by DNA methyltransferases (DNMTs). The maintenance of appropriate DNA methylation plays a significant role in the regulation of a variety of molecular pro-

cesses including the stability of chromosomal structure and the control of gene expression[116]. Due to changes in gene regulation caused by aberrant DNA methylation, this was the first epigenetic mark to be associated with cancer[111]. There are two main classifications of alterations to DNA methylation: hypermethylation, which refers to the gain of methylation at specific sites that are undermethylated under normal conditions and is associated with the stabilisation of transcriptional repression and loss of gene function; and hypomethylation, which is associated with the loss of DNA methylation in genome-wide regions and leads to genomic instability[110]. Interestingly, the two contrasting phenomena coexist in the cancer cell[117]; hypermethylation of CpG islands located in the promoters of tumour suppressor genes results in gene silencing[118, 119], and increased hypomethylation of gene-poor genomic areas such as repeat sequences lead to a higher rate of chromosomal rearrangements[120]. During tumour progression, the degree of hypomethylation of genomic DNA increases as the lesion progresses from benign to an invasive cancer[121]; indeed, the HPV16 genome — predominantly within the L1 and L2 genes — is highly methylated in cancer cells and the methylation status increased with progression from low grade disease to cancer[67]. The methylation state of the LCR is of particular interest as this region regulates the expression levels of the E6 and E7 oncogenes. Each E2-binding site contains two CpGs that can be mehtylated and *in vitro* experimnts have shown that methylation of E2BS CpGs localised withint the promoter prevents E2 binding. This results in the abrogation of E2-mediated inhibition of transcription at p97 promoter contributing to enhanced E6 and E7 expression[67].

1.7.2 MicroRNAs

The expression and activities of cellular proteins are regulated (among other factors) by cellular noncoding microRNAs (miRNAs). MicroRNAs are noncoding regulatory RNA molecules (18—25 nucleotides in length) that are derived from RNA polymerase II transcripts of coding or non-coding genes. MicroRNA expression is often tissue- or differentiation-specific; their temporal expression modulates gene expression at the post-transcriptional level by base-pairing with complementary nucleotide sequences

(seed matching) of target mRNAs[122]. Depending on the degree of sequence complementarity, the binding of miRNAs to a target mRNA inhibits protein translation and/or degrades the target mRNA. MicroRNAs are very often implicated in different stages of cell transformation during carcinogenesis and their altered expression in different cancer types has been considered a marker for diagnosis and therapy[123].

1.7.3 Chromatin structure

The structure of genomic DNA within the nucleus of a cell is intrinsically linked to the level of transcription and resultant gene expression[124]. Within the cell eukaryotic DNA is packaged into a DNA/protein complex called chromatin; it is a highly dynamic and organised structure composed of DNA, histones, and non-histone proteins. The nucleosome is the fundamental unit of chromatin and is composed of 147 base pairs of DNA wrapped around a core of two copies of each H2A, H2B, H3 and H4 histone proteins[125]. Nucleosomes are linked together by 20–80 base pairs of linker DNA. While the core histones are predominantly globular, highly basic histone amino (N)-terminal tails protrude from the nucleosome unit providing a site for enzymatic modification (Figure 1.7)[126].

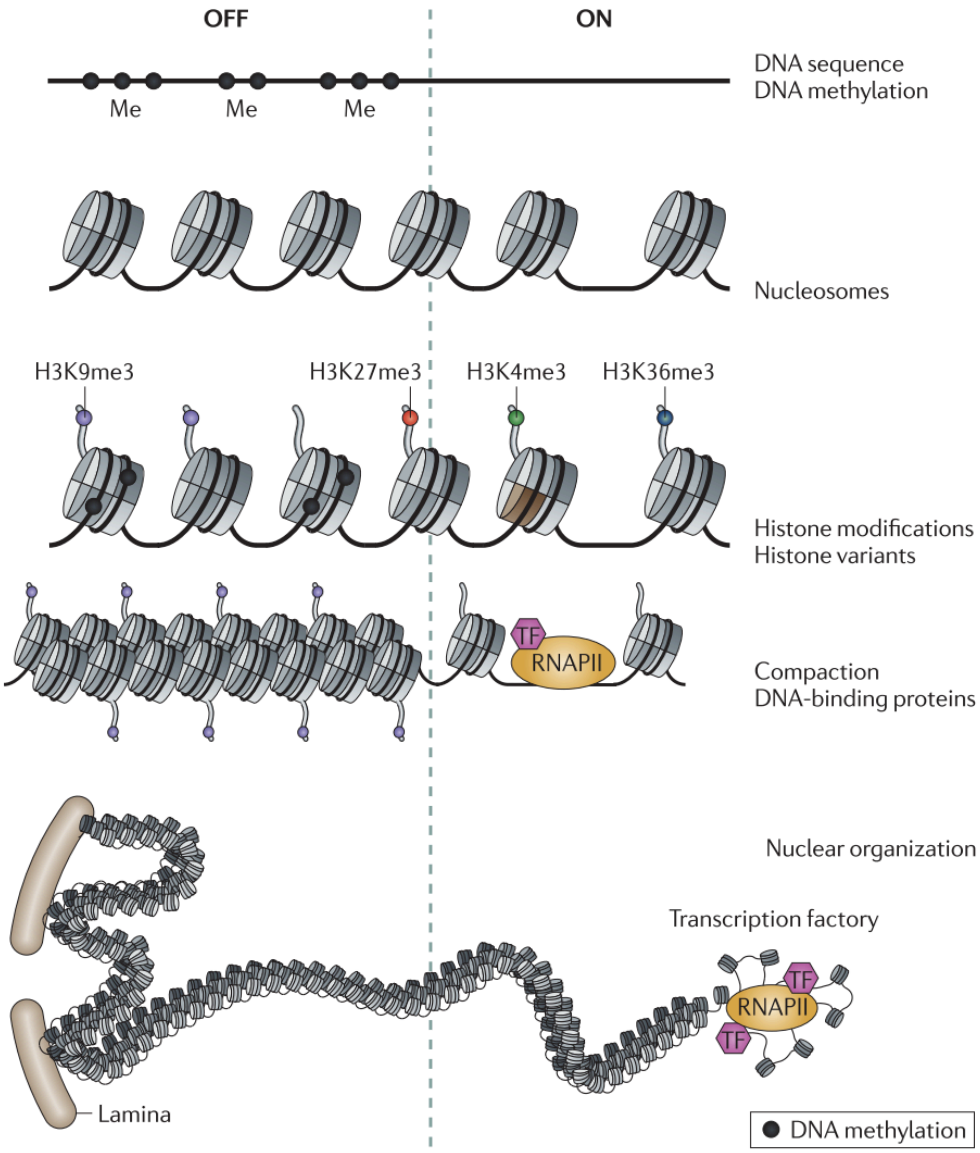

Figure 1.7: Hierarchy of chromatin organisation in the mammalian cell nucleus. From the top, genomic DNA is methylated (Me) on cytosine bases in specific contexts, e.g. promoters, and is packaged into nucleosomes. Nucleosomes vary in histone composition and post-translational modifications (e.g. histone H3 lysine 9 trimethylation (H3K9me3)). DNA in chromatin may remain accessible to DNA-binding proteins such as transcription factors (TFs) and RNA polymerase II (RNAPII) or may be further compacted dependent on the types of post-translational modification present on the histones (Taken from Zhou *et al.*, 2011[124]).

Histone modifications regulate numerous DNA-dependent processes including transcription, DNA replication and DNA repair[127], while the effects of histone modifications operate via two main mechanisms. The first, involving the direct modification(s) of amino acid residues on histone tails, directly influences the overall structure of chromatin. This, in turn, determines the accessibility of genomic DNA to the transcription machinery; there are at least eight distinct types of covalent modifications that can occur at different amino acid residues and at different levels of complexity (mono-, di-, or tri-) resulting in a vast array of modifications[41, 128]. Secondly, specific histone modifications can positively or negatively regulate the binding of effector molecules such as transcription factors[126].

1.7.4 Post-translational histone modifications

Histone modifications have crucial roles in the control of gene activity, DNA repair, DNA replication, chromosome condensation and alternative splicing[67]. Post-translational modifications (PTMs) are highly diverse and reversible, with distinct protein groups responsible for attaching (writers), recognising and binding (readers), and removing (erasers) each histone mark. The writers, readers and erasers of epigenetic marks contribute towards and drive disease via aberrant activity through the mediation of upstream signals as a result of mutation and/or altered expression of epigenetic factors[129]. The acetylation of lysine residues was the first PTM to be discovered[130] and has been shown to be intrinsic to the control of transcription. Acetylation occurs via histone acetyl transferase (HAT) enzymes; acetyl-CoA is used as a cofactor to transfer an acetyl group to the ϵ-amino group of lysine side chains in the histone protein. The covalent attachment of the acetyl group neutralises the positive charge on a lysine residue disrupting the electrostatic interactions between histone proteins and DNA, thus reducing the affinity between them. This results in a less compact chromatin structure allowing for increased access by transcriptional and replication machineries bringing about increased gene expression. Acetylation marks are found at the highest density surrounding transcription start sites (TSSs) of genes[131] as well as within actively transcribed regions of the genome[132, 133]. In ad-

dition, histone acetylation regulates gene expression indirectly through the action of so-called bromodomain and extra-terminal (BET) proteins that recognise and bind to acetylated residues at promoter regions[134]. These additional protein factors are able to activate transcription factors, destabilise nucleosome structure by recruiting additional chromatin remodellers and facilitate the binding of RNA polymerase II immediately upstream of the TSSs[135, 126, 129].

HATs fall into two major categories; type A, are mostly nuclear and have a diverse range of targets when compared with the type B enzymes that are found mostly in the cytoplasm of cells and only acetylate free histones. Type A HATs are classified into three families depending on amino acid sequence homology and conformational structure: GCN5 N-acetyltransferase (GNAT), MYST (MOZ/YBF2/SAS2/TIP60) and CBP/p300 families[136]. The overall levels of histone protein acetylation are controlled by the balance between the activities of HATs and histone deacetylases (HDACs). HDACs are enzymes that catalyse the removal of acetyl groups from histone lysine residues, thereby restoring the positive charge on the side chain of lysine residues and stabilising the local chromatin architecture[137].

In addition to acetylation, methylation — the addition of methyl moieties to the side chains of lysine or arginine residues — is the second most common histone modification that has been clinically associated with pathological epigenetic disruptions in cancer cells[138, 139]. Histone methylation does not alter the charge of the histone protein but rather serves as recognition marks that facilitate or prevent the binding of proteins and protein complexes[140]. Proteins containing PHD-finger domains[141] or the so-called Tudor 'royal' family of domains (Tudor, PWWP, MBT and chromodomains)[142] bind to specific methylation signatures and are able to remodel the structure of chromatin and/or facilitate the recruitment of additional chromatin modifiers that affect the rate of transcription[128, 126]. As with acetylation, the methylation of lysine or arginine residues at specific positions is carried out by enzymatic writers termed histone methyltransferases (HMTs), and are removed by demethylase enzymes; there is an additional level of complexity as these enzymes also modify the appropriate residue to a specific degree, i.e. mono-, di- and/or tri-methyl state.

1.8 Role of HPV16 viral oncogenes in modulating epigenetic mechanisms

The regulation of host gene transcription via epigenetic mechanisms complement those of non-epigenetic nature used by E6 and E7 HPV16 oncogenes in cellular transformation to promote cell proliferation, evade immune response and avoid cell death (see section 1.4.3). E6 and E7 oncoproteins interact with, and/or modulate the expression, of many proteins involved in epigenetic regulation. These include: DNA methyltransferases, histone-modifying enzymes and subunits of chromatin remodelling complexes and the expression of cellular microRNAs (Table 1.1)[67].

Table 1.1: Interactions of HPV oncoproteins with cellular epigenetic modifiers. Reproduced from Durzynska *et al.*[67].

Viral protein	Epigenetic interaction	Description
E6	Induces DNMT1 expression	Maintenance DNA methyltransferase
	Interactions with p300/CBP and inhibits HAT activity	KAT: H3 (K14, K18), H4 (K5, K8, H2A (K5), H2B (K12, K15)
	Binds to and inhibits methyltransferase activity of CARM1 and PRMT1	Protein arginine methyltransferases (PRMTs)
	Inhibits SET7 histone methyltransferase	KMT: H3 (K4)
	Destabilised histone acetyl transferase TIP60	KAT: H4 (K5, K8, K12, K16), H3 (K14)
E7	Binds DNMT1 and stimulates DNA methyltransferase activity	Maintenance DNA methylatransferase
	Induces DNMT3A and DNMT3B expression	*De novo* DNA methyltransferases
	Interacts with p300/CBP	KAT
	Interacts with PCAF histone acetyltransferase and reduces its activity	KAT: H3 (K9, K14, K18)
	Binds Mi2β and HDACs 1 and 2	Subunits of NuRD ATP-dependent remodelling complex
	Interacts with BRG1	A component of thehuman SWI/SNF complex
	Induces KDM6A and KDM6B histone demthylase expression	KDM: H3 (K27)
	Interacts with E2F6	A component of the PcG complexes
	Induces expression of histone methyltransferase EZH2	KMT: H3 (K27)

1.8.1 HPV16 and DNA methylation

Aberrant DNA methylation of the cellular genome has been observed in various types of HPV-associated cancers[67]; of particular interest to HPV-driven carcinogenesis if the aberrant DNA methylation in a number of tumour suppressor genes. Examples include: hypermethylation of the gene *CCNA1* which increases from 0% in the disease free control group to 93% in cervical carcinoma patients[143]; increases in a number of methylated sites in the promoter of the *hTERT* gene, where repressive sequences were blocked by methylation increasing the *hTERT* production in HPV-

dependent cells and cervical cancer cells[144]. In addition, following virus integration into the host genome, the virus is able to modulate host methylation machinery to regulate the expression of viral oncogenes genes in favour of the establishment of persistent infection by evading the host immune defence[145].

1.8.2 HPV16 and microRNAs

In HPV-dependent cancer cells, both E6 and E7 oncoproteins can influence cellular miRNAs, which are regulated in most cases by transcription factors eg. c-Myc, p53 and E2F. Examples of HPV16 oncogene-induced down-regulation of microRNAs include: E6 expression results in the down-regulation of miR-34a expression via the p53 protein, this leads to increased cell proliferation[137]; E7 expression down-regulates miR-203 expression — high levels of miR-203 are inhibitory to HPV amplification — facilitating productive viral replication in differentiating cells[146]. However, as a result of HPV16 infection the expression of microRNAs can also be increased. MicroRNA 21 — which targets the CCL20 gene — is significantly overexpressed to regulate cellular processes including proliferation, apoptosis and migration of HPV16-positive cervical squamous cells[147]. Not only does HPV16 alter the expression of cellular microRNAs, but the miRNAs may also target HPV transcripts and change the expression of papillomavirus genes in a differentiation-dependent manner[122].

1.8.3 HPV16 and histone-modifying enzymes and chromatin remodelling complexes

HPV16 E6 and E7 oncoproteins have been shown in interact with cellular epigenetic modifiers — a full list is reported in Table 1. Specifically the interaction of HPV16 oncoproteins with HATs has been shown to affect the activity of these enzymes toward non-histone substrates and to modulate their transcriptional coactivator activity; one example includes the HAT p300/CBP. HPV16 E6 oncoprotein inhibits the p300/CBP-mediated acetylation of p53 and through this mechanism represses p53-dependent gene activation[53], whereas HPV16 E7 forms a tertiary complex with both p300/CBP and pRb which promotes acetylation of pRb resulting in the decrease of

pRb expression[148]. E7 oncoprotein also interacts with p300/CBP-Associated Factor (PCAF) HAT and reduces its ability to acetylate free histones *in vitro*[149]. HPV16 E7 expression also causes epigenetic reprogramming of cells by modulating the activities of histone deacetylases (HDACs) and DNA methyltransferases (DNMTs). E7 expression has been shown to stimulate the activity of DNMT1 *in vitro*[150] as well as the lysine demethylase KDM6B[151]; induction of these enzymes results in a reduction of global levels of the repressive H3K27me3 mark and consequently the loss of polycomb repressive complex (PRC)-mediated repression. Additionally, E7 is able to bind and sequester HDACs, resulting in the activation of several cellular promoters. Examples include the enhanced expression of E2F by disruption of the repressive pRb-HDAC complex responsible for regulation of E2F S-phase specific genes, leading to increased cell proliferation[152], as well as increased activity of hypoxia inducible factor 1 (HIF-1), by inducing the dissociation of HDACs 1, 4, and 7 from the HIF-1a protein subunit, resulting in increased levels of angiogenesis[47].

1.9 Deregulated epigenetic regulation and cancer

Epigenetic modifications do not act in isolation but rather in combination with one another to determine the correct chromatin conformation and levels of accessibility to ensure the required levels of gene expression is achieved. For example, the repression of gene promoters by CpG hypermethylation is also associated with additional repressive PTMs such as the deacetylation of histones H3 and H4, loss of H3K4me3 and gain of H3K9me and H3K27me3[153, 154] adding further complexity to gene regulation.

Deviations and aberrant patterns of histone modifications are a hallmark of cancer. Alterations to the methylation and acetylation status of lysine residues on histone 3 (H3) have been indicated in prostate cancer[139] and have also been shown to have prognostic relevance for patients with non-small cell lung cancer[155]. Moreover, global levels of specific PTMs have been used to identify patients with low-grade bladder cancer[156] and used as independent predictors of renal cell carcinoma mortality[112]. In combination with the fact that epigenetic modifications are re-

versible, allowing the malignant cell population to revert to a more normal state, the epigenome has increasingly been targeted as a means of effective chemotherapy as well as chemoprevention of cancer. There are numerous epigenetic modifiers, namely DNA-methylation and HDAC inhibitors, in different stages of clinical trials[157]. Current examples already used in the clinic are HDAC inhibitors Zolinza® (virinostat) and Istodax® (romidepson) used for the treatment of T cell lymphoma[158] whilst Vidaza® (azacitidine) is used for the treatment of leukaemias[159]. For the treatment of cervical cancer, enzymatic inhibitors of DNA methylation and histone deacetylases have shown limited promise in phase I and II clinical trials, however their efficacy is increased when combined in addition to radiation or chemo-radiation therapy[41].

1.10 RNA polymerase II-dependent transcription

The transient modulation of chromatin structure by epigenetic mechanisms is required to facilitate the binding of the transcription machinery to DNA. In addition to epigenetic regulation, gene expression is also controlled at the level of transcription elongation by RNA polymerase II (RNAPII). Following binding of the general transcription factors (GTFs) to the transcriptional start sites (TSSs) of genes, RNAPII molecules are recruited and transcription is initiated upon phosphorylation of Serine 5 (Ser5) on the carboxy-terminal domain (CTD) of RNAPII by cyclin-dependent kinase 7 (CDK7)[160]. RNAPII molecules are able to transcribe a short distance (20–50 bp) before entering a paused state, which is controlled by the pause control elements: DRB-sensitivity-inducing factor (DSIF) and negative elongation factors (NELFs), both of which are physically associated with the paused RNAPII molecules[161]. Pause release and subsequent elongation requires the action of positive elongation factor b (P-TEFb), composed of cyclin-dependent kinase 9 (CDK9) and cyclin T1. The CDK9 subunit of P-TEFb mediates the release of paused RNAPII by phosphorylating the pause control factors as well as the second serine residue (Ser2) on the CTD of RNAPII; the latter mark is associated with active elongation of the polymerase and results in processive mRNA production (Figure 1.8)[162]. Studies carried out on other human viruses including HSV, CMV, EBV and HIV have shown

that the recruitment and function of P-TEFb to paused RNAPII molecules is a necessary step for active transcription[163] and, therefore, the components of the P-TEFb complex such as CDK9 are possible therapeutic targets. In recent years a number of small molecule CDK9 inhibitors such as flavopiridol, dinaciclib and seliciclib have been designed that demonstrate good antitumoral activity *in vitro*. However, thus far, when tested in phase I clinical trials, their lack of specificity against other CDKs and enzymes results in adverse events that make them unsuitable for clinical use[164]. Despite this, pharmaceutical companies designing CDK9 inhibitors have the tools required to improve the selectivity of CDK9 inhibitors and therefore they remain a promising future therapeutic option.

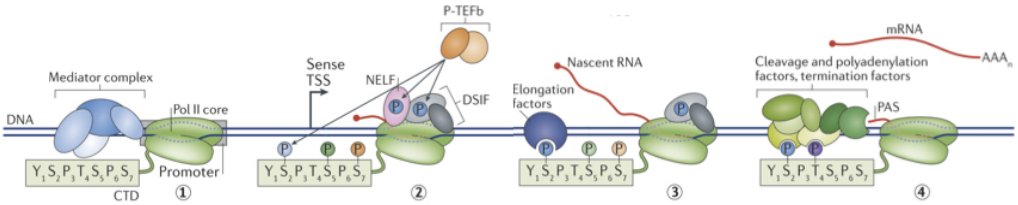

Figure 1.8: Transcription regulation by the RNAPII CTD code. Step 1) Recruitment of the core Pol II enzyme to the gene promoter with an unphosphorylated CTD through interaction with the Mediator complex. Pol II escapes the promoter during the initiation phase upon phosphorylation of Ser5 of the CTD by CDK7. Step 2) Pol II during promoter-proximal pausing; the arrival of P-TEFb leads to the phosphorylation of NELF, DISF and Ser2, which leads to productive transcription elongation shown in step 3. Step 4) Pol II transitions from transcription elongation to termination (Taken from Harlen *et al.*, 2017[165]).

In mammalian cells, more than half of all P-TEFb molecules are sequestered within an inactive complex containing the 7SK small nuclear ribonucleoprotein particle (snRNP) and the HEXIM1 protein[166]. However, in conditions that require rapid transcriptional induction, P-TEFb is released from the inactive complex[167] and is recruited to paused RNAPII molecules for subsequent activation. The recruitment of P-TEFb can occur via a number of different mechanisms; in the form of a large complex called the super elongation complex (SEC)[168], by specific DNA-binding transcription factors such as c-Myc[169] and/or through its interaction and association with the bromodomain-containing protein 4 (BRD4)[170]. In the case of c-Myc-mediated recruitment of P-TEFb, elevated levels of the transcription factor

accumulate in the promoter regions of most active genes. P-TEFb is subsequently recruited resulting in transcriptional amplification; thus, rather than binding and regulating a new set of genes when overexpressed, c-Myc amplifies the output of the existing gene expression program[171]. BRD4 is a member of the bromodomain and extra terminal domain (BET) family, and is a reader of acetylated lysine residues on histone tails. As previously discussed, H3K27ac is a mark of active chromatin and is found at the TSS regions of genes. BRD4 mediates the recruitment of P-TEFb to paused RNAPII molecules by binding the core CDK9/cyclin T1 complex through its C-terminal extra terminal domain region[172]. BRD4 and resultant P-TEFb recruitment is triggered by increased histone acetylation at, and near, the TSS. This leads to the active transcription of genes and once again illustrates the intricacy of epigenetic regulation. In addition to small molecule inhibitors of CDK9, pharmacological inhibition of BET proteins, including BRD4, have also been shown to have therapeutic activity in cancer models; inhibitors JQ1 and I-BET function by competitively binding to the bromodomain pockets of the BET proteins resulting in their displacement from the acetylated chromatin[173]. Adding to the complexity of the regulation of gene transcription, elongating RNA polymerase II is also able to trigger epigenetic cross talk between HMTs and DNA-methyltransferase enzymes to ensure the fidelity of gene transcription initiation[174].

1.11 Rationale and aims of the investigation

The overall aim of this study was to determine epigenetic mechanisms of HPV16 and host gene transcriptional deregulation following virus integration during early cervical carcinogenesis. The association between integration of the HPV16 genome into the host and the severity of disease has been widely commented on, and is known to be the major risk factor associated with disease progression. While a number of studies have identified HPV integration sites present in cell lines derived from advanced cancers, very little is known about how or why particular integration events are selected for amongst a pre-malignant polyclonal population of cells.

The first part of this investigation aimed to ascertain epigenetic mechanisms that directly control the level of transcription of the HPV16 genome, particularly the oncogenes E6 and E7. The epigenome is controlled by a multitude of enzymatic processes, the deregulation of which is a recognised hallmark of cancer[109, 110]. The development of small molecule inhibitors of enzymes that play an integral part in the regulation of the epigenome is currently an active area of pharmaceutical research. An increasing number of novel therapeutic strategies are being designed to reverse the abnormalities that are inherent to the cancer epigenome[175]. As such, findings that suggest that the expression level of the HPV16 oncogenes is determined by epigenetic mechanisms may indicate that cervical carcinomas could also be treated in a similar manner.

To address this aim I have used a panel of W12 integrant clones that exhibit different levels of HPV16 oncogene expression per template and have less than four copies of the virus genome to compare the abundance of epigenetic modifications — namely, post-translational modifications of histone tails and active RNA polymerase II — at the virus genomes. Furthermore, the functional significance of enzymes responsible for laying down the epigenetic marks has been evaluated and their impact on the levels of oncogene expression determined.

The second aim of the investigation was to determine whether HPV16 integration in pre-malignant, unselected cells results in the aberrant expression of host genes. A number of studies have used advanced cervical cancer cell lines to demonstrate

that HPV integration can cause a wide variety of somatic mutations, genomic amplifications and rearrangements resulting in the disruption of cellular genes ([176, 97]. However, it remains unknown whether this phenomenon occurs as a result of all integration events or whether it is seen only in cells that are ultimately selected for.

In order to address this aim, the virus-host breakpoints of each of the W12 five clone integrant panel were accurately mapped using capture-sequencing technology and, as a result, the mechanism of HPV16 integration hypothesised. Additionally, interactions between the integrated virus and the host were analysed to further elucidate regulation mechanisms of HPV16 transcription.

1.11.1 Hypothesis

With regard to the epigenetic landscape analysis conducted in this thesis, I hypothesise that the epigenetic marks asscociated with the integrated HPV16 genome — particularly at the 5' LCR and at the site of the p97 promoter — will be reflective of the levels of viral oncogene expression per template in the W12 clones tested in this thesis. Specifically, I hypothesise that the abundance of active PTMs will be more abundant at the LCR of highly expressing clones compared with low expression clones, and that the opposite will be true regarding repressive PTMs. Additionally, I hypothesis that the abundance of RNAPII and its various forms will also be predictive of viral oncogene expression in the W12 clones.

With regard to HPV16 integration, I hypothesise that in the W12 clones the virus integrates into gene-rich regions of the host and most likely within a host gene itself. In active regions of the genome the chromatin structure is relatively open when compared to gene-poor regions to facilitate cellular machinery for transcription and is therefore more susceptible to integration of the HPV16 genome. In addition, the integration into an active host gene body means that host transcription machinery will be in close proximity to the virus genome; dramatically increasing the likelihood of HPV16 oncogenes being transcribed. I hypothesise that HPV16 integration with have an effect on host gene expression, particularly of genes adjacent to the integrated virus genome but that the host genes disrupted by HPV16 integration will not make

a significant contribution to cell phenotype.

1.11.2 Published work

Results from Chapter 3 were published in Groves *et al.* Oncogene, 2016. Results presented in this thesis are my own and have not been contributed to or supplemented by any other authors on the published paper.

Chapter 2

Materials and Methods

2.1 Cell culture and cell treatments

2.1.1 Cell lines and cell culture maintenance

Three cells lines derived from cervical tissue were used in this study (Table 2.1), all were grown in monolayer culture to mimic the basal layer of the epithelium. All cell lines were grown in complete culture medium: Glasgow Minimum Essential Medium (GMEM) supplemented with 10% (w/v) Fetal bovine serum (FBS) (Sigma-Aldrich, St. Louis, U.S.), 2 mM L-glutamine (Thermo Fisher Scientific, Loughborough, UK) and 100 U/ml of penicillin and 100 μg/ml of streptomycin (Thermo Fisher Scientific). Cells derived from the W12 keratinocyte cell line were co-cultured with X-ray irradiated G3T3 feeder cells (Todaro and Green 1963) and grown in complete medium as previously stated with the addition of 10^{-10} M cholera toxin (Sigma-Aldrich), 0.5 μg/ml hydrocortisone (Sigma-Aldrich), with 10 ng/ml epidermal growth factor (EGF) (Sigma-Aldrich) added 24 hours after seeding. All cells were tested fortnightly for mycoplasma contamination using Mycoplasma Plus^{TM} PCR Primer Set (Agilent Technologies, Santa Clara, U.S.).

Table 2.1: Details of cell lines used in this study.

Cell Line	Origin	HPV status	Description	Culture medium
W12 parental cell line	Cervix	Episomal HPV-16[A] [~100-200 copies per cell]	Human ectocervical keratinocyte cell line generated from L-SIL.	W12 medium EGF- W12 medium EGF+
W12 integrant clones	Cervix	Integrated HPV16[B] [1-5 copies per cell]	Clones with different HPV16 integration sites isolated from a nonclonal population of W12.	W12 medium EGF- W12 medium EGF+
SiHa*	Cervix	Integrated HPV-16[CD] [~2 copies per cell]	Human keratinocytes cell line derived from tissue from SCC.	GMEM
NCx/6	Cervix	HPV-negative	Normal human ectocervical keratinocyte primary cell line.	GMEM
G3T3	Mouse fibroblasts	HPV-negative[E]	G418-resistant Swiss albino mouse fibroblasts	GMEM

A - M Stanley *et al.*, 1989[102]; B Dall *et al.*, 2008[104]; C - Friedl *et al.*, 1970[177]; D - Baker *et al.*, 1987[178]; E - Todaro & Green, 1963[179]; *Authenticated by short tandem repeat profiling by the American Type Culture Collection (Manassas, Va, USA)

2.1.2 Resuscitation of established cell lines from liquid nitrogen

Cells cryopreserved in freezing medium [90% (v/v) foetal calf serum (FCS), 10% (v/v) dimethyl sulfoxide (DMSO)] were thawed rapidly in a 37 °C water bath. The cryovial was decontaminated with 70% ethanol and transferred to a class II laminar flow hood. The cell suspension was transferred into a 15 ml falcon tube containing 10 ml of cell type-dependent culture medium equilibrated to 37 °C, 5% carbon dioxide (CO_2) and spun at 600 RCF for 5 minutes. The supernatant was aspirated and the cell pellet resuspended in 3 ml of the appropriate equilibrated medium. The suspension was then used to seed a 10 cm^2 dish (W12) or 75 cm^2 tissue culture flask (SiHa and G3T3) and cells were incubated at 37 °C, 5% CO_2. For resuscitated W12 and NCx/6 cells, 3 x 10^6 irradiated G3T3 feeder cells were added before incubation.

2.1.3 Subculture of cell lines

For the subculture of cell lines, cells were grown at 37 °C in 5% CO_2 and passaged at approximately 80–90% confluency. Culture medium was aspirated and the cells washed once with sterile phosphate-buffered saline (PBS) at room temperature (RT); the subculture of W12 and NCx/6 cells required an additional PBS wash step whereby the G3T3 feeder cells were removed by repeated spraying of the culture dish surface with PBS using a 5 ml glass pipette and then aspirated. Following the wash step, the PBS was aspirated before the addition of pre-warmed 0.5% Trypsin-Ethylenediaminetetraacetic acid (Trypsin-EDTA Thermo Fisher Scientific) to the cell monolayer; the volume of Trypsin-EDTA used was dependent on flask size (Table 2.2). Cells were then incubated at 37 °C in 5% CO_2 for approximately 5 minutes until they had detached from the culture surface. The trypsin-cell suspension was transferred to a 15 ml falcon tube and the culture surface rinsed with 5 ml of pre-warmed complete growth medium, which was also added to the cell suspension. The trypsin-cell suspension was fully neutralised by the addition of a further 5 ml culture medium and mixed by inversion. The suspension was then spun at 600 RCF for 5 minutes, the supernatant aspirated and the pellet resuspended using a 5 ml glass

pipette in 5 ml of culture medium to generate a single cell suspension. The number of cells per ml was determined using a haemocytometer and cells were passaged at the appropriate density into a fresh tissue culture vessel containing pre-warmed complete growth medium (see Table 2.2) and placed at 37 °C, 5% CO_2. If the culture was G3T3 supported, X-ray irradiated G3T3 feeder cells were added at appropriate concentrations (Table 2.2) before incubation. If W12 culture medium was used, 24 hours after initial plating, the medium was further supplemented with EGF at 10 ng/ml. Culture medium was subsequently changed every 2-3 days.

Table 2.2: Cell culture seeding densities and volumes

Culture vessel	Seeding densities		Vessel volume	Volume of Trypsin-EDTA
	Cell lines	Irradiated G3T3		
175 cm^2 flask	$9 \times 10^5 - 1.8 \times 10^6$	5×10^6	25 mL	5 mL
75 cm^2 flask	$4 \times 10^5 - 8 \times 10^5$	2×10^6	10 mL	3 mL
15 cm dish	$7.5 \times 10^5 - 1.5 \times 10^6$	6×10^6	25 mL	5 mL
10 cm dish	$5 \times 10^5 - 8 \times 10^5$	2×10^6 [3×10^6 for W12 resuscitation]	10 mL	3 mL
6-well plate	$5 \times 10^4 - 1.2 \times 10^5$	2.5×10^5	2 mL	1 mL
12-well plate	$1 \times 10^4 - 2 \times 10^4$	1×10^5	1 mL	0.5 mL

2.1.4 Subculture and X-ray irradiation of G3T3 cell line

The G3T3 cell line (mouse fibroblasts) was used to generate the feeder cells for monolayer culture of W12 and NCx/6 populations. G3T3 cells were grown in complete GMEM culture medium at 37 °C, 5% CO_2, and sub-cultured as described in section 2.1.3. Following sub-culture, the cells were lethally X-ray irradiated (18 minutes X-ray irradiation at 420 rads/ minute) before being used as feeder cells in monolayer culture.

2.1.5 Cryopreservation of cell cultures

Cells at 80–90% confluency were washed with PBS, trypsinised and centrifuged at 600 RCF for 5 minutes. The cell pellet was then resuspended in PBS and the

concentration of the cell suspension determined using a haemocytometer. The cell suspension was re-centrifuged at 600 RCF for 5 min and resuspended in freezing medium [90% (v/v) FCS, 10% (v/v) DMSO] to achieve a concentration of 2 x 10^6 cells/ml. Cryovials containing 1 ml of the resuspended mix were placed at -80 °C before being transferred to liquid nitrogen for long-term storage.

2.1.6 Cell treatment with small molecule inhibitors

Each drug (Sigma, unless specified) was dissolved in DMSO to generate a stock solution of desired concentration and stored at -20 °C. Before use, the drug stock solutions were diluted 1:500 in culture medium to give the required concentrations for treatment (Table 2.3). When cells seeded 1-2 days prior reached 50-60 % confluency the original cell medium was replaced with drug-supplemented medium. The plates were then incubated for 16 hours at 37 °C, 5% CO_2 before downstream processing. Negative control cells were treated with equivalent volumes of DMSO vehicle (vol/vol) to compensate for any DMSO-mediated effects on HPV16 transcript levels; all qPCR results were normalised to this control.

Table 2.3: Drugs used for pharmacological inhibition of chromatin modifiers

Drug	Cellular target	Supplier	Tested concentrations
DMSO *Dimethyl sulphoxide*	n/a	Sigma-Aldrich	Variable (control)
C646 *4-[4-[[5-(4,5-Dimethyl-2-nitrophenyl)-2-furanyl]methylene]-4,5-dihydro-3-methyl-5-oxo-1H-pyrazol-1-yl]benzoic acid*	p300	SML0002; Sigma-Aldrich	1 – 25 μM
MG149 *2-[2-(4-Heptylphenyl) ethyl]-6-hydroxy-benzoic acid*	TIP60	Axon 85; Axon Medchem	50 – 150 μM
Flavopiridol *Flavopiridol hydrochloride hydrate*	CDK9	F3055; Sigma- Aldrich	150 nM*
Trichostatin A (TSA) *[R-(E,E)]-7-[4-(Dimethylamino)phenyl]-N-hydroxy-4,6-dimethyl-7-oxo-2,4-heptadienamide*	HDAC class I/II	T95; Sigma- Aldrich	400 nM*

*Determined previously by the lab

2.1.7 Cell transfection

Targeted gene knockdown was carried out using small interfering RNAs (siRNAs); each target gene was depleted using human FlexiTube siRNAs (Qiagen, Crawley, UK). All siRNAs were used at 10 nM with cells being transfected using Lipofectamine RNAiMAX (Invitrogen, Paisley, UK) (Table 2.4).

Table 2.4: siRNA used for knockdown experiments

Gene target	Reference	Target sequence (5′ → 3′)
p300	EP300_7 SI02626267	TTGGACTACCCTATCAAGTAA
TIP60	KAT5_2 SI05120304	CCGGGCTCAGACCAACTCCAA
CDK9	CDK9_5 SI00605066 CDK9_6 SI00605073	TAGGGACATGAAGGCTGCTAA TGGGCACAGTTTGGTCCGTTA
Non-targeting control	AllStars Negative Control siRNA, 1027280	N/A

48 hours prior to transfection, cells were trypsinised and seeded into 6-well plates (for protein extraction) or 12-well plates (for RNA extraction) using cell-type dependent media (Table 2.5). 24 hours after seeding, medium was changed on all wells and replaced with the appropriate medium without antibiotics. After a further 24 hours, and with the cells at approximately 20–30% confluency, the cells were washed with PBS (in the case of W12 feeder cells were washed off thoroughly) and the appropriate volume of culture medium without antibiotics placed on the cells and incubated at 37 °C, 5% CO_2 before the cells were transfected with the gene specific or control siRNA. Briefly, a mastermix was made containing Lipofectamine RNAiMAX diluted in OptiMEM® (Thermo Fisher Scientific) and incubated at RT for 5 minutes (Solution 1). During the incubation, siRNA (Qiagen) was added at the desired concentration to OptiMEM® and incubated at RT for 5 minutes (Solution 2) (Table 2.5). The contents of both tubes were then gently mixed and incubated for 20 minutes at RT. The siRNA:Lipofectamine complex was pipetted slowly into each well and gently mixed with the culture medium covering the cells by rocking the plate. The cells were then incubated at 37 °C, 5% CO_2 for 12 hours.

Following the 12-hour transfection, the media:transfecting agent mix was aspirated from each well and the cells washed once with PBS. This was replaced with

fresh pre-warmed cell-type dependent culture medium without antibiotics, and for W12 cells medium supplemented with irradiated feeder cells (Table 2.5). 24 hours after transfection the media was changed again to cell-type dependent media supplemented with antibiotics and incubated at 37 °C, 5% CO_2. Cells were harvested 48 hours after initial transfection.

Table 2.5: Details for 10 nM siRNA transfection

	12-well	6-well
Cell Density		
W12 cell lines (G3T3)	2 x 10^4 (1 x 10^5)	5 x 10^4 (2.5 x 10^5)
SiHa cell line	1 x 10^5	2 x 10^5
Culture media	800 µl	2500 µl
Solution 1		
OptiMEM® (sol^n 1)	100 µl	250 µl
Lipofectamine RNAiMAX	4 µl	6 µl
Solution 2		
OptiMEM® (sol^n 2)	100 µl	250 µl
siRNA (20 µM)	0.5 uL	1.5 uL
Total volume	~1000 µl	~3000 µl

2.2 DNA analysis

2.2.1 DNA extraction

Genomic DNA (gDNA) was extracted from freshly pelleted cells from monolayer culture. Cell pellets were resuspended in an appropriate amount of lysis buffer (10 mM Tris-Cl pH 8, 10 mM EDTA pH 8, 150 mM NaCl, 0.4% SDS) and 1 mg/ml Proteinase K before being incubated for 10 minutes at 55 °C. The tubes were then incubated for 16 hours at 37 °C in a water bath. If necessary, PBS was added to viscous samples following Proteinase K digestion, before being transferred to phenol resistant tubes. An equal volume of 1:1 Tris-saturated phenol:chloroform was added and the samples and inverted to ensure thorough mixing. The tubes were then centrifuged at 12,000 x g for 10 minutes. The upper aqueous phase was decanted into fresh tubes and the phenol:chloroform extraction repeated. The upper phase

was then removed to fresh tubes and $1/10^{th}$ total volume of 3 M sodium acetate (pH 5.2) was added. Two times the volume of ice-cold 100% ethanol was added and the gDNAs were allowed to precipitate overnight at -20 °C. The gDNAs were pelleted by centrifugation at 12,000 x g for 30 minutes before being washed in 1 ml of cold 70% ethanol and being transferred to 2 ml Eppendorf tubes. Following a 10 minute centrifugation at 12,000 x g, the ethanol was removed and the gDNA pellet air-dried, before resuspension in 50 μl double distilled water (ddH_2O) and the concentration determined using a spectrophotometer by measuring absorbance at 260nm (Nanodrop 2000, Thermo Fisher Scientific) and stored at 4 °C.

2.2.2 Polymerase chain reaction of HPV16 long control region (LCR)

PCR reactions were conducted to obtain a suitable concentration of DNA to enable sequencing of the LCR DNA fragment of choice (20 ng/100bp in 10μl) and were performed using a GeneAmp® PCR System 9700. A range of primers spanning the HPV16 LCR were used in the following PCR reactions (Table 2.6). The reagents for PCR were combined in 0.2 ml non-flex PCR tubes (Starlab); 43μl SuperMix, 1 μl 10μM primer mix, 4 μl 25ng/μl DNA and 2μl 2.5mM $MgCl_2$. The reaction was then carried out using the following cycling conditions: 2 min at 94 °C; 50 cycles of 94 °C for 30 seconds, 55 °C for 30 seconds, and 72 °C for 1 minute; 72 °C for 5 minutes before cooling to 4 °C. The resultant PCR samples were kept at 4 °C until further use. Both positive and negative controls for the PCR reaction were conducted; the positive control mix contained E6/E7 primers, and a PCR mix without DNA and another without primers were used as negative control reactions. The PCR products were then mixed with 5 μl of Orange G (SigmaAldrich (O3756)) and run on a 1% agarose gel alongside 100 bp and 1 kb ladders (New England Biolabs). The resultant gel was visualised using a Quantity One UV machine (Bio-Rad).

Table 2.6: Primers used for PCR analysis of LCR of HPV16 positive gDNA samples

Name	Forward primer (5'→3')	Name	Reverse primer (5'→3')	Reference
E6 F	GCACCAAAAGAGAACTGCAA	E7 R	GATTATGGTTTCTGAGAACAGATGG	IJG design
7045 F	ACAAGCAGGATTGAAGGCCAAACCA	223 R	ACGTCGCAGTAACTGTTGCTTGCA	IJG design
7045 F	ACAAGCAGGATTGAAGGCCAAACCA	7681 R	CGTTGGCGCATAGTGATTTA	IJG design
7555 F	CCAAATCCCTGTTTTCCTGA	223 R	ACGTCGCAGTAACTGTTGCTTGCA	IJG design

2.2.3 Polymerase chain reaction of virus-host breakpoints

The PCR reactions were conducted to obtain a suitable concentration of DNA to enable sequencing of the hybrid virus-host fragment (20 ng/100 bp in 10μl) and were performed using an MJ Research® PTC-255 Thermocycler. A range of primers were used dependent on the W12 clone breakpoint as identified by the capture sequencing reaction (Table 2.7). The reagents for PCR were combined in 0.2 ml non-flex PCR tubes (Starlab); 2.5 μl Buffer 10X (Sigma), 2.5 μl 2 μM dNTP mix, 2.5 μl 2 μM of both the forward and reverse primers, 2 μl 12.5 ng/μl DNA, 0.5 μl Taq DNA polymerase with $MgCl_2$ (Sigma, D9307) and 12.5 μl ddH_2O. The reaction was then carried out using the following cycling conditions: 2 minutes at 94 °C; 50 cycles of 94 °C for 30 seconds, 50.2–60 °C for 30 seconds, and 72 °C for 2 minutes; 72 °C for 5 minutes before cooling to 4 °C. The resultant PCR samples were kept at 4 °C until further use. Both positive and negative controls for the PCR reaction were conducted; the positive control mix contained E6/E7 primers, and a PCR mix without DNA was used as a negative control. The PCR products were then mixed with 2.5 μl of Orange G (SigmaAldrich (O3756)) and run on a 1% agarose gel alongside a 100 bp ladder (New England Biolabs). The resultant gel was visualised using a Quantity One UV machine (Bio-Rad).

Table 2.7: Primers used for PCR analysis virus-host breakpoints

	Forward sequence (5'→3')		Reverse sequence (5'→3')		
	Species		Species		
3' F (A5)	HPV16	AAGGGCCCTAGCAGGTTTTA	Host	CACCGAAGAAACACAGACGA	EK design
5' F (A5)	Host	TGGTCACGTTGCCATTGACT	HPV16	GGGCAGTGTGGCAGTAGTTA	EK design
3' H	HPV	GCACCGAAGAAACACAGACG	Host	TCCTTCCCTCCCTAACAGCAT	EK design
5' H	Host	TGGGTCAGTGGTTTGATTGA	HPV	TGGGGATCCTTTGCCCCAGTGT	EK design
3' D2	HPV16	ACTGTGGTAGAGGGTCAAGT	Host	GGGGAAGGTGGCATCTCTTA	EK design
5' D2	Host	CTTTGCCACGGGACAAGTAT	HPV16	GGATCGGAAGGGCCCACAGGA	EK design
3' G2	HPV16	CAGTGCCCTGTTGGAACTACA	Host	GCTGTTGACCTCTTTGGGGT	EK design
5' G2	Host	TTACTCATGCCACCACACCT	HPV16	CAACTTGACCCTCTACCACAGT	EK design
Positive control	E6 F	GCACCAAAGAGAACTGCAA	E7 R	GATTATGGTTTCTGAGAAC	IJG design

2.2.4 DNA gel extraction

Specific DNA bands identified by 2.2.2 and 2.2.3 were cut out and gel purified in order to isolate the desired DNA fragment required for sequencing. Extraction of the DNA from the agarose gel was completed using the QIAquick® Gel Extraction Kit (Qiagen) according to the manufacturer's instructions. Briefly, the excised band was placed in a 1.5 ml Eppendorf tube and the mass of the gel fragment determined. 3x gel volumes of Buffer QG was added to each tube, where 100 mg = 100 μl, and the sample incubated at 50 °C for 10 min. Once dissolved, 1x gel volume of isopropanol was added and mixed. The sample mixture was then applied to a QIAquick column in a 2 ml collection tube and placed in a centrifuge and spun at 17,900 x g for 1 minute. The flow through was discarded and the column washed with 500 μl Buffer QG followed by 750 μl Buffer PE. The column was then placed into a clean microcentrifuge tube and DNA eluted in 30 μl ddH_2O. The concentration of DNA was then determined using a Thermo Scientific Nanodrop 2000 Spectrophotometer. The samples were then placed in a -20 °C freezer until further use. The DNA sequencing was completed by the Department of Biochemistry, University of Cambridge according to the specifications of the institute (DNA at 20 ng/100 bp in 10 μl).

2.2.5 DNA sequencing analysis

The resultant sequencing files were viewed using DNA sequencing software ChromasPro (Technelysium). Chromatograms were used to assess the quality of the sequence data and the alignment tool was used to compare samples as well as to detect any changes to the expected sequence.

2.2.6 quantitative-PCR (qPCR) analysis of gDNA samples

To quantify the level of specific hybrid virus-host genomic sequences between samples, qPCR was used. qPCR was performed using 2 μl of 12.5 ng/μl gDNA (25 ng of gDNA per qPCR reaction), 10 μl of qPCRBIO SyGreen Mix (PCRBiosystems) and 2 μl of 4 μM of forward and reverse primer pair mix (Table 2.8) in a final reaction volume of 20 μl/well. Test reactions and no template controls were run in triplicate on a realplex real-time PCR system (Mastercycler® Eppendorf). The cycling conditions consisted of a 2 min initial denaturation at 95 °C, followed by 40 cycles at 95 °C for 15 sec, 58 °C for 20 sec, 72 °C for 15 sec and 76 °C for 5 seconds. Fluorescence was measured at the last step of each cycle. Melting curve analyses were obtained after each qPCR run at 65 °C to 90 °C and showed a single PCR product, confirming specificity of amplification. Expression ratios of the genomic sequences were calculated using the comparative threshold cycle (Ct) method described by Pfaffl *et al.*, 2001[180], with normalisation to the housekeeping gene GAPDH:

$$\text{Relative expression} = \frac{(\text{Efficiency}_{\text{target}})^{\triangle \text{Ct}_{\text{target}}(\text{control} - \text{treated sample})}}{(\text{Efficiency}_{\text{reference}})^{\triangle \text{Ct}_{\text{reference}}(\text{control} - \text{treated sample})}}$$

Table 2.8: Primers used for qPCR analysis of gDNA samples

Virus-host junction	Forward primer (5'→3') Species		Reverse primer (5'→3') Species		Reference
3' F (A5)	HPV16	CAGCTCACACAAAGGACGGA	Host	TGGGACTTTTACCAAAGCATGT	EK design
5' F (A5)	Host	CTCTGCCTGCACATGACTTG	HPV16	TGTCCAATGCCATGTAGACG	EK design
3' H	Host	TGCAGAAAGCATAGCTGACTACT	HPV16	ACTGCAGTGTCGTCTACATGG	EK design
5' H	HPV16	TGCAAAGATGTTTTAATGTCCCA	Host	ACCAGCCGCTGTGTATCTG	EK design
3' D2	HPV16	TGTTCATGAAGGGATACGAACA	Host	CCCAAGTCCATTGAATCCTG	EK design
5' D2	Host	TGACGAGGATGTGGAAGTGAC	HPV16	ACCCGACCCTGTTCCAATTC	EK design
3' G2	HPV16	AAGTTTGCACGAGGACGAGG	Host	TCATCTGTGTTTTGAGCCACAT	EK design
5' G2	Host	ACCACACCTGGCTGAGAAAA	HPV16	AGTTGCAGTTCAATTGCTTGT	EK design
G2 splice jx	HPV16 2761	CAGTGCCCTGTTGGAACTACA	Host 799	GCTGTTGACCTCTTTGGGGT	EK design
G2 splice jx	HPV16 879	GGAATTGTGTGCCCCATCTG	Host 903	TGCTGTTGATTGGTCCTCCA	EK design
G2 splice jx	HPV16 879	TTGTGTGCCCCATCTGTTCTC	Host 662	TTCTCGTGGGAGGGAAAGCTA	EK design
G2 splice jx	HPV16 225	GCAACAGTTACTGCGACGTG	Host 662	GGTGTTAGGGCCATGTCAGG	EK design
E6/E7	HPV16 E6	TGTTTCAGGACCCACAGGAGC	HPV16 E7	CGCAGTAACTGTTGCTTGCAG	Herdman *et al.*, 2006
GAPDH	Host	TGCACCACCAACTGCTTAGC	Host	GGCATGGACTGTGGTCATGAG	Vandesompele *et al.*, 2002

2.2.7 Calculating primer amplification efficiencies

Primer amplification efficiency (E) is the multiple by which the target is amplified per PCR cycle, with a theoretical maximal value of 2. All primers used for qPCR analysis were assessed for efficiency of amplification. These efficiencies were calculated by setting up reactions as described by 2.4.3 with neat, 1:2, 1:4, 1:10, 1:50 and 1:100 dilutions of gDNA from cells known to express the target. All reactions were performed in triplicate in 96-well plates, using the SYBR® Green qPCR protocol described in 2.4.3. The raw fluorescent data was exported to Microsoft Excel, and the primer efficiency calculated using a program designed by Dr. Ian Roberts, Hutchinson/MRC Cancer Cell Unit, Cambridge. This program calculates primer efficiencies by plotting the CT values (the intersection between an amplification curve and a threshold line, i.e. the cycle threshold) for each reaction versus the log10 relative amount of target amplicon. Linear regression was performed to determine a line of best fit for the data points and the gradient of the line calculated. Primer efficiency was calculated using the following equation: $E = 10\text{^}(-1/gradient)$. In addition to the calculation of the primer efficiency, melting curves were run for new primer sets

to confirm the presence of only one product; to produce melt curves, the final PCR product was exposed to a temperature gradient from 60 °C to 95 °C with a heating rate of 0.2 °C per second while fluorescence readouts are continually collected. The melt curves are converted to distinct melting peaks by plotting the first negative derivative of the fluorescence as a function of temperature (-dF/dT)[181].

2.3 RNA analysis

2.3.1 RNA extraction from cell lines

Total RNA from cell lines was extracted using Tri Reagent (Sigma Aldrich). If present in the tissue culture vessel, feeder cells were washed off using 1X PBS and total RNA from the cells of interest extracted according to the manufacturer's instructions; 1 ml/well of Tri Reagent was used to lyse cells from a 6-well plate and 500 μl/well for a 12-well plate. Briefly, The lysate was mixed by pipetting, transferred to an RNase-free 1.5 ml Eppendorf tube and incubated for 5 minutes at RT to permit the dissociation of nucleoprotein complexes. Saturated chloroform (200 μl per 1 ml of Tri Reagent) was added to the lysate and extracts were inverted rapidly for 15 seconds and incubated at RT for 3 minutes before being centrifuged for 15 minutes at 12,000 x g at 4 °C (Eppendorf centrifuge 5415R). The upper aqueous phase, containing the RNA, was then transferred to a fresh 1.5 ml Eppendorf tube and 1 μl GlycoBlue^{TM} (Ambicon) and 500 μl isopropanol per 1 ml of Tri Reagent was added to precipitate the RNA. Samples were vortexed to mix, incubated for 10 minutes at RT, and then centrifuged at 12,000 x g for 10 minutes at 4 °C. The supernatant was carefully removed and the RNA pellet washed in 1 ml of 75% ethanol per 1 ml of Tri Reagent used. The supernatant was carefully removed and the pellet air-dried for approximately 10 minutes and re-suspended in 30 μl of nuclease-free water. The RNA samples were placed at 58 °C for 10 minutes before the concentration and purity of RNA was determined spectrophotometrically by measurements of absorbance at 260 and 280 nm using a Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific). RNA was then either immediately used for complementary DNA (cDNA) synthesis

or stored at -80 °C.

2.3.2 cDNA synthesis

Synthesis of complementary DNA (cDNA) from RNA was done using the QuantiTect Reverse Transcription Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. In a first step, designed to ensure the removal of any contaminating genomic DNA, 1 μg of RNA template and 2 μl of Genomic DNA Wipeout buffer were added to an RNase-free 0.2 ml microcentrifuge tube together with nuclease-free water up to a final volume of 14 μl. The reaction was incubated for 2 minutes at 42 °C and then samples placed immediately on ice. For the reverse transcription reaction step, 1 μl Quantiscript reverse transcriptase, 4 μl Quantiscript RT buffer and 1 μl RT primer mix (polyT plus random hexamer primers) were added to 0.2 ml microcentrifuge tubes containing RNA template. The mixture was then incubated at 42 °C for 30 minutes to generate complementary DNA (cDNA), followed by 3 minutes at 95 °C to inactivate the enzyme. Following the reverse transcription reaction, the synthesised cDNA was diluted 1:10 using ddH_2O, and was immediately used for qPCR amplification or placed at -20 °C for long-term storage.

2.3.3 qPCR analysis of cDNA samples

qPCR was used to quantify the level of relative cDNA levels between samples. qPCR was performed using 2 μl cDNA (5 ng cDNA per reaction), 12.5 μL SYBR® Green JumpStart™ Taq ReadyMix™ Green Master Mix (Sigma), 2.5 μL 3 μM of forward and reverse primer mix (0.3 μM per reaction, Sigma Aldrich, Table 2.9) in a final reaction volume of 25 μL/well. The qPCR reaction was carried out in triplicate for each primer pair in 96-well white PCR plates (Starlab) using an Eppendorf Mastercycler ep gradient S realplex2. The cycling conditions consisted of a 2 minute initial denaturation at 95 °C, followed by 40 cycles of 95 °C for 15 seconds, 58 °C for 20 seconds, 72 °C for 15 seconds and 76 °C for 5 seconds, and final extension 78 °C for 8 minutes; fluorescence was measured at the last step of each cycle. The primer efficiency of each qPCR primer was determined as 2.2.7.

2.3.4 Quantification of transcript level changes

Based on Vandesompele and co-workers findings (Vandesomplele *et al.*, 2002), a range of housekeeping genes were tested for each different experimental condition in order to obtain a baseline level of expression for comparative analysis of gene expression. Gene expression ratios were calculated using the comparative threshold cycle (Ct) method described by Pfaffl *et al.* 2004[181] (see 2.2.6). The house keeping genes used for normalisation were: GAPDH, YHWAZ, RPL13A and ACTB (Table 2.9).

Table 2.9: Primers used for qPCR analysis of cDNA samples

Target	Forward primer (5' to 3')	Reverse primer (5' to 3')	Reference
ACTB^	CTGGAACGGTGAAGGTGACA	AAGGGACTTCCTGTAACAATGCA	Vandersompele *et al.*, 2002
CDK9	CCATTACAGCCTTGCGGGAGA	CAGCAAGGTCATGCTCGCAGA	EK/IJG design, 2014
E6/E7 (HPV16 P_{97} transcripts)	TGTTTCAGGACCCACAGGAGC	CGCAGTAACTGTTGCTTGCAG	Herdman *et al.*, 2006
GAPDH^	TGCACCACCAACTGCTTAGC	GGCATGGACTGTGGTCATGAG	Vandersompele *et al.*, 2002
p300 (EP300, KAT3)	TCTGGTAAGTCGTGCTCCAA	GCGGCCTAAACTCTCATCTC	EK/IJG design, 2014
RPL13A^	CCTGGAGGAGAAGAGGAAAGAGA	TTGAGGACCTCTGTGTATTTGTCAA	Vandersompele *et al.*, 2002
TIP60 (KAT5, HTATIP)	CTTGGCCAAAAGACACAGGT	CATCCTCCAGGCAATGAGAT	EK/IJG design, 2014
YWHAZ^	ACTTTTGGTACATTGTGGATTCAA	CCGCCAGGACAAACCAGTAT	Vandersompele *et al.*, 2002

^house keeping gene

2.4 Protein analysis

2.4.1 Total protein extraction

Cells were washed with cold PBS, lysed in cold radioimmunoprecipitation assay (RIPA) lysis buffer (Thermo Fisher Scientific) (approximately 70 μl per well of a 6-well plate) supplemented with cOmplete^{TM} Protease Inhibitor Cocktail (Roche Diagnostics Ltd., Burgess Hill, UK) (1:25 dilution). Cells were then scraped off using a cell scraper and were transferred into a pre-cooled Eppendorf tube. After 15 minutes of continuous agitation on a vibrating shaking platform at 4 °C, cellular debris was removed by centrifugation (15 minutes, 4 °C, 14,000 x g). The supernatant was transferred to a fresh 1.5 ml Eppendorf tube and immediately quantified or placed at -80 °C for long term storage.

2.4.2 Protein quantification

Total protein concentration was determined using Pierce bicinchoninic acid (BCA) protein assay kit (Thermo Fisher Scientific) according to the manufacturer's instruction with some modification. Briefly, protein samples were diluted with PBS (2 μl of protein sample added to 12 μl of PBS, 1:7 dilution) and 3 μl of diluted sample added to the well of 96-well plate in triplicate. The BSA stock standard at 1 mg/ml concentration was used to generate a standard curve and was added to each well at 5 different volumes in triplicate: 0 μl (Blank), 1 μl, 3 μl, 5 μl, 7 μl and 10 μl. The provided Reagent A and Reagent B were mixed together in a ratio of 1:50 and 200 μl of the mixture was added to each well and incubated for 30 minutes at RT in the dark. Absorbance was measured at a wavelength of 570 nm using a Dynex Technologies plate reader and Revelation software. Protein concentrations were determined using a standard curve (Figure 2.1).

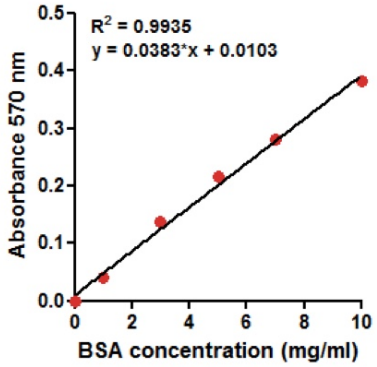

Figure 2.1: Example of a standard curve used for BCA assay to determine protein concentration based on Absorbance values.

2.4.3 Protein sample preparation and SDS-PAGE separation

Quantified protein extract (25 μg) was added to 5 μl of NuPAGE LDS Sample Buffer and 2 μl of NuPAGE Sample Reducing Agent (both from Novex, Thermo Fisher Scientific) made up to a final volume of 20 μl with 1X PBS. Samples were denatured for 10 min at 95 °C and run immediately or stored at -20 °C. The appropriate precast gel (all Thermo Fisher) (CDK9, TIP60, HDAC1 detection: 4–12% NuPAGE™ Bis-Tris Mini Gel; p300 detection: 3–8% NuPAGE™ Tris-Acetate Protein Gel) was placed in an XCell SureLock Mini-Cell and the inner tank filled with the appropriate running buffer supplemented with 500 μl NuPAGE Antioxidant (NuPAGE MOPS SDS Running Buffer or NuPAGE Tris-Acetate SDS Running buffer, respectively); and the outer tank filled to the level of the well base Running Buffer without antioxidant. The samples were loaded, along with 5 μl of See Blue molecular weight ladder (Novex, Thermo Fisher Scientific). Proteins were separated by electrophoresis at 150 V until the desired protein separation was achieved (approximately 1.5 hours).

2.4.4 Western blotting

A PVDF membrane (Thermo Fisher Scientific) and two pieces of 3MM Whatman® filter paper were cut to 7 x 8 cm. The PDVF membrane was then activated by incubation in 100% methanol for 15 seconds followed by a 2-minute rinse in ddH_2O. The activated membrane, along with the filter paper and four sponges were then placed

in ice-cold Transfer Buffer (Thermo Scientific) supplemented with 10% ethanol and 0.1% NuPAGE Antioxidant. Protein samples separated by SDS-PAGE were transferred to the PVDF transfer membrane using the XCell SureLock Mini-Cell and XCell II™ blot module. Firstly, XCell SureLock Mini-Cell was dismantled, gel separated from one of its plates and top lanes and very bottom part of the gel cut away. A sheet of 3MM filter paper was then adhered to the gel and gel carefully peeled away from the plastic plate. The equilibrated PVDF membrane was then carefully placed on top of the gel ensuring that no air bubbles were present. Another sheet of 3MM filter paper was placed on top of the PVDF membrane and the 'sandwich' placed on top of two pre-soaked two sponges that had been placed on top of the cathode core of the XCell II™ blot module. Another 2 pre-soaked sponges were then placed on top of the 'sandwich', the lid of the blot module placed on top and the whole assembly inserted into the XCell SureLock™ Mini-Cell. The blot module was topped up with Transfer Buffer, the outer reservoir filled with ice-cold ddH_2O and the whole device placed in a 4 °C room. The protein transfer was performed for 2 hours at 30 V. After completion of the transfer, the blot module was dismantled and the membrane placed in Blocking Buffer (5% milk, 1% Tween in 1X PBS) and incubated at 4 °C overnight with gentle rocking. After incubation in blocking buffer, the membrane was transferred to an Antisera Buffer (5% milk, 0.1% Tween in 1X PBS) containing the primary antibody at the appropriate dilution (Table 2.10) and incubated once more at 4 °C overnight with gentle rocking. The following day, the membrane was washed 3 times for 5 minutes each with Antisera Buffer at RT before incubation for 1 hour in species-specific HRP-conjugated secondary antibodies diluted in Antisera buffer at RT. Following three further washes in 1X PBS + 0.05% Tween, the membrane was developed using enhanced standard chemiluminesence (ECL) or ECL prime western blotting detection reagents (Amersham, GE Healthcare, Little Chalfont, UK) according to the manufacturer's instructions and exposed to autoradiographic films at various time-points. If required, the blots were stripped with Stripping Buffer (10 minute incubation at 37 °C in the dark; Thermo Fisher Scientific) and blots incubated with blocking buffer and re-probed with different primary antibodies. The images were scanned and densitometry analyses were

performed using FluorChem-9900 imaging system software (Alpha Innotech, San Leandro, U.S), with normalisation to a loading control (β-tubulin) and referencing to the chosen control samples set to 1.

Table 2.10: Antibodies used for Western Blotting

Target	Antibody catalogue no.	Host species	Supplier	Dilution
CDK9	sc-8338 X	Rabbit	Santa Cruz	1:1000
HDAC1	40967	Rabbit	Active Motif	1:500
p300 (C-20)	sc-585 X	Rabbit	Santa Cruz	1:1000
TIP60 (KAT5)	sc-5725X	Goat	Santa Cruz	1:1000
β-tubulin	ab6046	Rabbit	Abcam	1:10,000
Goat	P044901	Rabbit	Dako	1:1000
Rabbit	P044801	Goat	Dako	1:2000

2.5 Chromatin immunoprecipitation (ChIP) assays

2.5.1 Cell fixation

W12 cells were grown to 70–80% confluency in a 15 cm tissue culture dish as in 2.1.3. Once ready to harvest, the cells were cross-linked using 20 ml Fixation Solution (540 μgl 37% formaldehyde, 20 ml cell culture medium; 1% final formaldehyde concentration) on a shaking platform for 10 minutes at RT. The fixation reaction was quenched by removal of the fixation solution followed by a wash with ice-cold 1X PBS and the addition of 10 ml Glycine Stop-fix solution (1 ml 10X Glycine Buffer, 1 ml 10X PBS, 8 ml ddH_2O) and rocking for 5 minutes at RT. The cells were then washed for a second time with 1X PBS before being removed from the dish using a cell scraper and 2 ml ice-cold cell scraping solution supplemented with 10 μgl 100 mM PMSF. The cell:solution mix was then transferred to a 15 ml Falcon tube and placed on ice. The solution was spun at 720 RCF at 4 °C for 10 minutes and the cell pellet frozen at -80 °C after the addition of 1 μl 100 mM PMSF and 1 μl protease inhibitor cocktail (PIC) (both provided by Active Motif).

2.5.2 Shearing of chromatin

The cell pellet was thawed and resuspended in 1 ml ice-cold Lysis Buffer supplemented with 5 μl PIC and 5 μl PMSF, and incubated on ice for 30 minutes. Following incubation, the nuclei were released from cells by performing 30 strokes in an ice-cold dounce homogeniser. Samples were then centrifuged at 2,400 RCF for 10 minutes at 4 °C to pellet the nuclei, which were then resuspended in 400 μl Shearing buffer before incubation at 37 °C for 5 minutes. Next, the DNA was sheared using a sonicator (Active Motif, model no. Q120AM) at 25% power for 10 minutes of 30 seconds on 30 seconds off. The sheared chromatin samples were then centrifuged at 18,000 RCF at 4 °C for 10 minutes. The supernatant, containing the chromatin, was collected and place in a fresh tube 1.5 ml Eppendorf tube and used immediately (see 2.5.3) or stored at -80 °C until later use. At this stage, a 50 μl aliquot of sheared chromatin was removed for use as a control sample (input) for downstream processing. The DNA shearing efficiency was then checked by running a small aliquot on a 1.5% agarose gel and the concentration of the chromatin sample was measured using a Nanodrop 2000 Spectrophotometer.

2.5.3 Immunoprecipitation

Immunoprecipitation (IP) was performed by incubating 7 μg of chromatin per reaction with protein G magnetic beads (25 μl), ChIP buffer1 (10 μl), PIC (1 μl) and ChIP antibody (Table 2.11) to a final volume of 100 μl in siliconised microcentrifuge tubes which were placed on an end-to-end rotator overnight at 4 °C. After overnight incubation the beads were pelleted using a magnetic stand and the supernatant discarded. The beads were then washed with ChIP buffer 1 followed by two washes with ChIP buffer 2. After the wash steps, the beads were pelleted and the supernatant discarded once more.

2.5.4 Elution of chromatin and cross-link reversal

The beads were then resuspended in 50 μl Elution Buffer AM2 and incubated at RT on an end-to-end rotator. 50 μl Reverse Cross-linking Buffer was added to the eluted

chromatin and samples placed quickly on a magnetic stand to pellet the beads. The supernatant, containing the eluted chromatin, was transferred quickly to a fresh 1.5 ml Eppendorf tube. At this stage 10% of IP sample volume was taken from the input DNA sample (from 2.5.2), 2 μl 5M NaCl added, and the total volume made to 100 μl using ddH_2O. All samples were incubated at 95 °C for 15 minutes, then 2 μl Proteinase K added to each before incubation at 37 °C for 1 hour, followed by addition of 2 μl Proteinase K Stop Solution. The DNA samples were cleaned using a QIAquick PCR Purification Kit (Qiagen) prior to qPCR, and the purified DNA used for qPCR or stored at -20 °C.

2.5.5 ChIP-qPCR

qPCR of ChIP samples was performed as previously described (2 μl gDNA, see section 2.2.6). Multiple primer pairs were used to test for enriched sequences at different locations across the HPV16 genome (Table 2.12) and the primer efficiency for each determined as in 2.2.7. To calculate the relative levels of enrichment of each ChIP reaction for every primer set, the target Ct values were compared to input Ct values before normalisation to a control region of the host as described in Table 2.11.

By incorporating the average Ct values of the input DNA in both the numerator and the denominator of the equation, the number of viral copies in each clone is accounted for.

2.6 Statistical analysis

Statistical analyses were performed using GraphPad Prism 6 software (GraphPad Software, La Jolla, U.S.). For comparisons between groups, an unpaired, two-tailed Students *t*-test was used. P-values <0.05 were considered statistically significant. Data are presented as mean ± SEM.

Table 2.11: Antibodies used for ChIP experiments

Protein target	Supplier	Category no.	Volume (µl)	Species	Normalisation control target
CDK9 (P-TEFb)	Santa Cruz	sc-8338 X	5	Rabbit	*GAPDH*orf
Cyclin T1 (P-TEFb)	Santa Cruz	Sc-10750	5	Rabbit	*GAPDH*prom
H3ac	Active Motif	39139	10	Rabbit	*GAPDH*prom
H3K4me1	Active Motif	39297	10	Rabbit	*ACTB*prom
H3K4me3	Active Motif	39915	5	Rabbit	*GAPDH*prom
H3K9me2	Active Motif	39239	10	Rabbit	*γ-GLOBIN*prom
H3K27ac	Active Motif	39133	5	Rabbit	*ACTB*prom
H3K27me2	Active Motif	39245	10	Rabbit	*MYOG*prom
H3K27me3	Active Motif	39155	5	Rabbit	*MYOG*prom
HDAC1	Active Motif	40967	4	Rabbit	*TLR9*prom
IgG Goat	Dako	X0907	Variable	Goat	N/A
IgG Rabbit	Dako	X0902	Variable	Rabbit	N/A
IgG Rat	Sigma	R9759	Variable	Rat	N/A
p300 (EP300, KAT3B)	Santa Cruz	sc-585 X	5	Rabbit	*GAPDH*prom
RNAPII Total	Active Motif	61081	5	Rat	*GAPDH*prom
RNAPII Ser2P	Active Motif	61083	5	Rat	*GAPDH*prom
RNAPII Ser5P	Active Motif	61085	10	Rat	*GAPDH*prom
TIP60 (KAT5, HTATIP)	Santa Cruz	sc-5725 X	5	Goat	*GAPDH*prom

Table 2.12: Primers for ChIP-qPCR of HPV16 chromatin

HPV16 coordinates	Forward Primer (5' to 3')	Reverse Primer (5' to 3')	Reference
111F to 223R	AGGACCCACAGGAGCGACCC	ACGTCGCAGTAACTGTTGCTTGCA	Scarpini *et al.*, 2014
427F to 506R	GCCACTGTGTCCTGAAGAAAAGCA	GACCGGTCCACCGACCCCTT	Scarpini *et al.*, 2014
649F to 765R	GACAGCTCAGAGGAGGAGGA	GCACAACCGAAGCGTAGAGT	Gray *et al.*, 2010
1250F to 1368R	GCGAAGACAGCGGGTATGGCA	GCAACCACCCCCACTTCCACC	Scarpini *et al.*, 2014
2158F to 2316R	AGGGTAGATGATGGAGGTGATTGG	GATTTACCTGTGTTAGCTGCACCA	Groves *et al.*, 2016
2695F to 2757R	TCCTTTTTCTCAAGGACGTGGT	ACGTTGGCAAAGAGTCTCCA	Groves *et al.*, 2016
2853F to 2950R	GGAAACACATGCGCCTAGAATGTGC	TGATACAGCCAGTGTTGGCACC	Groves *et al.*, 2016
3102F to 3213R	CAGTGGAAGTGCAGTTTGATGG	CAACTTGACCCTCTACCACAGT	Groves *et al.*, 2016
3407F to 3514R	CACTCCGCCGCGACCCATAC	GGTGTGGCAGGGGTTTCCGG	Groves *et al.*, 2016
3936F to 4025R	ACGTCCGCTGCTTTTGTCTGTGT	ACCTAAACGCAGAGGCTGCTGT	Groves *et al.*, 2016
4419F to 4542R	CAGGGTCGGGTACAGGCGGA	GGATCGGAAGGGCCCACAGGA	Groves *et al.*, 2016
5175F to 5304R	TCGTAGTGGAAAATCTATAGGTGC	AAGGCTGCATGTGAAGTGGT	Groves *et al.*, 2016
5646F to 5726R	TGGCTGCCTAGAGGCCACTGT	TGCGTGCAACATATTCATCCGTGC	Groves *et al.*, 2016
6039F to 6157R	TGCAGCAAATGCAGGTGTGGAT	TGGGGATCCTTTGCCCCAGTGT	Groves *et al.*, 2016
7045F to 7121R	ACAAGCAGGATTGAAGGCCAAACCA	AGAGGTAGATGAGGTGGTGGGTGT	Groves *et al.*, 2016
7288R to 7441R	TGCTTGTGTAACTATTGTGTCATGCAA	AAATGGCCGCTGGCGCTAC	Groves *et al.*, 2016
7419F to 7552R	TTTGTAGCGCCAGCGGCCATTT	GCATGGCAAGCAGGAAACGTACAA	Scarpini *et al.*, 2014
7555F to 7681R	CCAAATCCCTGTTTTCCTGA	CGTTGGCGCATAGTGATTTA	Gray *et al.*, 2010
7854F to 65R	GCAAACCGTTTTGGGTTACA	ACTAACCGGTTTCGGTTCAA	Gray *et al.*, 2010
ACTBprom	CTGATGCCACAATCACCCCT	GTAATGTATTAACTTCCTGGCCATT	Groves *et al.*, 2016
*GAPDH*prom	CGGCTACTAGCGGTTTTACG	AAGAAGATGCGGCTGACTGT	Groves *et al.*, 2009
*GAPDH*orf	CTCATGCCTTCTTGCCTCTT	TTGATGGCAACAATATCCACTT	Palermo *et al.*, 2008
*γ-Globin*prom	GCCTTGACCAATAGCCTTGACA	GAAATGACCCATGGCGTCTG	Groves *et al.*, 2009
*MYOG*prom	GGAGAAAGAAGGGGAATCACAT	GATAAATATAGCCAACGCCACA	Groves *et al.*, 2016
TLR9prom	AGCAGGGCAGGACAGCCAGA	ACAGCCCACCGTCCCCATGT	Groves *et al.*, 2016

2.7 Fluorescent *in situ* hybridisation (FISH)

The method for performing 3D FISH with directly labelled DNA probes was based upon the methods outlined in Bolland *et al.* 2013[182], with a few alterations. See Figure 2.2 for workflow.

2.7.1 Preparation of FISH slides

W12 G2p11 cells (same passage number as used in SCRiBL) were resuscitated and grown in a 10 cm^2 tissue culture dish as previously described. Once at 70–80% confluency the cells were trypsinised, washed in ice-cold PBS and the concentration determined. The cells were then diluted to 5 x 10^5 /ml in 1X PBS. 20 μl of the cell suspension was then pipetted on to the centre of a polysine^{TM} slide (VWR) and incubated at RT for 30 minutes to allow the cells to settle and adhere to the slide. Following incubation, the cells were fixed on the slide by submerging into 4% paraformaldehyde (PFA) for 10 minutes. The fixation reaction was quenched by placing the slides in a 155 mM glycine solution for 10 minutes at RT in a coplin jar. The cells were then permeabilised in a 0.1% saponin/0.1% Triton-X/1X PBS solution at RT for 10 minutes. The slides were then washed once in 1X PBS and stored at -20 °C in 50% glycerol/1X PBS until further use (slide storage up to 1 month).

2.7.2 Growing BAC colonies

BAC clones were ordered from Thermo Scientific (Table 2.13). A pipette tip of each BAC clone was used to streak bacteria onto an agar plate supplemented with 20 μg/ml (100 μg/ml for HPV16) and the agar plates incubated at 37 °C overnight. Following incubation, a single colony was picked using a 200 μl pipette tip and used to streak on a fresh agar plate with the appropriate concentration of chloramphenicol, which was again incubated at 37 °C overnight and the original agar plate labelled, wrapped in parafilm and stored at 4 °C. Following the second overnight incubation, a pipette tip was used to pick a colony from the agar plate containing growth from a single colony and placed in 5 ml LB broth supplemented with chloramphenicol (4

μl), which was incubated at 37 °C with gentle shaking for 6-8 hours. Again, the agar plate was wrapped in parafilm and stored at 4 °C for future use. This was then transferred into 200 ml LB broth supplemented with chloramphenicol (160 μl) in a 1 L conical flask and incubated at 37 °C shaking at 2,500 rpm overnight (incubated shaker, New Brunswick Scientific Incorporated). The reaction was stopped when the solution appeared cloudy and was put on ice to cool. 700 μl of the overnight culture in LB broth was mixed with 300 μl sterile 50% glycerol, mixed, and stored at -80 °C for future use. The remaining bacteria were harvested from the LB culture by centrifugation at 6,000 RCF in an ultra centrifuge (Beckman AvantiTM J-20 series) for 15 minutes at 4 °C, the pellet was then frozen at -20 °C until further use.

Table 2.13: details of BAC clones used in 3D FISH experiment

BAC ID	Species, chromosome	Coordinates	Supplier	Fluorescent dye
RP11-467N14	Human, Chr 5	51,676,020 – 51,873,551	Thermo Fisher	Alexa Fluor 647
CTD-2015C9	Human, Chr 5	53,473,886 – 53,584,235	Thermo Fisher	Alexa Fluor 555
pSP64-HPV16	HPV16	0 – 7904	Cinzia Scarpini	Alexa Flour 488

2.7.3 Extraction of BAC DNA

The extraction of BAC DNA was done using the NucleoBond® BAC 100 plasmid DNA purification kit (Macherey-Nagel) according to the manufacturer's instructions. Briefly, the bacterial pellet (from 2.7.2) was defrosted on ice and the cells lysed by resuspending the pellet in 24 ml Buffer S1 supplemented with RNase A (100 μg/ml). 24 ml Buffer S2 (pre-warmed to 30 °C) was added to the suspension mixed by inversion, and incubated at RT for 3 minutes. 24 ml of pre-cooled (4 °C) was then adding to the suspension, mixed by inversion and incubated on ice for 5 minutes. During this time the Nucleobond® BAC 100 (Maxi) Column was equilibrated with 6 ml Buffer N2. The bacterial lysate was then cleared by filtration; a Nucleobond® Folded Filter was placed in a plastic funnel and dampened with a few drops of Buffer N2. The bacterial lysate was then loaded onto the filter and the flow-through collected. The cleared lysate was then loaded on to the equilibrated column and the

column emptied by gravity flow (DNA bound to the column). The column was then washed twice with 18 ml Buffer N3 and the flow-through discarded. The BAC DNA was then eluted from the column with 15 ml Buffer N5 pre-warmed to 50 °C and collected in a 50 ml centrifuge tube (VWR). 11 ml room-temperature isopropanol was then added and mixed to precipitate the eluted plasmid DNA and centrifuged at 18,000 RCF (Beckman Coulter Avanti J-25) for 30 minutes at 4 °C. The supernatant was then carefully discarded and the DNA pellet washed with 1 ml 70% ethanol and the pellet transferred to a 1.5 ml Eppendorf tube. The DNA was pelleted again by centrifugation at RT and washed once more with 500 μl 70% ethanol. The ethanol was then removed and the pellet dried at RT. Once dry, the pellet was re-suspended in 200 μl ddH_2O and the DNA concentration determined by Nanodrop.

2.7.4 Extraction of HPV16 plasmid DNA

The extraction of HPV16 plasmid DNA was carried out using the GenElutetextsuperscriptTM Plasmid Miniprep kit (Sigma Aldrich) according to the manufacturer's instructions. Briefly, the bacterial pellet (from 2.7.2) was completely re-suspended in 200 μl Resuspension Solution before the addition of 200 μl Lysis Solution. The solution was mixed by inversion and incubated on ice for up to 5 minutes. 350 μl Neutralisation/Binding Solution was then added and the solution mixed by inversion before being spun in a table-top microcentrifuge (Eppendorf 5415 D) at maximum speed for 10 minutes. The GenElute Miniprep Binding Column was then prepared by insertion into a provided microcentrifuge tube, followed by the addition of 500 μl Column Prep Solution, spinning at maximum speed for 1 minute, and the flow-through discarded. The cleared lysate (supernatant) from the previous step was then transferred to the column and spun at maximum speed for 1 minute. The column was then washed with 750 μl Wash Solution. The DNA was eluted from the column in 50 μl pre-warmed (65 °C) ddH_2O and the DNA concentration determined by Nanodrop.

2.8 Generating directly labelled DNA probes by nick translation

2.8.1 Nick translation

The nick translation reaction contained: 2 μg DNA (in 54 μl), 10 μl 10x NTB (1 M Tris-HCl, pH 7.5, 1 M $MgCl_2$, 10 mg/ml BSA fraction V), 10 μl DTT (0.1 M) (Thermo Fisher Scientific), 8 μl d(GAC)TP mix (0.5 mM each), 2 μl dTTP (0.5 mM), 12 μl aminoallyl-UTP (0.5 mM, Sigma), 2 μl DNA polymerase I (10 U/μl, New England Biolabs), 2 μl DNase I (1:25 dilution with 1X buffer, Roche). The reaction mixture was incubated at 16 °C for 2 hours and then placed on ice. Following incubation, 1 μl was run on a 2% agarose gel and the DNase I inactivated by heating the reaction mixture to 75 °C for 5 minutes. The amine-modified DNA was then purified using a QIAquick PCR Purification kit (Qiagen) and eluted in 100 μl ddH_2O. The DNA was precipitated by the addition of 10 μl NaOAc (3M, pH 5.2) and 250 μl 100% ethanol and incubation at -20 °C overnight. After, the sample was spun at maximum speed for 30 minutes and 4 °C and the resulting DNA pellet washed with 100 μl 70% ethanol. The pellet was then re-suspended in 6 μl ddH_2O and 1 μl used to determine the concentration by UV-Vis spectroscopy (Nanodrop).

2.8.2 Coupling fluorescent dye

Fluorescent labelling of the probe is achieved by chemical coupling of dye (see table 2.13). 5 μl amine-modified DNA (max. 2 g) was heated to 95 °C for 5 minutes and snap cooled on ice. 3 μl NaB labelling buffer (0.2 M sodium bicarbonate pH 8.3, Sigma) was added and mixed. To this, the amine-reactive dye reconstituted in 2 μl DMSO was added, the solution mixed and incubated in the dark at RT for 1 hour. Following incubation, 40 μl ddH_2O was added and the labelled probe purified using the QIAquick PCR purification kit, with two column washes with buffer PE and the DNA eluted in 50 μl ddH_2O. The probe concentration and labelling efficiency was determined by UV-Vis spectroscopy (Nanodrop) and use of an algorithm written by Dr Daniel Bolland, Babraham Institute.

2.8.3 Probe precipitation

To precipitate the probes required for DNA FISH, 20 ng of each directly labelled probe required per slide was mixed with 2 μl human cot-1 DNA (6 μg), 1 μl single stranded DNA from salmon sperm testes (9.7 μg) and the volume adjusted to 100 μl with ddH_2O. 10 μl 3 M NaOAc and 275 μl ethanol was then added and the solution mixed by inversion and incubated at -20 °C for at least 1 hour. After incubation, the precipitation reaction was spun at maximum speed (14,500 RCF) for 30 minutes at 4 °C and the pellet washed with 70% ethanol. The dried pellet was then re-suspended in 5 μl formamide and incubated at 37 °C, shaking at 1,000 rpm protected from light. After 30 minutes, 5 μl pre-made dextran sulphate mix (20% dextran sulphate in 2x SSC) buffer was added and put back on the heated shaker for a further 10 minutes. The probe was mixed thoroughly before pipetting onto the coverslip.

2.9 DNA FISH

The slides (from 2.7.1) were taken from storage at -20 °C and put into 20% glycerol/1X PBS to equilibrate to RT for at least 20 minutes. 3x freeze/thawing cycles were then performed in liquid nitrogen. One slide at a time was submerged in liquid nitrogen using forceps for 5–6 seconds or until the characteristic popping sound, and placed on a paper towel to defrost. The opaque frozen glycerol thawed before repeating the freeze/thaw cycle for a total of 3 times in liquid nitrogen. Following completion of all the necessary slides, all were washed twice in 1X PBS for 5 minutes each before incubation in 0.1 M HCl for 30 minutes at RT. Following this, the slides were again washed twice in 1X PBS for 5 minutes at RT before the slides were equilibrated in 50% formamide/2X SSC for 10 minutes at RT. The slides were put in 1X PBS immediately before adding the probe to the cells.

10 μl probe was pipetted on to the centre of a 22 x 22 mm coverslip and protected from any light source. One at a time, slides were removed from 1X PBS and any excess liquid around the cell spot was dried using a paper towel. The coverslip was then quickly inverted onto the slide covering the cell spot and sealed with rubber cement (Fixogum, Marabu) and the slides protected from light at all subsequent

stages of the protocol. Once all coverslips had been added, the slides were placed on a heat block at 78 °C for exactly 2 minutes before being transferred to a light-tight humidified chamber and incubated at 37 °C overnight (at least 16 hours).

The following day the rubber cement was peeled off and the slides placed in 2X SSC for 15 minutes to enable the coverslips to loosen. The slides were then washed in the following series of washes; 50% formamide/2X SSC at 45 °C for 15 minutes, 0.2X SSC at 63 °C for 15 minutes, 2X SSC at 45 °C for 5 minutes, 2X SSC for 5 minutes and finally 1X PBS for 5 minutes, both at RT. The slides were then stained in a DAPI solution (1 μg/ml) for 2 minutes at RT and de-stained in 1X PBS for 5 minutes. The slides were once again fixed in a 3.7 % formaldehyde/1X PBS solution for exactly 5 minutes at RT before being quenched in 155 mM glycine for at least 30 minutes to remove the autofluoresence of the formaldehyde. The slides were washed in 1X PBS for 5 minutes at RT before mounting a 22 x 50 mm coverslip using 30 μl ProLong® Diamond Antifade Mountant (Life Technologies) and sealing with nail varnish.

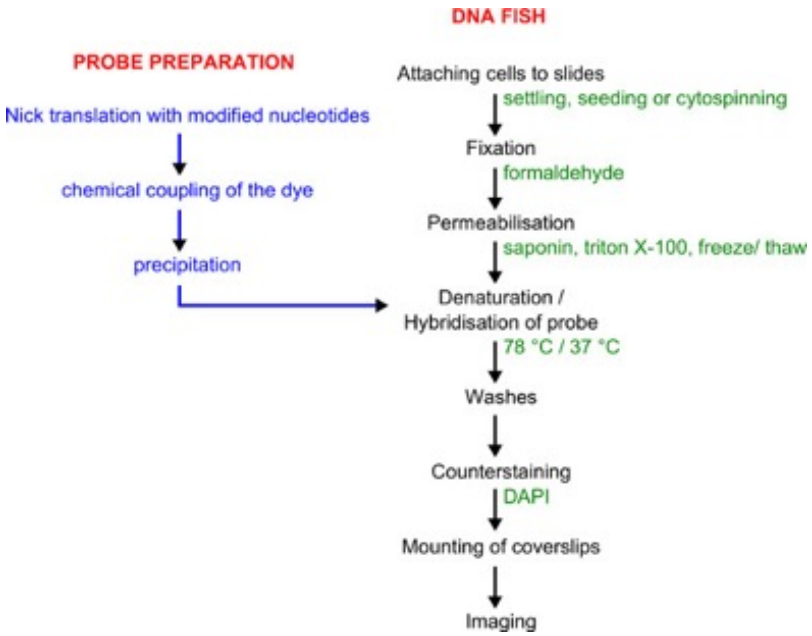

Figure 2.2: Workflow for probe labelling and DNA FISH (Bolland *et al.* 2013)[182]

2.10 DNA FISH analysis

2.10.1 Microscope analysis

All FISH slides were analysed at the Babraham Institute using their facilities, with supervision from Olga Mielczarek (PhD student, Corcoran Lab, Babraham Institute). Successful probe hybridisation was determined by first looking at the slides using an Olympus FV1000 confocal microscope. Once a successful FISH reaction had been confirmed, the slides were transferred to a MetaSystems Metacyte connected to Zeiss Axio Imager Z2 microscope for more detailed analysis; the Metacyte scans the slide and automatically images fields of view with multiple cells, capturing the fluorescent signals across twenty focal planes. For this experiment a 3-probe assay of wavelengths 488 (green), 555 (red) and 647 (far red) was programmed.

2.10.2 Determining probe distance

Following Metacyte analysis the data was transferred to a computer where the data was analysed using Metafer 4 v3.11.2 software. Metafer performs an automated analysis of fluorescence signals including the identification of the number of fluorescent spots per cell and distance between signals. A total of ~1500 cells per slide were captured using the Metacyte; cells containing human BAC probes ≠ 2 and HPV probe ≠ 1 were discarded from the subsequent analysis of the 3D distances between probes. The x, y and z coordinates between all 3 probes were exported and analysed using a customised Perl script (Felix Krueger, Babraham Institute). The Perl script identified the 3D distances between each probe in each individual cell (8 combinations in total), i.e. red1/green, red2/green, far red1/green etc. The chromosome of interest was identified by the presence of a green signal (HPV probe), and the distance between the red signals (ARL15 probe) to the HPV integrant was determined by identifying the smallest distance between red1/green and red2/green. The smallest distance represents a *cis* interaction between the HPV probe and the ARL15 site on the integrated allele whilst the longer distance is in *trans* in relation to the ARL15 site on the unintegrated allele. The same analysis was performed on the far

red signals (control probe) and the distance between HPV-ARL15 and HPV-control compared.

Chapter 3

HPV16 oncogene expression levels during early cervical carcinogenesis are determined by the balance of epigenetic chromatin modifications at the integrated virus genome

3.1 Introduction

Previous work carried out by the Coleman group characterised a panel of seventeen W12 integrant clones that were isolated from a mixed cell population in a non-competitive manner. In contrast to the current dogma of cervical carcinogenesis, the study showed that in comparison to episome-containing cells, HPV16 integration does not necessarily lead to a competitive growth advantage, nor does it always increase the expression levels of the virus oncogenes[183]. Additionally, the genome copy number was compared to the individual HPV16 expression levels of each of the integrant clones to determine the level of transcription per DNA template copy. Across the seventeen clones examined, the range of E6 and E7 transcript level per template varied by approximately 17-fold and 16-fold, respectively[183]. This data indicated that the difference in the level of transcription from integrated HPV genomes is not dependent on virus copy number; clones containing a high DNA copy number displayed relatively low levels of expression per template and vice versa. However, there were a number of examples where clones containing the same and/or similar virus DNA copy number showed significant differences in the levels of E6/E7 expression per template (Figure 3.1).

Initial investigations into possible reasons for the significant variation in the level of virus oncogene transcripts per template showed that epigenetic differences in the virus chromatin influenced the level of transcription from the virus early promoter. The study indicated that post-translational modification (PTM) of histone tails is a mechanism by which the expression of the viral oncogenes is controlled[183].

In order to provide a tractable system, and to avoid any complexity caused by heterogeneous changes in cells containing concatemerised integrants, clones with genome copy number less than or equal to four were analysed. From this group, two clones with high levels of expression per template (F and A5), two with medium levels (D2 and H), and one with low levels of expression per template (G2) were selected. The three groups of clones, each with statistically different levels of oncogene expression, were used as comparators to one another in the analysis of epigenetic regulation of the integrated HPV16 DNA genome.

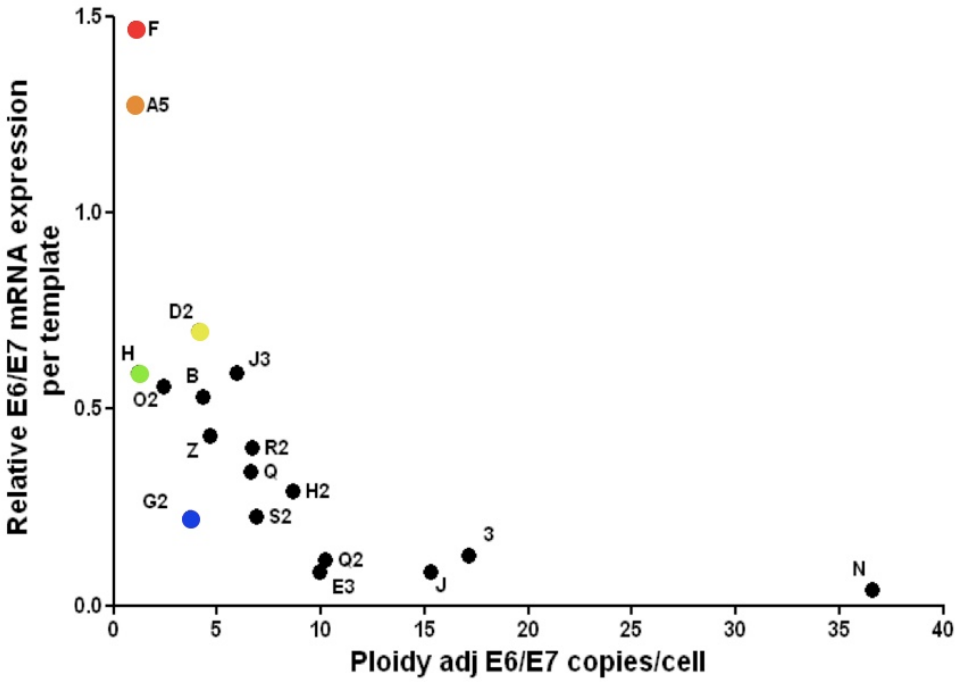

Figure 3.1: A plot of mean E6/E7 expression per template versus template copy number per cell across the W12 integrant clones[183]. Individual clones are represented by black circles.

As previously discussed, individual marks as well as patterns of PTMs of histone tails are associated with varying levels of transcription from gene promoters. By investigating whether the abundance and distribution of hallmark PTMs correlate with the levels of HPV16 transcription, this investigation aimed to elucidate whether this type of epigenetic regulation has a direct influence on expression of the virus oncogenes in pre-malignant cervical keratinocytes. Both characteristic methylation and acetylation marks have been examined and include marks that are associated with active transcription: H3K4me3, predominantly found at gene promoters; H3K4me1, found at all enhancer elements; and H3K27ac, a PTM associated with the subset of active enhancers[184]. To balance this, hallmark PTMs that result in transcriptional repression were also examined, namely H3K9me2, H3K27me2 and H3K9me3.

In addition to the analysis of the abundance of hallmark PTMs associated with the integrated HPV genome, this study sought to determine functional significance of enzymes responsible for laying down these marks with particular focus on acetylation of histone tails (H3ac, H3K27ac, H2AK5ac). As previously mentioned, the covalent attachment of an acetyl-moiety onto lysine residues present on histone tails results in specific epigenomic patterns that correlate with active transcription. This study will

focus of the activity of two type A HATs: p300 (EP300) and TIP60 (Tat-interacting protein 60 kDa; also known as KAT5). p300 is a global HAT that is able to efficiently acetylate the amino-terminal tails of all four-core histones, whereas TIP60 is more selective, primarily targeting the tails of histone H4 and H2A for acetylation[185, 186]. Additionally, it has been shown that somatic mutations in the p300 and TIP60 ORFs occur in a number of malignancies[187, 188, 189]. Mutations that result in augmented HAT activity lead to the increased expression of the viral oncogenes, which, in turn, accelerate disease progression and as such HATs p300 and TIP60 represent potential targets for the control of transcription from the integrated HPV16 early promoter.

This part of the investigation aimed to determine epigenetic modifications to the integrated HPV16 genome that might affect the levels of transcription from the virus early promoter. Fundamental to the level of transcription of any gene is the activity of RNAPII. As previously discussed, the conversion from the paused/poised version of the enzyme to the actively elongating form is mediated by the kinase CDK9 within the P-TEFb complex; dysfunctions in the CDK9-related pathway have been shown to be related with several malignancies including advanced solid tumours, leukemia and lymphomas[190].

In the first part of this chapter, the abundance and distribution of hallmark PTMs and RNAPII, and its associated forms, across the viral genome were characterised by performing chromatin immunoprecipitation (ChIP) assays followed by RT-qPCR analysis of the integrated viral genome. Secondly, the functionality of p300 and TIP60 HATs as well as CDK9 in the context of P-TEFb were probed by depletion and inhibition with siRNAs and small molecule inhibitors, respectively.

3.2 Results

3.2.1 Genetic mutation of the HPV16 LCR is not responsible for the differential oncogene expression of the W12 integrant clones.

Since the aim of this study was to investigate whether changes to the epigenetic landscape of the HPV16 genome are responsible for the differential viral expression in clones from the polyclonal W12 cell line, genomic sequencing of the LCR of all W12 clones was carried out to determine any mutations and dismiss the possibility of aberrant transcription factor binding and consequent differential virus gene expression. Polymerase chain reaction (PCR) of the LCR genomic region and subsequent sequencing was carried out on twelve different integrant clones derived from the W12 Series2 cell line, two early passage W12 episomal cell lines as well as SiHa cells and compared to the published W12E DNA sequence (GenBank: AF125673.1). Two forward and two reverse orientation primers spread across the HPV16 LCR genome were used to amplify viral DNA to the required concentration for DNA sequencing (20 ng/100 bp in 10 μl) (Figure 3.2 A). Optimisation of PCR conditions for the amplification of 7045F-223R was carried out using a range of $MgCl_2$ concentrations and primer annealing temperatures. This indicated that 62.5 °C and 2.5 mM $MgCl_2$ were the most favourable conditions based on product band intensity and these conditions were carried forward (Figure 3.2 B). Full length PCR products (7045F-223R) were only produced at the required concentration for sequencing for W12 cell lines Par1, Par2, A5 and G2 (Figure 3.2 C; lanes 1, 2, 4 and 8, respectively); hence it was decided to use primer pairs 7045F-7681R (Figure 3.2 D) and 7555F-223R (Figure 3.2 E) to re-amplify the LCR of all lines. DNA samples obtained by gel extraction were then sequenced and the data overlaid to compare the sequences of each clone. Analysis indicated there were no genetic mutations in the W12 clones, apart from a single nucleotide deletion in one of the seven integrant copies of the viral genome in clone R2 (Figure 3.3). In addition, differences in the SiHa genomic sequence compared to W12E correlated with the published SiHa sequence (GenBank:

AF001600.1/AF001599.1) (Appendix 1). The lack of a genomic mutation after analysis of the LCR sequence of the W12 clones indicated that genetic factors did not differentially influence viral expression across the clones.

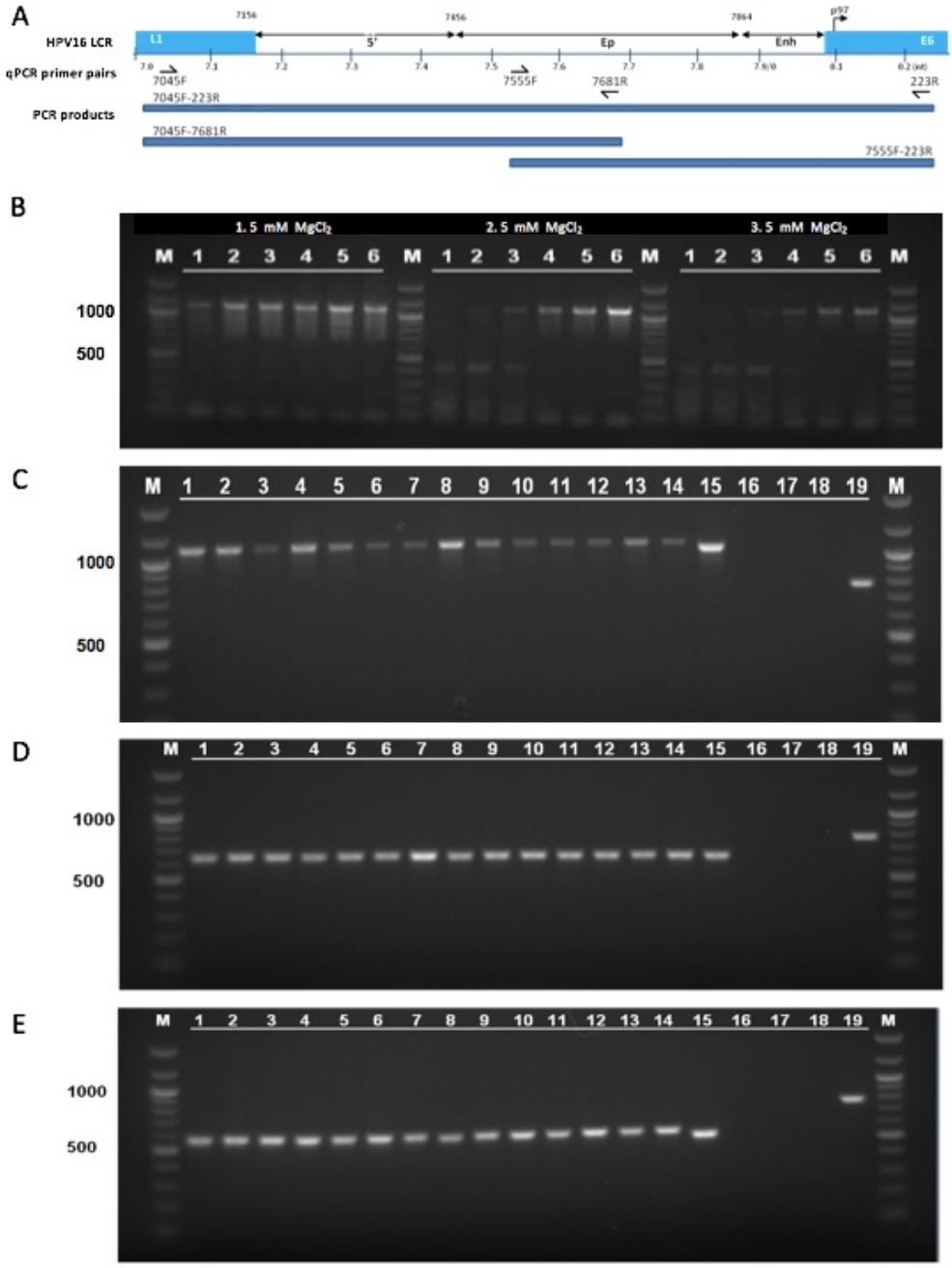

Figure 3.2: Amplification of the long control region (LCR) by PCR across a panel of 14 W12 integrant clones. A) Diagram showing the location of PCR primer sets specific for the HPV16 LCR region and the amplified product length. Numbers refer to the location of the 5 end for the forward primer relative to the start of the annotated HPV16 genome. B) Optimisation of conditions for PCR of HPV16 LCR using primers 7045F-223R. PCR reactions were conducted at a range of temperatures; 1=50.0°C, 2=52.5°C, 3=55.0°C, 4=57.5°C, 5=60.0°C, 6=62.5°C at three different concentrations of $MgCl_2$ (1.5, 2.5 and 3.5 mM) for 50 PCR cycles. M is a marker (100 bp ladder), and each 50 μl PCR sample was run on a 1 % agarose gel. PCR reaction of the LCR of numerous W12 cell lines using primers (C) 7045F-223R, (D) 7045F-7681R and (E) 7555F-223R. PCR reactions were conducted at 62.5°C with a 2.5 mM concentration of $MgCl_2$ using different W12 cell lines; 1=W12 Par1, 2=W12 Par2, 3=W12 Cl. 3, 4=W12 Cl. A5, 5=W12 Cl. D2, 6=W12 Cl. E3, 7=W12 Cl. F, 8=W12 Cl. G2, 9=W12 Cl. H, 10=W12 Cl. H2, 11=W12 Cl. J3, 12=W12 Cl. Q, 13=W12 Cl. R2, 14=W12 Cl. S2. Control reactions included; 15=SiHa, 16=NCx/6, 17=W12 Par1 (no primers) [neg ctrl], 18=reaction specific primer pair (no DNA) [neg ctrl], 19=W12 Par1 (E6-E7) [pos ctrl]. Targets were amplified using primer pairs for 50 PCR cycles. 20 μl of each sample was loaded onto a 1 % agarose gel.

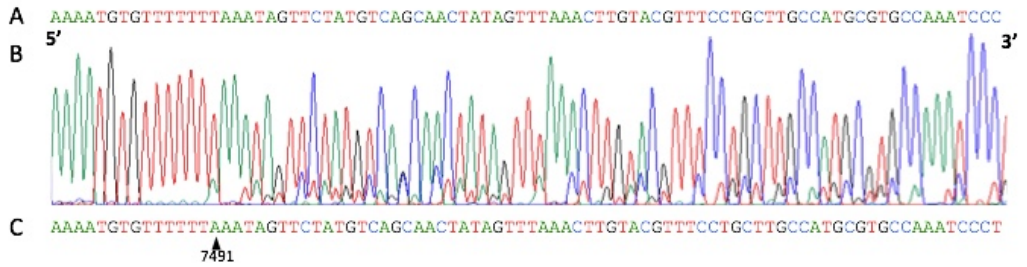

Figure 3.3: Genomic sequencing of the long control region (LCR) of W12 clone R2. A) DNA sequence called computationally by the program Chromas Pro; bases determined by the height of each peak. B) Chromatogram produced by Sanger sequencing the PCR product generated across the LCR of the HPV16 genome. C) DNA sequence of the frame shift caused by the apparent deletion of a thymine residue at position 7491. A - green, T- red, C - blue, G - black.

3.2.2 HPV16 oncogene expression from the W12 integrant clones depends upon the level of association of activating or repressive chromatin marks.

Having identified that the cause of differential HPV16 expression in the W12 integrant clones was not due to a genetic mutation in the LCR of the viral genome, the aim was then to assess the epigenetic environment of the W12 integrant clones with viral genome copy number less than four (F, A5, D2, H, G2), which have a wide variation in E6/E7 transcript abundance per template between them. In the first instance, chromatin immunoprecipitation reactions were performed using antibodies specific to active histone post-translational modifications (PTMs) to quantify levels at the integrated virus genome. ChIP analysis showed that higher virus oncogene expression per template was associated with greater levels of H3K4me3 across the early genes E6, E7 and E1, a hallmark of transcriptional activity. Although the association of H3K4me3 with the HPV genome appeared to occur that the same position in each of the W12 clones — the central LCR — the profiles of W12 clones F and A5 altered slightly with the highest peak in F located over gene E6 and the highest peak in clone A5 over E7. In W12 clone G2 the peak of H3K4me3 abundance over E6 was much smaller than the highly expressing clones, and, in contrast to W12 clones F, A5, D2 and H the abundance of H3K3me3 appeared to be maintain across E7 and E1 portions of the HPV16 genome. Additionally, the H3K4me3 PTM was

absent from L1 gene in all five clones. (Figure 3.4 A).

The cells with high expression per template also showed enrichment of the histone PTMs associated with gene enhancer/promoter regions H3K4me1 and H3K27ac (Figure 3.4 B and C). The H3K4me1 ChIP profiles of the highly expressing W12 clones F and A5 were very similar with peaks over the central/3' LCR and mid-way through the early gene E1, however, the profiles for the medium expressing clones D2 and H were very different. As with F and A5, two main peaks of H3K4me1 were observed on W12 clone H, one over the LCR and the other over E7 — this is upstream of the second peak observed in clones F and A5. In W12 clone D2 however, H3K4me1 was more abundant over the 3' end of the L1 gene remained high over the 5' end of the LCR. Abundance then sharply decreased at the central LCR and slowly increased until a broad peak over E1. The profile of clone G2 was similar to that of F/A5, however the level of the activating H3K4me1 mark was much reduced.

Whereas there were peaks of H3K4me1 distributed across the length of the virus genome tested, the H3K27ac mark was found predominantly over the LCR and early genes of each virus genome (Figure 3.4 C). The main deviation from this trend was seen in clones A5 and D2, both of which had greater peaks of the H3K27ac at the 5' end of E1, over the region of the E8 promoter (splice donor SD1302)[191].

After determination of the association of active PTMs with the integrated HPV16 genome, ChIP analysis was also carried out using antibodies against hallmarks of repressive chromatin. In this instance the opposite observation was made; lower levels of expression per template were associated with higher levels of each PTM tested, namely H3K9me2, H3K27me2 and H3K27me3 (Figure 3.5 A-C).

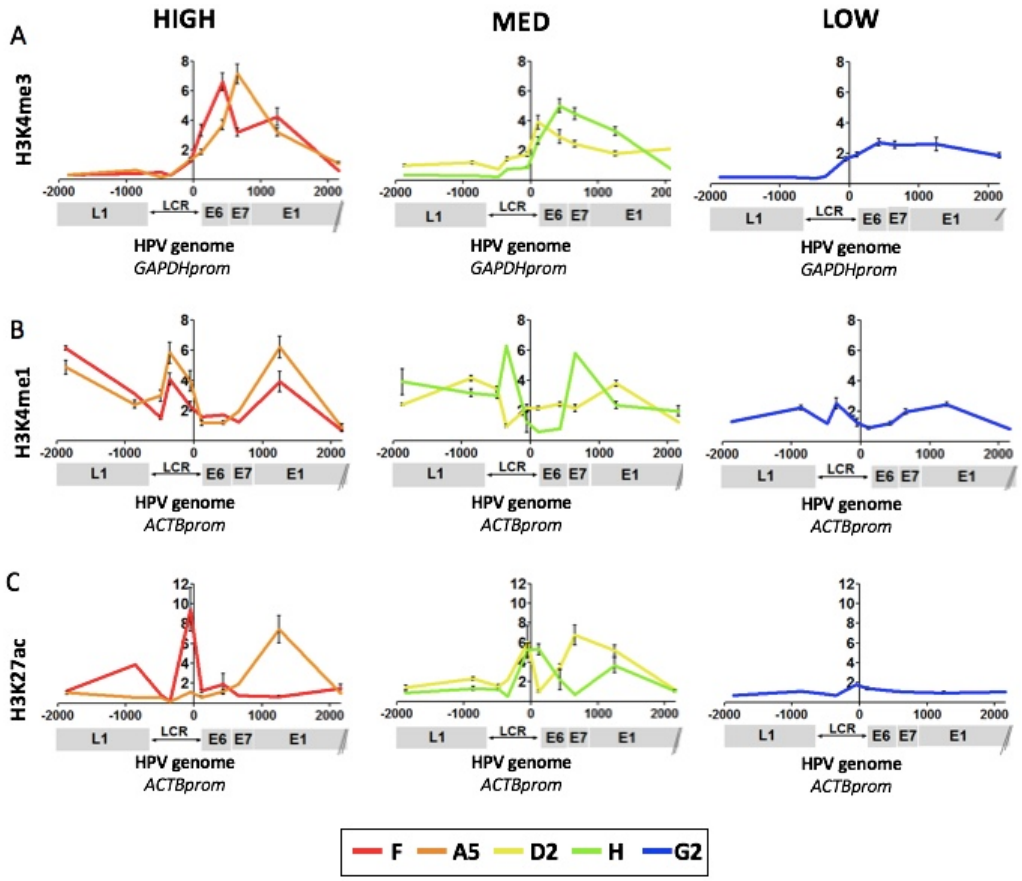

Figure 3.4: Integrated HPV16 genome associations with active histone post-translational modifications (PTMs). Level of association of the histone PTM (A) H3K4me3, as well as the transcriptional enhancer marks (B) H3K4me1 and (C) H3K27ac. In each graph, the y-axis shows the relative levels of enrichment normalised to host control target regions as indicated in italics under each panel. The x-axis and underlying schematic show the region of the HPV16 genome analysed. Data presented as mean ± SEM (n=2). In all panels, data for each of the five clones are colour coded according to the key at the foot of the figure.

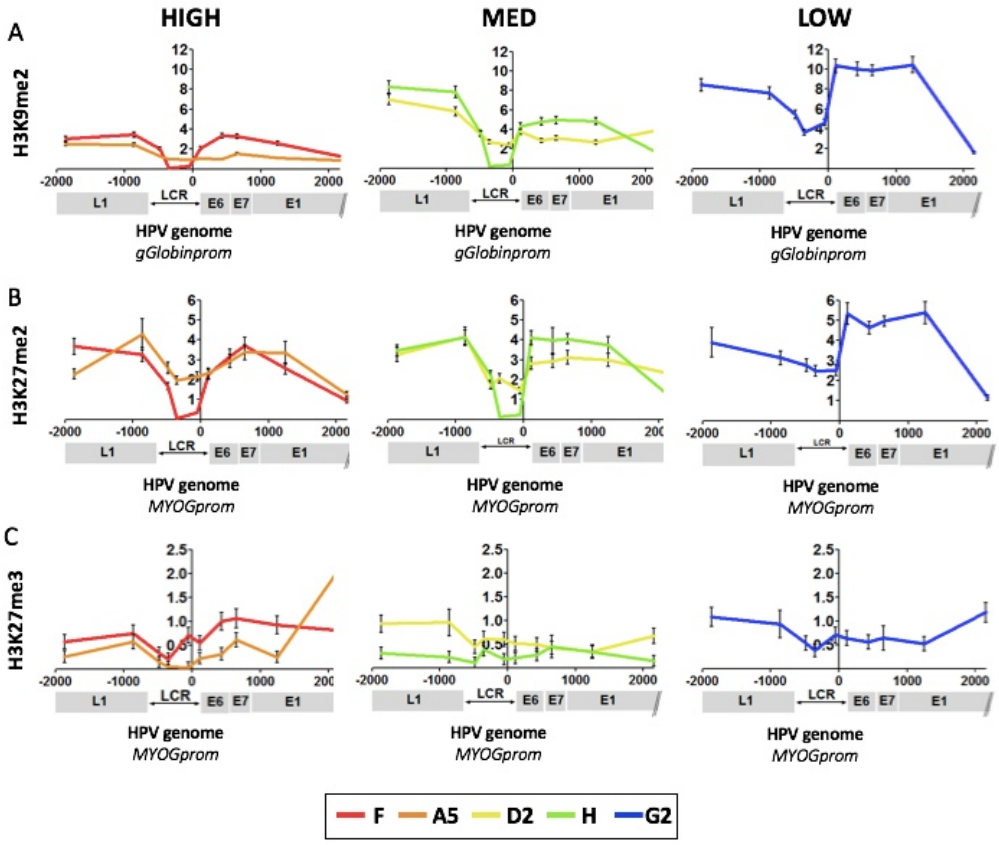

Figure 3.5: Integrated HPV16 genome associations with repressive histone PTMs. Level of association of the histone PTM (A) H3K9me2, (B) H3K27me2 and (C) H3K27me3. In each graph, the y-axis shows the relative levels of enrichment normalised to host control target regions as indicated in italics under each panel. The x-axis and underlying schematic show the region of the HPV16 genome analysed. Data presented as mean ± SEM (n=2). In all panels, data for each of the five clones are colour coded according to the key at the foot of the figure.

3.2.3 Level of HPV16 integrant transcription per template correlates with association of histone acetylation modifying enzymes

Acetylation of histone tails as a mark of transcriptional activation of the integrated HPV16 genome was next focussed upon. In addition to previous H3K27ac, ChIP analysis of general histone 3 acetylation level (H3ac) was performed (Figure 3.6 A). Again, clones with higher expression per template were associated with higher levels of H3ac, with the greatest abundance seen downstream of the transcription start site over the E6 and E7 genes in all of the W12 clones. To begin investigations into the enzymes responsible for writing acetyl-PTMs, the association of global histone acetyltransferase (HAT) enzymes p300 and TIP60 with the virus genome was also determined by ChIP. Analysis showed that the association of p300 was greatest at the genomes of the high expressing clones, with lesser levels at the low expressing clone G2 genome (Figure 3.6 B). Specifcally, in W12 clones F and A5 there was a sharp peak of enzyme associated with the 3' end of the LCR of the HPV16 genome — virus early promoter (p97) — as well as smaller increases of p300 association over the E6 and E7 loci and the 5' end of E1 (E8 promoter). In contrast, the medium- and low-expressing clones D2, H and G2 had reasonably consistent levels of p300 associated with the HPV16 genome, with a slight elevation of abundance across the viral early promoter.

ChIP analysis of association of HAT TIP60 with the virus genome again showed that clones with hisger expression per template were associated with higher levels of the activating enzyme (Figure 3.6 C). The distribution profiles of the two highly expressing clones F and A5 were quite different. Both clones had a peak of TIP60 abundance over the 3' LCR, however in clone A5 this was slightly shifted and also included the 5' end of E6. Additionally in clone A5 there was a broad peak spanning from the 3' end of E7 to the centre of the E1 gene, encompassing the E8 promoter. This peak was absent in the medium- and low-expressing clones D2, H and G2, respectively.

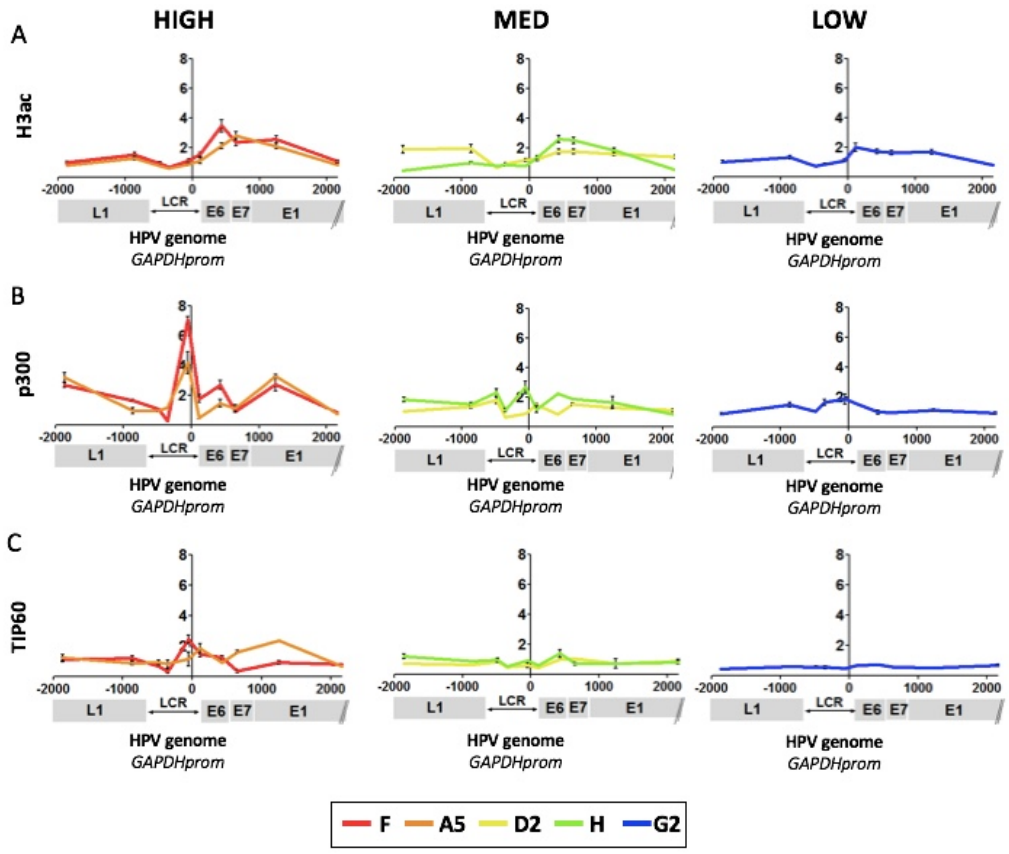

Figure 3.6: Integrated HPV16 genome associations with histone acetylation and HAT abundance. Level of association of the histone PTM (A) H3ac, and associated HAT enzymes (B) p300 and (C) TIP60. In each graph, the y-axis shows the relative levels of enrichment normalised to host control target regions as indicated in italics under each panel. The x-axis and underlying schematic show the region of the HPV16 genome analysed. Data presented as mean ± SEM (n=2). In all panels, data for each of the five clones are colour coded according to the key at the foot of the figure.

As such a clear association between enzymes that write activating chromatin marks and the high-expressing clones had been identified, the opposite scenario was then tested. ChIP analysis was performed on all five clones using an antibody specific for histone deacetylase 1 (HDAC1), an enzyme that removes acetyl groups from histone lysine residues. The resultant deactelyation makes DNA less accessible to RNA polymerase machinery; HDAC enzymatic activity is associated with repressed chromatin and a lack of transcription[126]. In this instance, cells with low virus transcript levels per template (G2) showed much higher abundance of HDAC1 (Figure 3.7 A). To determine the effect of inhibiting enzymatic action of HDACs on the level of viral transcription across the panel of W12 clones, cells were treated with the class I/II HDAC-specific small-molecule inhibitor Trischostatin-A (TSA). After 16

hours of treatment the relative levels of E6/E7 expression had increased in all of the clones. The trend showed that de-repression of transcription from the viral genome increased from the high-, medium-, low-expressing clones with a significantly greater increase in E6/E7 transcript levels in clone G2 than in F (Figure 3.7 B).

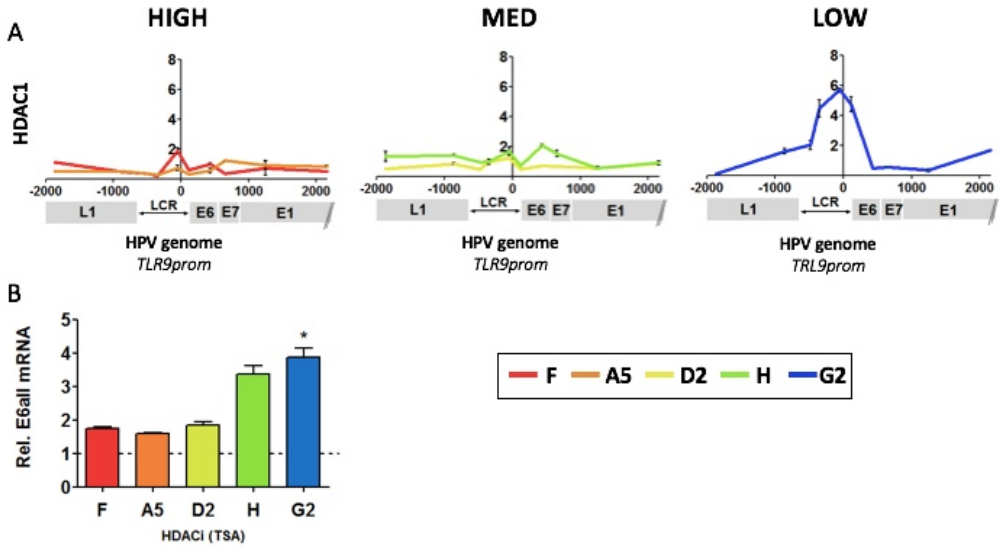

Figure 3.7: Integrated HPV16 genome associations with HDAC abundance/activity.(A) Level of association of the enzyme HDAC1. The y-axis shows the relative levels of enrichment and normalised to host control target regions as indicated in italics under each panel. The x-axis and underlying schematic show the region of the HPV16 genome analysed. Data presented as mean ± SEM (n=2). (B) Changes in HPV16 E6/E7 transcript levels following type I/type II HDAC inhibition with Trichostatin A (TSA) (n=3). Asterisks refer to comparisons with control vehicle-treated cells set to 1. *P*-values (Students t-test): $*P<0.05$, error bars=SEM. In all panels, data for each of the five clones are colour coded according to the key at the foot of the figure.

The abundance of HAT and HDAC enzymes at the integrated W12 genome bore no relation to the total levels of enzyme in each clone. Western blot analysis of protein abundance was conducted (Figure 3.8 A-C) and densitometric quantitation indicated that, in the majority of cases, enzyme abundance in each individual clone was similar to the five-clone average, which was set at one (Figure 3.8 E-G). Notable deviations were the levels of p300 in clone G2, which was significantly lower than the high- and medium-expressing clones; however this trend was the opposite for the HAT TIP60. It is pertinent to note that alternative splicing of the gene encoding TIP60 results in at least four different protein isoforms; isoform 2 (KAT5_2) encodes a 513 amino acid protein (58.5 kDa) and is accepted as the canonical form of the protein, it is this that was measured by Western Blot analysis. Microarray analysis compared the total transcript abundance of each enzyme in the individual clones compared with the episomal parental W12 cell line (Figure 3.8 I-K). Together this showed that the total cellular amount of each enzyme was similar across the panel of W12 clones regardless of levels of viral transcript per template, indicating specific enzyme loading onto integrated virus chromatin.

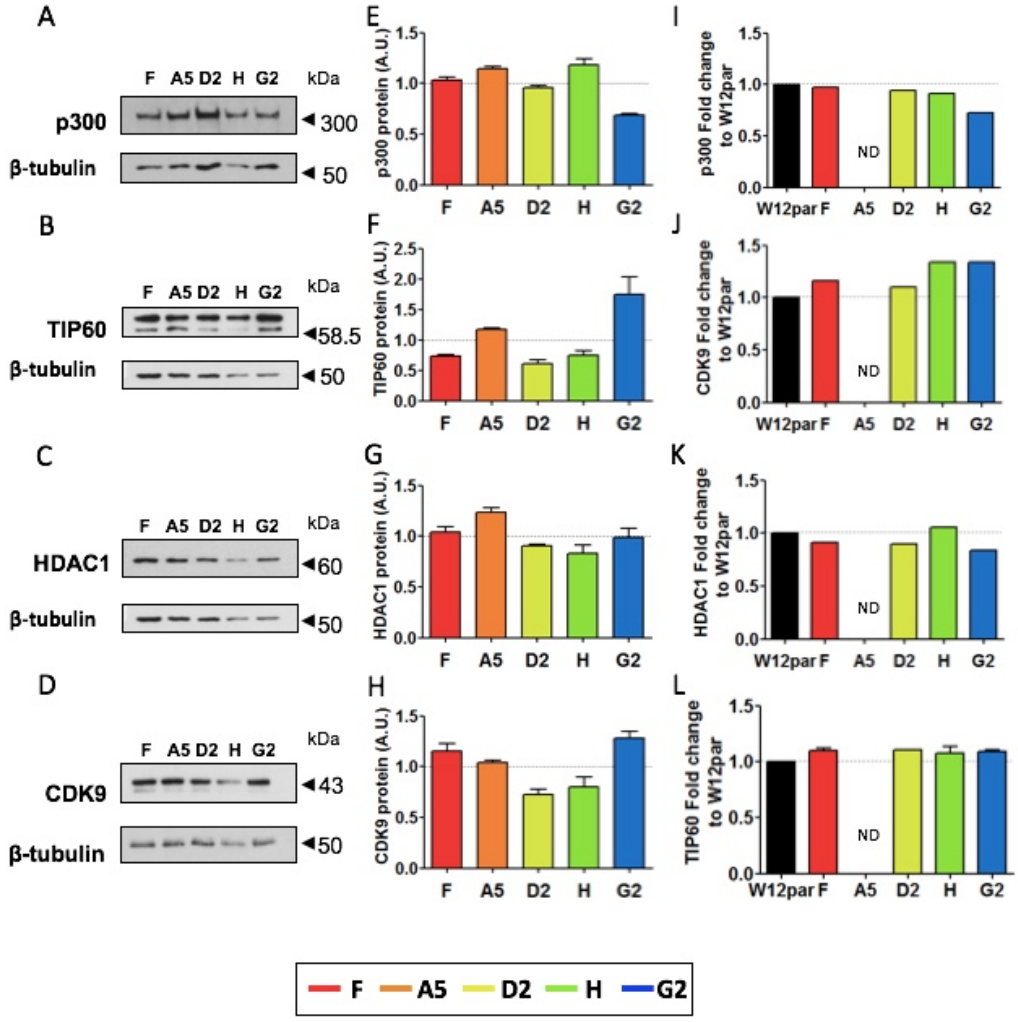

Figure 3.8: Overall protein levels per cell of enzymes detected by Western Blot. Baseline levels were determined by (A-D) Western blot, (E-H) quantified using Image J and (I-L) analysed by microarray. The enzymes assessed were: p300 (n=1) (A, E, I), TIP60 (n=2) (B, F, J), HDAC1 (n=1) (C, G, K) and CDK9 (n=2) (D, H, L). For each blot, levels of the target and loading control were quantified at two or three different exposures and the mean values determined. The results were normalised to the loading control and referenced to 5/(sum of individual values). Where more than one blot was performed, a representative image is shown. (I-L) Baseline levels were also determined by microarray analysis (A5 data not available). For each panel the expression level fold change of each integrant clone was compared to episomal W12par, which was set to 1. A.U. = arbitrary units. In all panels, data for each of the five clones are colour coded according to the key at the foot of the figure.

To test the functional significance of HAT recruitment in determining levels of HPV16 oncogene expression from integrated genomes, specific small molecule inhibitors and siRNAs were used to inhibit and knockdown the enzymes, respectively. Cells with the highest and lowest level of virus early gene expression per template (clones F and G2, respectively) were used in these analyses to simplify interpretation of the resulting data. In the first instance a serial dilution of p300-specific small molecule inhibitor C646[192] was carried out in order to find the lowest drug concentration that produced a significant change to the E6/E7 mRNA levels of both clones, F and G2, compared to vehicle-treated control (Figure 3.9 A and B). The use of the lowest concentration of each drug limits the number of off-target effects caused by small molecule inhibitor treatment. At the lower concentrations tested, viral oncogene expression of clone F was similar if not higher (although not significantly) than control but was significantly reduced at 25 μM. Lower concentrations of C646 treatment on G2 cells had a stronger effect on viral oncogene expression with 10 μM and 25 μM treatment both resulting in a significant decrease. A series dilution was also performed for the TIP60-specific small molecule inhibitor MG149[193] (Figure 3.9 C and D). A drug concentration of 150 μM resulted in a significant decrease in E6/E7 mRNA levels in both clone F and G2, and as a result this concentration was carried forward to future experiments.

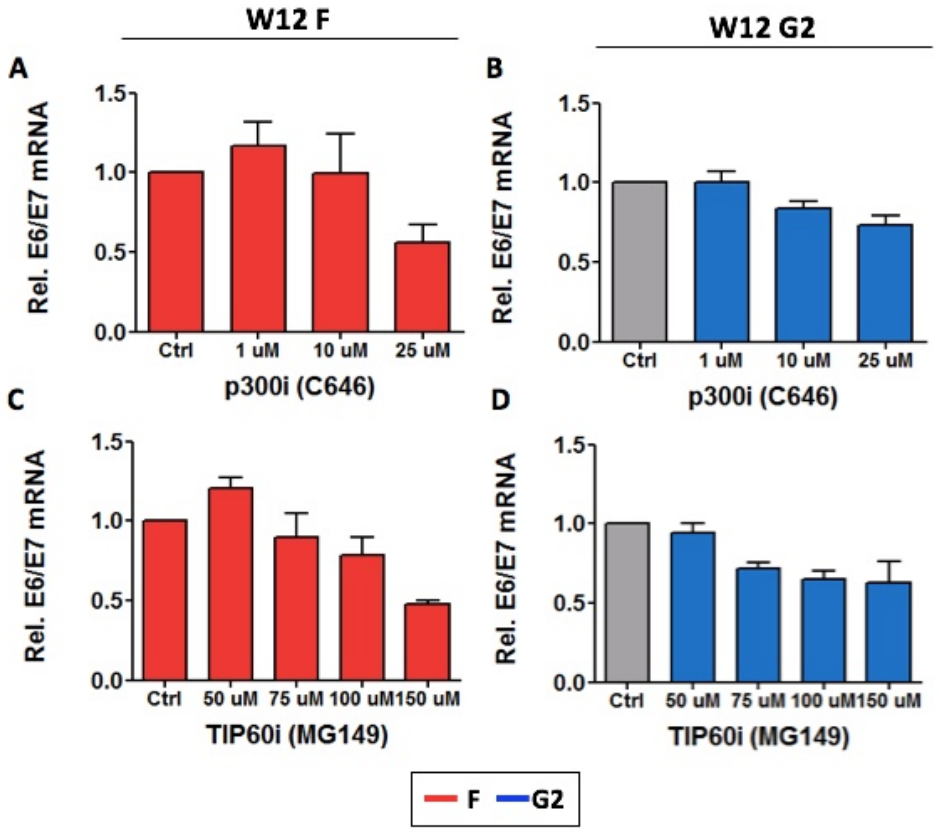

Figure 3.9: Histone acetyltransferase (HAT) small molecule inhibitor optimisation. qRT-PCR analysis of HPV16 mean E6/E7 transcript levels in W12 clone F (A, C) and W12 clone G2 (B, D) cells following treatment for 16 hours with (top row) p300i (C646) and (bottom row) TIP60i (MG149), respectively, at the concentrations indicated. Resultant gene expression was referenced to vehicle-treated cells (Ctrl bar), which was set to 1. Error bars=SEM, n=2. In all panels, data for each of the two clones are colour coded according to the key at the foot of the figure.

To support the small molecule inhibitor studies, enzyme depletion through siRNA transfection was employed. Using the specific protocol for W12 keratinocytes (as detailed in section 2.1.7 in the Materials and Methods), a dilution series of p300- and TIP60-specific siRNA was carried out. From the series, the lowest concentration of siRNA (utilised to abrogate off-target effects) that produced consistent, significant gene knockdown by way of mRNA reduction was carried forward to future experiments. W12 clones A5 and H were used in this experiment as F and G2 were unavailable for use at this time. The dilution series ranged from 10-100 nM siRNA and affects were determined using mRNA quantification of each enzyme, p300 (Figure 3.10 A and B) and TIP60 (Figure 3.10 C and D), compared to a non-targeting control sample at the same concentration. Transfection results were most consistent for W12 clone H; however, for both clones transfection with 10 nM siRNA resulted

in significant knockdown of both targets (P<0.0001).

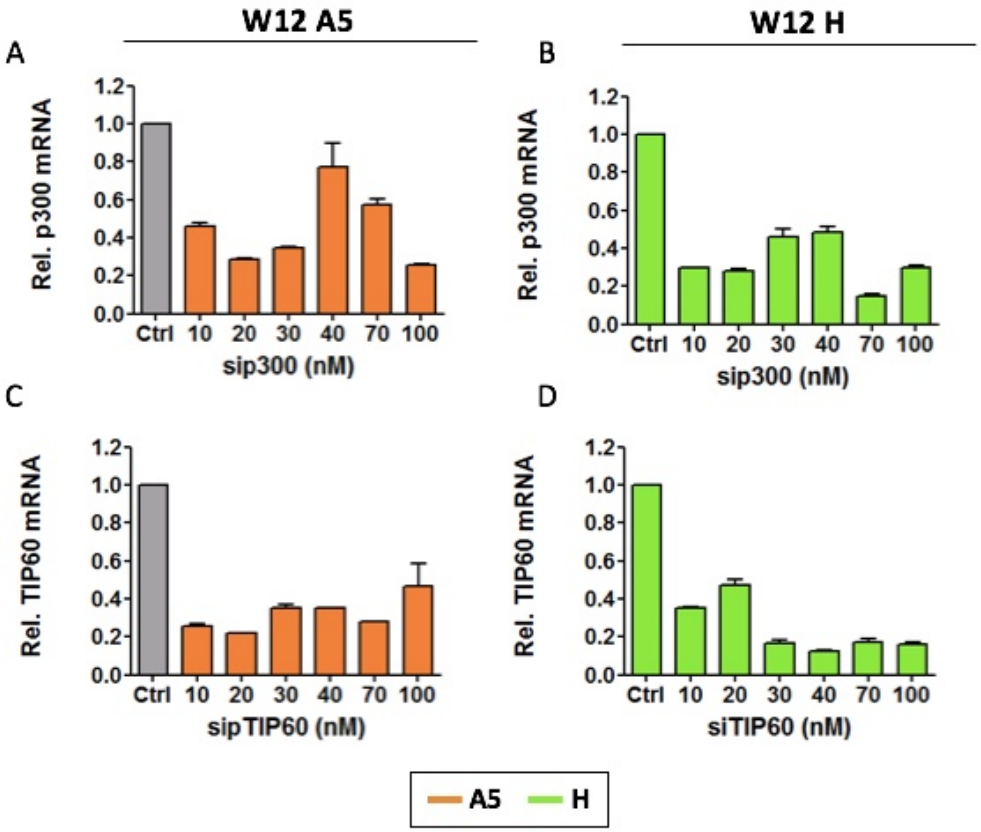

Figure 3.10: Optimisation of siRNA knockdown of enzyme target genes. Target mRNA levels determined by qRT-PCR in W12 clones A5 and H following treatment with (A, B) sip300 and (C, D) siTIP60, at the concentrations indicated. The level of gene knockdown was referenced to cells treated with siNTC (non-targeting control) at the appropriate concentration, which was set to 1 (Ctrl bar). Error bars = SD, n=1. In all panels, data for each of the two clones are colour coded according to the key at the foot of the figure.

Using siRNA at 10 nM, levels of p300 and TIP60 mRNA were depleted to similar levels in W12 clones F and G2 (Figure 3.11 A and E). The protein abundance of each enzyme following siRNA treatment was determined by Western blot analysis and quantified using ImageJ software (Figure 3.11 B and F). siRNA treatment of clone F did not produce as great a knockdown of enzyme protein compared with clone G2; in addition, sip300 was slightly less efficient at reducing the amount of protein than siTIP60. However, the reduction of each enzyme in both clones had a significant impact on viral oncogene expression (Figure 3.11 C and G). Knockdown of the HAT enzymes resulted in decreased levels of E6/E7 mRNA in both W12 F and G2 compared to non-targeting control samples. Additionally, there were significantly

greater reductions in viral oncogene expression in clone F compared with G2, despite greater protein knockdown in clone G2. The differential sensitivity between clones F and G2 was also observed following p300- and TIP60-specific small molecule inhibitor treatment at 25 μM and 150 μM respectively (Figure 3.11 D and H).

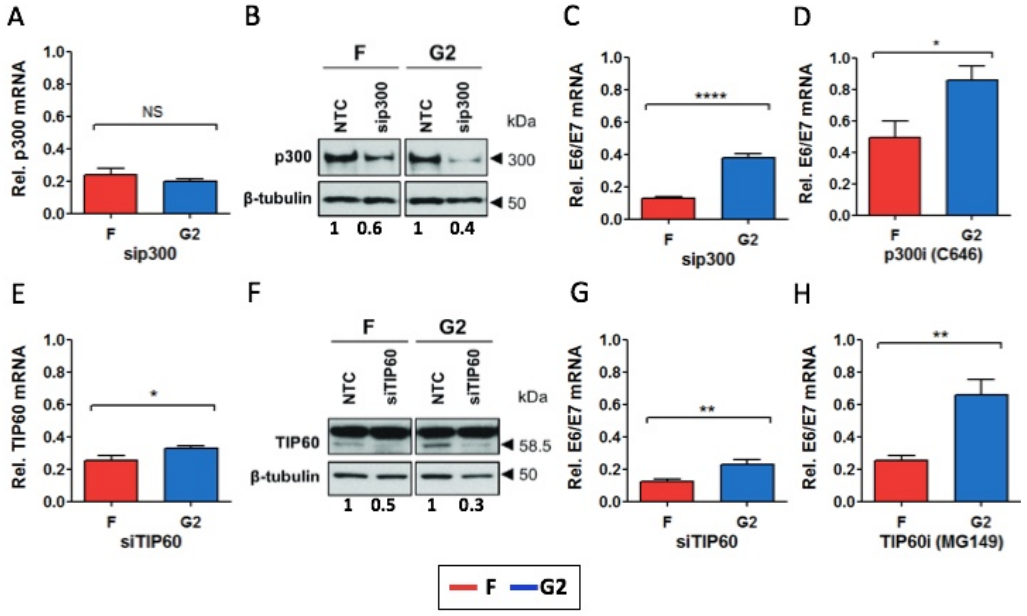

Figure 3.11: Effects of histone acetyltransferase (HAT) depletion/inhibition on HPV16 oncogene expression. Depletion/inhibition in W12 clones F and G2 of HAT enzymes p300 (n=4) (upper row) and TIP60 (n=6) (lower row). The panels show levels of depletion of target mRNAs (A, E), target protein (B, F) and HPV16 E6/E7 transcripts (C, G) in siRNA-treated vs. non-targeting control (NTC)-treated cells. Panels (D and H) show HPV16 E6/E7 transcript levels in cells treated with specific small molecule inhibitors, vs. cells treated with vehicle only. (B, F) Protein samples from all replicate experiments were combined for use in Western blot and specific protein bands were normalised to the loading control (-tubulin) and referenced to NTC set to 1. Error bars=SEM. *P*-values (Students t-test): $*P<0.05$, $**P<0.01$, $***P<0.001$, $****P<0.0001$. In all panels, data for each of the two clones are colour coded according to the key at the foot of the figure.

To ensure that the effects of small molecule inhibitors on virus transcription were driven by direct changes to the genome chromatin structure, ChIP analysis of inhibitor treated samples was performed. W12 clone F was chosen for this experiment as enzymatic inhibition had a resulted in the greatest decrease in viral oncogene expression. The use of small molecule inhibitors was necessary as the chromatin yield from siRNA treated samples was too low. p300 is known to specifically mediate the acetylation of H3K27, hence this PTM was used to detect changes as a result of enzyme inhibition. The abundance of H3K27ac associated with the HPV16 genome in the C646 treated sample was very significantly reduced compared with control (Figure 3.12 A and B). In slight contrast to Figure 3.4 C, the H3K27ac profile of the control treated W12 clone F cells was broader over the E6 and E7 loci and less restricted to the LCR. There was also significant removal of the activating histone mark H2AK5ac, specifically associated with the TIP60 enzyme, following inhibitor treatment (Figure 3.12 C and D). These data indicate that the effects on viral transcription as a result of enzyme inhibition were a direct effect of changes in chromatin structure.

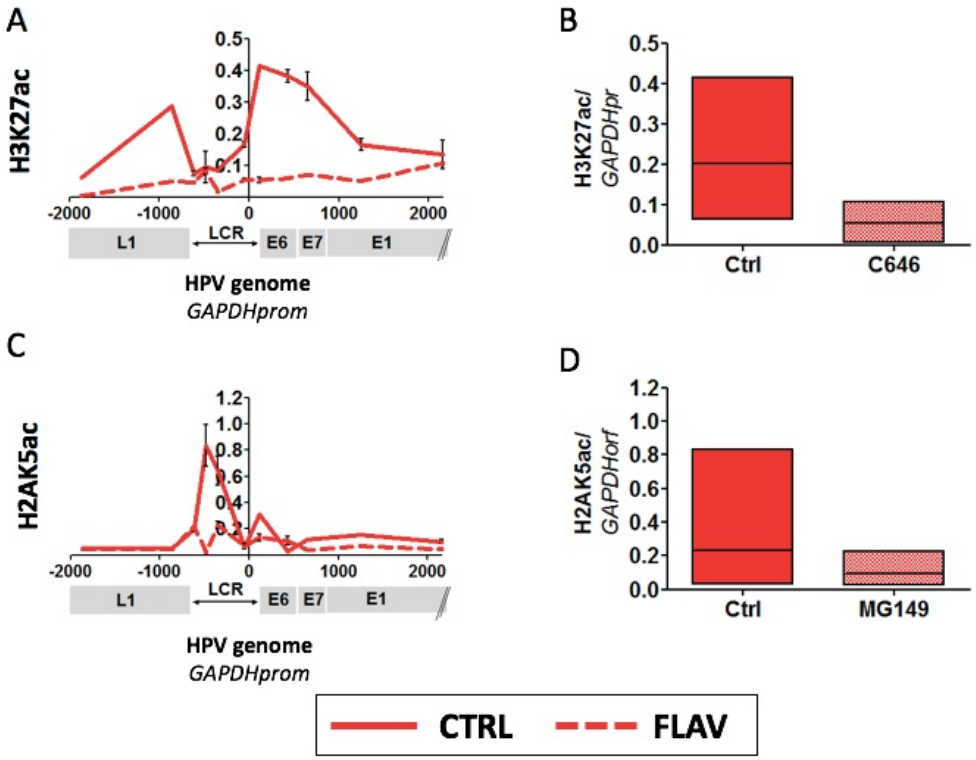

Figure 3.12: Effects of histone acetyltransferase (HAT) inhibition on the abundance and distribution activating chromatin PTMs in W12 clone F. Levels of association of H3K27ac (A, B) following p300 specific small molecule inhibition (C646), and H2AK5ac (C, D) following TIP60 small molecule inhibition (MG149), both for 16 hours. In each graph (A and C), the y-axis shows the relative levels of enrichment normalised to host control target regions as indicated in italics under each panel. The x-axis and underlying schematic show the region of the HPV16 genome analysed. Data presented as mean ± SD (n=1). Panels B and D show an alternative representation of data from A and C; high-low bar graphs with the mean value represented by a centre line.

3.2.4 HPV16 transcript levels per template correlate with active RNA polymerase II (RNAPII) level and activating complex P-TEFb

The role of RNA polymerase II (RNAPII) was next assessed in determining the differential viral oncogene expression seen in the integrated W12 clones with copy number less than four. Initially, the abundance and association of total RNAPII with the HPV16 genome was analysed by ChIP and no overall differences were seen across the high-, medium-, and low-expressing clones in total loading (Figure 3.13 A). However, cells with lower virus expression per template showed a significantly greater association with the poised/paused or stalled form of RNAPII, Serine 5 phosphorylated (Ser5P), particularly across the early genes (E6, E7 and E1) (Figure 3.13 B). Conversely, cells with higher expression per template showed higher amounts of the active/elongating form of RNAPII, Serine 2 phosphorylated (Ser2P), across the virus LCR and early genes (Figure 3.13 C).

The positive transcription elongation factor (P-TEFb) complex is essential to switch paused RNAPII to its actively transcribing form. P-TEFb is comprised of Cyclin T1 and its kinase partner CDK9, which is responsible for the phosphorylation of the C-terminal domain repeats at Ser2. ChIP analysis of these components showed that there were also higher levels of both proteins associated with W12 clones, with high E6/E7 transcription per template (Figure 3.14 A and B). CDK9 protein quantification revealed that, as with acetyl-associated enzymes p300 and TIP60 and HDAC1, basal levels were similar across the panel of W12 clones (Figure 3.8 D).

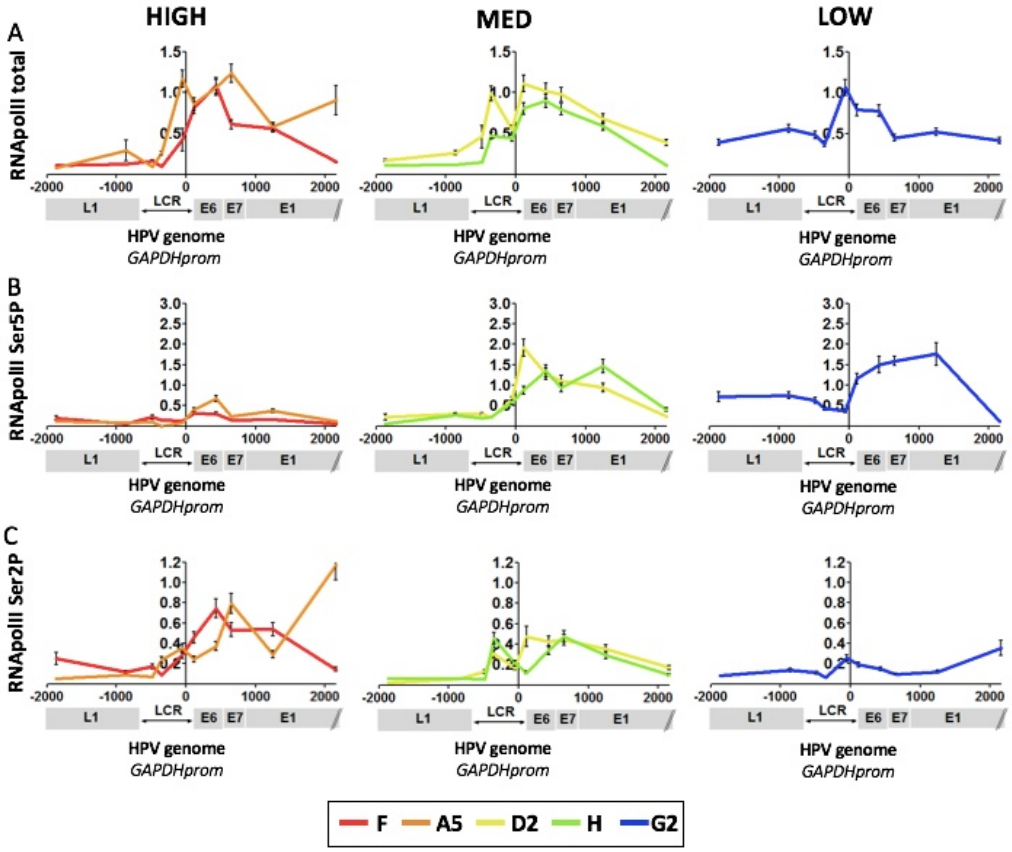

Figure 3.13: Integrated HPV16 genome associations with RNA polymerase II (RNAPII). Level of association of (A) total RNAPII, (B) RNAPII-Ser5P (poised/paused) and (C) RNAPII-Ser2P (active/elongating). In each graph, the y-axis shows the relative levels of enrichment normalised to host control target regions as indicated in italics under each panel. The x-axis and underlying schematic show the region of the HPV16 genome analysed. Data presented as mean ± SEM (n=2). In all panels, data for each of the five clones are colour coded according to the key at the foot of the figure.

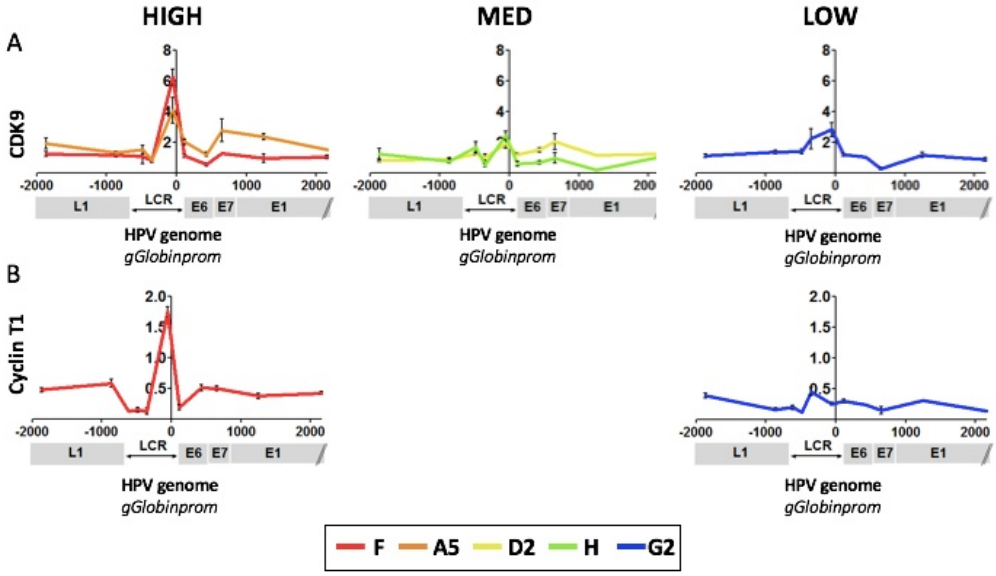

Figure 3.14: Integrated HPV16 genome associations with the components of the P-TEFb complex. Level of association of (A) CDK9 (n=3), and (B) cyclin T1 (not determined for clones A5, H and D2) (n=1). In each graph, the y-axis shows the relative levels of enrichment normalised to host control target regions as indicated in italics under each panel. The x-axis and underlying schematic show the region of the HPV16 genome analysed. Data presented as mean ± SEM. In all panels, data for each of the five clones are colour coded according to the key at the foot of the figure.

The functional significance of CDK9 recruitment in determining levels of HPV16 oncogene expression was tested by depletion and inhibition of the enzyme in W12 clones F and G2. Transfections were carried out with CDK9-specific siRNAs at 10 nM (Figure 3.15 A-D) and resulted in similar knockdown of CDK9 mRNA to approximately 20% compared to the non-targeting control (Figure 3.16 A). Residual CDK9 protein following siRNA treatment was analysed by Western blot and quantified (Figure 3.16 B). Treatment with siRNA CDK9-6 caused slightly greater reduction of CDK9 protein than CDK9-5. Depletion of CDK9 produced significantly greater reductions in viral oncogene expression in clone F (higher levels of virus transcript per template) than in clone G2 (lower levels of virus transcript per template) (Figure 3.16 C). The differential sensitivity between clone F and G2 as a result of CDK9 enzyme depletion mirrors that seen upon depletion of the HATs p300 and TIP60.

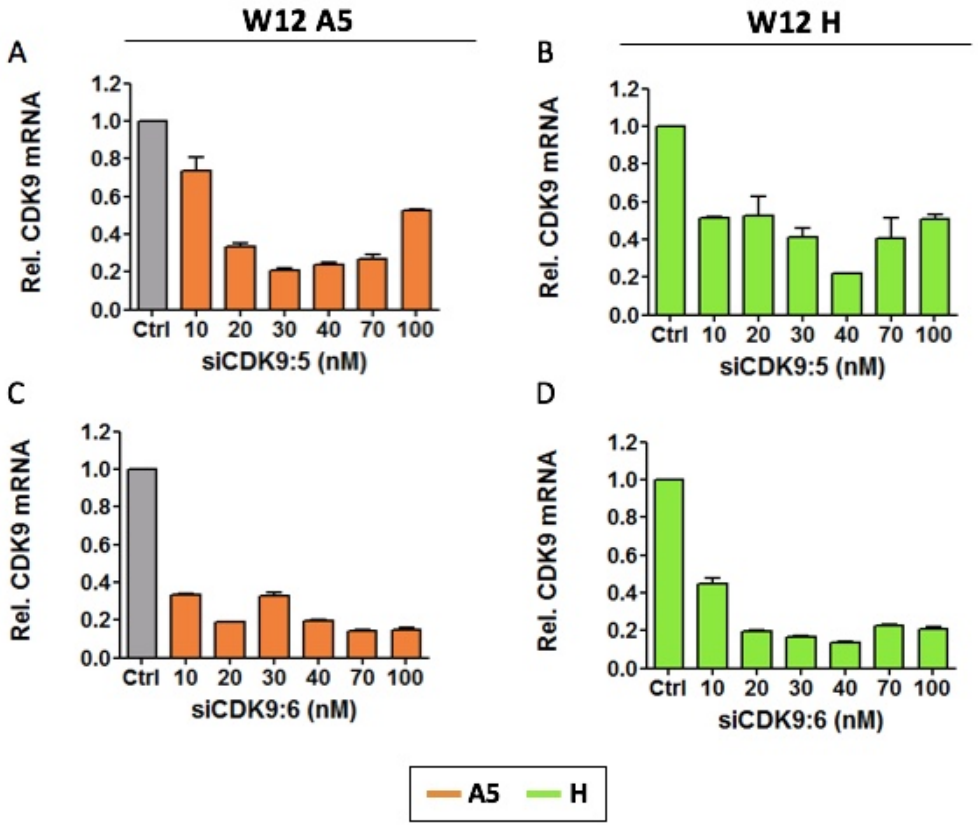

Figure 3.15: Optimisation of siRNA knockdown of CDK9 enzyme. Target mRNA levels determined by qRT-PCR in W12 clones A5 and H following treatment with (A, B) siCDK9_5, (C, D) siCDK9_6 at the concentrations indicated. The level of gene knockdown was referenced to cells treated with siNTC (non-targeting control) at the appropriate concentration, which was set to 1 (Ctrl bar). Error bars = SD, n=1. In all panels, data for each of the two clones are colour coded according to the key at the foot of the figure.

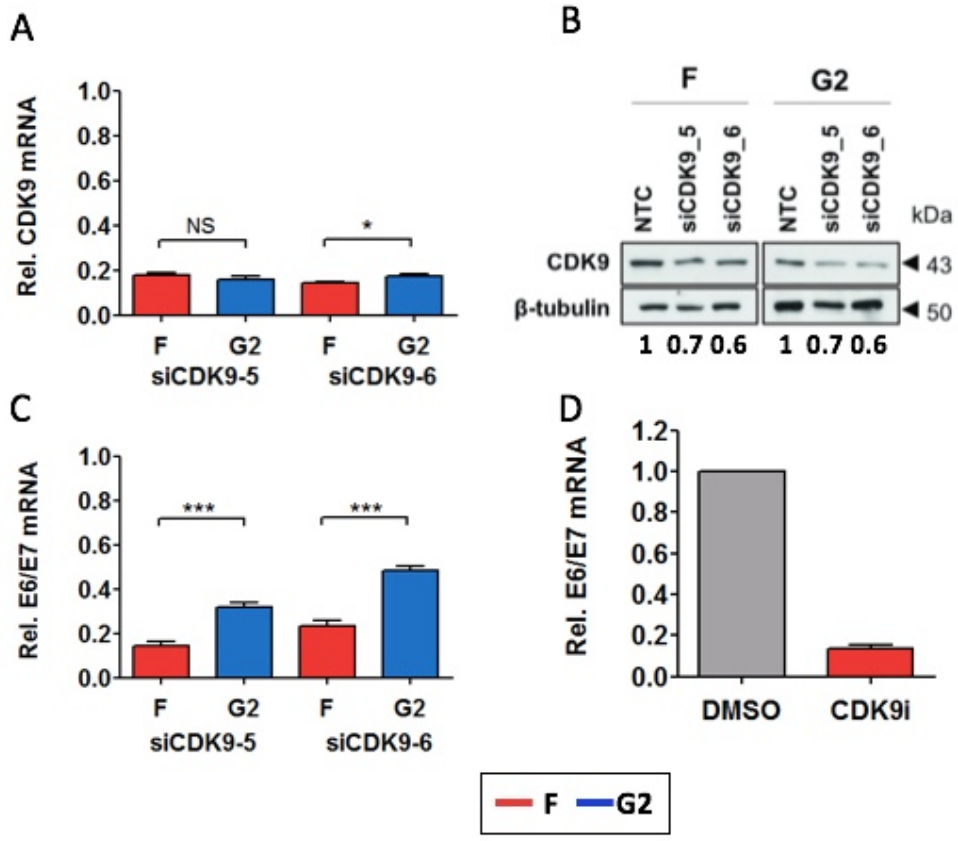

Figure 3.16: Effects of CDK9 depletion/inhibition on HPV16 oncogene expression. (A-C) Depletion of CDK9 using siRNAs (n=2), showing the levels of target mRNA (A) and protein (B), together with the changes in HPV16 E6/E7 transcript levels (C), in siRNA-treated vs. non-targeting control (NTC) -treated cells. (D) The effect of CDK9 inhibition on the E6/E7 transcript levels of W12 clone F cells following treatment with Flavopiridol for 16 hours vs. cells treated with vehicle only (set to 1) (n=2). Protein samples from all replicate experiments were combined for use in the Western blot and specific protein bands were normalised to the loading control (β-tubulin) and referenced to NTC, which was set to 1. Asterisks refer to comparisons with control vehicle-treated cells set to 1. Error bars = SEM. P-values (Students t-test): $^{*}P<0.05$, $^{***}P<0.001$. In all panels, data for each of the two clones are colour coded according to the key at the foot of the figure.

To ensure the transcript level changes were a direct effect of CDK9 inhibition, the ChIP profiles of RNApolII (and its associated forms) as well as hallmarks of activating and repressive chromatin structure were compared. CDK9-specific small molecule inhibitor Flavopiridol generated an 87% reduction in E6/E7 transcript levels in W12 clone F compared with a vehicle-only treated control (Figure 3.16 D) and chromatin from both samples was used for ChIP analysis. Flavopiridol treatment produced no significant change in overall levels of CDK9 associated with the integrated HPV16 genome (Figure 3.17 A and B). Despite no overall change in abundance, the association profile of CDK9 with the HPV16 genome changed as a result of flavopiridol inhibition; the peak over the viral early promoter was reduced and an additional peak over the early gene E6 was present when compared to the control sample.

Flavopiridol treatment of W12 clone F cells resulted in a significantly reduced level of total RNAPII and this was particularly pertinent downstream of the virus early promoter (Figure 3.17 C and D). Interestly, the abundance of total RNAPII across the integrated LCR region was very different when the two samples were compared. In the control sample, the major peak of total RNAPII over the 3' of the LCR (p97) was t totally absent in the flavopiridol treated sample, with a peak over the 5' LCR (E2BS3 and 4) seen instead.

CDK9 inhibition resulted in a significant reduction in the active form of RNAPII, Ser2P; levels were near baseline over the promoter region of the integrated virus LCR and early genes (Figure 3.17 E and F). As with total RNAPII, a peak over the 5' LCR was seen in the flavopridol treated sample indicating that the polymerase is becoming stalled in this area upon CDK9 inhibition.

The abundance and distribution of PTMs were also affected by CDK9 inhibition. The amount of histone PTM of transcriptional activation, H3K4me3, was dramatically reduced, most noticeably over the p97 and E8 promoter (Figure 3.16 G and H). This was mirrored by a significantly increased level of H3K9me2, a mark of constitutive heterochromatin and transcriptional repression, across the integrated genome (Figure 3.17 I and J).

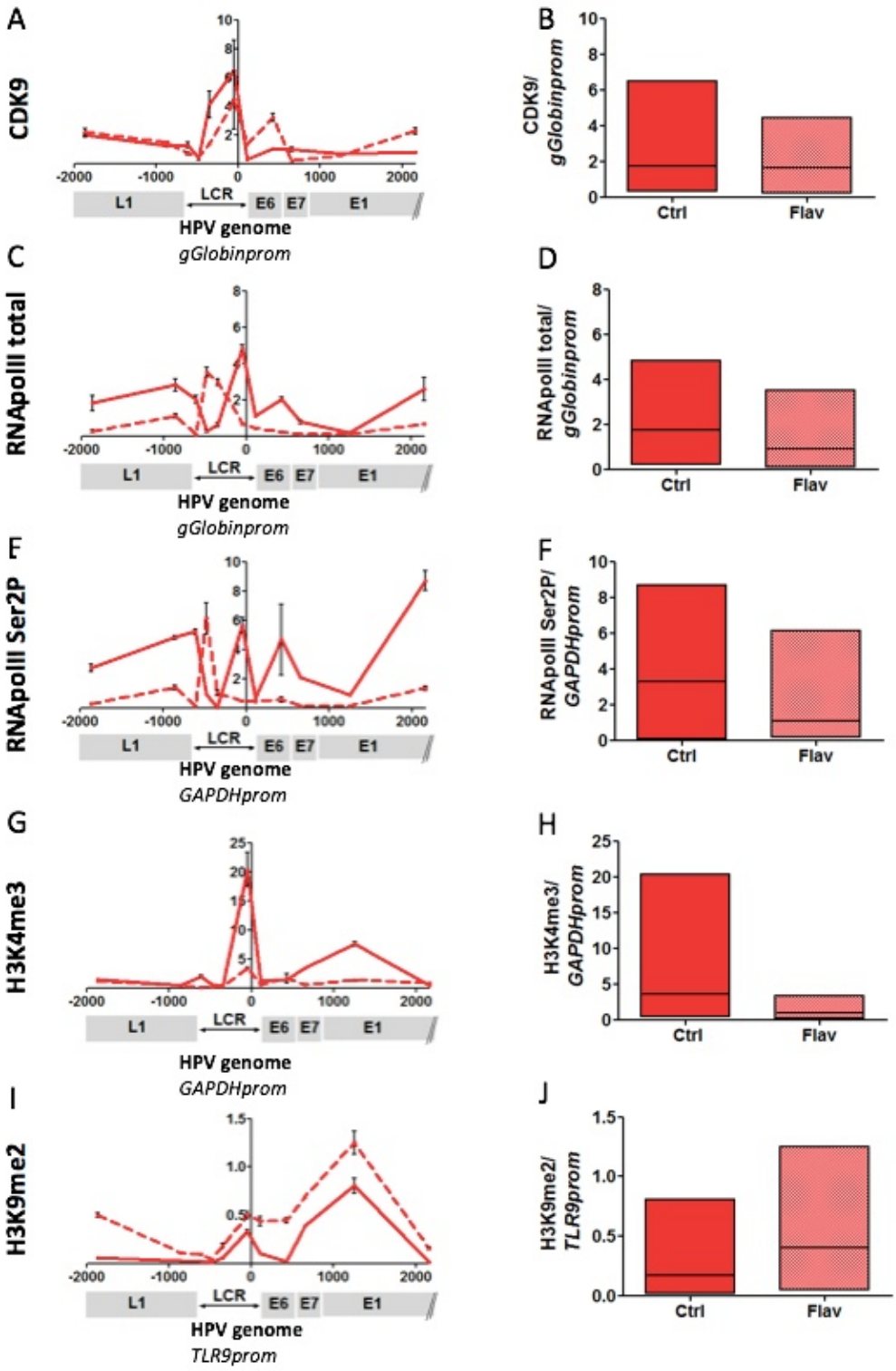

Figure 3.17: Effects of CDK9 inhibition on the abundance and distribution of hallmark PTMs in W12 clone F. Level of association of the enzymes CDK9 (A-B), total RNAPII (C-D), RNAPII Ser2P (E-F), as well as the active PTM H3K4me3 (G-H) and repressive PTM H3K9me2 (I-J) following treatment with specific CDK9 small molecule inhibitor Flavopiridol for 16 hours. In each graph (A, C, E, G, I), the y-axis shows the relative levels of enrichment normalised to host control target regions as indicated in italics under each panel. The x-axis and underlying schematic show the region of the HPV16 genome analysed. Data presented as mean SD (n=1). Panels B, D, F, H, J (right column) show an alternative representation of data from left column; high-low bar graphs with the mean value represented by a centre line. Abbreviations: Flav = Flavopiridol, Ctrl = control.

3.3 Discussion

The LCR of the HPV16 genome is known to be the region with the most genetic variation[194]. Numerous studies have identified a high number of nucleotide variations between cell lines and tissue samples from different geographical locations and, most notably, with clinical severity[195, 196]. As such, investigations into the genomic sequence of a panel of fourteen W12 integrant clones was conducted to determine whether genomic mutations could be responsible for the ~16-fold difference in viral oncogene transcript levels[183]. The genome sequences of the episome-containing W12 (W12E) and the HPV16-positive cervical cancer (SiHa) cell lines are published (GenBank: AF125673.1 and AF001599.1/AF001600.1, respectively) and were used as comparators for the LCR sequence of the integrant clones. Known mutations of the LCR of SiHa include a single nucleotide substitution from adenosine to thymine at position 7,519 and a 38 base pair deletion of nucleotides 7754-7791[197]; both were detected by the PCR amplification of the LCR and subsequent sequence analysis indicating adequate accuracy and robustness of the methods employed. Analysis of the LCR sequence showed extensive sequence homology between the W12 integrant clones. A single mutation was found in clone R2, which contains seven copies of the viral genome[183]. Although the computerised base call of a thymidine residue at position 7,491 matched the W12E reference sequence, further scrutiny of the chromatogram for this clone indicated a potential deletion of this nucleotide in $1/7^{th}$ of the population resulting in a frame shift of the affected population (Figure 3.3). Mutations within the many transcription factor-binding sites located in the LCR of the HPV16 virus can impact viral expression. For example, single nucleotide polymorphisms (SNP) in the binding sites of the repressive transcription factor YY1 have been shown to disrupt the binding of associated proteins[198, 199]. In contrast, a SNP in the binding site of the activating transcription factor AP-1 increases the binding affinity of associated proteins; both mutations result in increased activity of the HPV16 p97 promoter[200]. However, the mutation in the sub-population of W12 clone R2 does not occur within a transcription factor binding site, nor does it lead to a functional change in the transcript, therefore is unlikely to affect the viral

expression of this clone. It is clear that genetic mutation and resultant changes to the binding affinity of transcription factors to the regulatory region of viral DNA is not responsible for the differential expression of viral oncogenes E6 and E7 across the W12 integrant clones, certainly those with a genome copy number less than 4, the focus of this study (F, A5, D2, H, G2).

Having established that changes to the genetic sequence of the regulatory region of the viral genome are not responsible for the wide range of viral oncogene expression per template, investigations into epigenetic mechanisms of gene regulation were conducted. Previous work carried out by the Coleman group indicated that the post-translational modification of histone tails was a mechanism by which the expression of the viral oncogenes E6 and E7 are controlled[183]. In the present study, an increased number of hallmark PTMs representative of active and repressed chromatin have been analysed to generate a broader picture of the epigenetic landscape at the integrated HPV16 genome. Moreover, the design and use of additional primers along the length of the HPV16 genome mean that the abundance and distribution of PTMs have been analysed at a greater resolution to produce more in-depth data and facilitate more robust conclusions.

Levels of HPV16 expression were positively associated with higher abundance of histone PTMs that marked transcriptionally active chromatin. However, the this analysis found that the distribution profiles of H3K4me3 and H3K4me1 differ. For H3K4me3, there is a broad association over the early region of the viral genome with the greatest abundance at the oncogenes E6 and E7; this is compared with discrete peaks over the early LCR and E1 regions of the viral genome for the histone modification H3K4me1. It is likely that the relative distribution or abundance of the enzymatic writers (histone methyltransferases) and their cofactors responsible for each mark results in the alternate profiles[201].

Conversely, viral oncogene expression levels were negatively associated with repressive chromatin marks. The abundance of repressive marks were particularly low across the LCR region of the HPV16 genome in each of the five clones; this is to be expected as the regulatory region remains structurally more open compared with the rest of the genome to facilitate transcription factor binding and the recruitment

of transcriptional machinery to the virus early promoter (p97). For W12 clone G2 the highest levels of repressive chromatin marks were found over the E6 and E7 gene bodies indicating particularly firm repression of transcription of the viral oncogenes. Interestingly, the di-methylated form of H3K27 was much more abundant than the equivalent tri-methylated mark. This is likely a result of the global reduction of H3K27me3 in high-risk HPV-infected cells; expression of virus oncogene E7 leads to transcriptional induction of demethylase enzymes KDM6A and 6B[151]. In addition to removing methyl-moieties from histone tails, the demethylases delocalise the polycomb repressive complex 2 (PRC2), the writer of the H3K27me3 mark[151, 202]. In combination with the increased abundance of repressive PTMs, investigations conducted by Dr. Cinzia Scarpini showed that the overall levels of endogenous CpG DNA methylation were also greatest at the virus genome of the low-expressing W12 clone[203]. DNA methylation is an additional layer of epigenetic regulation and results in the stable repression of transcription through direct and indirect mechanisms[204].

Additionally, histone acetylation associated positively with levels of HATs examined, namely p300 and TIP60. A link between the HAT p300 and cervical malignancy has previously been identified with a positive correlation between p300 expression levels in HPV16 infected cells, and disease progression from CIN1 to 3 has been demonstrated; levels of p300 in cervical cancer cell lines SiHa and CaSki are greater when compared to HPV negative keratinocytes NHEK and RT3SB[205]. Additionally 16 % of SCCs are associated with somatic mutations in the p300 coding sequence[97]. In our system we found that p300 was particularly abundant at the LCR of the viral genome in the highest expressing clones per template, F and A5, indicated by a sharp peak in the region; this is likely a result of AP1 dependent, contact-driven recruitment of p300 required for the transcription of the HPV genome[206]. As with p300, there is greater association of TIP60 with the viral genome of the highly expressing W12 clones F and A5, represented by a broad peak over the LCR and viral early genes. The increased level of association is likely due to the preferential binding of the TIP60 chromodomain to H3K4me1[207], which is also found associated with these regions of the HPV16 genome and is also reflective of the elevation of TIP60 in transcribed regions and at the promoters of active genes[137].

It has been previously shown that p300 can activate HPV gene expression[206, 205, 208]; however, in addition to this, across the panel of W12 clones a functional, dose-dependent relationship with p300 was observed (Figure 3.9). Cells with high virus expression per template (clone F) showed a significantly greater sensitivity to equivalent levels of p300 depletion or inhibition compared with those with low virus expression per template (clone G2) (Figure 3.11 A-D). Observations following the depletion and inhibition of TIP60 mirrored those seen with p300 indicating the importance of functional HATs for effective transcription of the HPV16 genome (Figure 3.11 E-H).

HAT inhibition not only had a significant effect on viral oncogene expression but also resulted in a dramatic change in the abundance and distribution of activating chromatin marks along the HPV16 genome. The H3K27 and H2AK5 residues are preferentially acetylated by p300[209] and TIP60[210], respectively; small molecule inhibition of each enzyme reduced the amount of each mark associated with the HPV16 genome to near baseline levels compared with vehicle-treated control (Figure 3.12).

Interestingly, despite HATs usually being associated with the activation of transcription, there is evidence that TIP60 acts as a transcriptional repressor of integrated HPV18. In Hela cells, TIP60 binds to the HPV18 LCR in a YY1-dependent manner and subsequently recruits the bromodomain containing cellular repressor Brd4 to the enhancer/promoter; as such TIP60 is targeted by the HPV18 viral oncogene E6 for degradation[211]. However, in contradiction to this model, there was no association between levels of TIP60 and the transcriptional repressor protein YY1 observed in HPV16-positve W12 cells[203]. Further to this, previous work carried out by the Coleman group indicated an activating role for TIP60 in the HPV16 system through determination that viral expression decreased as a result of inhibiting TIP60 in W12 episome containing cells (data not shown). Although the reasons for opposing observations of TIP60 function between HPV18 and HPV16 are unclear, we have shown that this is not due to the structural form of the virus genome (i.e. episomal vs integrated) and it is likely that the mechanism of TIP60 recruitment is relevant. In addition to binding to the HPV LCR in a YY1-dependent manner, TIP60 can also be recruited to this region through a number of alternate mechanisms including interactions with phosphorylated RNAPII[137] or direct binding to the chromatin through its

chromodomain. This can occur via the repressive PTM H3K9me3 at double stranded breaks[212] but also via the active marks H3K4me3[213] and H3K4me1[207], both of which are found at greater levels at the HPV16 LCR in the highest expressing clones F and A5 (Figure 3.4). Additionally, when combined with H3K27ac, as is the case in the high expressing W12 clones, H3K4me1 is an indicator of active enhancers.

Moreover, whilst Brd4 commonly acts as a cellular repressor it has also been shown to have dual functionality through its ability to interact with and recruit P-TEFb, the cofactor required for RNAPII elongation[214]. As such, Brd4 plays a positive role in RNAPII-dependent transcription through enhancing the recruitment of the active P-TEFb complex to acetylated chromatin in the promoter region[76]. The novel finding that TIP60 depletion and inhibition results in decreased viral expression in HPV16-containing W12 cells is likely caused as a result of decreased levels of TIP60 binding to the viral promoter/enhancer; this in turn results in reduced levels of both Brd4 and the subsequent P-TEFb recruitment causing diminished transcription of the viral genome.

In direct contrast to p300 and TIP60, the abundance of HDAC1 at the HPV16 genome was negatively correlated with virus expression per template. HDAC1 was detectable at the virus genome in all clones with slightly elevated levels at the LCR; this is consistent with the previously described necessity for HDACs at gene promoter regions to reset chromatin by the dynamic turnover of acetyl groups without which, despite increased acetylation, gene induction is inhibited[215, 216, 137].

In addition to alterations in chromatin structure as a result of post-translational modifications to histones, transcriptional activation of a gene is predominantly dependent on the activity of RNAPII[217]. As previously noted, the recruitment of P-TEFb is required to phosphorylate the serine 2 residue on the C terminal domain (CTD) of RNAPII, which is necessary for active transcription. The CDK9 enzyme was functionally significant, as evidenced by a greater sensitivity to depletion in cells with higher HPV16 gene expression per template. As with the HATs, selective enzyme inhibition of CDK9 with Flavopiridol resulted in striking changes in the distribution of RNAPII and hallmark chromatin PTMs. Interestingly, in addition to depleted levels of RNAPII associated with the viral genome, the distribution pro-

files of total and Ser2P RNAPII altered in a similar fashion with the predominant peak translocating away from the viral promoter (control) to the 5'-end of the LCR (CDK9i). This suggests that CDK9 inhibition results in the stalling of RNAPII at the E2BS4 — which loops and interacts with the E2BS1/2 at the viral early promoter — resulting in the dramatic reduction of transcription from the viral early promoter.

Conversely, although exhibiting very similar distribution profiles, the amount of the repressive PTM H3K9me2 was increased following CDK9 inhibition. These data illustrate that the genome is much less accessible to the transcription machinery and, consequently, viral transcription is diminished. The importance of P-TEFb/CDK9 for transcription of the HPV16 genome adds to previous observations for the necessity of this complex for the transcription of other viral genomes including Epstein-Barr virus[218] and the human immunodeficiency virus (HIV)[219].

Observations presented in this chapter indicate that genetic mutation in the non-coding LCR of the virus genome is not responsible for differential levels of HPV16 oncogene expression across the W12 integrant clones; rather the level of E6/E7 expression from the virus early promoter (P_{97}) is determined by the balance of hallmark histone PTMs characteristic of transcriptionally active or repressed chromatin. It is interesting to consider the expression of HPV16 oncoproteins in terms of clinical progression. Whilst it is known that the levels of HPV16 oncogenes are statistically similar in lesions ranging from normal through LSIL to HSIL[21], levels of dysplasia and progression towards carcinoma increases. This can be attributed to the loss of cell cycle control and the accumulation of genetic mutations as a result of pRb and, even more importantly, p53 degradation. In this study, the W12 cells were derived from a pre-malignant, low-grade cervical lesion. At this early stage of disease, the levels of E6 and E7 expression — influenced by epigenetic modifications to the integrated HPV16 genome — presumably contribute to the likelihood of progression of a low-grade cervical lesion.

Chapter 4

Adapting SCRiBL Hi-C methodology to capture the integrated HPV16 genome

4.1 Introduction

Regulation of gene expression and epigenetic control of the transcriptome extends far beyond that of post-translational histone modifications. The human genome does not exist as a one-dimensional polymer or function in a sequential fashion; rather, it is folded in three-dimensional (3D) space. There is an increasing body of evidence showing that the 3D genomic organisation of the nucleus has an important role in determining gene expression patterns[220]. Gene expression is controlled by regulatory elements that can be located several mega bases (Mb) away from their corresponding gene promoters; this indicates that communication between distal gene enhancers and promoters is essential for regulated gene expression[221, 222].

Microscopic study of the nucleus, primarily with fluorescence in situ hybridisation (FISH) revealed many of the basic principles underpinning genomic organisation. Early light and electron microscopy studies showed the separation of active euchromatin and inactive heterochromatin to distinct regions of the nucleus,[223] and more recently revealed the existence of distinct sub-nuclear organelles[224, 225, 226, 227, 228]. FISH with chromosome paints revealed that chromosomes occupy distinct territories within the nucleus throughout interphase, with limited intermingling[229, 230, 231]. Additionally, the positioning of each chromosome is not random but related to gene density; small, gene-dense chromosomes (e.g. human chromosome 17, 19, 20) occupy the 3D nuclear interior compared with gene-poor chromosomes (e.g. human chromosome 18) that are located towards the nucleus periphery[232, 233, 231]. A correlation between nuclear location and transcriptional output has also been determined by microscopy studies; upon activation, individual gene loci have been shown to move from the nuclear periphery and to preferentially associate with euchromatin in the nuclear interior[234]. Despite important contributions to the understanding of genome architecture, microscopic techniques are limited by their low-throughput nature. The advent of chromatin conformation capture (3C) technologies has enabled a more systematic, genome-wide and high-throughput approach to study nuclear architecture.

Multiple chromosome conformation capture assays have been developed and provide a population-averaged impression of contact frequencies between genomic sites[235].

Each assay is based upon the principle that regions of interacting chromatin can be cross-linked, cut using restriction enzymes, and re-ligated so that genomic sequences in close physical proximity in the nucleus become linked to one another. The resultant ligation junctions reflect the 3D organisation of the genome at the time of fixation and can be used to infer chromatin structure (Figure 4.1)[236, 235]

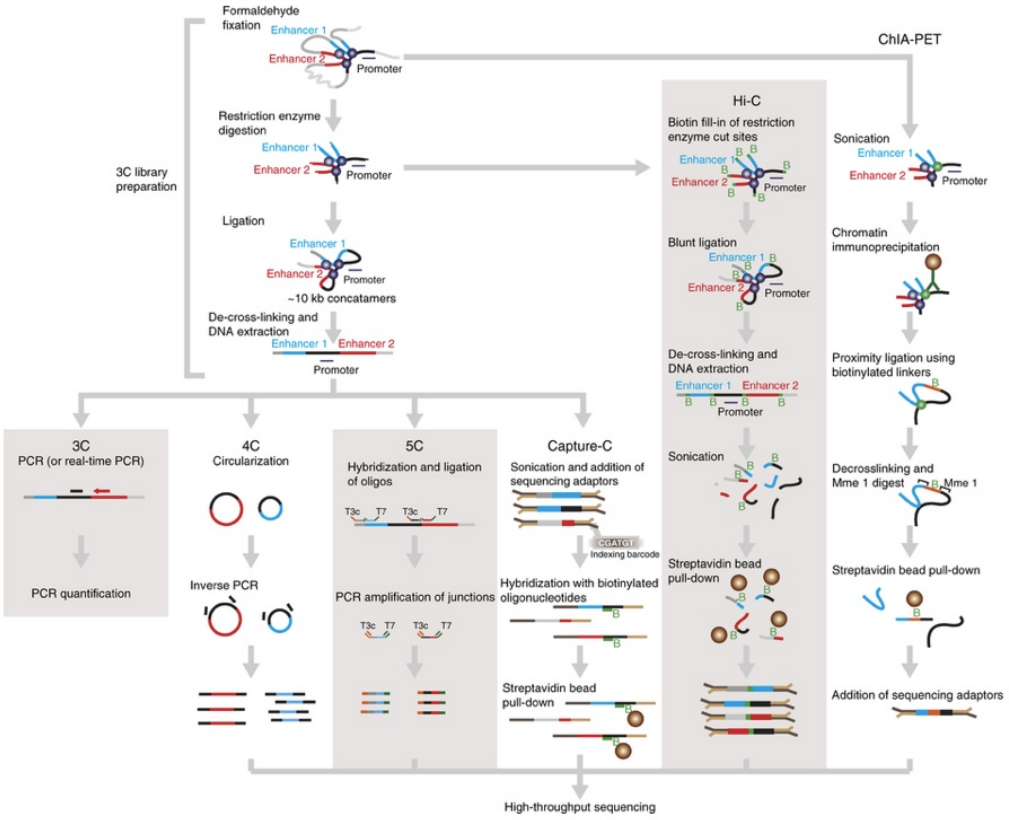

Figure 4.1: Comparison of different 3C-based methodologies taken from Davies, J. *et al.*, 2017[235]. 3C libraries share the indicated steps and can then be interrogated by the specific steps in the 3C protocol shown subsequently.

The original 3C method described in Dekker *et al.*, 2002 is a powerful technique to assess whether a region of interest interacts with a series of pre-specified genomic fragments, and, as such, is termed a one-to-one approach. Following the re-ligation of DNA fragments and the removal of crosslinks, contacts are analysed between selected pairs of sequences. Prior knowledge, or strong hypotheses, of interacting genomic regions is required to generate the loci-specific primers required to amplify and quantify ligation junctions. Initially, 3C technology was used to define the spatial organisation of yeast chromosomes[236] and later adapted to demonstrate long-range gene regulation by the physical looping of an enhancer to its target gene at the

β-globin locus[237, 238]. The technique has since facilitated the identification of long-distance *cis* and *trans* physical contacts at multiple gene loci[239, 240, 241]; however, this approach is limited to the confirmation of suspected interactions rather than identifying novel ones in an unbiased way. The improved access and affordability of next generation sequencing (NGS) meant that future generations of progressively unbiased, high-throughput, 3C-based methods were developed with increasing genomic resolution; these include 4C ('one-to-all' approach)[242, 243], 5C ('many-to-many' approach)[244] and Hi-C ('all-to-all' approach)[245, 246] (Figure 4.1).

Whilst 4C and 5C have been instrumental in furthering our knowledge and understanding of nuclear architecture and its role in genome regulation,[247, 248, 249, 242, 221, 250] Hi-C is unique in its ability to generate contact maps between all parts of the genome. The modifications to the original protocol used to generate 3C libraries include: the fill-in of restriction enzyme digested DNA fragments with biotin-labeled nucleotides; blunt-end ligation; further DNA fragmentation using sonication; and a streptavidin pull down of ligated fragments. Adaptations to 3C library generation result in the concentration of the informative ligation junctions representative of the 3D interactions across the genome, which are subsequently sequenced from both ends by paired-end sequencing[246]. The resolution of resultant Hi-C maps is determined by the restriction site density as well as depth of sequencing, and has increased from a scale of 1 Mb[246] to single kilobase resolution[251]. As such, Hi-C has been used to refine our understanding of genome-wide compartments of open and active and closed and inactive genomic regions[246]; to extensively describe the principles of chromosome looping[251]; as well as to identify a further layer of nuclear organisation termed topologically associating domains (TADs)[252]. Typically, TADs are one megabase in size and represent chromosomal units within which DNA sequences preferentially contact one another and form the framework within which promoters can find their respective enhancers and *vice versa*[253]. However, the production of quality Hi-C data for large mammalian genomes requires the sequencing and mapping of several billion reads per sample. In order to significantly reduce the number of sequencing reads required to generate contact maps of equivalent resolution, capture Hi-C was generated. An additional hybridisation selection of chosen loci, e.g. gene promoters,

specifically enriches Hi-C libraries for the chosen loci and the DNA elements that they contact in 3D[254, 255, 256].

Based upon similar methodology to capture Hi-C, Peter Fraser and Stefan Schoenfelder (Babraham Institute) invented sequence capture of regions interacting with bait loci (SCRiBL) Hi-C. Using SCRiBL Hi-C, all interactions between elements within a capture region as well as between the capture region and the rest of the genome are identified (Figure 4.2).

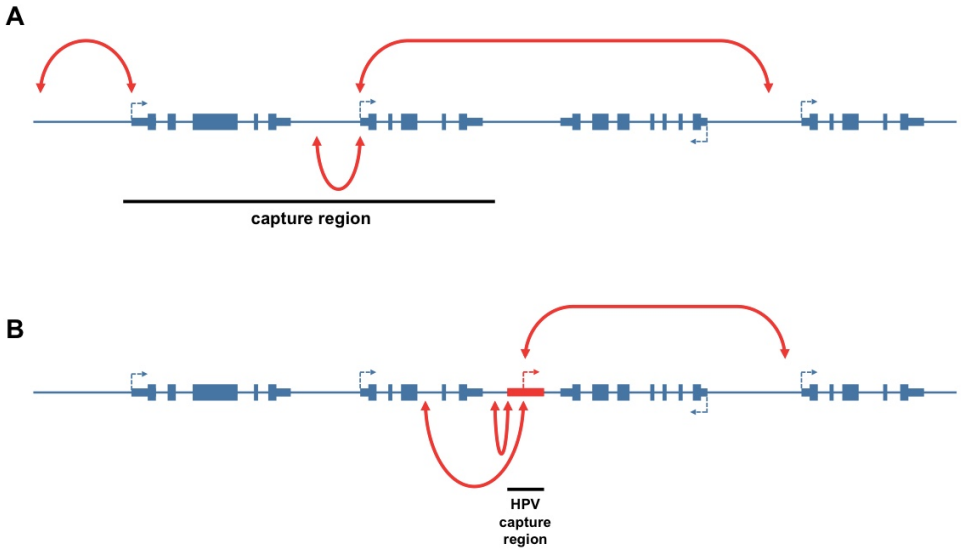

Figure 4.2: Cartoon illustrating the principles of Sequence Capture of Regions interacting with Bait Loci (SCRiBL) Hi-C A. The 3D interactions (red double headed arrow) within a capture region and short- and long-range interactions between the capture region and the rest of the genome. B. HPV16 genome (bold red line) integrated into the host is the designated capture region. 3D interactions between the virus and host are identified.

This chapter outlines how SCRiBL Hi-C has been adapted to enrich for the HPV16 genome in the W12 integrant clone system. The need for significant enrichment of Hi-C libraries is particularly pertinent due to the relatively tiny size of the viral genome in comparison to the host. In the first part, Hi-C libraries from which genome-wide contacts can be determined are generated; the second part, describes the design and production of HPV16-specific baits that are used to capture only the ligation fragments containing the virus genome. The successful generation of SCRiBL Hi-C libraries will enable a much broader analysis of the epigenetic regulation of viral transcription of the W12 clones.

4.2 Materials and Methods

4.2.1 Hi-C protocol (SCRiBL)

An undigested library (used in capture-seq experiment) and Hi-C libraries (SCRiBL) were generated concurrently with any alterations for the generation of the undigested library indicated in the text below. At each stage the reaction volumes and the processes for both the Hi-C and undigested libraries were kept equal so that as far a possible the samples were treated the same. For each W12 clone used in this experiment there are two biological replicates; the cells for each replicate were fixed and frozen on the same day but the generation of subsequent libraries occurred separately for each replicate. This work was carried out in collaboration with Marco Michalski from the Babraham Institute.

4.2.2 Part I: Generation of Hi-C libraries

Cell culture and crosslinking of chromatin

W12 clones were grown in monolayer culture as previously described in Chapter 2. Cells were analysed at the lowest available passage (p) after cloning (F p6, G2 p12, NCx/6 p5, A5 p5, D2 p8 and H p6) in order to minimise any effects of genomic instability caused by deregulated HPV16 oncogene expression. W12 cells of each clone were seeded into four 15 cm^2 plates and grown to 80–90% confluence (approximately 30 million cells in total), and fixed in EGF positive culture media supplemented with methanol-free formaldehyde (Agar Scientific) to a final concentration of 2% for 10 minutes at room temperature with gentle agitation. Crosslinking was quenched by the addition of ice-cold glycine (VWR International) to a final concentration of 125 mM. Plates were incubated at room temperature for 5 minutes on a rocker followed by 10 minutes at 4 °C. The adherent W12 cells were then scraped from the culture plate using disposable cell scrapers and the resultant cell suspensions from two plates were pooled into one 50 ml Falcon tube i.e. there were two Falcon tubes per cell line, representing two biological replicates with 15 million cells each (Figure 4.5). The Falcon tubes were then spun at 720 RCF in an Eppendorf 5810R centrifuge for 10

minutes at 4 °C. The supernatant of each 50 ml Falcon tube was discarded, the pellet of each was carefully resuspended in 10 ml ice-cold PBS and transferred to a 15 ml Falcon tube. The 15 ml Falcon tubes were then spun at 720 RCF for 10 minutes at 4 °C, and the supernatant was discarded. The pellets were then flash frozen in liquid nitrogen and stored at -80 °C.

Cell lysis and chromatin digestion

The cell pellets were thawed on ice before resuspension in 1 ml ice-cold lysis buffer (10mM Tris-HCl pH 8 (Sigma), 10 mM NaCl, 0.2% IGEPAL® CA-630 (Sigma), protease inhibitor cocktail tablet (EDTA-free) (Roche). Cells resuspended in 1 ml lysis buffer were then added to 49 ml ice-cold lysis buffer in a 50 ml Falcon, and the tubes were inverted four times to mix. The tubes were then incubated on ice for 30 minutes with occasional mixing. Following lysis, nuclei were pelleted by centrifugation at 650 x *g* for 5 minutes at 4 °C and the supernatant was discarded. Next, the pellets were resuspended in 1.25x NEBuffer 2; briefly, 500 μl 1.25x NEB2 were used to resuspend the pellets, this was followed by the addition of a further 400 μl 1.25x NEBuffer 2 to fully disperse the pellets by gentle pipetting and cells were transferred to a 1.5 ml Eppendorf tube. The final volume was then measured and made up to 1.25 ml with the addition of 1.25x NEBuffer 2 to the cell suspension, and divided between five 1.5 ml Eppendorf tubes (250 μl and 5-6 million cells in each). Each of the aliquots were made up to a final volume of 358 μl with 1.25x NEBuffer 2, the chromatin was solubilised by the addition of 11 μl 10% SDS (0.3% final concentration) (Promega) and incubated at 37 °C for 60 minutes with shaking at 950 rpm. SDS-solubilisation removes non-crosslinked proteins and opens the chromatin, making it accessible for restriction endonuclease cleavage. SDS was quenched by adding 75 μl 10% Triton X-100 (1.7% final concentration) (Sigma) to each tube and the nuclei were incubated at 37 °C for 60 minutes with shaking at 950 rpm. Restriction digestion was performed by the addition of 800 units of the restriction endonuclease MboI (32 μl 25 U/μl, New England Biolabs) per tube (5 million cells). One Eppendorf tube was kept as undigested material and 32 μl dH_2O were added to this tube instead of MboI. All tubes were incubated at 37 °C overnight with shaking at 950 rpm.

Biotin marking of DNA ends and blunt end ligation (Hi-C samples only)

Following overnight incubation the four Eppendorf tubes containing digested DNA were placed on ice. The restriction fragment ends were filled in and the DNA ends marked with biotin by adding: 1.5 μl of each 10 mM dCTP, dGTP, dTTP, 37.5 μl 0.4mM biotin-14-dATP (all Thermo Fisher Scientific) and 10 μl 5U/μl Klenow (DNA polymerase I large fragment, New England Biolabs) to each tube. Each solution was then mixed and incubated for 75 minutes at 37 °C with shaking at 700 rpm. Following incubation all tubes were placed on ice.

Blunt-ended DNA fragments cross-linked together in chromatin complexes were then ligated following the in-nucleus ligation protocol previously described[257], with minor modifications. Prior to ligation, excess salts and enzymes were removed by centrifugation (600 x g for 5 minutes at 4 °C) and the supernatant was discarded. Each cell pellet was then resuspended in 995 μl 1x T4 DNA ligase buffer (New England Biolabs) supplemented with BSA (100 μg/ml final concentration). The ligation was carried out using 2000 units (5 μl) T4 DNA ligase (400U/μl, New England Biolabs) per 5 million starting material of cells. The reaction was then placed at 16 °C for 4 hours followed by 30 minutes at room temperature. Following the ligation of DNA fragments in close 3D proximity, DNA crosslinks were reversed and proteins were degraded by the addition of 25 μl 10 mg/ml Proteinase K (Roche) per tube and the chromatin was incubated at 65 °C overnight with shaking at 900 rpm. At this stage, the undigested DNA samples were treated the same way, after adjusting the volume by adding 556 μl TLE (10 mM Tris pH 8.0, 0.1 mM EDTA).

DNA purification

Each reaction mixture was cooled to room temperature before the addition of 10 μl of 10 mg/ml RNaseA to each Eppendorf and incubated at 37 °C for at least 60 minutes, shaking at 600 rpm. The DNA was then purified by a phenol and two phenol-chloroform extractions. Material from the four digested samples (for each replicate) were pooled into a 50 ml Falcon tube and the undigested libraries made up to 4 ml with TLE in a 50 ml Falcon. To both Falcon tubes, 4 ml (1:1 ratio) phenol

pH 8.0 (Sigma) was added and vortexed for 1 minute to mix. The tubes were spun for 10 minutes at 3500 rpm at room temperature and then the aqueous phase was transferred into a new tube. A back extraction of the original phenol was used for optimum recovery of DNA; briefly, 2 ml TLE were added to any remaining aqueous liquid and phenol left behind in the original Falcon. Again, the tubes were vortexed to mix and spun for 10 minutes. The entire aqueous phase was then added to the Falcon tube from the previous step. The DNA extraction was repeated twice using phenol pH 8.0: chloroform (Sigma) following the steps as previously described. The DNA was then precipitated by adding $1/10^{th}$ the total volume of 3 M sodium acetate pH 5.2 and 2.5 x the total volume of ice-cold 100% EtOH (VWR Chemicals), and the tubes were incubated overnight at -20 °C.

The tubes were then spun for 30 minutes at 4 °C at 3500 rpm, DNA pellets were washed three times with 70% ethanol and then resuspended in 100 μl (Hi-C DNA) and 25 μl (undigested DNA) TLE respectively (25 μl per 5 million cells starting material). The DNA concentrations of a dilution series (1:500, 1:1000 and 1:2000 dilutions) were measured using the Quant-iT PicoGreen assay (Life Technologies) as per the manufacturer's instructions.

Quality control of Hi-C libraries

Quality control reactions were carried out to analyse the Hi-C marking and ligation efficiency. To determine the efficiency of digestion, fill-in and subsequent ligation of DNA in the Hi-C libraries a PCR reaction was carried out using two forward primers (IDT) designed around MboI restriction sites of the RPL13A locus. The PCR reaction mix included: 200 ng template DNA, 0.125 μl Hot Start *Taq* DNA polymerase (New England Biolabs), 2.5 μl 10X *Taq* reaction buffer (New England Biolabs), 2 μl dNTP mix (2.5 mM), 1 μl RPL13A B forward primer (10 M), 1 μl RPL13A G forward primer (10 μM) (Table 4.1) and dH_2O to 25μl. The PCR cycling conditions were the following: 95 °C for 15 minutes, followed by 38 cycles of 94 °C for 30 seconds, 53.9 °C for 30 seconds, 72 °C for 1 minute followed by a final amplification at 72 °C for 7 minutes. The reaction products were then run on a 1.5% agarose gel. Subsequently the PCR products from four reactions were purified

(Qiagen purification kit) and spilt into two; one sample digested with the restriction enzyme ClaI and the other one was left untreated. For this digestion reaction, 750 ng PCR product were incubated with the restriction enzyme ClaI (20 U), 2 μl 10X Cut Smart buffer made to 20 μl with dH_2O and incubated for 2 hours at 37 °C. Both samples were then run on a 1.5% agarose gel.

The detection of expected interactions in the Hi-C libraries was also conducted by designing primers that would generate a specific 3C product. PCR reactions were set up as before, replacing the primers, which were used before, with RPL13A D2J forward and RPL13A D2J reverse (10 mM) (Table 4.1). The PCR cycling conditions were the same as before, but the reaction consisted of 35 cycles and had an annealing temperature of 62 °C. Again, the products were run on a 1.5% agarose gel and the DNA bands in undigested and Hi-C libraries were compared.

Biotin removal from un-ligated ends (Hi-C libraries only)

To remove biotin-dATP from any non-ligated fragments ends, 40 μg of DNA from each Hi-C library was incubated with T4 DNA polymerase (NEB). Briefly, eight reactions containing: 5 μg DNA, 1 μl 10mg/ml BSA, 10 μl 10x NEBuffer 2, 2 μl 10 mM dATP, 2 μl 10 mM dGTP and 5 μl (15U) T4 DNA polymerase (New England Biolabs) in a total volume of 100 μl with dH_2O were set up, mixed and incubated at 20°C for 4 hours. As there was no biotin to be removed in the undigested libraries and a smaller quantity of DNA was required, 3 x 5 μg DNA aliquots (15 μg total) were made to an equivalent volume with dH_2O and also placed at 20 °C for 4 hours. Following incubation, biotin removal in the Hi-C samples was then stopped by the addition of 2 μl 0.5 M EDTA pH 8.0 (Invitrogen) to each aliquot. Two Hi-C reactions were then pooled together to make a total amount of DNA ~10 μg per sample, and all three undigested reactions (15 μg) were pooled. DNA was extracted using phenol pH 8.0:chloroform (1:1) and was subsequently precipitated with sodium acetate pH 5.2 and 100% ethanol as described previously (see DNA Purification). Following overnight incubation at -20 °C samples were spun and the DNA was washed in 70% EtOH. Each pellet was then resuspended in 130 μl dH_2O.

DNA fragmentation and size fractionation

130 μl of each sample was added to a Microtube AFA fibre pre-slit tube (LGC Genomics), and DNA was sheared to an average size of 400 bp using the Covaris E220 Sonicator (settings: duty factor 10%, peak incident power 140 (W), cycles per burst 200 and time 55 seconds).

End repair and 'A' tailing

Sonication randomly breaks the DNA into fragments and the subsequent end repair step refills the ends of these broken DNA fragments. Following sonication the 130 μl samples were transferred to fresh 1.5 ml Eppendorf tubes and the volumes were adjusted to 180 μl by the addition of the following: 18 μl 10x T4 DNA ligase buffer, 18μl dNTP mix 2.5 mM, 6.5 μl T4 DNA polymerase, 6.5 μl T4 polynucleotide kinase (NEB), 1.3 μl Klenow DNA polymerase (5 U/μl, NEB). All samples were then incubated at room temperature for 30 minutes. The DNA was then purified using the QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's instructions. The DNA was eluted from each column in a fresh 1.5 ml Eppendorf with 30 μl TLE.

The addition of a single dATP to the repaired 3' end was required to facilitate the ligation of sequencing adapters that are required for downstream Next Generation Sequencing. To add dATP, 5 μl 10x NEB2, 11.5 μl 1 mM dATP, 3.5 μl Klenow (exo-) (New England Biolabs) were added to the DNA and mixed (50 μl total). The reactions were incubated at 37 °C for 30 minutes after which the Klenow (exo-) was inactivated by incubating at 65 °C for 20 minutes before cooling on ice. The two Hi-C samples were then pooled (100 μl total) and the undigested samples were made up to 100 μl with dH_2O, both in Eppendorf LoBind 1.5 ml microcentrifuge tubes (Sigma).

Double-sided SPRI-selection

DNA fragments ranging in size between 250-500 bp were isolated by a double-sided size selection using AMPure XP beads (Beckman Coulter) by performing sequential

Solid Phase Reversible Immobilisation (SPRI) bead selections. Briefly, 0.6x (60 μl) SPRI beads were added to the 100 μl DNA, mixed thoroughly by pipetting and vortexing and incubated at RT for 15 minutes to allow DNA with high molecular weight to bind to the beads. The tubes were then placed on a magnetic separator and the supernatant was transferred into a fresh 1.5 ml LoBind tube. During the first incubation the SPRI beads were concentrated for the next size selection. 80 μl of beads per reaction were placed in a LoBind tube, put on a magnetic separator and all but 35 μl of the supernatant was removed. The 35 μl concentrated SPRI beads were added to the 160 μl supernatant from the previous step (1.14x), the sample was vortexed and incubated for 15 minutes to allow all DNA >200 bp to bind. The beads were then washed twice with 500 μl fresh 70% EtOH and dried at 37 °C for 2–3 minutes. The DNA was eluted from the beads with 50 μl TLE. A further 250 μl TLE was subsequently added to the Hi-C samples only to make a total of 300 μl (required for the next step). Samples were stored at -20 °C until required.

Streptavidin pull-down of biotinylated HiC ligation products (Hi-C samples only)

In order to minimise loss of DNA during the streptavidin pull-down of biotinylated Hi-C ligation products, all steps were carried out in LoBind tubes and with LoBind pipette tips. The pull down of biotin-marked ligation junctions was carried out using Dynabeads MyOne Streptavidin C1 beads (Life Technologies), which were prepared in the following way before use; for each wash the beads were resuspended in the appropriate buffer, transferred to a new tube and place on a rotating wheel for 3 minutes. Briefly, 150 μl C1 beads per sample were placed in a fresh Eppendorf, placed on the magnetic rack and the supernatant was removed. The beads were then washed twice with 400 μl Tween Buffer (5 mM Tris-HCl pH 8.0, 0.5 mM EDTA, 1M NaCl, 0.05% Tween). The beads were then washed with 300 μl of 2x No Tween Buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 2 M NaCl) and the supernatant was removed. The beads were then combined with 300 μl Hi-C DNA and incubated at room temperature for 15 minutes with rotation. Following this, the beads were reclaimed and resuspended in 400 μl 1x NTB (5 mM Tris-HCl pH 8.0, 0.5 mM

EDTA, 1M NaCl). The suspension was then placed at 55 °C with shaking at 500 rpm for 5 minutes to improve the specificity of the pull down before the supernatant was removed and the beads were washed in 100 μl 1x ligation buffer (New England Biolabs). The reclaimed beads were then resuspended in 50 μl 1x ligation buffer and transferred to a new tube.

Paired-end adapter ligation and library amplification

The undigested sample was defrosted on ice and 5.8 μl 10x ligation buffer were added to the DNA. Custom SCRiBL adapters were ligated to the Hi-C and undigested libraries by the addition of 4 μl SCRiBL TruPE adapter mix (15μM) (Table 4.1) and 4 μl NEB T4 Ligase (400 U/μl). Reactions were incubated at room temperature for 2 hours on a rotator; SCRiBL TruPE adapters were generated by annealing SCRiBL_TruPE_adapter_1 and SCRiBL_TruPE_adapter_2 by decreasing the temperature from 90 °C to 4 °C with Δ -1 °C /minute. The undigested material was placed at 65 °C for 15 minutes to inactivate the ligase enzyme and subsequently frozen at -20 °C. The beads from the Hi-C samples were recaptured, the supernatant was removed and the beads were washed twice with 400 μl TB. The beads were subsequently washed with 200 μl 1x NTB, 100 μl 1x NEBuffer 2 and then 50 μl 1 x NEBuffer 2. Finally the beads were resuspended in 50 μl 1x NEBuffer 2 and transferred into a new tube. Samples were stored at -20 °C.

Final quality control and library quantification

A test PCR amplification was run on both the Hi-C and undigested samples to determine the number of PCR cycles required to generate enough material for capture and sequencing. The following PCR reactions were set up; 2.5 μl DNA, 0.075 μl 100 μM PE PCR primer 1.0.33, 0.075 μl 100 μM PE PCR primer 2.0.33, 0.7 μl each 10 mM dNTP (x4), 0.3 μl Phusion polymerase (New England Biolabs), 5 μl 5X Phusion HF buffer (New England Biolabs) and 14.25 μl dH_2O. The PCR conditions were 98 °C for 30 seconds, followed by either 6, 9 or 12 cycles of 98 C for 10 seconds, 65 °C for 30 seconds, 72 °C for 30 seconds followed by 72 °C for 7 minutes. The PCR products were then run on a 1.5% agarose gel against MassRuler DNA Ladder using 6X DNA

loading dye (both Thermoscientific). The appropriate number of cycles for the final amplification PCR was taken as the number immediately below that at which the smear is just visible.

The final PCR amplification was performed as described above, but using the whole Hi-C and undigested DNA libraries (20–30 25 μl PCR aliquots). Following the PCR reaction, library products were pooled and the beads were reclaimed on a magnetic rack. The PCR products in the supernatant were purified with 1.8x volume AMPure XP beads and washed twice with 1 ml 70% EtOH before DNA was eluted with 50 μl TLE. Samples of each library were sent to the Bioanalyser for concentration and size distribution quantitation.

4.2.3 Part IIa: Generation of biotinylated RNA for target enrichment (SCRiBL)

RNA baits were designed for the enrichment of the HPV16 genome in the Hi-C samples generated in Part I.

Capture RNA bait library design

120-mer capture RNA baits were bioinformatically designed with custom perl script from Simon Andrews (Babraham Institute) to both ends of MboI restriction fragments overlapping the HPV-16 genome. Requirements for target sequences were as follows: GC content between 25% and 65%, no more than two consecutive Ns within the target sequences, and maximum distance to a MboI restriction site 330bp. For short MboI fragments, where 120-mer RNA baits originating from both ends would have overlapped (potentially interfering with optimal hybridization to Hi-C libraries), only the coding strand was used for capture RNA bait design; if necessary the baits were trimmed to minimum length no shorter than 97 nt. This resulted in the design of 16 RNA bait sequences (Table 4.3) covering the MboI restriction fragment ends of the entire HPV16 genome, with the exception of two fragments too short (18 and 63 bp, respectively) for capture RNA bait design.

Zero Blunt® TOPO® cloning of gBlocks

DNA sequences encoding for the 16 RNA bait sequences, with different restriction enzyme sites at each fragment end (5' BglII and 3' HindIII or SpeI, Figure 4.12) were ordered as two gBlocks (Integrated DNA Technologies) and cloned into plasmid vectors using the Zero Blunt® TOPO® cloning kit with One Shot® TOP10 chemically competent cells according to the manufacturer's instructions. Following overnight incubation, plasmid DNA from gBlock clone 1.1 and 2.1 was isolated using the QIAprep spin Miniprep kit (Qiagen) according to the manufacturer's instructions. DNA was eluted in 50 μl Elution Buffer.

Digestion to cut out gBlock fragments from pCR-Blunt II-TOPO cloning vector

To isolate the gBlock fragments from the cloning vector, digestion reactions with restriction enzyme EcoRI were performed. 8.5 μg of each gBlock 1 and 2 were incubated with 12 μl 10x CutSmart Buffer and 6 l EcoRI HiFi enzyme (both New England Biolabs) for 2 hours at 37 °C. After incubation the mixtures were run on a 1% agarose gel and the bands containing the desired insert (~1 kb) were cut out and purified using the QIAquick Gel extract kit (Qiagen) according to the manufacturer's instructions. The DNA was eluted in 50μ l H_2O per column and quantified by Nanodrop.

Digestion of gBlock fragments and ligation of T7 sequencing adapters

DNA from both gBlocks was then further digested to release the fragments specific to the HPV16 genome. The DNA from gBlock1 was incubated with restriction enzymes BglII (10 U/μl) and HindIII (20 U/μl), whereas DNA from gBlock2 was incubated with BglII (10 U/μl) and SpeI (10 U/μl). Each fragment contained a single BglII cut site and enabled side specific ligation of a T7 promoter sequence adapter (Table 4.1). T7 adapters were generated by annealing T7_promoter_adapter_1 and T7_promoter_adapter_2; 20 μl each of both forward and reverse primers (100 μM) were mixed with 60 μl oligo annealing buffer (10 mM Tris pH 8.0, 50 mM NaCl, 1

mM EDTA) and placed in a PCR machine at 95 °C for 5 minutes. Following this initial incubation the temperature was decreased at a rate of 1 °C per minute to 4 °C. Digestion with both restriction enzymes and adapter ligation was carried out in one reaction simultaneously in the presence of BamHI (20 U/μl) to each reaction in order to cut any unspecific adapter-adapter products that may be present. Two reactions containing 700 ng gBlock1 or 850 ng gBlock2 DNA, 30 units BglII each, 100 units BamHI each, 5-fold molar excess of pre-annealed T7 promoter adapters, and either 80 units HindII (NEB) or 40 units SpeI (NEB) were incubated at 37 °C for 2 hours in 1x T4 DNA ligase buffer (NEB). Following this incubation 1200 units T4 DNA ligase (NEB) were added to each reaction and incubated at 25 °C for 3 hours. The samples were then run on a 1% agarose gel and specific bands at 180 bp were cut out and gel purified.

In vitro transcription

In vitro transcription was carried out using the T7 MegaScript kit (Ambion) with biotin-labelled dUTP (Roche). Equimolar amounts of gBlock1 and gBlock2 purified DNA was combined and used in *in vitro* transcription reaction; 2 μl 10X buffer, 5.5 μl DNA template (280 ng), 5 μl biotin-UTP (Roche), 1 μl unlabelled rUTP (100 mM), 1.5 μl rATP (100 mM), 1.5 μl rCTP (100 mM), 1.5 μl rGTP (100 mM) and 2 μl T7 enzyme mix. The reaction mixture was incubated at 37 °C overnight.

Purification of biotinylated RNA

To remove any remaining template DNA the samples were treated with 1 μl Turbo DNase (Life Technologies) for 15 minutes at 37 °C. The RNA was then purified using the MEGAclear kit (Ambion) following the manufacturer's instructions. The RNA was eluted in 50 μl elution solution. The size and integrity of the RNA was tested by running 2 μl on a 2% agarose gel and the final concentration of RNA baits was quantified by Nanodrop.

4.2.4 Part IIb: Generation of biotinylated RNA for target enrichment (Capture-seq)

The generation of RNA bait for the capture of undigested libraries was carried out using a plasmid approach. The pSP64 HPV16 plasmid (Figure 4.11 A) contains the linearised HPV16 genome, from which four consecutive, non-overlapping DNA fragments spanning the length of the genome were produced. These were then *in vitro* transcribed to RNA and fragmented to produce RNA baits for the capture reaction.

Plasmid preparation

An overnight culture from a glycerol stock made from a single colony of plasmid pSP64_HPV16_1.1 (Cinzia Scarpini) was prepared with 5 ml LB broth and 100 μg/ml ampicillin and incubated at 37 °C with gentle shaking overnight. Following incubation, plasmid DNA was extracted using QIAprep® Spin Miniprep Kit (Qiagen) according to the manufacturer's instructions. The quantity and purity of the DNA was determined by Nanodrop analysis.

Plasmid integrity check

To check the integrity of the plasmid DNA a digest assay was performed using restriction enzymes BamHI (20 U/μl, New England Biolabs) and EcoRI HIFI (20 U/μl, New England Biolabs), which cut the plasmid at two and three sites respectively. For each restriction enzyme reaction 1 μg DNA, 1 μl restriction enzyme, 2.5 μl 10x buffer (NEBuffer 3.1 for BamHI or Cutsmart for EcoRI, both New England Biolabs), and 17.8 μl dH_2O. Both reactions were incubated at 37 °C for 1 hour before being run on a 1% agarose gel.

PCR amplification of HPV16 genome

To amplify the HPV16 genome four primer sets were designed across the entire W12E genome. Plasmid DNA was diluted to 10 ng/μl and four PCR reactions were set up, each with a separate primer pair (Table 4.1); 1 μl plasmid DNA (10ng/ul), 2 μl

dNTP mix (2.5 mM), 10 μl Expand High Fidelity buffer (10x), 0.5 μl forward primer (100 μM), 0.5 μl reverse primer (100 μM), 0.75 μl Expand High Fidelity enzyme mix, 82.5 μl dH_2O. A touchdown PCR was run to ensure specific amplification of the DNA. The following PCR conditions were used: 94 °C for 5 minutes, followed by 13 cycles of 94 °C for 1 minute, 74–62 °C for 1 minute, 68 °C for 8 minutes, followed by 22 cycles of 94 °C for 1 minute, 62 °C for 45 seconds, 68 °C for 8 minutes, and then 68 °C for 10 minutes before cooling to 4 °C. PCR products were then run on a 1.5% agarose gel.

In vitro transcription with biotin-UTP

T7 promoter sequences were added to one side of the PCR product during the PCR amplification described above, which enabled subsequent directional in vitro transcription. Sequences were *in vitro* transcribed in the presence of biotin-UTP and purified as described previously.

RNA fragmentation

An equimolar mix of the four full length (2,000 nt) RNA products was then fragmented to ~150 nt for use in the capture reaction using a modified version of the FGRS protocol for RNA-seq library preparation protocol (Iyer Lab, Lauren Fairchild). Briefly, chemical fragmentation of the RNA occurred by combining 250 ng of RNA, 100 mM Tris pH 8.0 and 4 mM $MgCl_2$ in 10 μl and incubating at 95 °C for 8 minutes in a PCR machine. A total of 5 μg RNA was fragmented using this method.

Precipitation of RNA

The fragmentation reactions were then pooled into 2 x 100μl samples and precipitated using 1x volume (100 μl) ice-cold isopropanol and $1/10^{th}$ volume (20 μl) ammonium acetate Stop Solution (Ambion) and incubated at -20 °C for 2 hours. Following incubation, the RNA was spun at 14,000 rpm at 4 °C for 20 minutes to pellet the RNA. The pellet was then washed twice with 500 μl 75% ethanol and resuspended in 20 μl dH_2O. The resulting RNA was then quantified using the Nanodrop.

4.2.5 Part III: Solution hybrid capture of Hi-C library

Hybridisation of Hi-C library with biotin-RNA target bait

The amount of Hi-C library DNA or genomic DNA library captured was determined by the concentration of each library and ranged from 500–2000 ng. To prepare the Hi-C library (pond) for capture with RNA baits the appropriate volume was transferred into a 1.5 ml LoBind Eppendorf tube and concentrated using a vacuum concentrator (Savant SPD 2010, Thermo Scientific). After evaporation of all liquid, the Hi-C DNA pellet was resuspended in 5 μl dH_2O and transferred into a fresh LoBind Eppendorf. 2.5 μg mouse cot-1 DNA (Invitrogen) and 2.5 μg sheared salmon sperm DNA (Ambion) were added as blocking agents. To prevent concatemer formation during hybridisation 1.5 μl blocking mix (300 μM) were added (equimolar mix of four oligo blockers: (P5_b1_for_33, P5_b1_rev_33, P7_b2_for, P7_b2_rev) (Table 4.1), resulting in a 10 μl reaction mixture. This was resuspended thoroughly, transferred into PCR strip tubes (Agilent 410022), closed with a PCR strip tube lid (Agilent optical cap 8x strip) and kept on ice until use.

A master mix of the 2.23x hybridisation buffer was then prepared; 167.25 μl 20x SSPE (Gibco, 11.15x final), 66.9 μl 50x Denhardts (Invitrogen, 11.15x final), 6.69 μl 500 mM EDTA (Gibco, 11.15 mM final), 6.69 μl 10% SDS (Promega, 0.223% final) and 52.47 μl H_2O. The hybrisation buffer was mixed thoroughly and heated to 65 °C for at least 5 minutes. 30 μl were then aliquoted per capture reaction into a PCR strip, closed with a PCR strip tube lid and kept at room temperature.

Biotinylated RNA baits were used in a ratio of 1:12 to Hi-C libraries (25 ng biotinylated RNA baits per 300 ng Hi-C library). The baits were prepared by transferring 25 ng of biotinylated RNA into a 1.5 ml LoBind Eppendorf tube, and made up to a volume of 5.5 μl with H_2O. Subsequently, 30 units (1.5 μl) SUPERase-In (Ambion, 20 U/μl) were added (7 μl total), mixed, transferred into a PCR tube, closed with a strip tube lid and kept on ice. Biotinylated RNA baits for capture DNA-Seq were used in a ratio of 1:3.33 (300 ng RNA baits per 1,000 ng genomic DNA library); these baits were prepared in the same way as above.

The PCR machine (PTC-200, MJ Research) was set to the following program; 95

°C for 5 minutes, 65 °C forever. The PCR strip containing the pond Hi-C libraries was transferred to the PCR machine in the position marked red (Figure 4.3 A), the PCR program was started and the DNA denatured. Once the temperature returned to 65 °C the PCR strip containing the hybridisation buffer was transferred to the PCR machine in the position marked in blue (Figure 4.3 B) and incubated for 5 minutes. Following this, the final PCR strip containing the biotinylated RNA bait was transferred to the PCR machine in the position marked in green (Figure 4.3 C) and incubated for 2 minutes. After 2 minutes, 13 μl of hybridisation buffer were pipetted into the 7 μl RNA baits (blue into green). This was immediately followed by pipetting 10 μl of the Hi-C library into the hybrisation buffer:RNA mix (red into green). The remaining PCR strip was closed with a new strip tube lid and the reactions were incubated for 24 hours at 65 °C, in a total reaction volume of 30 μl.

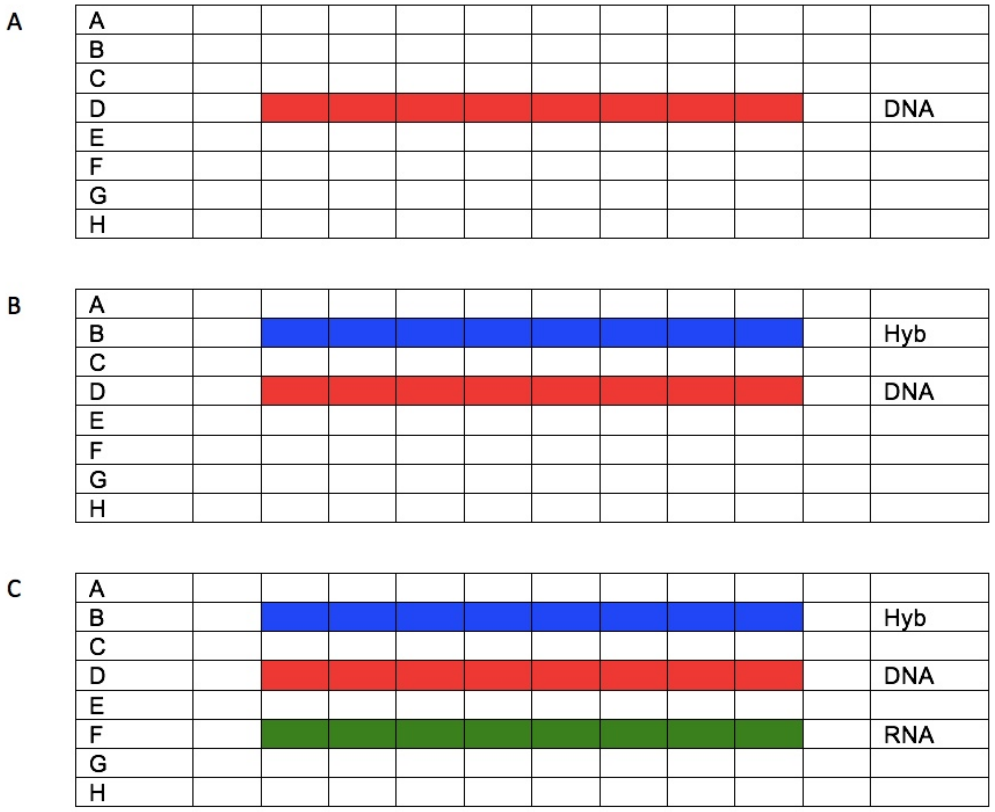

Figure 4.3: PCR machine set up for the hybridisation of RNA baits to DNA libraries. PCR strip containing DNA (red) is placed in row D at 95 for 5 minutes. Once the temperature has cooled to 65 the PCR strip containing hybridisation buffer (blue) is put in row B in the machine and incubated for five minutes. After the PCR strip containing RNA (green) is added to row F and incubated for 2 minutes before 13 hybridisation buffer is added to the RNA (blue into green) immediately followed by the transfer of 10 μl DNA library into the RNA mix (red into green).

Streptavidin-biotin pull-down and washes

Captured DNA/RNA hybrids were enriched using Dynabeads MyOne Streptavidin T1 beads (Life Technologies). 60 μl T1 beads per captured library were aliquoted into a LoBind Eppendorf and washed three times in 200 l binding buffer (BB: 1 M NaCl, 10 mM Tris-HCl pH 7.5, 1 mM EDTA). With the streptavidin beads in 200 μl BB the entire hybridisation reaction from the 24-hour incubation was transferred into the beads and mixed. This was then incubated at room temperature on a rotating wheel. After 30 minutes the beads were reclaimed on a magnetic separator and the supernatant was discarded. The beads were then resuspended in 500 μl wash buffer I (WBI: 1x SSC, 0.1% SDS) and incubated at room temperature for 15 minutes with agitation every 2-3 minutes. Following incubation, the beads were once more reclaimed and the supernatant was discarded. The beads were then resuspended in 500 μl wash buffer II (WBII: 0.1x SSC, 0.1% SDS) pre-warmed to 65 °C, and then incubated at 65 °C for 10 minutes with agitation every 2-3 minutes. This was repeated for a total of 3 washes in WBII. The beads were next reclaimed, the supernatant was discarded and the beads were resuspended in 200 μl 1x NEBuffer 2 and immediately transferred into a fresh LoBind Eppendorf tube.Tubes were placed immediately back on the magnetic rack and the supernatant was removed. Finally the streptavidin beads (with bound captured DNA/RNA) were resuspended in 30 μl 1x NEBuffer 2 and transferred into a fresh tube.

Post-capture PCR amplification of SCRiBL Hi-C libraries

To determine the optimal number of PCR cycles for SCRiBL Hi-C library amplification, test PCRs were set up as previously described with PCR cycle numbers tested 9, 12 and 15. The amount of amplified DNA was then checked by running the entire reaction on a 1.5% agarose gel. Again, the number of cycles chosen for the final PCR amplification of the SCRiBL library was determined by choosing the number of PCR cycles immediately before the appearance of a smear on the gel. For the final PCR amplification, multiple reactions were set up so that the entire volume of SCRiBL Hi-C library was used (2.5 μl in each PCR reaction). In the final reaction primer

pairs consisted of one TruSeq adapter reverse compliment and the TruSeq universal adapter (Table 4.1). Each primer pair introduced a library-specific barcode required for accurate sequencing of each library.

The samples from the complete PCR reaction were pooled, placed on a magnetic separator and the supernatant transferred into a fresh LoBind Eppendorf tube. The beads were resuspended in 20 μl 1x NEBuffer 2 and kept at -20 °C as a back up. The volume of supernatant containing the captured library was determined and was then purified by performing sequential SPRI bead selections at 1x and 1.8x volume (Beckman Coulter). The captured libraries were finally eluted from the beads in 20 μl TLE. Before sequencing, the quality and quantity of all libraries were checked by Bioanalyzer (Agilent) and Kapa Q quantitative PCR (Kapa Biosystems).

Paired-end Next Generation Sequencing

Two biological replicate Hi-C and capture Hi-C libraries were prepared for each of the cell lines. Sequencing was performed on Illumina HiSeq 2500 generating 50bp paired-end reads (Sequencing Facility, Babraham Institute). CASAVA software (v1.8.2, Illumina) was used to make base calls and reads failing Illumina filters were removed before further analysis. Output FASTQ sequences were mapped to the human reference genome (GRCh37/hg19) containing the HPV16 genome as an extra chromosome and were filtered to remove artefacts using the Hi-C User Pipeline (HiCUP)

Table 4.1: Details of oligonucleotides used in the synthesis of SCRiBL Hi-C libraries

Oligo	**DNA sequence (5' → 3')**
RPL13A B forward	AGGCGTGTTACTGGAAGTCG
RPL13A G forward	GAGCCTTGCTGGTCTTCGTT
RPL13A D2J forward	GGTGCGATCGATCCTCATGAAA
RPL13A D2J reverse	GTGACTGACAGCTGGGCATA
SCRiBL TruPE adapter 1	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
SCRiBL TruPE adapter 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
PE PCR primer 1.0 33	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
PE PCR primer 2.0 33	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
T7 promoter adapter forward	TCTAGTCGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAG
T7 promoter adapter reverse	CTAGGAGCGGGATATCACTCAGCATAATGTTAAGTGACCGGCAGCTGATCT
HPV_RNA_block1_for	AATTGTAATACGACTCACTATAGGGAGACCCATGTACCAATGTTGCAG
HPV_RNA_block1_rev	ATCCCGAAAAGCAAAGTCAT
HPV_RNA_block2_for	AATTGTAATACGACTCACTATAGGGAGATGACTTTGCTTTTCGGGATT
HPV_RNA_block2_rev	TTGCTTCCAATCACCTCCAT
HPV_RNA_block3_for	AATTGTAATACGACTCACTATAGGGAGAAGATGTGATAGGGTAGATGATGGA
HPV_RNA_block3_rev	TGTAATTAAAAAGCGTGCATGTG
HPV_RNA_block4_for	AATTGTAATACGACTCACTATAGGGAGACATACACATGCACGCTTTTT
HPV_RNA_block4_rev	TTCCCCTATAGGTGGTTTGC
P5b1for33	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCdd
P5b1rev33	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P7b2for	ACACTCTTTCCCTACACGACGCTCTTCCGATCdd
P7b2rev	AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
TruSeq universal adaptor SCRiBL	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
T7 adapter sequence	TAATACGACTCACTATAGGG
Barcoded primers	**Sequence 3' → 5'**
TruSeq adapter, index 1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 8	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 9	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 13	CAAGCAGAAGACGGCATACGAGATTGTTGACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 14	CAAGCAGAAGACGGCATACGAGATACGGAACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 16	CAAGCAGAAGACGGCATACGAGATCGGGACGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq adapter, index 19	CAAGCAGAAGACGGCATACGAGATCGTTTCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

4.2.6 Bioinformatic analysis performed by Jack Monahan

HiCUP & SeqMonk

Sequence data was obtained from Illumina HiSeq paired-end sequencing. Using the HiCUP Pipeline[258], paired-end capture Hi-C (cHi-C) fastq files were mapped with Bowtie 2[259] to a human GRCh37 reference containing an HPV16 pseudo-chromosome. HiCUP removes invalid and artefactual di-tags by overlaying the di-tags on an *in silico* restriction digest of the reference. The resulting BAM files contained putative di-tags for use in subsequent analyses. SeqMonk (Somin Andrews, Babraham Bioinformatics) was used to quantitate and visualise the density of di-tags contained in the BAM files. The HPV16 sequence and annotation files were downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/ena/data/view/K02718). ENCODE Annotation for NHEK[260] was obtained from Ensembl release 75[261].

Circos

The raw cHi-C fastq files were converted to fasta format and BLAST[262] was used to search the HPV16 genome for reads mapping to it. The partner human reads were determined and the 2 sets of reads were mapped to the GRCh37 reference containing the HPV16 pseudo-chromosome using Bowtie 2. The BAM outputs were converted to BED format and modified to be compatible with the circular visualisation tool *Circos*[263]. The HPV16 genome was split into bins of 500 bp and the count per bin determined from the chimaeric human-HPV16 di-tags. The counts, the HPV16 MboI restriction map and gene coordinates were annotated on the Circos plots.

GOTHiC

The HiCUP output was converted to format compatible with the Bioconductor package *GOTHiC*[254, 264]. To find significant interactions between distal locations GOTHiC implements a cumulative binomial test based on read depth. This was used to identify regions of the human genome in contact with the HPV16 pseudo-chromosome at a resolution of 1kb. Di-tag mappings were visualised with Circos after filtering the previous Circos input by the GOTHiC determined interactions.

Breakpoint Mapping with USearch

The precise sites of HPV16 integration in the W12 cell lines were identified by sequencing undigested Hi-C libraries. The raw fastq files were converted to fasta format and BLAST was used to search for reads mapping to the HPV16 genome. From these, the corresponding human tag were determined. Fast clustering of the reads with *USearch*[265], based on an sequence identity score of 0.65, identified clusters of sequences in the human and HPV16 derived reads. Consensus sequences from non-singleton clusters were obtained by aligning the clustered reads to each other using *Clustal Omega*[266]. The breakpoints were inferred from these consensus sequences and validated by Sanger Sequencing[267].

From the validated integration sites, custom chimaeric references were generated for each W12 line. Due to the existence of tandem amplifications in some of the regions of integration, two versions of the chimaeric human-HPV16 chromosomes were generated. In the first case, the HPV16 provirus was 5' of a single amplified human sequence. For the second, the provirus was placed 3' of the amplified human sequence. For another W12 line, 'H', there is a deletion in the region of integration and this was reflected in the chimaeric chromosome.

Juicer and Juicebox

Using the specific chimaeric references, Hi-C contact maps at different resolutions were generated from raw Hi-C fastq files using the *Juicer* Pipeline[268]. Juicer constructs a compressed contact matrix from pairs of genomic positions located in close proximity in 3D space. The Hi-C contact maps were imported into *Juicebox*[269] for visualisation.

HiCUP (2^{nd} time) & FourCSeq

HiCUP alignment and filtering was repeated on the cHi-C using the chimaeric references. The di-tags were subsequently filtered to remove those that did not have a tag in the captured region (i.e. the HPV16 provirus). cHi-C approximates a multiplexed Circularized Chromosome Conformation Capture (4C) experiment with

multiple viewpoints. However due to the relatively low number of captured di-tags per HPV16 MboI fragment, the entire provirus had to be treated as a single 4C viewpoint. The tag counts per HPV16-interacting fragment were determined and these counts were supplied to the Bioconductor package FourCSeq[270]. To find significant interactions between the viewpoint and fragments, FourCSeq applies a variance stabilising transformation to the counts and calculates a distance-dependent monotone fit.

Z-scores are derived from the fit residuals of the fragments selected by the fit. Significance is dependent on genomic distance from the viewpoint and takes into account the agreement between replicates. An asymmetric fit, allowing for differences between the regions upstream and downstream of the provirus was used. Significant fragments were those with z-scores greater than 2 in both replicates and an adjusted p-value of less than or equal to 0.05 in at least one replicate.

4.3 Results

4.3.1 Modifications to the 'Sequence Capture of Regions Interacting with Bait Loci' (SCRiBL) protocol for production of Hi-C and captured libraries from the W12 clones.

The aim of the collaboration with Peter Fraser's group (Babraham Institute) was to adapt their protocol for producing SCRiBL Hi-C libraries toward use on the panel of integrated HPV16 W12 clones. The experimental aims were two fold and ran concurrently; to determine areas of the host genome that the integrated HPV16 genome was in contact with in three dimensions (3D), and, to identify both 5' and 3' virus-host junctions in each of the W12 clones tested. The key experimental stages of generating SCRiBL libraries are outlined in Figure 4.4, and are described in detail in the Methods section (Chapter 4.2).

Important adaptations to the original protocol for producing SCRiBL libraries for the W12 integrant clones (with viral genome copy number less than four: F, A5, D2, H, G2) were the use of the 4-cutter restriction enzyme MboI, with cut site GATC, and the generation of biotinylated RNA baits specific to the HPV16 genome. The major distinction between the generation of libraries required for SCRiBL Hi-C and those used for breakpoint identification was the digestion of chromatin with a suitable restriction enzyme. Digestion with MboI was not required for the generation of libraries used for breakpoint identification; as such these libraries were named undigested throughout the course of this work. For each W12 clone tested, two biological replicates of both libraries were produced (Figure 4.5).

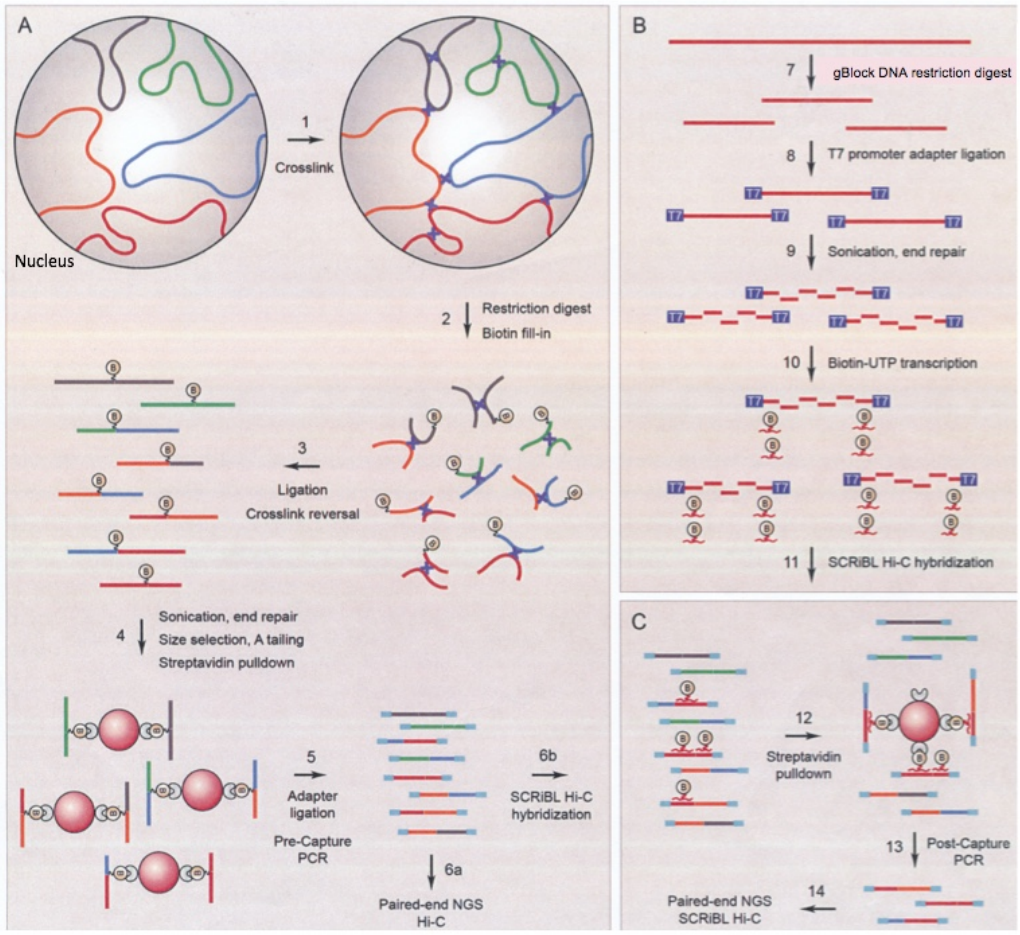

Figure 4.4: Schematic overview of the experimental processes required to produce Sequence Capture of Regions interacting with Bait Loci (SCRiBL) libraries suitable for next generation sequencing (NGS) on an Illumina platform. A) The generation of a Hi-C library is divided into six main steps. (1) Initially cells were grown to 90% confluency and fixed using methanol-free formaldehyde, preserving the 3D nuclear architecture of DNA. (2) The subsequent chromatin was subject to a restriction enzyme digest and the resulting sticky ends filled in with generic dNTPs and biotin labelled dATP. (3) Fragments in close proximity were ligated together via blunt-end, in-nucleus ligation and the crosslinks between DNA fragments reversed. (4) Ligated DNA was fragmented further by sonication centred at 400 bp and resultant DNA ends were repaired. Size selection was performed to exclude fragments of the wrong size and a single dATP added to the repaired 3' end to allow sequencing adapters to be ligated at a later stage. Streptavidin pull down of the remaining fragments from the Hi-C library ensured that only those containing biotin-incorporated DNA were carried forward. (5) Illumina sequencing adapters were then ligated to the 3 dATP and a pre-capture test PCR carried out to determine the number of amplification cycles required to generate the final Hi-C library. (6a) Hi-C libraries were then sent for paired-end sequencing or (6a) used in a hybridisation reaction with specifically designed biotin-labelled RNA baits. B) Generation of RNA baits. (7) DNA complementary to the capture region was designed using the IDT gBlock approach. DNA fragments were then released via DNA restriction digest. (8, 10) T7 promoter adapters were then ligated to the DNA fragment ends to facilitate *in vitro* transcription with biotin-labelled dUTP. (11) The RNA baits were then incubated with the DNA Hi-C library in a hybridisation reaction. C) Generation of the final SCRiBL library. (12) Biotin-labelled hybrid DNA:RNA was captured via a pull down with streptavidin beads, and (13) a post-capture PCR reaction performed to determine the number of ampflication cycles required to generate the final SCRiBL library. (14) Each library was then placed on the Illumina platform for 100 bp, paired-end NGS.

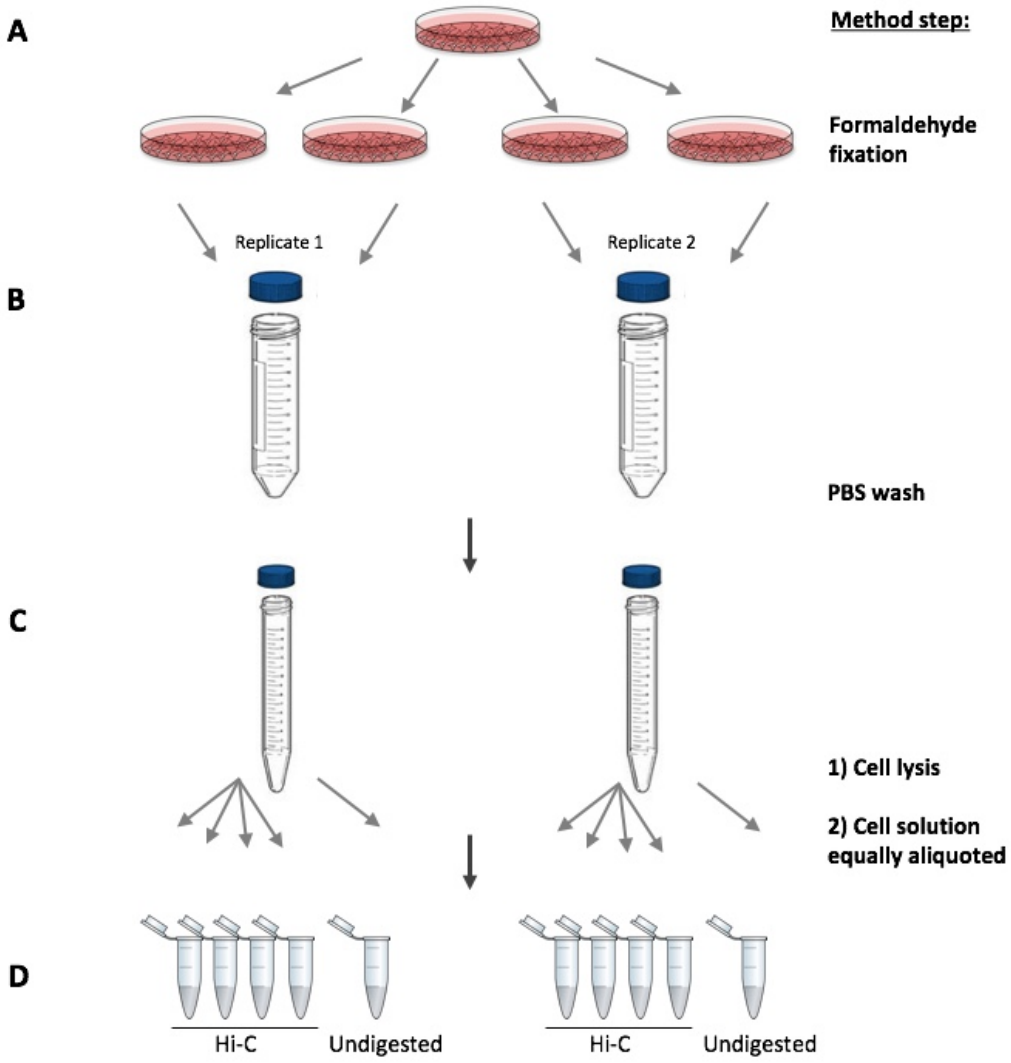

Figure 4.5: Diagram indicating the experimental set up of Hi-C library generation for the W12 clones. A) A single 15 cm^2 plate of either F p6, G2 p12, NCx p5, A5 p5, D2 p8 or H p6 at 80–90% confluence was used to seed four additional 15 cm^2 plates. B) At 90% confluence, cells were fixed with formaldehyde and two 15 cm2 plates were combined into one 50 mL Falcon tube to provide cellular material for one biological replicate. Cellular material was then separated from the supernatant by centrifugation. C) The chromatin pellet was washed with PBS and transferred into a 15 mL Falcon tube before a second centrifugation step. D) Following cell lysis, the cell solution was divided equally between five 1.5 mL Eppendorf tubes. Four were used for the generation of a Hi-C library for use in SCRiBL, and one was kept as an undigested library and was used in the capture-seq experiment.

4.3.2 Short-range interactions of W12 clone Hi-C libraries are detected using chromosome conformation capture (3C) assays.

Quality control checks give a strong indication of whether a Hi-C library will be successful; as such, forward primers were designed across the RPL13A host genomic control locus to detect short-range interactions generated by unique ligation events in the Hi-C libraries (Figure 4.6 A). The RPL13A genomic locus was divided according to MboI restriction sites and the fragments labeled alphabetically. Forward primers within a number of fragments were designed, and two forward primers were paired in a subsequent PCR reaction. The use of two primers in the same orientation meant that PCR products were only detectable in libraries previously digested with MboI where ligation had occurred (Hi-C). PCR reactions with primer pairs B:I and B:G resulted in expected 394 bp and 319 bp products respectively using W12 clone F Hi-C template DNA (Figure 4.6 B). An annealing temperature of 53.9 °C was carried forward and the number of PCR cycles increased to amplify the strength of the band. Use of the B:I forward-forward primers resulted in Hi-C specific products; the strongest band was produced using W12 G2 Hi-C DNA template compared to a minimal amount of 394 bp-specific product from the NCx or W12 F Hi-C libraries (Figure 4.6 C). In contrast, use of the B:G forward-forward primers in the PCR reaction resulted in an abundance of the 319 bp-specific ligation product in all three Hi-C libraries (Figure 4.6 D). Although there were additional 3C ligation products generated by both primer pairs, fewer were made using the B:G primers; moreover, any additional products were further from the band of interest making detection of the desired product more reliable. Consequently, the B:G forward-forward primers designed across the RPL13A locus were used for the detection of short-range interactions in all subsequent W12 Hi-C libraries.

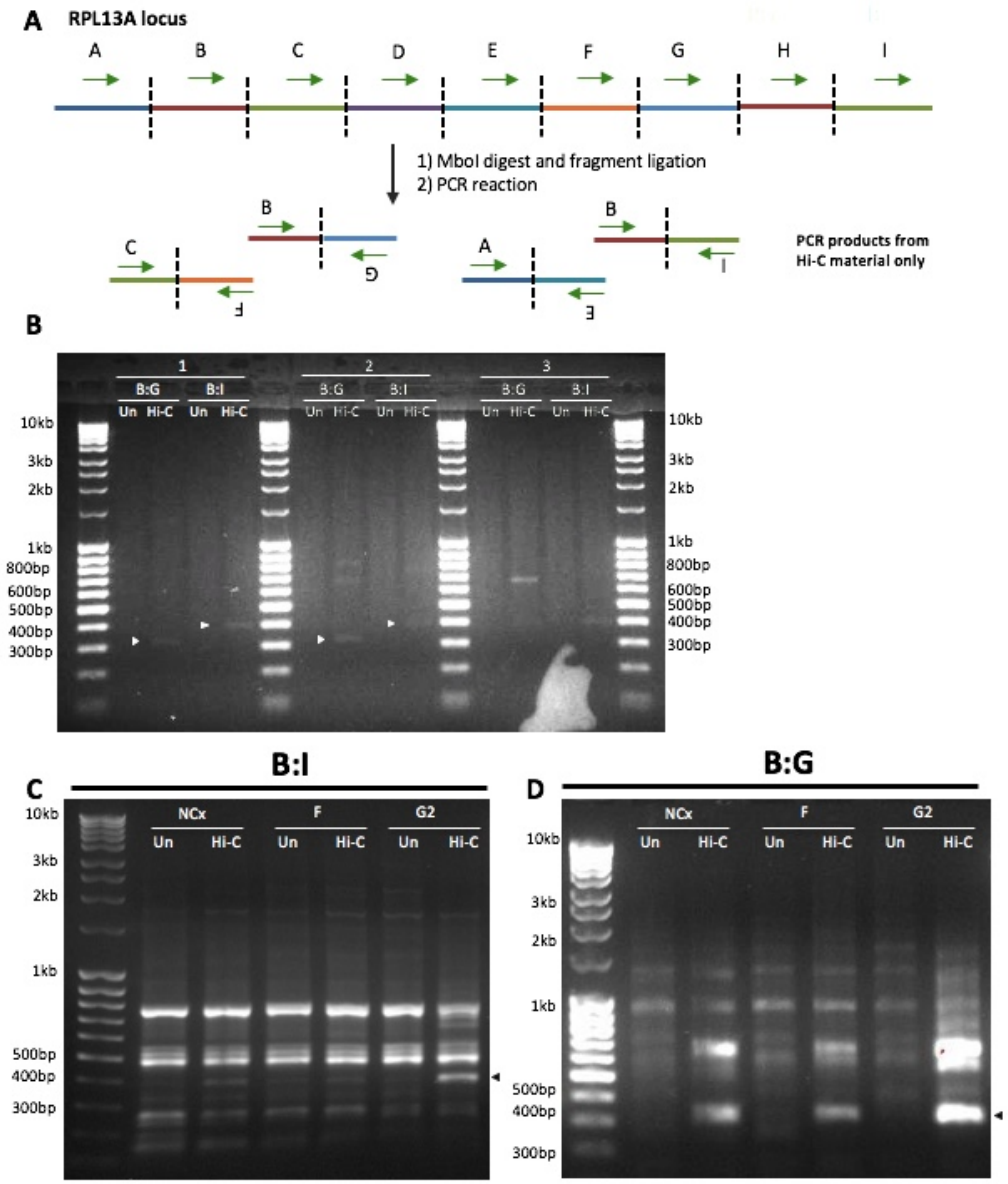

Figure 4.6: Detection of short-range interactions in the W12 clone F Hi-C library. A) Diagram showing the generalised principle of using two forward primers designed across the RPL13A locus (control region) according to MboI restriction sites (indicated by dashed line) to determine short-range interactions in the W12 clone F Hi-C library. Primers are depicted as green arrows. B) Optimisation of conditions for PCR of Hi-C library using primer pairs B:G and B:I. PCR reactions were carried out at a range of temperatures; 1=53.9 °C, 2= 52.7 °C and 3=52 °C for 35 cycles. Each 25 μl PCR sample was run on a 1.5% agarose gel. White arrowheads indicate the PCR products of interest; B:G = 319 bp, B:I = 394 bp. PCR reaction of NCx undigested and Hi-C, W12 F undigested and Hi-C and W12 G2 undigested (Un) and Hi-C libraries (Hi-C) using (C) B and I forward primers and (D) B and G forward primers. PCR reactions were carried out at 53.9 °C for 38 cycles. Each 25 μl PCR sample was run on a 1.5% agarose gel. Black arrowheads indicate the B:I and B:G specific PCR products.

For robustness, an alternative method for detecting short-range interactions in Hi-C libraries was utilised (Figure 4.7 A). The RPL13A genomic locus was again divided into fragments according to MboI restriction sites and labeled alphabetically. Forward primers spanning two restriction fragments were designed, each containing a 5'-GATCGATC-3' sequence. This eight-nucleotide sequence was generated from blunt-end, in-nucleus ligation of MboI restriction fragments and was present in Hi-C libraries. A reverse primer in the same fragment as the 3'-end of the forward primer was designed and the combined pair used to generate specific PCR products. No PCR product should be generated using an undigested library as the forward primer will not be complementary to template DNA and should not anneal.

PCR reactions were carried out testing three alternative primer pairs; similar amounts of PCR product were produced regardless of W12 F DNA template with primer pairs B:H and C:H; 369 bp and 373 bp, respectively. However, although some product was formed using undigested material with primers D:J, product formation was much more specific to the Hi-C sample (Figure 4.7 B). An abundance of PCR product in both the undigested and Hi-C libraries for the B:H and C:H primer pairs was likely due to the unspecific binding of the forward primer as a result of a low annealing temperature (55 °C) and a high GC content (64%) compared with D:J primer pair (38%). Optimisation of PCR conditions for Hi-C specific amplification of the D:J 401 bp product was carried out at a range of primer annealing temperatures (Figure 4.7 C). An annealing temperature of 62 °C and 35 cycles of amplification was carried forward when testing for short-range interactions in all W12 clone Hi-C libraries (Figure 4.8 A-D). Specific, short-range interactions were detected in the Hi-C libraries of all clones and no product was formed in the undigested counterpart; this was a clear indication of successful library generation.

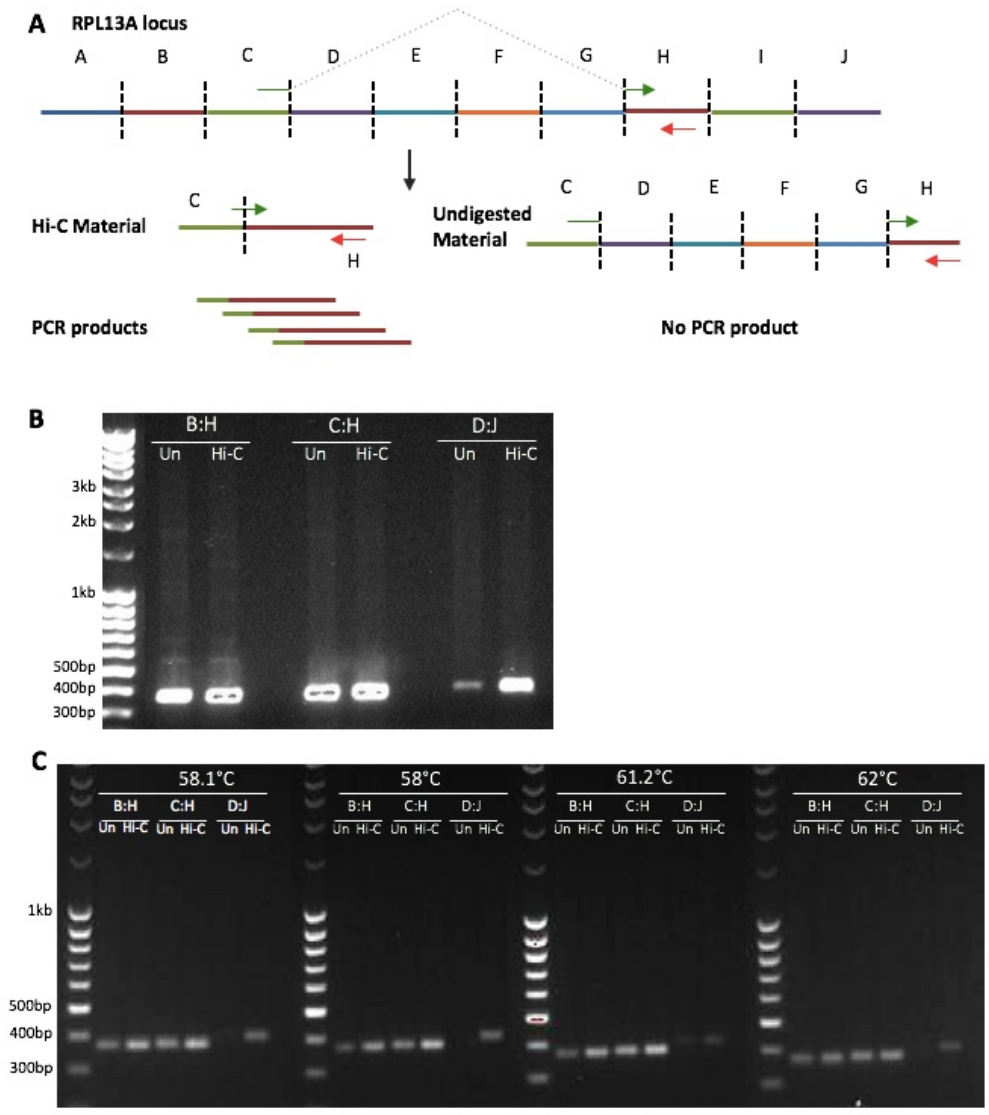

Figure 4.7: Alternative method of detecting short-range interactions in the W12 F Hi-C library. A) Diagram showing the generalised principle of pairing a forward primer, designed to span an MboI restriction site (indicated by dashed line) of the RPL13A locus (control region), with a reverse primer in a different segment to determine short-range interactions in a Hi-C library. Resultant PCR products are specific to digested Hi-C material. Primers are depicted as arrows; forward orientation = green, reverse = red. B) PCR of short-range interactions in W12 F undigested (Un) and Hi-C samples using primer pairs: B forward and H reverse (B:H), C forward and H reverse (C:H), and D forward and J reverse (D:J). PCR reactions were conducted at 55 °C for 35 cycles; 20 μl of the resultant PCR sample was run on a 1.5% agarose gel. C) Optimisation of conditions for PCR using primers B:H, C:H and D:J. PCR reactions were conducted at a range of temperatures; 1=58.1 °C, 2=60 °C, 3=61.2 °C and 4=62 °C for 35 cycles; 10 μl of the resultant PCR sample was run on a 1.5% agarose gel.

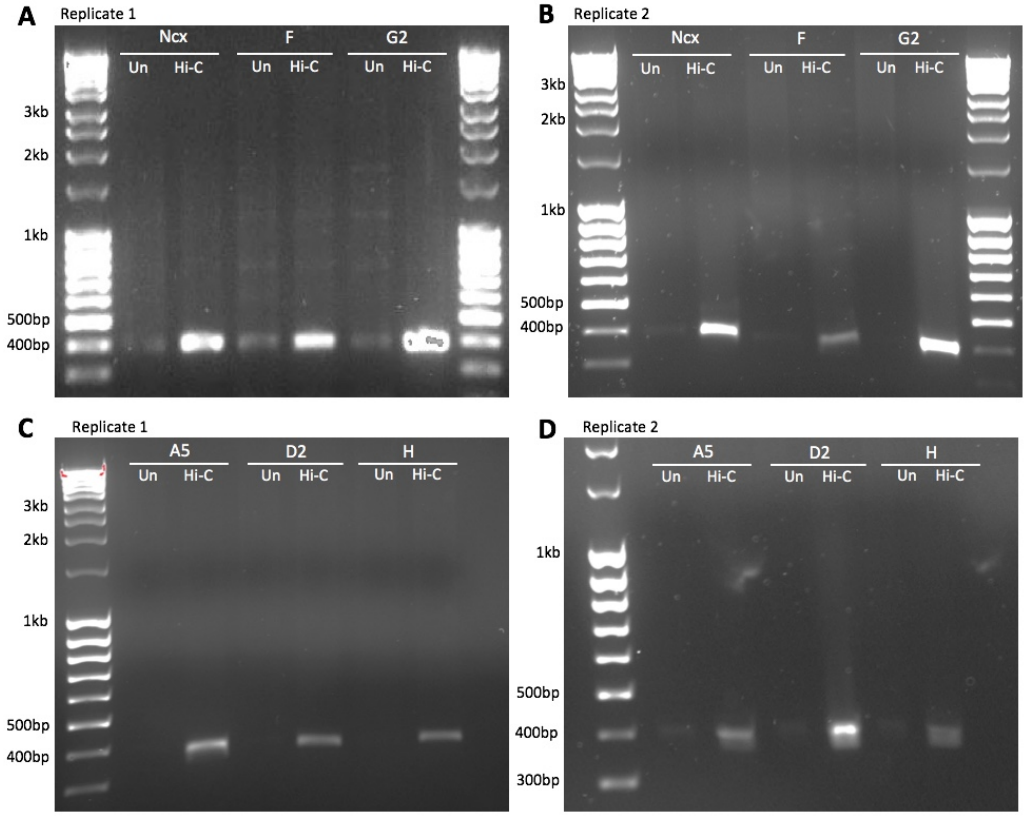

Figure 4.8: Detection of short-range interactions in each W12 clone Hi-C library.PCR reactions of undigested (Un) and Hi-C libraries with D:J primers; NCx, W12 F and W12 G2 replicate 1 (A) and replicate 2 (B), W12 A5, W12 D2 and W12 H replicate 1 (C) and replicate 2 (D). Each PCR reaction was conducted at 62 °C for 35 cycles; 10 μl of the resultant PCR sample was run on a 1.5% agarose gel.

4.3.3 Verification of in-nucleus ligation efficiency of W12 clone Hi-C libraries by PCR digest assay.

An additional quality control was carried out before proceeding with the generation of Hi-C libraries; the ligation efficiency of in-nucleus ligation was verified by PCR digest assay. Successful fill-in and ligation of two MboI sites (GATCGATC) creates a site for the restriction enzyme ClaI (ATCGAT) (Figure 4.9 A). Hi-C DNA was used as a template for PCR reactions using the previously optimised B:G forward-forward primers (Figure 4.9 D). Following amplification, half of the PCR reaction mixture was incubated with ClaI for 2 hours at 37 °C. The products of the PCR and digestion reactions were compared by agarose gel electrophoresis. Every Hi-C sample showed the specific B:G PCR product (319 bp). In contrast, in the ClaI digested samples this product band was cut into two alternative bands of 220 bp and 100 bp (Figure

4.9 B-D). The ligation efficiency was estimated by quantifying the intensity of the cut and uncut bands using Image J software. The ligation efficiency was high in all W12 Hi-C libraries ranging from 71.1% (W12 clone A5 rep I) to 94.2% (W12 clone G2 rep I); this was also suggestive of successful library generation.

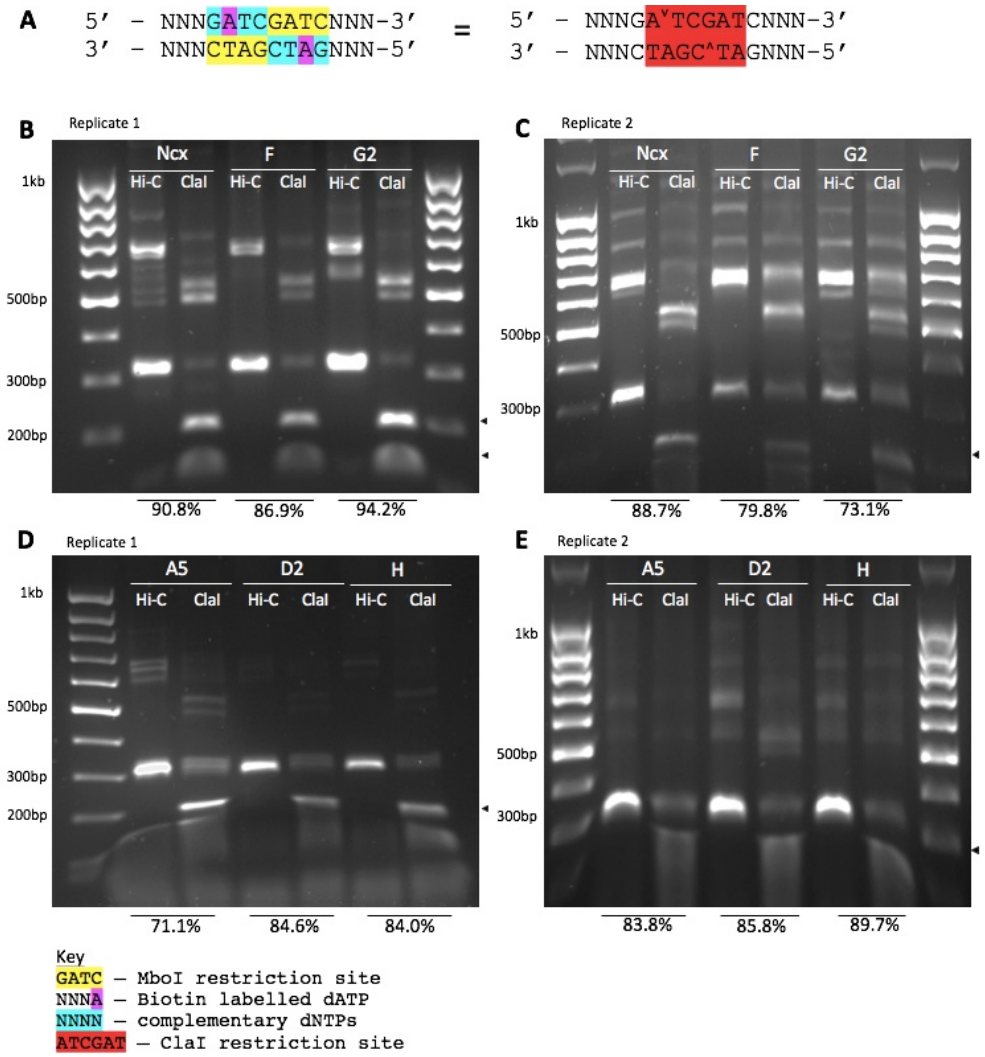

Figure 4.9: Determination of ligation efficiency in each W12 clone Hi-C library. A) Diagram showing that the product of blunt-end ligation of MboI restriction fragments constitutes a ClaI restriction site. PCR reactions with BG primers and ClaI digestion reactions of NCx, W12 F and W12 G2 replicate 1 (B) and replicate 2 (C), W12 A5, W12 D2 and W12 H replicate 1 (D) and replicate 2 (E). For each clone, the Hi-C lane is loaded with the 750 ng PCR product produced using the B:G forward primers at 59.3 °C for 38 cycles. The ClaI lane contains digestion products following incubation of 750 ng PCR product with 20 U ClaI restriction enzyme for 2 hours at 37 °C. All samples were run on a 1.5% agarose gel. The ligation efficiency for each library is shown under each gel.

4.3.4 Determination of conditions for final PCR amplification of W12 clone Hi-C libraries.

Following stages 4 and 5 of Hi-C library generation (Figure 4.4 A), a test PCR was carried out to determine the number of amplification cycles required to generate sufficient DNA material for the subsequent hybridisation reaction and NGS as well as an appropriate level of library complexity. PCR reactions were carried out using PE primers 1.0 and 2.0 (Illumina) at 65 °C for 6, 9 or 12 cycles (Figure 4.10 A-D). A smear of 300-500 bp was produced using NCx, W12 F and W12 G2 material (Figure 4.10 A and B), whereas the smear produced from W12 A5, W12 D2 and W12 H samples was slightly higher, centered around 500 bp (Figure 4.10 C and D). The DNA concentration of the undigested libraries was higher than the Hi-C equivalent; this was due to the loss of material of the Hi-C samples as a result of the extra processing steps required for library generation. For example, the smear produced following PCR with 6 amplification cycles of undigested NCx template DNA was similar to that produced by 12 cycles of amplification of NCx Hi-C DNA template. The necessary number of PCR amplification cycles for Hi-C libraries was decided by choosing one fewer number of cycles than that at which a smear was visible; this was constant across samples generated at the same time, i.e. NCx, F and G2 = 8 cycles and A5, D2 and H = 9 cycles.

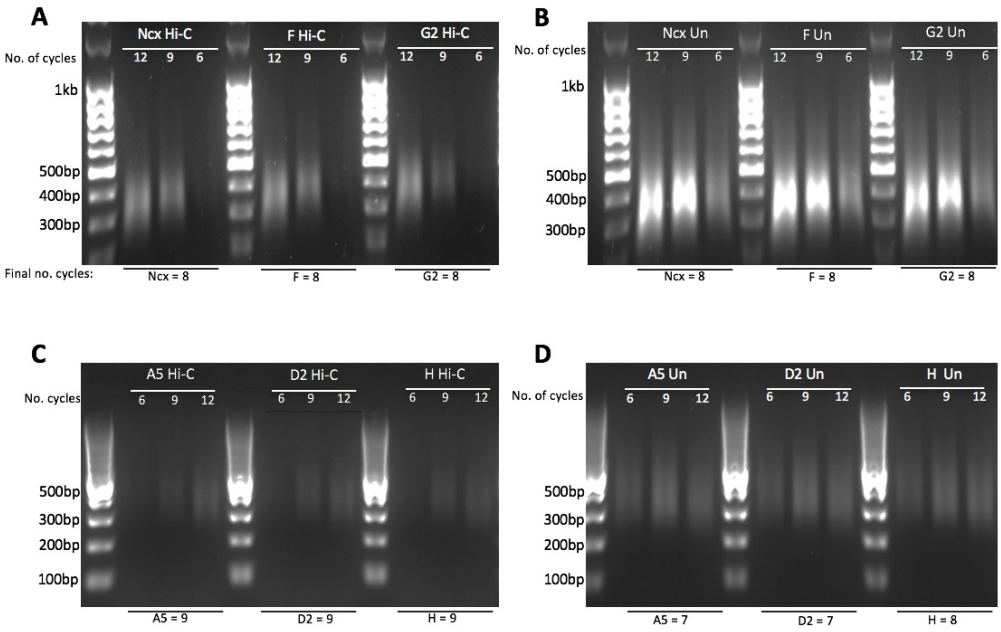

Figure 4.10: Test amplification of Hi-C and undigested libraries to determine the number of PCR amplification cycles required to generate sufficient material and library complexity. 2.5 μl DNA from the Hi-C (A and C) or undigested libraries (B and D) of each clone were combined with Illumina primers PE 1.0 forward and PE 2.0 reverse for either 6, 9 or 12 cycles (as indicated) at 65 °C. Each 25 μl PCR sample was run on a 1.5% agarose gel against a 100 bp ladder. The number of cycles chosen to generate the final Hi-C or undigested libraries are shown under each gel.

4.3.5 Generation of RNA baits for capture-sequencing to detect virus-host breakpoints in the W12 clones.

A key experimental aim was to accurately identify the virus-host breakpoint junctions in the W12 clones. To do this, enrichment of the HPV16 genome sequence from the undigested libraries, which had been fragmented by sonication only, was first used. HPV16-specific biotinylated-RNA baits were made, and used to hybridise to W12 undigested DNA libraries (Figure 4.4 B, steps 7-11). Subsequent streptavidin bead pull-down resulted in the sequencing of DNA fragments comprised at least partially of HPV16 genome. Aligning host-virus DNA reads to the human genome identified two peaks of reads that indicated both the 5' and 3' virus-host junctions.

The pSP64 plasmid, in which the W12E HPV16 genome is cloned (Cinzia Scarpini) (Figure 4.11 A), was used to generate fragments of HPV16 DNA that were later in vitro transcribed to RNA for use in the hybridisation reaction with the W12 clone

undigested libraries. Initially the integrity of the plasmid DNA was checked by digesting with either EcoRI or BamHI restriction enzymes (Figure 4.11 B). Digestion with EcoRI resulted in three DNA fragments whereas BamHI resulted in two, both indicative of the number of restriction sites in the pSP64 plasmid respectively. Four sets of primers were designed to evenly cover the entire HPV16 W12E genome (see Table 4.2).

Table 4.2: DNA primers spanning W12E genome

Name	HPV16 coordinates	Product length (bp)
HPV_DNA_block1	4 – 2,004	2000
HPV_DNA_block2	1,985 – 3,941	1956
HPV_DNA_block3	3,903 – 5,852	1950
HPV_DNA_block4	5,826 – 7,891	2066

These were used to amplify the virus genome in a touchdown PCR reaction and accurately produced the desired DNA products (Figure 4.11 C, Table 4.2). PCR products of amplified HPV16 genome were combined in equimolar amounts to produce HPV16 template. *In vitro* transcription with unlabeled ribonucleotides (rUTP, rATP, rCTP, rGTP) was trialed for a range of different reaction lengths (Figure 4.11 D). Samples were run against two different RNA molecular weight ladders to accurately determine product length. A strong band centered on 2 kb with a smear of different molecular weight products was produced in each of the *in vitro* transcription reactions of both the HPV and control DNA templates. The greatest concentration of HPV template RNA was produced after a reaction time of 18 hours, therefore this condition was taken forward. *In vitro* transcription of each HPV_RNA_block as well as an equimolar mix of all four was then carried out using biotin-UTP. HPV_DNA_block3 was not loaded onto the gel properly due to an air pocket and therefore reaction products are not clear. Despite this, a smear of RNA product from 2 kb – 200 bp was produced for each individual block as well as for the 'HPV template'.

Following the generation of full-length biotin-labeled RNA complementary to the HPV16 genome, chemical fragmentation with Tris pH 8.0 and 4 mM $MgCl_2$ was used to generated RNA baits approximately 150 bp in length (Figure 4.11 F). A

fragmentation time of 8 minutes resulted in appropriately sized RNA baits and this condition was used to generate a total of 5 μg biotinylated HPV16 RNA bait. After purification, an aliquot of the fragmented RNA baits was run against another RNA sample of known molecular weight (Marco Michalski) as well as two RNA ladders. This ensured that the baits were an appropriate size for even coverage and optimal hybridisation to the integrated HPV16 genome in the W12 clone undigested DNA libraries (Figure 4.11 G).

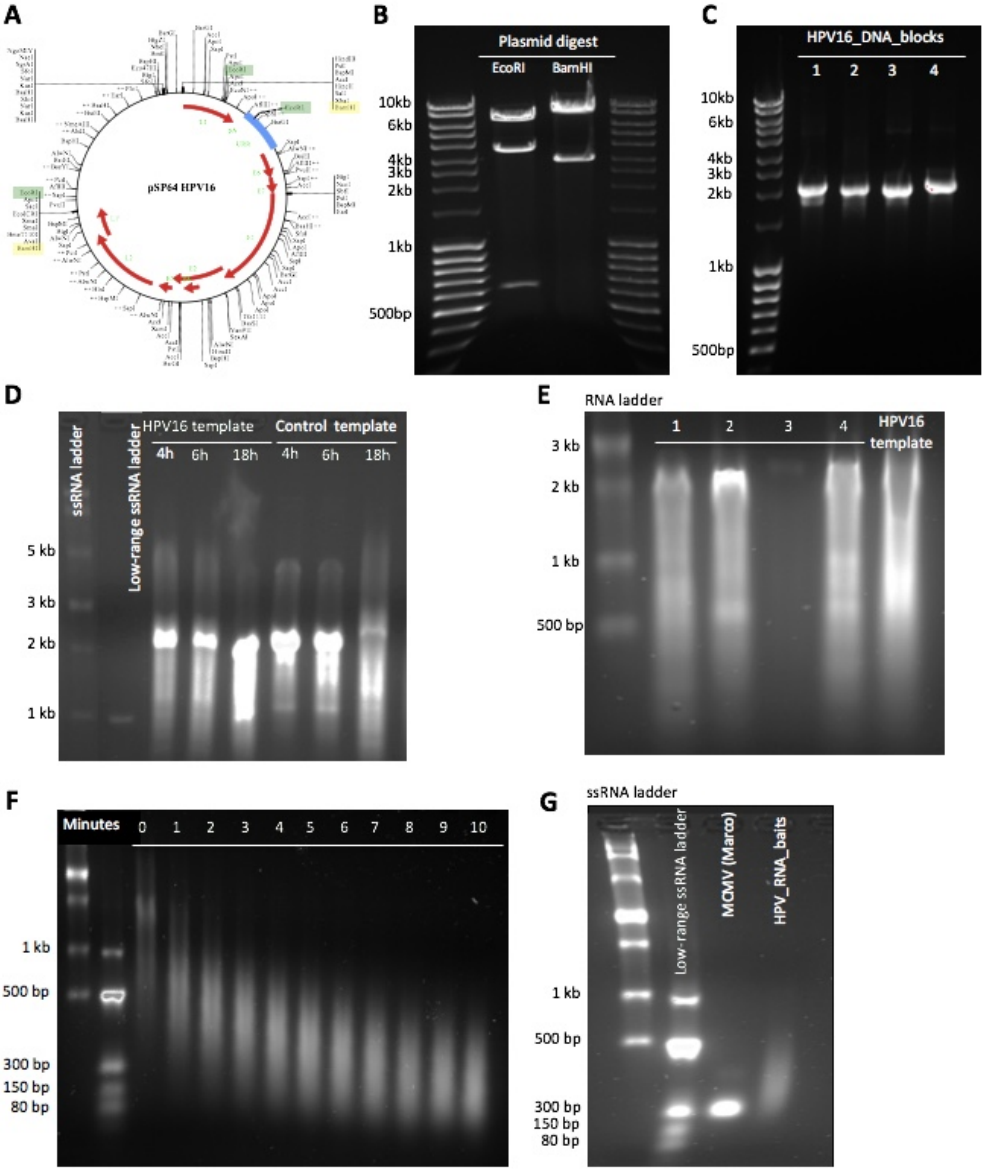

Figure 4.11: Generation of RNA baits for use in the capture-seq experiment with undigested W12 clone DNA libraries. A) Diagram showing pSP64 HPV16 plasmid design. HPV16 genes are indicated with red arrows and the restriction enzyme cut sites labelled on the outside of the circularised genome (EcoRI = green; BamHI = yellow.) B) Plasmid integrity check. Purified plasmid DNA was incubated with EcoRI or BamHI for 1 hour at 37 °C and the products run on a 1% agarose gel. C) PCR amplification of the HPV16 genome. Four sets of primer pairs were used to amplify the viral genome in a touch down PCR reaction. The PCR reaction was conducted at 74-62 °C for 13 cycles followed by 22 cycles at 62 °C. PCR products were run on a 1.5% agarose gel; HPV16_DNA_block1: 2000 bp, HPV16_DNA_block2: 1956 bp, HPV16_DNA_block3: 1950 bp, HPV16_DNA_block4: 2066 bp. D) *In vitro* transcription of the HPV16 genome trial. An equimolar mix of the full length DNA generated from each primer pair was combined to generate the HPV16 template, ptri-Xef RNA (Ambion) was used as a control. Various incubation times of *in vitro* transcription with unlabelled rUTP were tested; 4, 6 and 18 hours at 37 °C. The reaction products were run on a 1.5% agarose gel against a 100 bp DNA ladder, single-stranded (ss) RNA ladder and a low-range RNA ladder. E) *In vitro* transcription of the HPV16 genome with biotin-labelled dUTP. In addition to the equimolar mix, each HPV16_DNA_block was individually transcribed with biotin-labelled dUTP at 37 C overnight (18 hours). Resultant RNA was run on a 1.5% agarose gel; HPV16_RNA_block3 was not loaded properly and can therefore not be seen on the gel. F) HPV16 RNA bait fragmentation. 250 ng of full length HPV16 template RNA from each primer pair was incubated with 4 mM $MgCl_2$ and 100mM Tris pH 8 at 95 °C for 0 to 10 minutes. Reaction products were run on a 1.5% agarose gel. G) The size of the HPV16 purified RNA baits was checked compared to a known reference (MCMV, supplied by Marco Michalski) by running both on a 1.5% agarose gel.

4.3.6 Design and generation of RNA baits for the production of W12 clone SCRiBL libraries.

To identify 3D interactions between genomes of the integrated virus and the host, it was necessary to enrich the W12 clone Hi-C libraries for the HPV16 genome. This was done by hybridising biotinylated RNA baits specific for the HPV16 genome to Hi-C DNA libraries. The approach taken to generate appropriate RNA baits was different to that used in the capture-seq reaction; it was essential for the RNA baits used for SCRiBL to hybridise to the ends of MboI digested DNA fragments of the Hi-C libraries. As such, the virus genome was also fragmented according to MboI restriction sites (GATC) (Figure 4.13 A).

gBlock® Gene Fragments from Integrated DNA Technologies (IDT) were designed to generate biotinylated RNA baits for the capture of HPV16 genome within the Hi-C DNA libraries, the principle of which is illustrated in Figure 4.12 A-D. In the first instance ~120 bp long sequences of DNA complementary to the 5'-end of HPV16 genome MboI digested fragments were identified (Table 4.3). These DNA sequences were constructed into gBlocks® flanked by two restriction enzymes, either BglII/HindIII (gBlock1) or BglIII/SpeI (gBlock2) (Figure 4.12 E and F). Both gBlocks® were isolated following Zero-Blunt® TOPO® cloning and the reaction products visualised (Figure 4.13 B). The 1.2 kb and 1 kb products of gBlock1 and gBlock2, respectively, were cut from the gel and purified; the band at 4 kb in each sample represented the cloning vector and was not required. Double digestion of both isolated gBlocks® with the appropriate restriction enzymes released the ≈120 bp DNA sequences, to which T7 promoter adapters required for in vitro transcription were ligated. Reaction products from the digestion and ligation reaction were visualised and the specific 180 bp product (≈130 bp DNA fragment + 50 bp T7 adapter) extracted and gel purified for *in vitro* transcription (Figure 4.13 C). As well as the desired 180 bp product band, there were additional bands at 360 bp and 100 bp; these represented products comprised of two DNA fragments and two T7 adapters, and two adapters ligated together, respectively. In vitro transcription with biotin-UTP generated a tight band of RNA baits at 130 bp, which were used for the

capture of HPV16 fragments in the W12 clone Hi-C libraries (Figure 4.13 D).

Table 4.3: MboI restriction sites within the HPV16 genome and RNA baits for SCRiBL

	MboI restriction fragments in HPV16 genome			RNA baits				
#	Ends	Coordinates	Length (bp)	HPV16 gene	RNA bait Coordinates	Direction	Length (bp)	gBlock
1	MboI-MboI	525-621	97	E6/E7	529-622	F	93	1
2	MboI-MboI	622-870	249	E7	624-744	F	120	1
				E7/E1	872-752	R	120	1
3	MboI-MboI	871-3479	2609	E1	874-933	F	59	1
				E2	3480-3361	R	119	1
4	MboI-MboI	3480-4360	881	E2	3482-3602	F	120	1
				L2	4363-4242	R	121	1
5	MboI-MboI	4361-4519	159	L2	4364-4483	F	119	1
6	MboI-MboI	4520-4537	18					
7	MboI-MboI	4538-5071	534	L2	4541-4660	F	119	1
				L2	5073-4953	R	120	2
8	MboI-MboI	5072-6150	1079	L2	5075-5194	F	119	2
				L1	6150-6030	R	120	2
9	MboI-MboI	6151-6950	800	L1	6152-6271	F	119	2
				L1	6950-6830	R	120	2
10	MboI-MboI	6951-7013	63					
11	MboI-MboI	7014-524	1415	L1	7014-7134	F	120	2
				E6	525-406	R	119	2

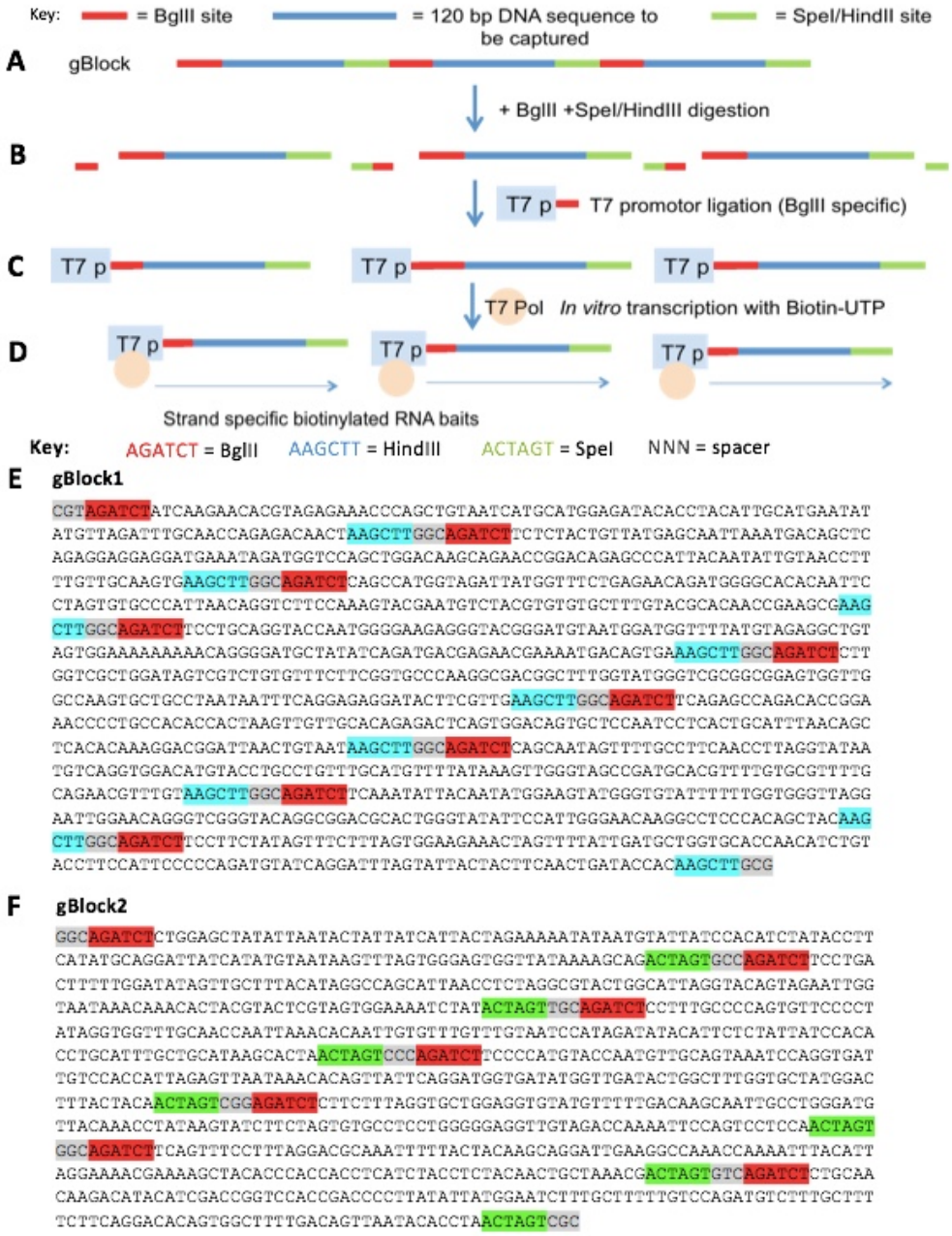

Figure 4.12: HPV16-specific RNA bait generation using the IDT gBlock approach for use in capture reaction to produce W12 SCRiBL libraries. A) Diagram depicting the gBlock design of 120 bp HPV16 DNA fragments encompassed by enzyme-specific restriction sites; gBlock1 contained restriction sites for BglII and HindIII whereas restriction sites for BglII and SpeI surrounded HPV16 DNA fragments in gBlock2. B) Digestion of each gBlock with the two specific restriction enzymes resulted in release of the HPV16 DNA fragments. C) T7 promoter adapters, required for *in vitro* transcription of template DNA, were ligated to the BglII site of the digested gBlock fragments. D) An equimolar mix of the two gBlock fragments were combined and the HPV16 DNA fragments *in vitro* transcribed to RNA using T7 RNA polymerase and biotin-UTP. E) Genomic sequence for gBlock1 design, F) Genomic sequence of gBlock2 design (coordinate details given in Table 4.2).

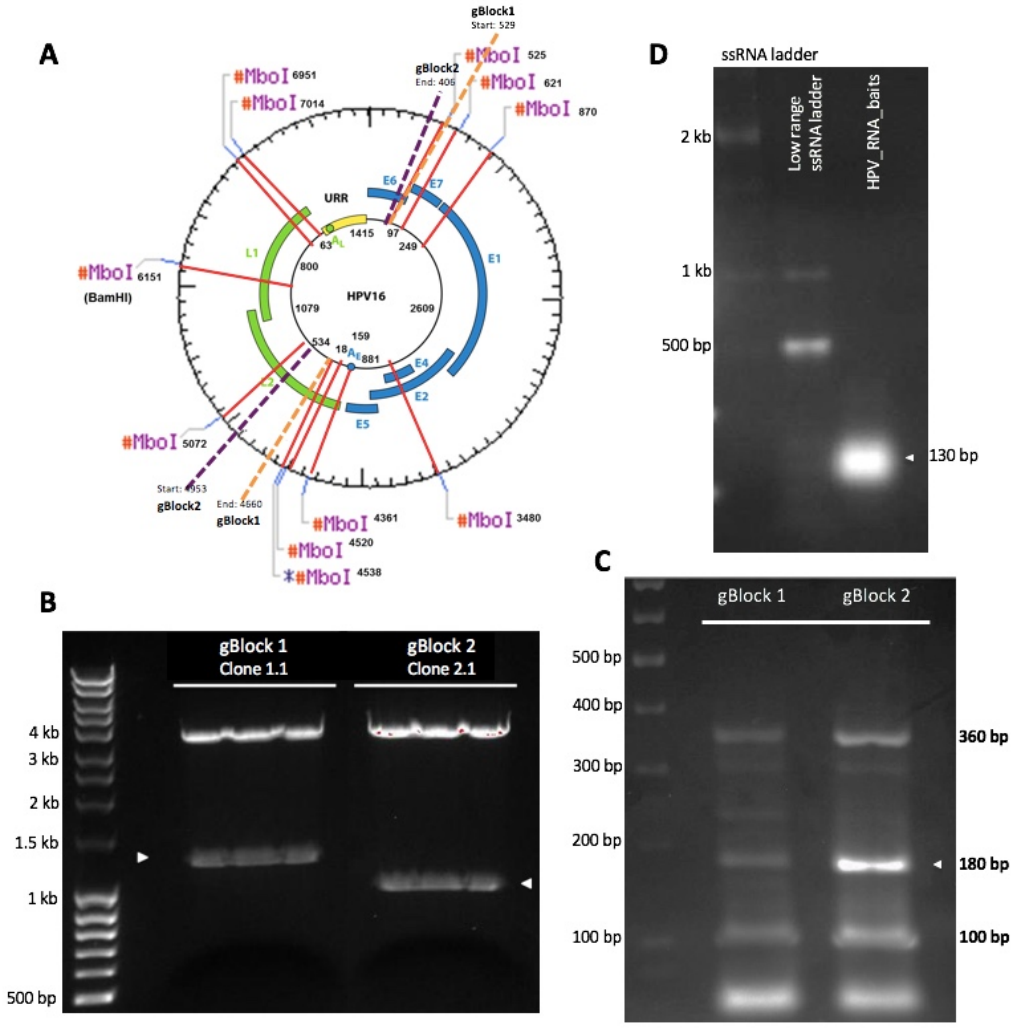

Figure 4.13: Generation of HPV16 RNA baits for use in SCRiBL experiments. A) Diagram showing the MboI restriction sites (red) and the coordinates of gBlock1 and gBlock2 (orange and purple dashed lines, respectively) marked on the circularised HPV16 genome. B) Isolation of gBlock fragments from the Zero Blunt® TOPO® cloning vector. Cloning vectors containing gBlock1 and gBlock2 were incubated with EcoRI for 2 hours at 37 °C and the reaction products run on a 1% agarose gel. White arrowheads indicate the excised gBlock product. C) gBlocks 1 and 2 were digested with two specific restriction enzymes to release individual HPV16 DNA sequences; gBlock1 was incubated with BglII and HindIII and gBlock2 with BglII and SpeI for 2 hours at 37 °C. This was followed by ligation of preannealed T7 sequencing adapters to the individual HPV16 fragments by incubation at 25 °C for 3 hours. Reaction products were then run on a 1% agarose gel. D) Equimolar amounts of each gBlock were combined, and the HPV16 fragments *in vitro* transcribed with biotinylated-dUTP. Following RNA purification, 2 μl of the reaction product was run on a 2% agarose gel to check the size of the RNA fragments (baits). The white arrowhead indicated the expected RNA product at 130 bp.

4.3.7 Enrichment of HPV16 genome from W12 clone Hi-C libraries through capture with biotinylated RNA baits.

After the hybridisation reaction and streptavidin pull down on RNA/DNA hybrid complexes (steps 11 and 12, Figure 4.4 C), a test PCR was carried out to determine the number of amplification cycles required to generate enough material for genomic sequencing but not to introduce excessive library complexity. PCR reactions of the SCRiBL libraries were carried out using PE primers 1.0 and 2.0 (Illumina) at 65 °C for 9, 12 or 15 cycles (Figure 4.14 A and C). Smears of ~300–800 bp were produced after 12 and 15 amplification cycles, although these were very faint in the NCx, F and G2 replicates; the final number of amplification cycles chosen was one directly below that at which a smear was visible, hence the final F and G2 libraries had an additional amplification cycle compared with A5, D2 and H (11 vs. 10). The PCR conditions for the captured undigested libraries were also determined using PE primers 1.0 and 2.0 (Illumina) at 65 °C for an increased number of cycles; 17, 20, or 23 cycles (Figure 4.14 B and D). A smear centered on 400 bp was produced using all W12 clone undigested DNA as PCR template and the final number of PCR amplifications determined similarly to the SCRiBL libraries. Both the Hi-C and undigested NCx libraries were not sequenced and a final library not generated, hence the amplification cycle number determination was not necessary.

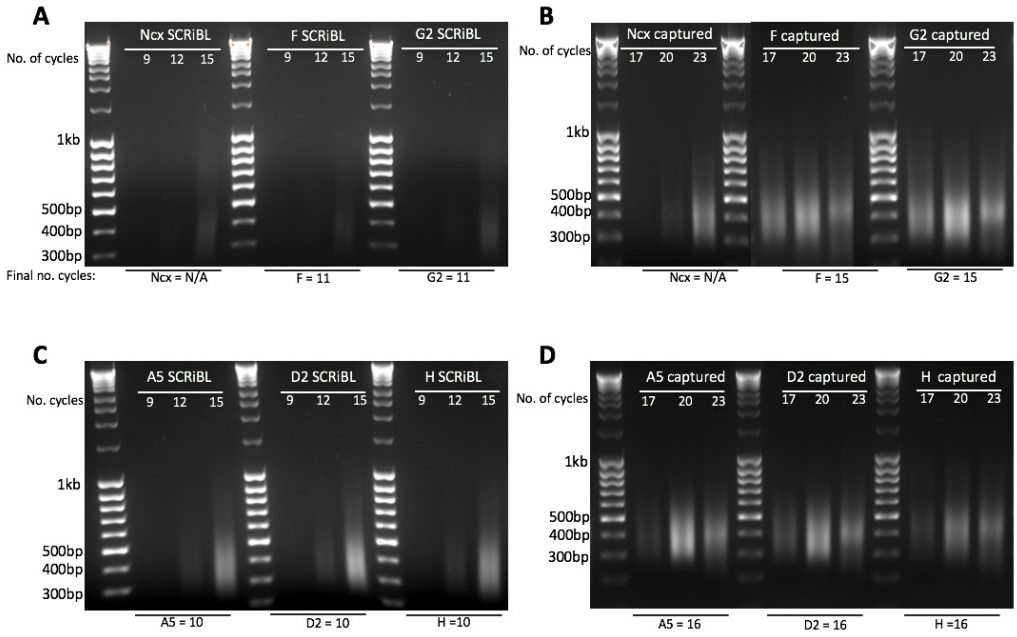

Figure 4.14: Post-capture test amplification of W12 SCRiBL and undigested libraries to determine the number of PCR amplification cycles required to generate sufficient material for NGS and library complexity. The PCR conditions required for the generation of the final SCRiBL Hi-C libraries and captured undigested libraries was determined. The RNA/DNA hybrid 'catch' was amplified by PCR with 9, 12 and 15 (A and C) or 17, 20 and 23 cycles (B and D) amplification cycles at an annealing temperature of 65 °C. Reaction products were run on a 1.5% agarose gel. The number of PCR amplification cycles chosen to generate the final libraries is shown under each gel.

4.3.8 Quality assessment of SCRiBL library NGS using the HiCUP pipeline.

After completion of the 100 bp, paired-end sequencing run of each of the SCRiBL libraries, mapping and quality control checks were conducted using the Hi-C User Pipeline (HiCUP) (Bioinformatics, Babraham Institute). HiCUP is a series of Perl scripts which perform different tasks including the truncation, mapping, filtering and de-duplication of the forward and reverse sequences generated by sequencing (Figure 4.15 A). Initially the raw number of sequence reads were compared (Figure 4.15 B). The number of reads varied from 5.25 x 10^7 (A5 SCRiBL rep I) to 9.81 x 10^7 (G2 SCRiBL repI); this rudimentary check indicated that each sequencing run had been successful. The total number of mapped pairs was determined as a percentage of the total number of reads; additional filtering of the 'mapped pairs'

identified 'valid pairs', and this was also shown as a percentage of the total number of reads (Figure 4.15 C). The percentage of total mapped reads was fairly consistent across all of the libraries (25.7–39.9%), however the percentage of valid reads was much higher in SCRiBL libraries of D2 rep II, G2 repI and G2 repII (24.9, 34.0 and 26.1%, respectively) compared to the 6.0% average of the seven other libraries. Identification of 'valid' pairs meant that 'invalid' pairs present in each library could be further characterised (Figure 4.15 D). Whilst no pairs containing a contiguous sequence of the wrong size were identified in any library, the percentage of read pairs containing religation events (8.1–27.5%) and dangling ends (2.7–52.0%) were the most common cause of invalidity. A final quality control check was to determine the cis:trans ratio of the valid reads (Figure 4.15 E). For all libraries the cis:trans ratio was below 10% (5.4–9.5 %); this indicated that the proportion of valid reads to be used for subsequent analysis were of high quality. Finally, a pilot SCRiBL experiment was run to determine the optimal ratio of HPV16 genome-specifc RNA bait to Hi-C library required in the (hybridisation step of the protocol) to maximise the enrichment of HPV16-specific DNA and subsequent sequencing reads (Figure 4.16).

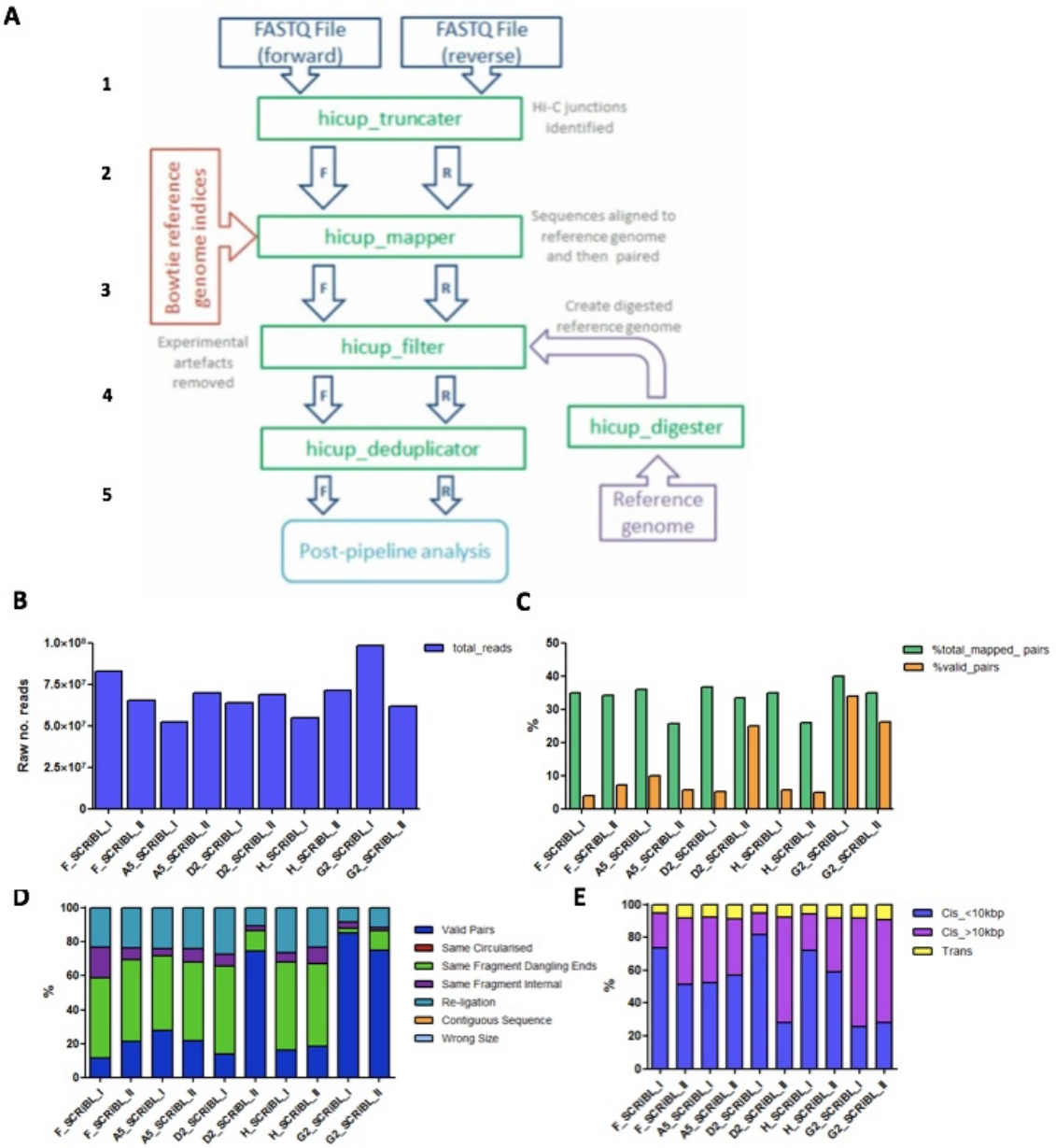

Figure 4.15: Quality assessment of W12 clone SCRiBL libraries processed using the Hi-C user pipeline (HiCUP). A) Schematic overview of the HiCUP pipeline. For each library a FASTQ file is generated for the forward and reverse sequencing strands (1). Hi-C ligation junctions are identified by locating reads comprising the GATCGATC sequence and truncating the reads at this point (2). Truncated reads are then mapped to the GRCh37 human, and W12E HPV16 reference genomes (3). Mapped reads are filtered to remove common artefacts (invalid reads) (4). Putative PCR duplicates are then removed from the valid reads to produce a final selection of reads that are used for subsequent analysis (5). B) Bar-chart showing the number of raw reads obtained from each individual sequencing run. C) Bar-chart showing the percentage of total mapped pairs (green) and valid pairs (orange) compared with the total number of reads obtained for each W12 clone SCRiBL library. D) 100% stacked column chart showing the percentage of valid and invalid pairs (same circularised, dangling ends, same fragment internal, re-ligation, contiguous sequence and wrong size) compared with the total number of mapped pairs for each W12 clone SCRiBL library. E) 100% stacked column chart showing the percentage of close cis (<10 kbp), far cis (>10 kbp) and trans reads in each W12 clone SCRiBL library.

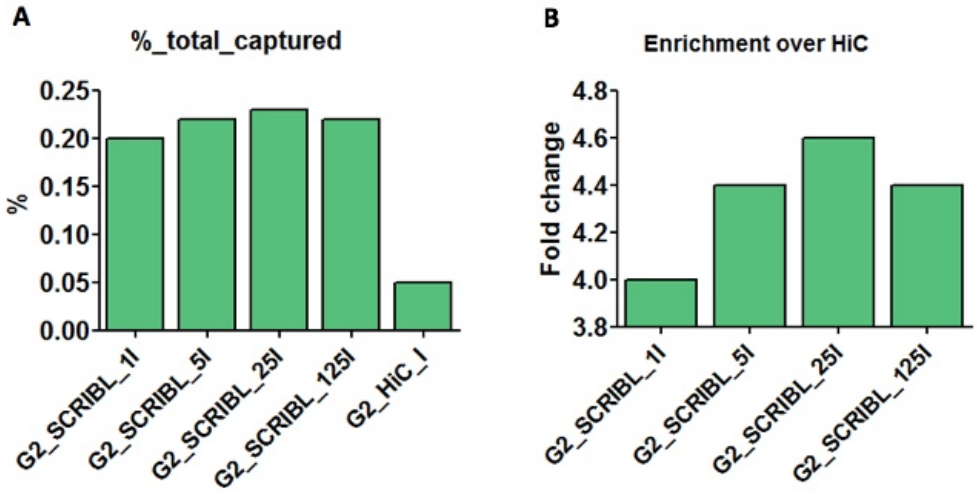

Figure 4.16: Pilot SCRiBL experiment to determine the optimal RNA bait concentration required to enrich the Hi-C libraries for the HPV16 genome A) Bar-chart showing the percentage of HPV16-specific sequencing reads following hybridisation of samples of the W12 clone G2 Hi-C library with 1 ng, 5 ng, 25 ng and 125 ng RNA bait. B) Bar-chart showing the fold change enrichment of HPV-specific reads following hybridisation compared to the W12 clone G2 Hi-C library.

4.4 Discussion

The results presented in this chapter indicate that the SCRiBL Hi-C protocol (Peter Fraser & Stephan Schoenfelder, Babraham Institute) can be successfully adapted to generate Hi-C libraries, and ensure the capture and subsequent enrichment of the HPV16 genome within the W12 integrant clones.

In the first part of this chapter, the methodology and validation techniques used to generate W12 clone Hi-C libraries were described. The resolution of a chromosome conformation capture assay is dependent on the frequency with which the restriction enzyme fragments the genome[271]; this, in conjunction with the small ~8 kb HPV16 genome meant that the choice of appropriate enzyme was limited. As such, the HPV16 genome was assessed for restriction enzyme cut sites; the 4-cutter MboI (GATC) generated 11 fragments within the HPV16 genome and has previously been shown to be used in the successful generation Hi-C libraries[272]. As a result, MboI was chosen for the production of W12 clone Hi-C libraries. The use of a 4-cutter restriction enzyme compared with more commonly used 6-cutter restriction enzymes (HindIII[273, 274, 256], BglII[237] BamHI[275] or EcoRI[238] resulted in the generation of li-

braries with a much higher level of complexity. Library complexity relates to the number of possible interactions within a library produced for conformation capture. Digestion of the human genome with a restriction enzyme with a 4 bp recognition site results in approximately 16 million ~256 bp fragments with the possibility of up to 100 trillion unique pairwise interactions. This is a 100-fold increase in the number of possible pairwise interactions between the ~4 kb fragments generated by a 6-cutter restriction enzyme[245, 276]. Given the enormity of theoretical interactions within a library generated using a 4-cutter enzyme, it is impossible to capture all interactions by sequencing alone, as such, the capture and enrichment of regions of interest is required. Capturing the HPV16 genome fragments markedly enriches their interacting fragments. Also, as a consequence, the overall library complexity is reduced compared with the corresponding pre-capture Hi-C library, and the identification of significant virus-host interactions at the restriction fragment level is increased[274]. Additionally, it was important to take into account increased library complexity when producing final Hi-C libraries by PCR as over-amplification resulted in the production of a greater percentage of invalid sequencing reads that were discarded before downstream analysis.

Quality control analyses were based upon detecting short-range interactions in Hi-C libraries following MboI restriction enzyme digest and in-nucleus ligation of DNA fragments in close spatial proximity. Following the principles of 3C conformation capture technology, primers were designed close to, and towards, the ends of MboI restriction fragments of the known genomic locus RPL13A. The RPL13A genomic locus was used as a control region as it had previously been shown to be abundant and stably expressed in the W12 integrant clones[183]; alternative genomic loci such as GAPDH and YHWAZ could also have been considered as these have also been used as housekeeping genes to compare HPV16 transcripts across the panel of W12 clones[183, 203]. In addition to the single control locus, optimal 3C PCR conditions were determined using W12 clone F Hi-C material; clone F contains just one copy of the integrated HPV16 genome and was considered the simplest system for testing. The same primer pair and PCR conditions were carried forward for equivalent quality control assays and were equally effective for each W12 integrant clone

library. During the optimisation of the forward-forward primer 3C PCR assay, products other than the specific product were generated. These represent additional 3C interaction products whereby the two specified DNA fragments are not directly ligated together but have extra DNA fragments ligated between them. DNA fragments generated by a 4-cutter restriction enzyme (e.g. MboI) have an average length of ~256 bp; as such, a ladder of products spaced by 256 bp is predicted and is observed following the PCR of ligated fragments.

After restriction enzyme digest of DNA, fragments in close proximity were ligated together within preserved nuclei (in-nucleus ligation); the efficiency of which was determined by PCR digest assay. Until recently, re-ligation of digested fragments has occurred in dilute solution to prevent non-specific ligation due to chance inter-molecular collisions[277, 246]. However, more recent investigations indicate that carrying out the ligation step within preserved nuclei lead to superior results such as reducing technical noise as represented by the decrease in trans-chromosomal interactions, as well as improving reproducibility[257]. To determine the efficiency of the in-nucleus ligation step for each of the W12 clone Hi-C libraries, the specific B:G product of blunt-end ligation was digested with the restriction enzyme ClaI. The efficiency varied from 71.1–94.2% across the panel of W12 libraries, however this variation does not indicate superiority, and will not have inferred a technical advantage of any library over the other. It is possible to observe a maximum of four interactions from one individual fragment per cell (both ends of the fragment per allele) and as previously mentioned there are 16 million potential interactions per fragment in a library generated using a 4-cutter restriction enzyme. Therefore, assuming 100% ligation efficiency, it would be possible to detect all ligation interactions starting with just four million cells (4^4 = 16). For the generation of each W12 Hi-C library, the ligation step was carried out with a starting material of ~fifteen million cells; despite a maximum loss of efficiency of 28.9% all possible interactions are still recovered ((15 × 0.711)^4>16). In reality, the number of real interactions is much fewer than the theoretical maximum due to such factors as distance decay whereby the probability of an interaction occurring decreases the further the linear distance between the two genomic fragments[276].

Subsequent sonication of ligated DNA products produced fragments of approximately 400 bp in length. Size selection to remove fragments with much larger or smaller lengths resulted in libraries with a more uniform size distribution (300–500 bp); this minimised any bias that can arise in subsequent amplification steps of the protocol[271]. The final stage of the generation of Hi-C libraries was to determine the appropriate number of amplification cycles required; a balance between producing a sufficient concentration of DNA for the capture step and subsequent sequencing and not introducing additional library complexity was struck. Testing a range of amplification cycles allowed for each library to be assessed individually and the number of amplification cycles chosen was one fewer than that at which a smear was first visible.

The second part of this chapter describes how HPV16-specific RNA baits were derived to enrich the W12 clone Hi-C libraries for the virus genome (generation of SCRiBL Hi-C libraries). Capturing DNA fragments comprised of hybrid HPV16 and host DNA facilitated the identification of viral-host interactions, the results of which are far-reaching and include insights into the effect of HPV16 integration on host gene expression, or vice-versa, via long-range interactions from gene enhancer or promoter sequences in 3D.

The capture of Hi-C libraries according to predetermined regions of interest, such as cancer risk loci and gene promoters, has been conducted in an increasing number of studies[273, 256, 255, 274, 278]; however enriching a Hi-C library for an integrated viral genome is novel. The enrichment of the HPV16 genome represents a uniquely small capture region (~8 kb); previous capture Hi-C (cHi-C) experiments have captured much larger whole chromosomal regions ranging from 350–750 kb in length[273]. As a result of the small capture region and increased library complexity generated by using MboI, the ratio of Hi-C library to biotinylated RNA bait could not simply be replicated from published studies. Consequently, the optimal DNA:RNA bait ratio for HPV16 genome enrichment was determined using the W12 clone G2 Hi-C library hybridised to biotinylated RNA baits at a range of different concentrations. It was shown that a quantity of 25 ng of biotinylated-RNA baits per 300 ng Hi-C library produced the most favourable enrichment of HPV16-containing reads — 4.6

fold change — compared with uncaptured Hi-C library (Figure 4.16).

gBlock® Gene Fragments from IDT were designed to generate biotinylated RNA baits for the capture of the HPV16 genome within W12 clone Hi-C libraries. RNA bait fragments were designed to have similar GC content (25–65%) and were of uniform size (~120 bp); this prevented hybridisation and amplification bias, respectively, in subsequent reaction stages of the protocol. In order to produce RNA baits for the hybridisation reaction it was necessary to *in vitro* transcribe the isolated DNA gBlock® fragments; this process required the ligation of a T7 promoter adapter to each DNA fragment. To ensure controlled and specific ligation, T7 promoter adapters were designed with a compatible, cohesive end (BamHI overhang) to that of the 5'-end of DNA fragment (BglII overhang). Ligation of the two DNA molecules generated a new restriction site (5'-GGATCT-3') that could not be cleaved by either enzyme used in the digestion reaction.

Following the capture of HPV16:host hybrid sequences from W12 Hi-C libraries, the PCR conditions for the generation of the final SCRiBL libraries were determined using custom PE PCR primers 1.0.33 and 2.0.33. Test PCRs indicated that the post-capture DNA concentration of the HPV16-negative NCx library was lower compared with the W12 integrant clones regardless of equal starting concentrations, illustrating that the biotinylated-RNA baits successfully enriched for DNA fragments containing the HPV16 genome. Additionally, the NCx cell line was used as a negative control throughout the process of making the SCRiBL libraries; however, due to the costs involved this library was not sequenced. The final W12 SCRiBL libraries were generated using Illumina sequencing adapters, required for the binding of DNA libraries to a flow cell for next generation sequencing (NGS) on the Illumina HiSeq 2500 machine. Primer pairs consisted of one TruSeq Indexed adapter (reverse compliment) and the TruSeq universal adapter (both Illumina); the resulting DNA fragments ligated between the two sequencing adapters are termed 'di-tags'. Consideration was given to the TruSeq Indexed adapter used to generate each library to allow for the multiplexing of three SCRiBL libraries per sequencing lane[279]. Libraries were sequenced to produce 100 bp paired-end reads. Paired-end sequencing enables both ends of the DNA fragments to be sequenced — as the distance between each paired

read is known, alignment algorithms can map the reads to a reference genome more precisely[280].

Resulting output FASTQ sequences were mapped to the human reference genome (GRCh37/hg19) containing the HPV16 genome as an extra chromosome, and were also filtered to remove experimental artefacts using HiCUP by Jack Monahan (EBI-EMBL). Assessing the number of raw reads obtained for each library indicated that each sequencing run had been successful given a theoretical maximum of 150–180 million read-pairs per lane as well as three libraries multiplexed per lane (Illumina 2013). Slight differences between the numbers of raw reads per SCRiBL library may be as a result of variable binding efficiencies to the sequencing flow cell. Following read mapping, the percentage of valid and invalid pairs was identified and the libraries filtered accordingly; even a small number of invalid di-tags could lead to incorrect conclusions being drawn concerning genomic structure. Invalid reads include sequences representing Hi-C artefacts and other uninformative di-tags and are excluded from downstream analyses; invalid reads comprised mainly of artefacts produced when a sequenced read pair maps to a single restriction fragment ('same circularised', 'same fragment dangling ends', 'same fragment internal'), di-tags of the 'wrong size' identified by the mapped reads positioned too far away from the putative restriction enzyme cut-site than allowed by the experimental size-selection step, and 'contiguous sequences' generated by the re-ligation or incomplete digestion of fragments (Figure 4.17)[258].

HiCUP analyses of individual W12 SCRiBL libraries exposed variations in the percentage of valid reads; the percentage of valid reads for the D2_SCRiBL_II, G2_SCRiBL_I and G2_SCRiBL_II libraries were much higher (28.3% mean) than the remaining seven (6.0% mean). Interestingly, the separation of W12 D2 and G2 from F, A5 and H coincides with integrated viral copy number; an increased percentage of valid reads is seen in the W12 clones containing three or four copies of the HPV16 genome whereas the percentage of valid reads remains stable among all W12 clones tested with just one integrated copy of the viral genome. It is probable that the increased numbers of di-tags containing HPV16 in W12 G2 and D2 SCRiBL

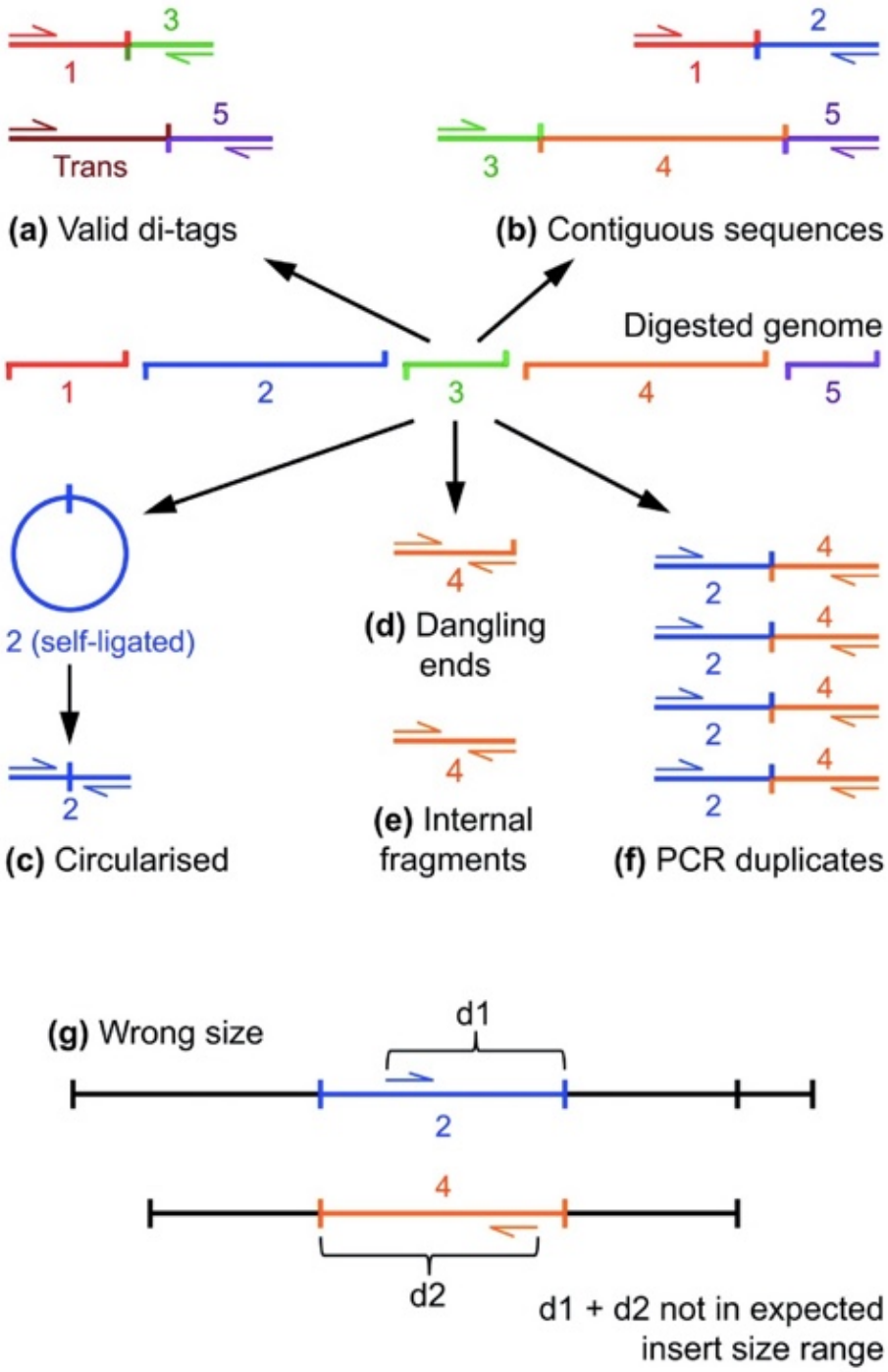

Figure 4.17: Overview of experimental artefacts generated by the Hi-C experimental protocol adapted from HiCUP: pipeline for mapping and processing Hi-C data, Wingett S. *et al.*, 2015[258]. Schematic shows the genome digested into 5 restriction fragments. These fragments may subsequently ligate to each other, or fragments derived from another chromosome, forming valid cis or trans di-tags respectively (a). In contrast, re-ligation or incomplete digestion leads to the generation of invalid contiguous sequences (b). Another common artefact occurs when the sequenced read-pair maps to a single restriction fragment (c), (d) & (e). Further, PCR may result in a fragment being copied multiple times (f). Di-tags are also rejected when the mapped reads are positioned too far away from the putative restriction enzyme cut-site than allowed by the experimental size-selection step (g).

libraries map to more unique regions of the host genome which, as a percentage of total reads, increase validity compared to the smaller number of HPV16 containing di-tags generated in F, A5 and H. Additionally, the discrepancy of valid reads between biological replicates in W12 clone D2 highlights the necessity for at least two replicates to produce robust data. Downstream analysis of the W12 SCRiBL Hi-C libraries was conducted using data from valid read-pairs only; to ensure that libraries could be directly compared, stringent normalisation of read numbers per library was applied.

To address the second experiment of identifying the 5' and 3' virus-host junctions in each of the W12 integration clones (with viral genome copy number less than four) a HPV16 capture system, additional to that for the generation of SCRiBL Hi-C libraries, was devised.

In order to identify the virus-host breakpoint, whole-genomic W12 DNA libraries were captured with HPV16-specific biotinylated RNA baits that evenly covered the whole virus genome; mapping the DNA sequence of viral-cellular junctions to the human reference genome (GRCh37/hg19) containing the HPV16 genome as an extra chromosome directly indicates the HPV integration site. Multiple studies have conducted similar studies to determine the viral integration sites using tissue samples, at different stages of carcinogenesis, for a range of HPV genotypes[281, 282, 283, 101]. In each study, extracted DNA was prepared into a sequencing library and enriched for the HPV genome via the hybridisation of HPV genome-specific probes; however, the design and manufacture of HPV probes used was carried out by external companies including MyGenostics Inc.[281, 282, 283] or Roche NimbleGen Inc.[101]. In this investigation the HPV probes were designed in-house; the generation of HPV16-specific biotinylated RNA baits for use in our capture system was based upon isolating and amplifying the W12E DNA genome from the pSP64 plasmid using primer pairs evenly spaced across the HPV16 genome. The forward primer of each of the primer pair was designed to incorporate the T7 promoter adapter sequence that enabled the *in vitro* transcription of DNA to RNA. Gel electrophoresis analysis of the HPV16 RNA showed that, despite a pronounced band representing full length RNA (2 kb),

numerous other RNA products were generated as represented by 300 bp–2 kb smear; as such, the baits could only be used in qualitative experiments. The HPV16-specific RNA was suitable for use in the capture-seq experiment but consideration should be given as to whether they are appropriate for use in future, alternative experiments.

The methodology presented in this chapter lays the foundation for furthering our understanding of HPV16 virus integration and associated selection of cells in the field of papillomavirus biology. Viral genome integration represents a crucial step in tumorigenesis[27] and elucidation of integration events is an essential requirement for understanding HPV-induced carcinogenesis. Coupled with SCRiBL Hi-C analyses, further levels of virus and host genome regulation can be identified. Changes to gene expression as a result of virus integration and long-range interactions may begin to explain the mechanisms behind the growth advantage of particular cells present across the cells of a polyclonal LSIL, and will be investigated in the next chapter.

Chapter 5

Integrated HPV16 genomes interact with host chromosomes three-dimensionally (3D) modulating nuclear architecture and host gene expression.

5.1 Introduction

The successful generation of both capture-seq and SCRiBL-Hi-C DNA libraries for each of the W12 clones meant that the process of HPV16 integration and its effects could be explored at both a genomic and epigenetic level.

As previously stated, HRHPV integration is seen in ~85% of cervical SCCs and is viewed as a key driver of squamous carcinogenesis[2]. Previous studies of clinical samples and cell lines have used PCR-based approaches to map integration sites; namely restriction site PCR (RS-PCR), rapid amplification of cDNA ends PCR (RACE-PCR), including amplification of papillomavirus transcripts assay (APOT). However, these virus-host breakpoint identification methods are limited by inherent technical bias and, as such, the validity of results obtained by these techniques is questionable. A number of different studies have now employed next generation RNA and/or DNA sequencing to more accurately determine the presence and integration sites of virus genomes in HPV-associated malignancies across the human genome[97, 98, 87, 176, 101, 100]. Studies have demonstrated that HPV commonly integrates into genomic 'hotspots', a number of which are associated with common fragile sites (CFS)[95, 284, 96] and, in addition, have shown that HPV can integrate directly into a gene — into both introns and exons — and can lead to varying changes in host gene expression level[87, 97, 98]. Several mechanisms for host gene disruption have been proposed, including: deletion or intragenic disruption of potential tumour suppressor genes and upstream virus promoter insertion or amplification of oncogenes[97], amplification of the local region resulting in copy number variation (CNV)[176], or rearrangement and translocation of the integration locus elsewhere in the host genome[97, 100].

In addition, other studies have suggested that higher level transcriptional control may be at play — there is a growing appreciation for how the three-dimensional (3D) organisation of the genome contributes to the control of gene expression. HPV integration into the flanking regions of genes, sometimes as far away as 500 kb, has been found associated with large increases in gene transcription, specifically the

MYC proto-oncogene which is encoded for at the 8q24.21 locus. Haplotype resolved RNA-seq data of the HeLa cell line showed that *MYC* is highly overexpressed from the HPV18 integrated allele, which is also associated with higher levels of transcriptionally active chromatin marks, transcription factors and RNAPII[285]. Analysis of 'chromatin interaction analysis with paired-end tag' (ChIA-PET) sequencing data demonstrated a long-range *cis* interaction between the integrated HPV18 promoter/enhancer and the MYC gene[285]. Indeed, the HPV18-*MYC* 3D interaction has been verified by chromosome conformation capture (Hi-C) analyses; in HeLa cells, integration of HPV18 modifies normal chromatin loops present at, and around, the 8q24.21 locus such that virus and *MYC* sequences are in direct contact[251]. Moreover, CRISPR/Cas9 knockout of the integrated HPV18 fragment resulted in a 30% decrease in MYC gene expression, providing experimental evidence supporting the hypothesis that integrated HPV influences host gene expression via long distance chromatin interactions[286].

To date, next generation sequencing studies have primarily focused on the analysis of advanced cervical SCCs that consist of clonal HPV integrant cells derived under selective pressure. This study aimed to investigate naturally occurring HPV16 integration events that occur in premalignant cells isolated under non-competitive conditions using the W12 clones; these data will elucidate whether characteristic virus and host changes as a result of integration, such as long-range 3D interactions and resultant host gene expression changes, are a typical feature of all HPV integrants or whether they are restricted to cells with a selective growth advantage.

5.2 Results

5.2.1 Integrated HPV16 genomes interact in 3D with host chromosomes

Regions of the integrated HPV16 genome that interact in three dimensions with the human genome were determined using GOTHiC software and visualised using the Circos tool. Bioinformatic analysis including the mapping and aligning of sequence reads was carried out in collaboration with Jack Monahan (EBI-EMBL). In each panel (Figure 5.1 A-E), a single line within the circle represents a virus-host read indicating a 3D interaction between the HPV16 genome and the host, and is coloured according to the virus gene from which the read originates. The frequency of *cis* interactions between the virus and the host is known to be greatest for host sequences at the site of integration and to decrease with distance[276]. As such, analysis of the SCRiBL Hi-C data simultaneously enabled the HPV16 integration locus in each of the clones to be identified. For each clone a representative image of the biological replicate libraries is shown.

In W12 clone G2 repII reads from across the virus genome were shown to interact with the host, and came from all RNA baits designed around MboI restriction sites. Although the distribution of the reads was fairly uniform, the greatest percentage reads interacting with the host genome came from virus gene E7 (Figure 5.1 A left panel). The integrated HPV16 genome interacts exclusively with chromosome 5; when analysing the single chromosome view, there is a split between the bulk of reads from the virus — indicating the integration site — and a subset of reads that mapped to a separate region of the host. Given the large scale of the plot, this indicates a long-range 3D interaction between the integrated virus in G2 and the host (Figure 5.1 A right panel).

The number of virus-host reads that were captured and mapped for W12 clones D2 (84.3%), F (11.2%), A5 (12.4%) and H (10.3%) were lower than for clone G2 arbitrarily set at 100%. As a result, there are reduced numbers of virus-host reads in the circos plot analysis. In clone D2 the HPV16 genome also integrated into

chromosome 5 of the host; the virus-host reads converge on a single point at the host chromosome (Figure 5.1 B). The greatest percentage of virus-host reads in D2 come from the restriction fragment that covers the 5' half of the virus gene L1. For clone H, the captured reads indicate that interactions between the virus and the host occur from the early genes E6 and E7, E2 and L1, with the majority of reads coming from the E2 portion of the virus genome (Figure 5.1 C left panel). Here, HPV16 integrates into chromosome 4 and results in a large deletion of the host (~170 kbp). The deletion is illustrated by the separation of the virus-host reads in the chromosome-only Circos view (Figure 5.1 C right panel). For each W12 clone there is consistently an absence of reads originating from the virus genome between genes E1 and E2; this is a result of the RNA bait design, which was based upon MboI restriction sites in the HPV genome, rather than a true biological finding (see 5.3). Interestingly, we found that W12 clones F and A5 had the same integration site, with virus-host reads converging to the same region of chromosome 4 (Figure 5.1 D and E). Although virus-host reads were identified across all possible regions of the virus genome, in both clones the greatest percentage came from the 3' section of E2. In all the W12 integrant clones tested, the HPV16 genome was shown to interact with regions of host chromosomes in *cis*; there were no examples of the virus interacting in *trans*.

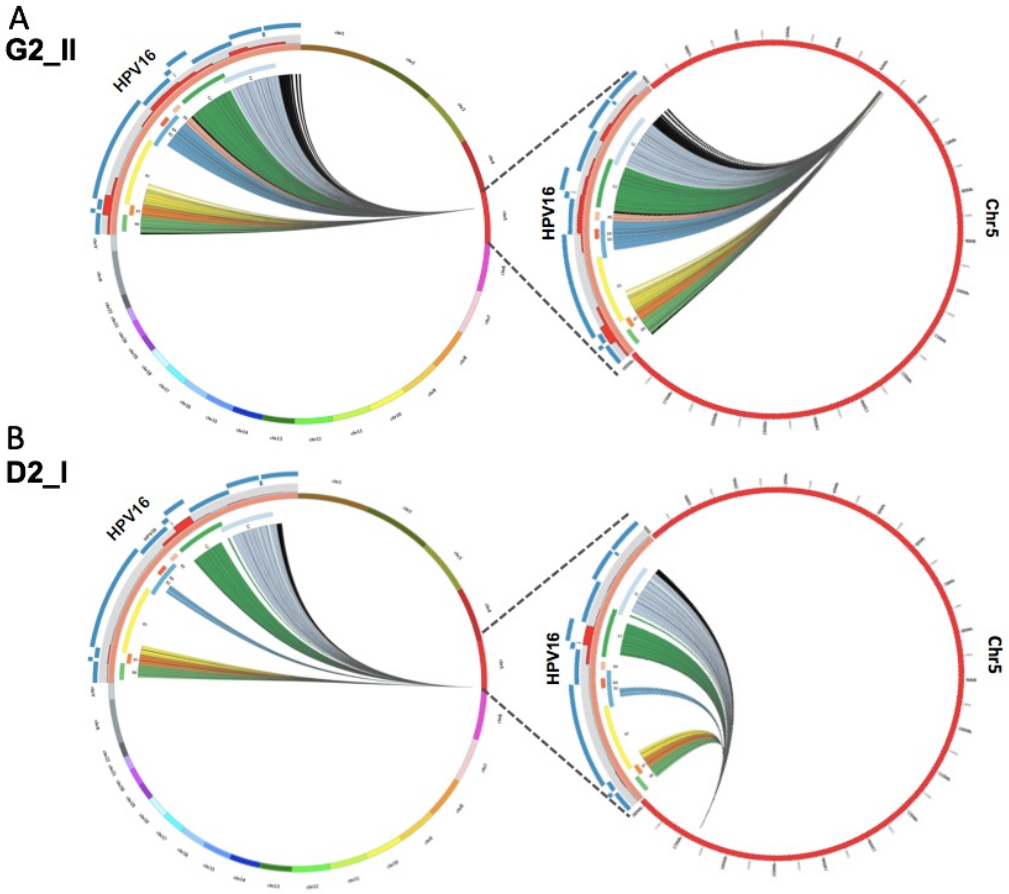

Figure 5.1: Circos plots indicating 3D interactions between the integrated HPV16 genome and host chromosomes across a panel of five W12 integrant clones. Each line within each circle represents a virus-host read indicating a 3D interaction between a region of the HPV16 genome and the host. Reads are coloured to match the individual genes of the HPV16 genome: E6 = green, E7 = orange, E1 = yellow, E2 = blue, E4 = red, E5 = pink, L1 = dark green, L2 = light blue and the non-coding region = black. The left column contains circos plots that comprise the HPV16 genome (orange) and the entire host genome, individual chromosomes identified by different colours. The right column contains circos plots that comprise the HPV16 genome (orange) and the single host chromosome where 3D interactions occur. The percentage of reads coming from different regions of the virus is indicated by the histogram on the outside of the HPV16 genome, which is split into 500 bp windows coloured red. The HPV16 RNA bait fragments using in the capture-Hi-C experiment are indicated in blue on the outside of the circos plot. In each Circos plot the absence of reads originating from HPV16 E1/E2 is due to the RNA bait design, rather than a true biological finding. A) W12 G2 replicate II, B) W12 D2 replicate I. Plots were generated using the Gothic program and are not to scale.

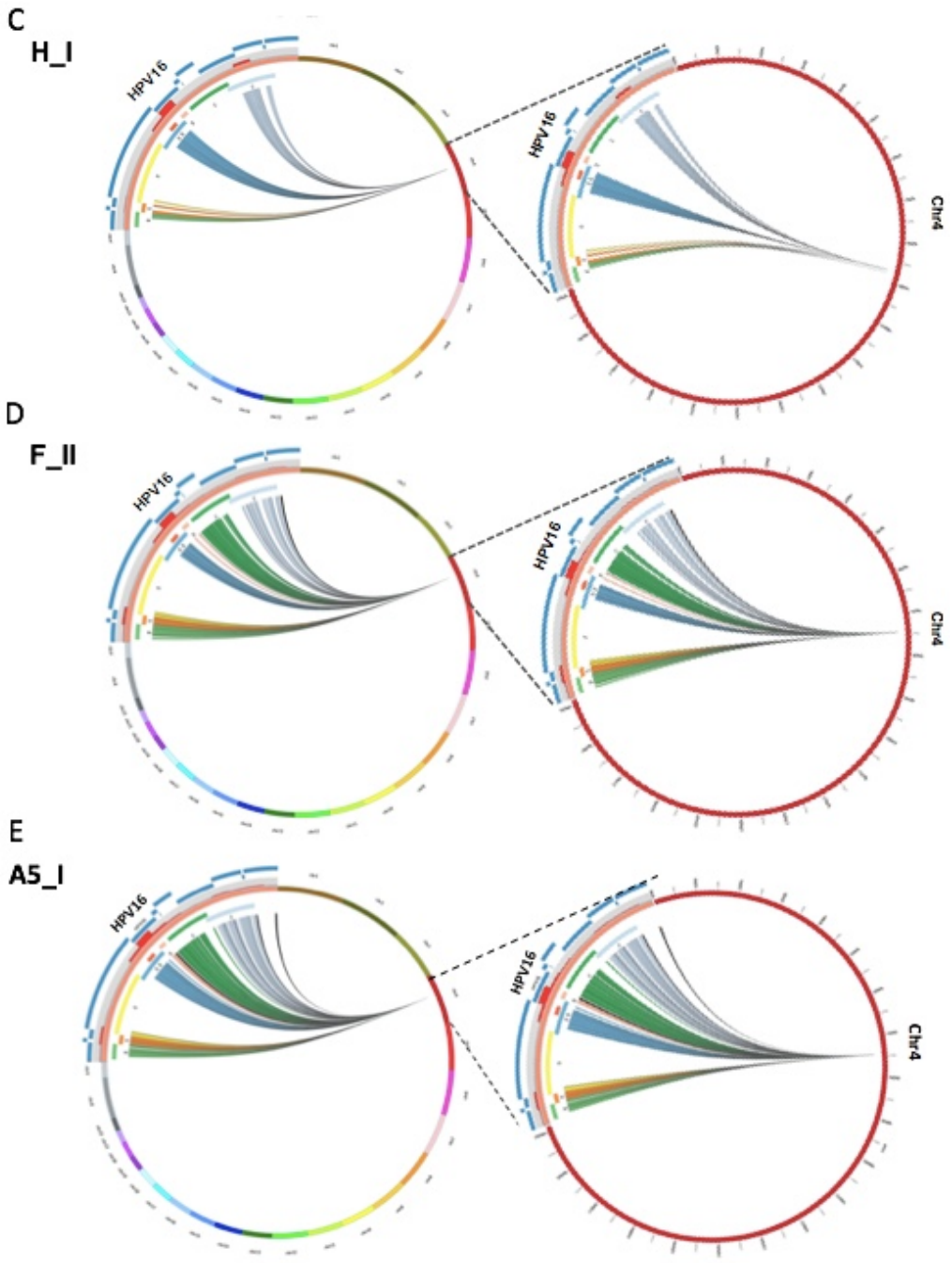

Figure 5.1: Continued. Circos plots indicating 3D interactions between the integrated HPV16 genome and host chromosomes across a panel of five W12 integrant clones. C) W12 H replicate I, D) W12 F replicate II and E) W12 A5 replicate I. Plots were generated using the Gothic program and are not to scale.

5.2.2 W12 integrant clone 5' and 3' virus-host breakpoints identification at nucleotide resolution.

Analysis of the capture-seq experiment determined that for each clone, peaks of virus-host reads mapped to two distinct sites of the host genome (data not shown; these correlated with the results described in section 5.2.1). The identification of two peaks, regardless of HPV16 genome copy number, demonstrated that there is only one 5' and one 3' virus-host breakpoint in each of the W12 clones.

In order to validate the findings of the capture-seq experiment, primer pairs consisting of one primer complementary to the virus genome and one to the host genome flanking the virus-host breakpoint loci were designed in order to amplify the virus-host chimaeric DNA for subsequent Sanger sequencing. For each W12 clone the 5' and 3' breakpoints were amplified and the size of the PCR products corresponded to the original primer design (Figure 5.2 A-E left panel). Given successful PCR amplification of the virus-host breakpoints, RT-qPCR primers were designed using the same chimaeric DNA sequence used for PCR primer design. Each clone specific 5' and 3' breakpoint primer pair was tested on gDNA from the specific clone, and gDNA from W12par1 (episomal) and the HPV-negative cell line NCx/6 as negative controls to test the specificity of each breakpoint. For W12 clones G2, D2 and H, both the 5' and 3' fusion transcripts were unique to the individual clones (Figure 5.2 A-C middle and right panels). Interestingly, however, while primer pairs specific to breakpoints in W12 clones F and A5 confirmed the sequencing data, these primers also produced a positive result, although relatively small amounts of product, in the parental W12 cell line (Figure 5.2 D and E middle and right panels).

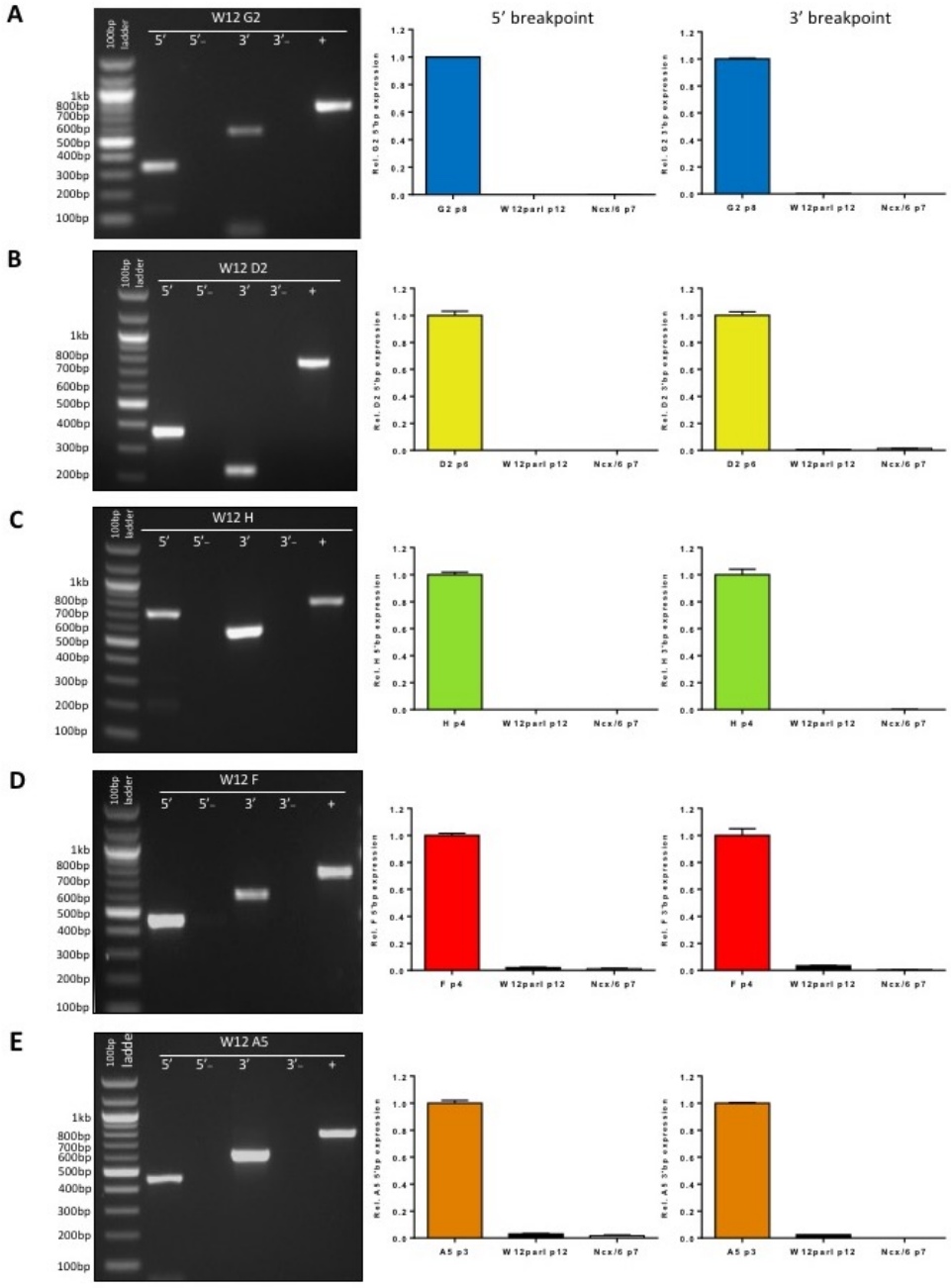

Figure 5.2: Identification and specificity of the virus-host breakpoints across a panel of five W12 integrant clones. Gels in the left column show the PCR amplification products of W12 integrant clone-specific 5' and 3' virus-host breakpoints (lane 1 and 3, respectively). PCR reactions were carried out at a range of temperatures (see below) for 50 cycles with a 1.5 mM concentration of $MgCl_2$. (A) W12 clone G2; 5' = 60 °C, 3' = 52.1 °C, (B) W12 clone D2; 5' = 55.4 °C, 3' = 55.4 °C, (C) W12 clone H; 5' = 56 °C, 3' = 60 °C, (D) W12 clone F; 5' = 59.3 °C, 3' = 55.1 °C, (E) W12 clone A5; 5' = 59.3 °C, 3' = 55.1 °C. Control reactions included; negative controls which contained the PCR mix without DNA (lane 2 and 4) and a positive control mix that contained E6-E7 primers (60 °C) (lane 5). The central and right columns indicate RT-qPCR analysis of the 5' and 3' virus-host breakpoints, respectively. Virus-host junctions in the specific integrant clones were compared to the episomal (W12par1 p12) cell line and an HPV-negative cell line (NCx/6).

To pinpoint the exact break in both the virus and host genome at the site of HPV16 integration in each of the W12 integrant clones, PCR amplified DNA was sent for Sanger sequencing (Biochemistry, University of Cambridge). The resultant 5' and 3' breakpoint sequences for each clone were aligned using Nucleotide Blast against both the W12E and the Hg19 human genome sequences to find the breakpoints in the virus and the host at nucleotide resolution, respectively.

In W12 clone G2 the HPV16 genome is linearised by breaking the virus genome in the E2 ORF [5': 2,940 and 3': 2,768] resulting in a loss of 173 bp and placing the majority of the E2 gene (913 bp) upstream of the virus early promoter. Additionally, the three copies of the HPV16 genome integrate into chromosome 5 — chromosome band 5q11.2 — in an intergenic region of the host genome [5': 52,681,626 and 3': 52,655,805] (Figure 5.3 A).

The four HPV16 genomes in clone D2 are also integrated into chromosome 5, in the region 5q34 [5': 167,112,984 and 3': 167,141,612]. Linearisation of the HPV16 genome occurs via breakage in the L2 [5': 4,361] and E2 [3': 3,272] ORFs, resulting in a 1,089 bp deletion of the virus genome. Furthermore, the orientation of the virus promoter opposes that of the transcription of the host gene TENM2 into which the HPV16 genome has integrated (Figure 5.3 B).

In W12 clone H only a single copy of the HPV16 genome has integrated into chromosome 4 — chromosome band 4q21.3 — within host gene MAPK10; moreover transcription from the virus early promoter occurs in the same direction as transcription of the host gene. Virus integration results in a large deletion of the host genome, with the 5' and 3' host breakpoints separated by more than 170 kbp [5': 86,983,196 and 3': 87,153,458]. In addition, in comparison with the other W12 integrant clones included in this study, a large proportion of the virus genome is also deleted; the HPV16 genome is broken in the L1 [5': 5,883] and E2 [3': 3,751] ORFs, meaning that the length of the integrated virus is 5,773 bp (Figure 5.3 C).

The Sanger sequencing data confirmed our initial finding that clone F and A5 have the same site of HPV16 integration. In both clones, a single copy of the virus integrates into chromosome 4 — chromosome band 4q13.3 — within the host gene RASSF6 [5': 74,549,681 and 3': 74,480,662]. Sanger sequencing also revealed an 18

bp truncation and rearrangement of the HPV16 genome at the 3' breakpoint whereby a region of 54 bp (3,637 to 3,690) is inverted. The linearisation of the HPV16 genome and resultant rearrangements occur within the E2 ORF [5': 3,677 and 3': 3,637] and as a result places a portion of the E2 ORF upstream of the virus early promoter (Figure 5.3 D).

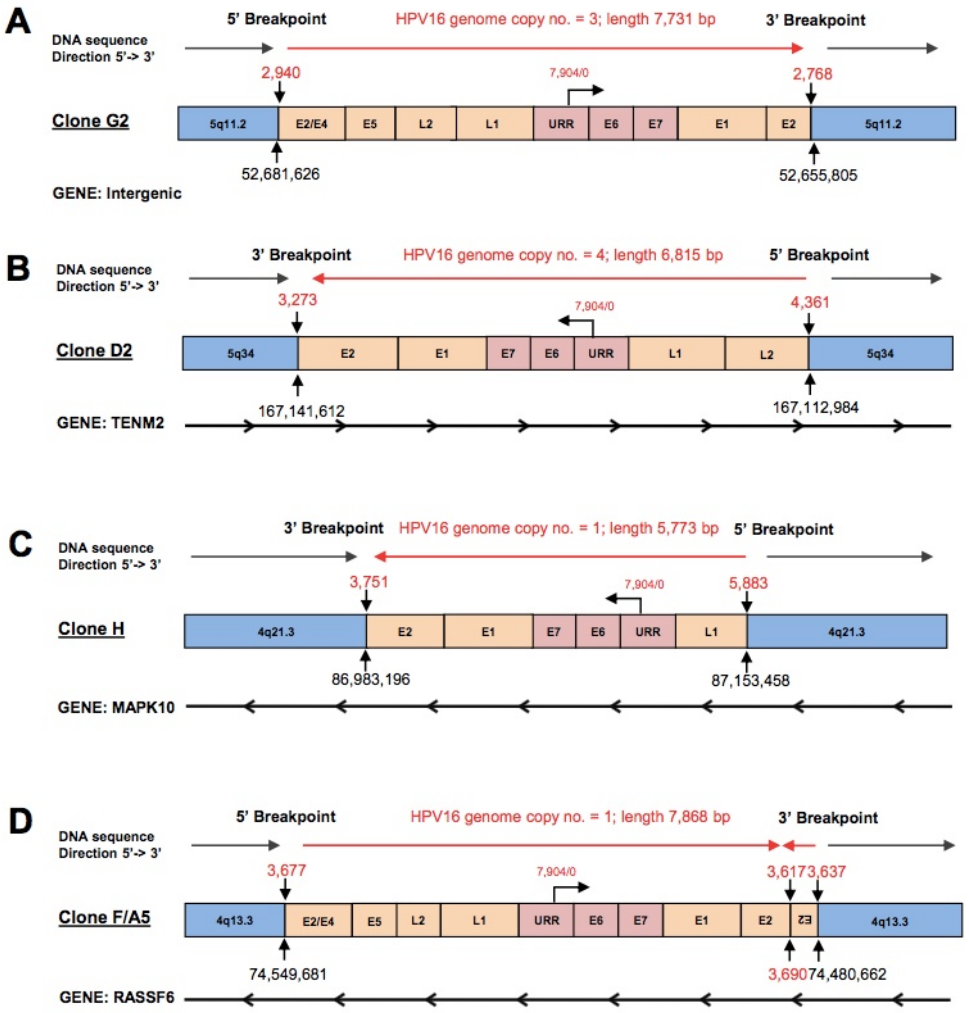

Figure 5.3: Schematics showing host-virus junctions at the different integration sites. In all schematics, host chromosomal DNA is shown in blue and the orientation indicated by the grey arrow above (5' to 3'). Integrated HPV16 DNA is shown in orange, with the viral oncogenes and LCR highlighted in red, and the direction of transcription from the viral early promoter shown by an arrow from the URR. The location of the viral breakpoint in base pairs is given above the junction, whereas the cellular DNA breakpoint in base pairs is given below the junction. The genome copy number and length of the integrated HPV16 genome is indicated in red above the schematic. When HPV16 has integrated into a host gene, the orientation is shown beneath the schematic. A) W12 clone G2, B) W12 clone D2, C) W12 clone H and D) W12 clones F/A5. (Virus genome copy number taken from Scarpini *et al.*, 2014.)

Identification of the exact virus-host breakpoints enabled the chimaeric sequence to be aligned against the host (Hg19) and HPV16 (W12E) genome sequences across this region. Regions of microhomology between the two sequences were highlighted; often, nucleotides of both genomes directly adjacent to the breakpoint were homologous (Figure 5.4 A-D). In clone G2, at both the 5' and 3' breakpoints there was a region of five homologous nucleotides, a feature that was also seen at the 5' virus-host breakpoint in clone D2 (Figure 5.4 A and B). In W12 clone H, a three-nucleotide sequence was found at the 5' and 3 breakpoints (Figure 5.4 C). In contrast, at the 5' breakpoint in F/A5 there were no adjacent homologous nucleotides; however nucleotides at the 3' breakpoint exhibited greater homology. A region of eight nucleotides, separated into two groups of four by a single nucleotide was homologous between the HPV16 and Hg19 DNA sequences (Figure 5.4 D). When the microhomology of 10 nt either side of the breakpoint was compared to that generated from 10,000 random shuffles of each sequence extended to 1,000 nt, the regions of microhomology that were significant included the G2 3' breakpoint, H 5' breakpoint and F 5' breakpoint.

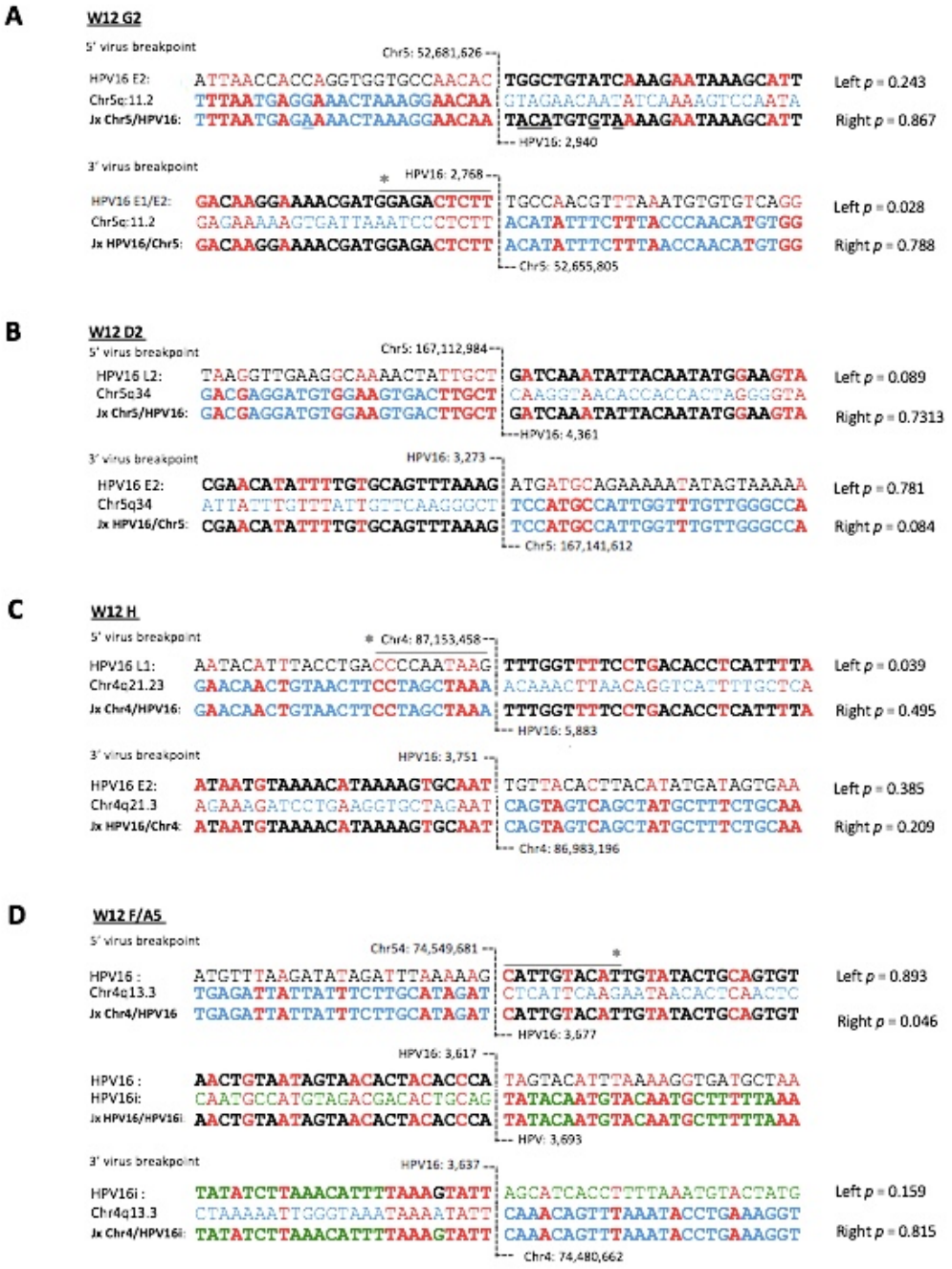

Figure 5.4: Regions of HPV16 and host sequence homology at the integration site. Figures show comparisons between the virus-host sequences obtained by Sanger sequencing and the normal host and HPV16 genomic sequences, 25 nucleotides either side of the breakpoint (indicated by a central dotted line). HPV16 DNA sequence = black, inverted HPV16 (HPV16i) DNA sequence = green, human DNA sequence = blue, homologous nucleotides = red, underlined nucleotides = do not match virus or host sequence, DNA fusion junctions denoted as 'Jx'. Significant levels of microhomology between host and HPV16 sequences were calculated by comparing the homology seen at the 10 nt directly either side of the breakpoint compared to 1000 nt of extended sequence which was shuffled 10,000 times. $^*p < 0.05$.

5.2.3 HPV16 integrates into regions of open and active host chromatin.

For each clone the site of HPV16 integration was mapped onto the host (Hg19) genome and aligned with ChIP-seq marks from the normal human epidermal keratinocyte (NHEK; ENCODE datasets) cell line across a 5 Mb window from the integration locus (Figure 5.5 A-D). Marks of actively transcribed chromatin included H3K27ac, H3K4me1, H3K4me2 and H3K4me3, whereas transcriptional repression was denoted by the H2K27me3 mark. Areas of open chromatin were illustrated by regions of DNaseI hypersensitivity and, additionally, binding sites of the boundary element CTCF were also aligned across the integration locus.

The HPV16 genome in G2 integrates into an intergenic region that is abundant in the activating histone modifications, particularly enhancer marks H3K27ac and H3K4me1. Additionally concentrated DNaseI hypersensitivity marks indicate an open chromatin structure at the site of HPV16 integration. Moreover, the lack of repressive marks within this genomic region is striking (Figure 5.5 A). Similar observations were made when evaluating the integration sites of clone D2 and F/A5 (Figure 5.5 B and D). Additionally, in clone D2, HPV16 integrates into the very large host gene TENM2; the presence of a number of repressive marks downstream of the integrated virus coincides with a large gene intron. In contrast to the aforementioned W12 clones, in clone H the virus integrates into a region of the host genome absent of either activating or repressive marks. However, the structure of the genome at this site appears to be reasonably open, given alignment of the DNaseI hypersensitivity sites (Figure 5.5 C). Interestingly, in all clones, the virus appears to integrate into regions of the genome that contain or are close to CTCF sites.

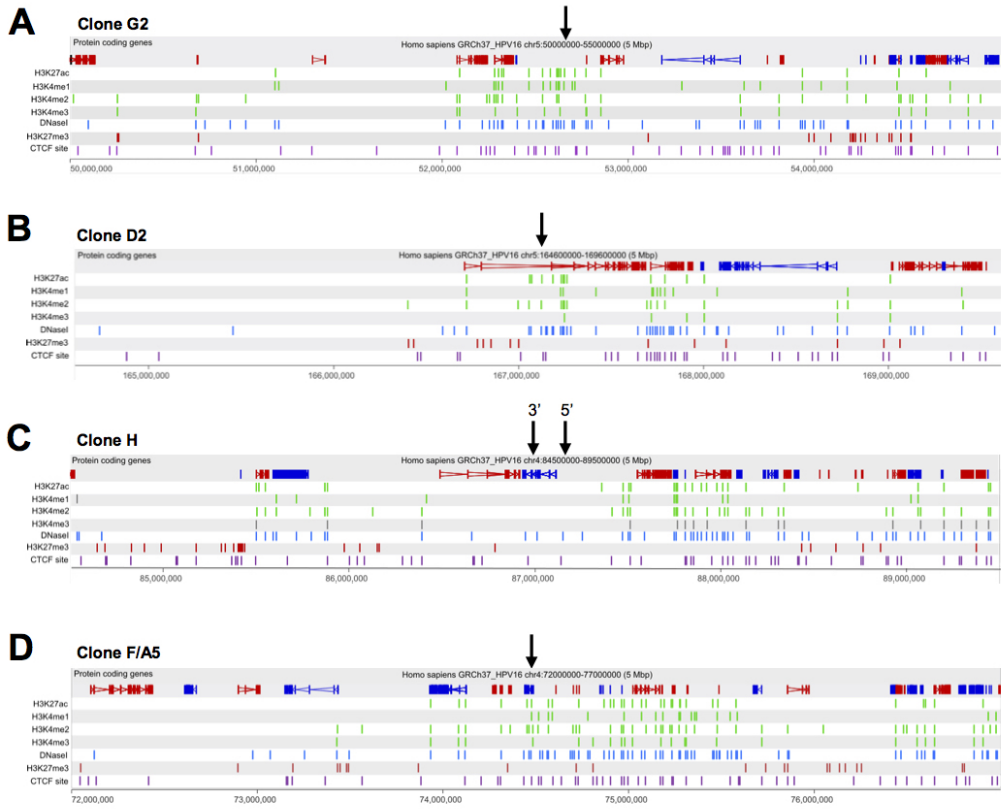

Figure 5.5: Analysis of host chromatin structure at the site of HPV16 integration. Each panel shows 5 Mb of the host genome across the integration loci, with the virus integration site indicated by a black arrow (5' and 3' separated in clone H due to deletion). Protein coding genes are shown in the first track and the direction of each gene indicated by colour; red = 5' to 3', blue = 3' to 5'. ChIP-seq data from normal human epidermal keratinocyte (NHEK) cell line is aligned with the host genome (taken from ENCODE). PTMs of active chromatin: H3K27ac, H3K4me1, H3K4me2, H3K4me3 are coloured green; DNaseI hypersensitivity sites are coloured blue; the repressive H3K27me3 mark is coloured red and CTCF sites are coloured purple. A) W12 clone G2, B) W12 clone D2, C) W12 clone H, D) W12 clone F/A5.

5.2.4 Short- and long-range 3D interactions occur between the HPV16 and host genomes regardless of cell selection during early cervical carcinogenesis.

SCRiBL-Hi-C mapped sequence data were analysed and visualised using the SeqMonk program (Babraham Bioinformatics). Each peak in the viewer represents a 3D interaction between the integrated HPV16 genome and the host. The different peak heights and colours refer to the normalised number of reads that correspond to a particular interaction. The most populated reads in each of the data sets consistently represented both the 5' and 3' virus-host breakpoints identified by capture seq; close analysis of the HPV16 integration locus (100–200 kbp window) at high normalised read depth revealed clearly defined peaks that matched the integration breakpoints identified by Sanger sequencing (Figure 5.6 A (W12 G2), 5.8 A (W12 D2) and 5.9 A-C (W12 H, F, A5, respectively).

By increasing the size of the window across the integration locus (to 700 kb) and decreasing the normalised read depth, additional distinct peaks of reads were visible in clone G2 (Figure 5.6 B) and clone D2 (Figure 5.8 B). Multiple short-range (<500 kbp) 3D interactions occur between the integrated virus and the surrounding host genome; at this scale interaction distances vary from 34–238 kbp. In figure 5.6 B the short-range interaction loops in G2 were drawn from the HPV16 integration site to regions of the host genome where peaks contained more than 44,000 normalised reads. Of note, the interaction peaks in clone G2 appeared to align with CTCF sites, with the majority additionally overlapping marks of enhancer regions (H3K27ac and H3K4me1) and regions of DNaseI hypersensitivity (Figure 5.6 B). Expanding the window further (5 Mb) illustrated a number of long-range (>500 kbp) 3D interactions between the virus and the host. In figure 5.6 C the long-range interaction loops were drawn to regions of the host genome where peaks contained more than 16,000 normalised reads. The furthest and most prominent peak was located at 53,520,000 within the first intron of host gene ARL15, approximately 900 kbp from the site of HPV16 integration (Figure 5.6 C).

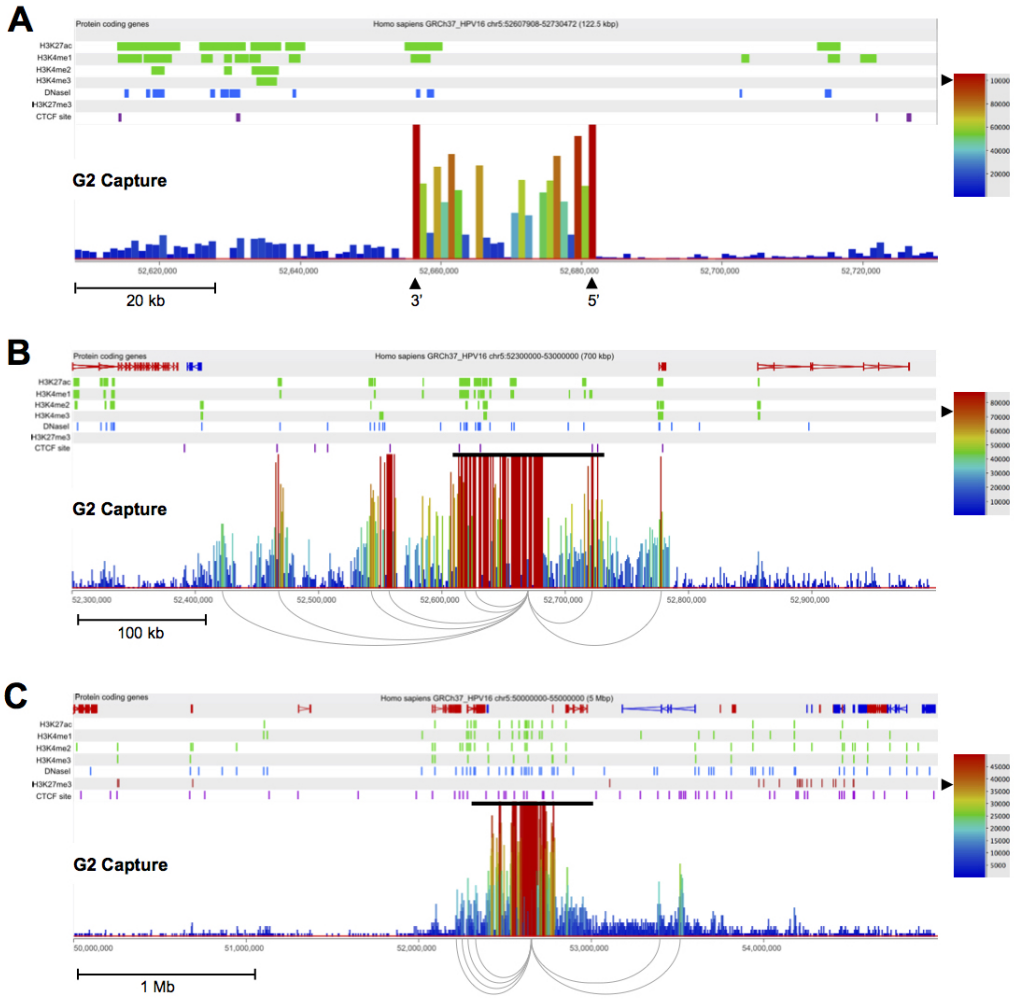

Figure 5.6: Identification of short and long-range interactions between integrated HPV16 and the host genome in W12 clone G2. A) SCRiBL-Hi-C data 122.5 kbp across the HPV16 integration locus. The 5' and 3' breakpoints of the virus are indicated by the tallest red bars and are labelled with black arrowheads. B) SCRiBL-Hi-C data 700 kbp across the HPV16 integration locus. The black line above the read peaks indicates the genomic window seen in panel A. Peaks of reads indicate regions of the host interacting with the integrated virus in 3D. Loops from the integration site to interacting regions of the host are shown beneath the panel and have been drawn from the approxiate HPV16 integration site within the amplified region to peaks containing more than 44,000 normalised reads. C) SCRiBL-Hi-C data 5 Mbp across the HPV16 integration locus. The black line above the read peaks indicates the genomic window of seen in panel B. Long-range 3D interactions between the integrated virus and the host are indicated by peaks of reads and highlighted by loops drawn beneath the panel and have been drawn from the approximate HPV16 integration site within the amplified region to peaks containing more than 16,000 normalised reads. In each panel the key indicates the normalised read count. Additionally, protein-coding genes are shown in the first track of each panel (genes coloured according to their orientation: red = 5' to 3', blue = 3' to 5'), followed by the alignment of ChIP-seq data from the NHEK cell line (ENCODE). PTMs of active chromatin: H3K27ac, H3K4me1, H3K4me2, H3K4me3 are coloured green; DNaseI hypersensitivity sites are coloured blue; the repressive H3K27me3 mark is coloured red and CTCF sites are coloured purple.

To validate the finding of this long-range interaction to ARL15 in clone G2, 3D fluorescent *in situ* hybridisation (FISH) was carried out. Three fluorescent DNA probes were produced to hybridise to either the integrated HPV16 genome, ARL15, or a control region of the genome that was the same linear distance in the opposite direction from the integrated virus as the ARL15 probe (Figure 5.7 A). Only cells containing one HPV16 signal and two copies of both the control and ARL15 probes were analysed. A representative image is shown in Figure 5.7 B. Analysis of the 3D distances (x, y and z plane) indicated that in the integrated chromosome the HPV16 probe and ARL15 interacting probes were significantly closer together than the HPV16 probe and the control probe (Figure 5.7 C and D). Additionally, when comparing the distances between the control probe and the ARL15 probe in both the integrated and unintegrated chromosomes the two probes were significantly closer together in the chromosome with HPV16 integrated (Figure 5.7 E and F). This suggests that HPV16 integration affects host genome architecture, and that the long-range interaction to gene ARL15 results in the two regions of DNA coming closer together.

3D interactions between the integrated HPV16 genome and the host were also identified in clone D2 (Figure 5.8 A and B). The majority of the interactions occurred upstream of the integrated virus and are all within the large host gene TENM2, over distances ranging from ~49 to 527 kbp. The virus-host interaction loops in D2 were drawn to regions of the host genome where peaks contained more than 6,000 normalised reads (Figure 5.8 B). Whilst the virus-host interactions closer to the integrated virus align with marks of active and open chromatin, those further from the virus appear to correlate less well with individual marks. Despite this, the long-range interaction at 167,647,500 (the furthest from the integrated virus) does align with a CTCF binding site.

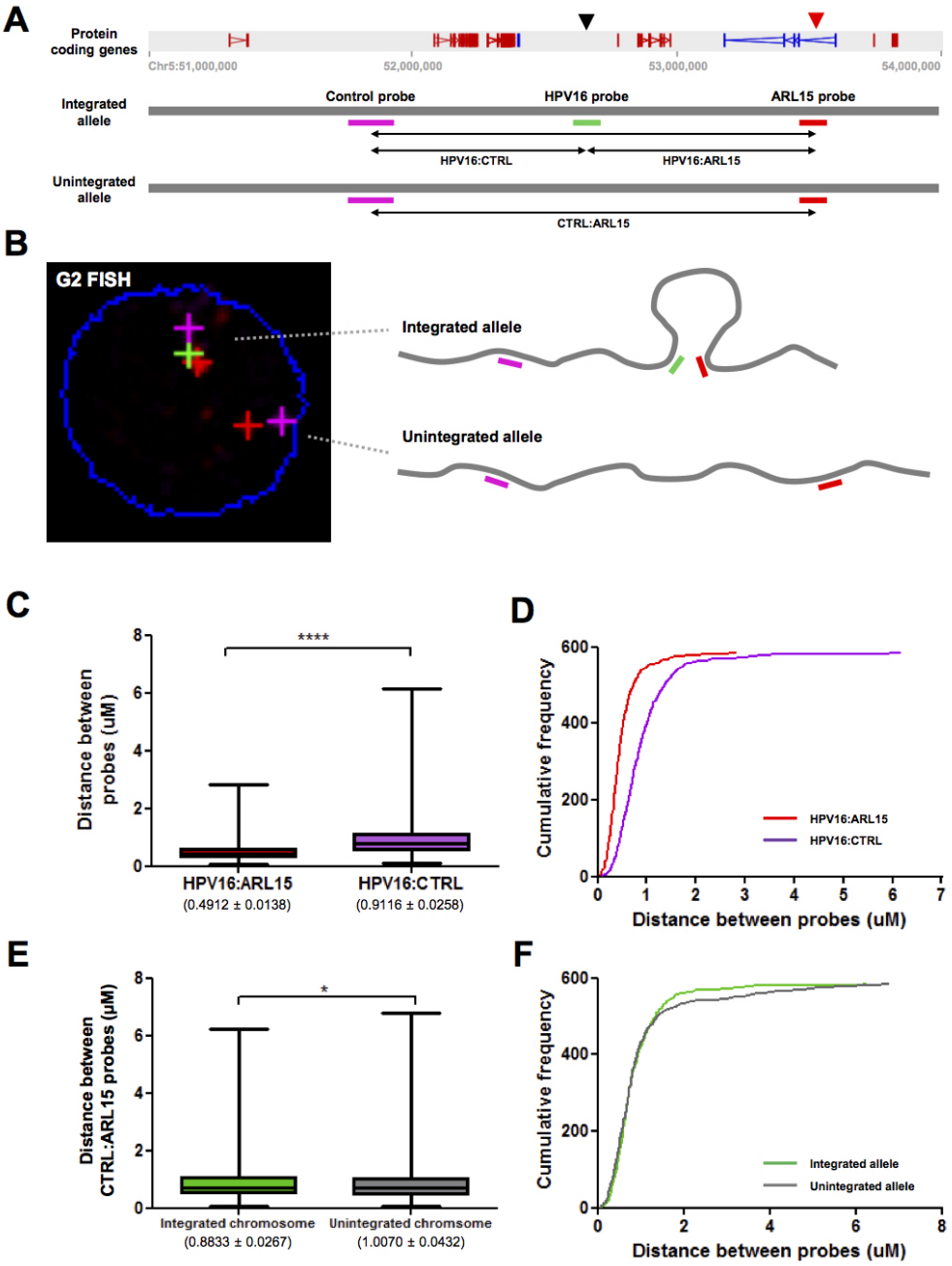

Figure 5.7: Validation of HPV16-host 3D chromatin interactions in W12 clone G2 by fluorescence *in-situ* hybridisation (FISH). A) Schematic detailing the positions of the DNA probes used on the integrated and unintegrated alleles of a portion of chromosome 5 in W12 clone G2. The 'control probe' hybridises to a region of the host genome (51,676,020–51,873,551) and is coloured in purple, the HPV16 probe is green, and the 'interacting probe' maps to the host gene ARL15 (53,473,886–53,584,235) is coloured in red. B) Representative image of the probes hybridised to W12 G2 genome in a 3D FISH experiment. C and D) Analysis of the 3D distance between both sets of FISH probes: HPV16:control (purple) and HPV16:ARL15 (red) in the copy of chromosome 5 that contained the integrated HPV16. Comparisons are shown in a box and whisker diagram (C) and a frequency distribution graph (D). E and F) Analysis of the 3D distance between the 'control' and 'interacting' probes in both the integrated (green) and unintegrated (grey) alleles. Comparisons are shown in a box and whisker diagram (E) and a frequency distribution graph (F). n=585; data presented as mean ± SEM; using unpaired, two-tailed Students T-test: * $p<0.05$, **** $p<0.0001$.

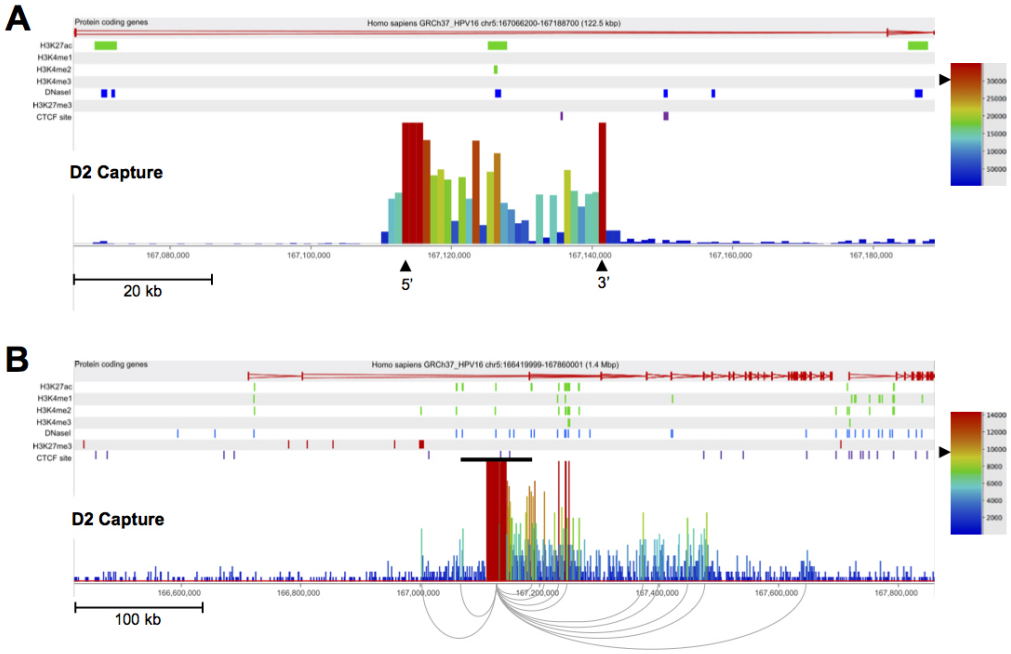

Figure 5.8: Identification of short and long-range interactions between integrated HPV16 and the host genome in W12 clone D2. A) SCRiBL-Hi-C data 122.5 kbp across the HPV16 integration locus. The 5' and 3' breakpoints of the virus are indicated by the tallest red bars and are labelled with black arrowheads. B) SCRiBL-Hi-C data 1.4 Mbp across the HPV16 integration locus. The black line above the read peaks indicates the genomic window seen in panel A. Peaks of reads indicate regions of the host interacting with the integrated virus in 3D. Loops from the integration site to interacting regions of the host are shown beneath the panel and have been drawn from the approximated HPV16 integration site within the amplified region to peaks containing more than 6,000 normalised reads. For both panels the scale bar indicates the normalised read count. Additionally, protein-coding genes are shown in the first track (genes coloured according to their orientation: red = 5' to 3', blue = 3' to 5'), followed by the alignment of ChIP-seq data from the NHEK cell line (ENCODE). PTMs of active chromatin: H3K27ac, H3K4me1, H3K4me2, H3K4me3 are coloured green; DNaseI hypersensitivity sites are coloured blue; the repressive H3K27me3 mark is coloured red and CTCF sites are coloured purple.

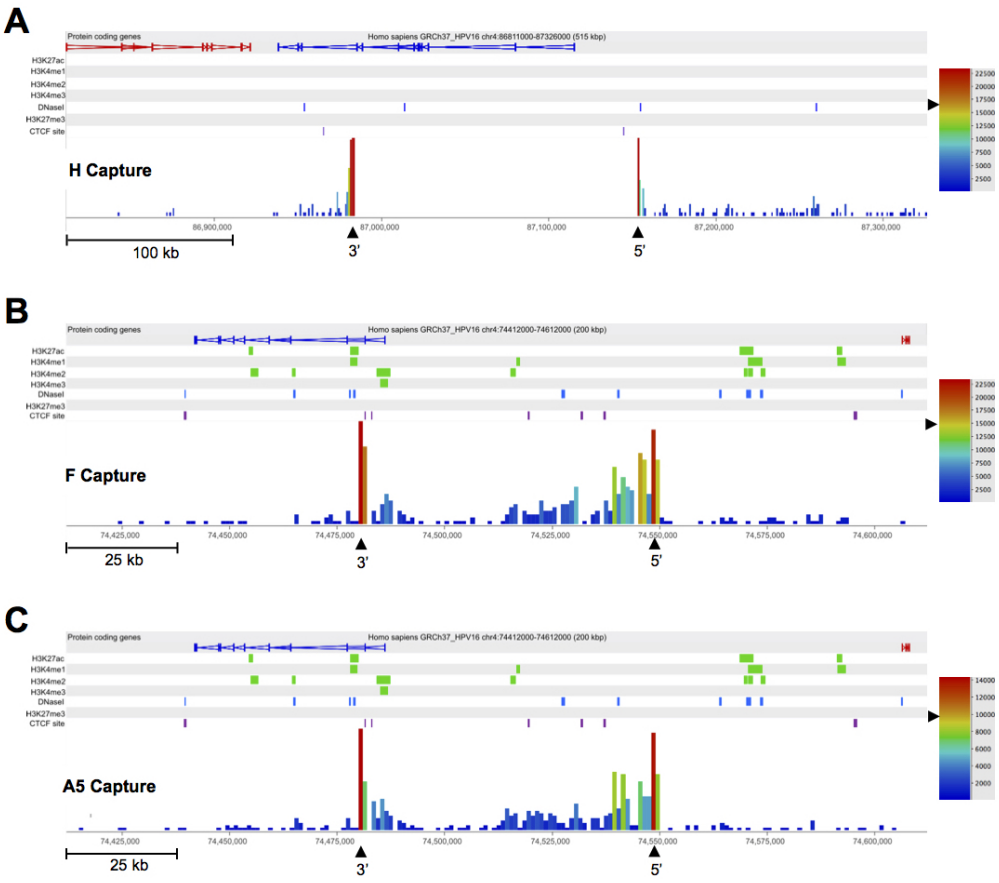

Figure 5.9: Virus-host breakpoints identified in clones H, F and A5 by SCRiBL Hi-C. A) W12 H SCRiBL-Hi-C data 200 kbp across the HPV16 integration locus. B) W12 F SCRiBL-Hi-C data 200 kbp across the HPV16 integration locus. C) W12 A5 SCRiBL-Hi-C data 515 kbp across the HPV16 integration locus. In each panel the 5' and 3' virus-host breakpoints are indicated by the tallest red bars and are labelled with black arrowheads, while the scale bar indicates the normalised read count. Additionally, protein-coding genes are shown in the first track (genes coloured according to their orientation: red = 5' to 3', blue = 3' to 5'), followed by the alignment of ChIP-seq data from the NHEK cell line (ENCODE). PTMs of active chromatin: H3K27ac, H3K4me1, H3K4me2, H3K4me3 are coloured green; DNaseI hypersensitivity sites are coloured blue; the repressive H3K27me3 mark is coloured red and CTCF sites are coloured purple.

5.2.5 HPV16 integration can disrupt local host genome architecture and affects the expression of host genes adjacent to the integration site.

The Hi-C libraries (generated as part of the SCRiBL protocol – prior to the HPV16 sequence capture element) of clone G2 and D2 were sequenced and mapped against the human Hg19 genome in order to evaluate the nuclear architecture of the host and to determine whether any changes to host chromosome interactions are caused as a result of HPV16 integration. The integration site of clones G2 and D2 are distinct and, as such, they act as a control for one another.

A 5 Mb map of Hi-C data from clone G2 across the HPV16 integration locus reveals clearly defined regions of interacting host DNA up to ~1 Mb (approximately the size of a topologically associating domain (TAD)) (Figure 5.10 A). Analysis of D2 Hi-C data across the same genomic locus reveals a predominantly similar structure of host architecture (Figure 5.10 B and C). Moreover, upon aligning the publically available TAD boundary information for IMR90 and hESC cell lines (Dixon *et al.*, 2012[252]), it is clear that all 3D interactions between integrated HPV16 and the host in clone G2 occur within a TAD (Figure 5.11).

The nuclear architecture of clones G2 and D2 was additionally evaluated 2.5 Mb either side of the D2 HPV16 integration site (Figure 5.10 D-F). As with G2, both clones exhibit similar host architecture across the region and the majority of clearly defined 3D interactions between the virus and the host occurred within a TAD boundary (Figure 5.12).

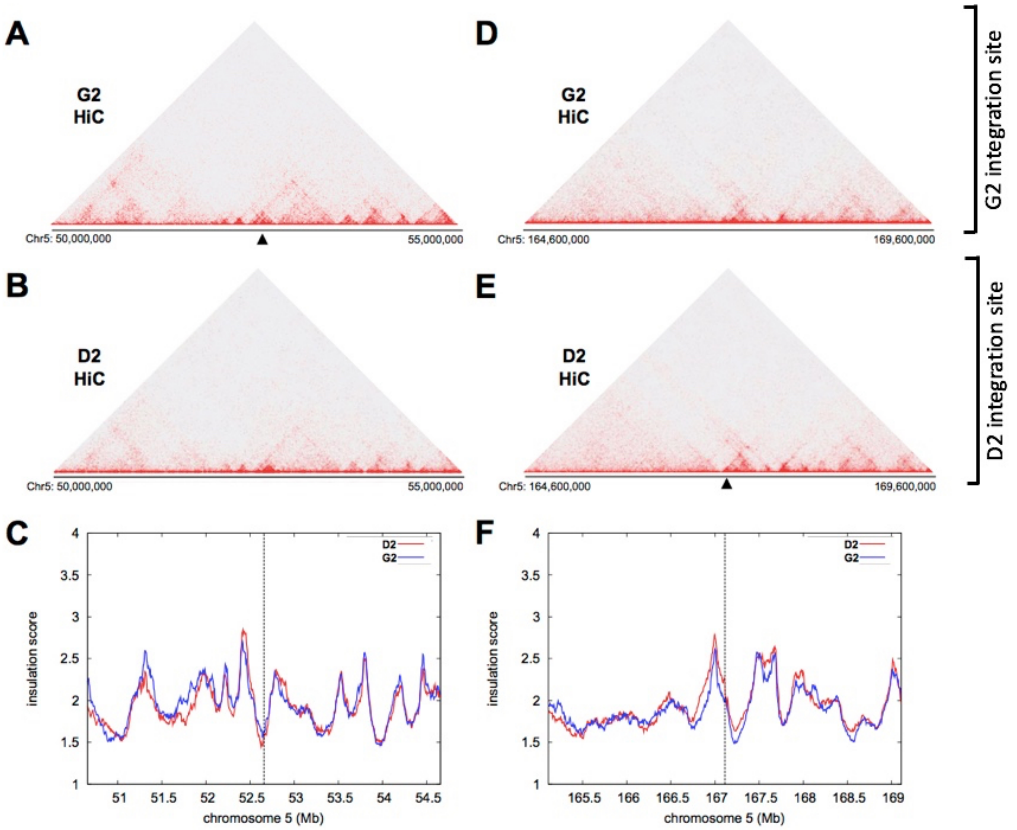

Figure 5.10: Changes in host genome architecture and gene expression as a result of HPV16 integration in W12 clones G2 and D2. A) W12 G2 Hi-C map, 5 Mb across the G2 integration locus (Chr5: 50,000,000–55,000,000). The HPV16 integration site is annotated with a black arrowhead. B) W12 D2 Hi-C map, 5 Mb across the G2 integration locus (Chr5: 50,000,000–55,000,000). C) Chart plotting the insulation scores of W12 G2 (blue line) and W12 D2 (red line), 5 Mb across the G2 integration locus (Chr5: 50,000,000–55,000,000). D) W12 G2 Hi-C map, 5 Mb across the D2 integration locus (Chr5: 164,600,000–169,600,000). E) W12 D2 Hi-C map, 5 Mb across the D2 integration locus (Chr5: 164,600,000–169,600,000). The HPV16 integration site is annotated with a black arrowhead. F) Chart plotting the insulation scores of W12 G2 (blue line) and W12 D2 (red line), 5 Mb across the G2 integration locus (Chr5: 164,600,000–169,600,000).

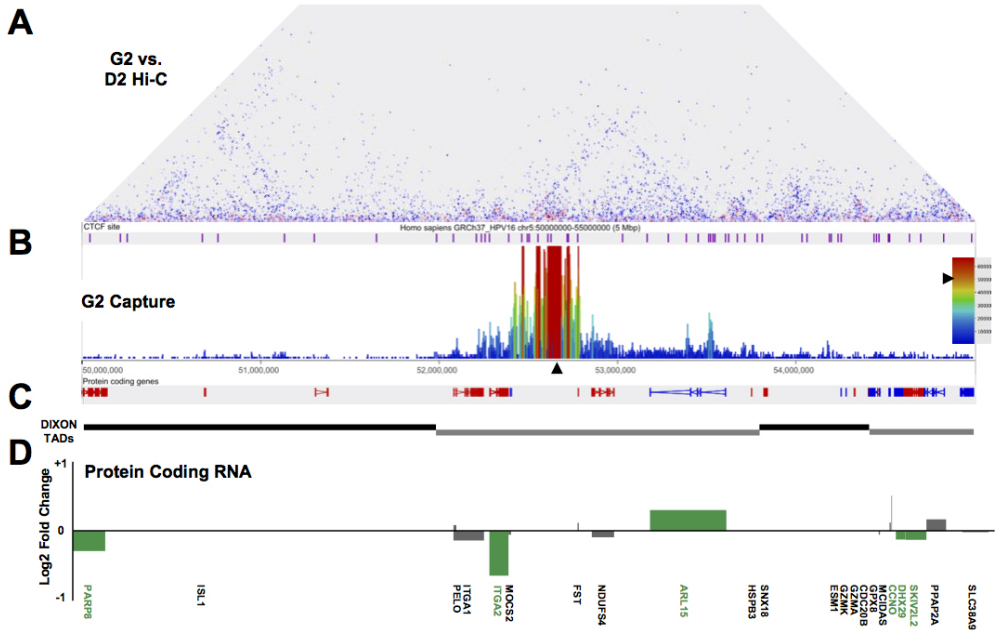

Figure 5.11: Changes in host genome architecture and gene expression as a result of HPV16 integration in W12 clone G2. A) Comparative Hi-C map of G2 vs. D2 control. Blue dots represent host-host interaction sites that occur less frequently in G2, while red dots represent interactions that are stronger in G2 compared with D2. B) W12 G2 SCRiBL Hi-C data showing 3D interactions between the integrated virus and the host, CTCF sites (purple) are aligned in the top track. The scale bar indicates the normalised read count and a black arrowhead indicates the HPV16 integration site. C) Aligned protein coding track (genes coloured according to their orientation: red = 5' to 3', blue = 3' to 5') above an alignment of TAD boundaries from the publically available IMR90 (human lung fibroblast) cell line (Dixon *et al.* 2012). D) Chart indicating the relative expression of W12 G2 protein-coding genes 2.5 Mb either side of the HPV16 integration site. The chart indicates the log fold-change of host gene expression in the clone of interest compared with a 6-clone integrant average. Significant changes in host gene expression are indicated by genes coloured green, non-significant expression changes are coloured grey (p<0.05, negative binomial Wald test).

Changes to host gene expression were evaluated by using available duplicate RNA-seq data for seven W12 integrant clones: A5, B, D2, F, G2, H and R2. Bioinformatic analysis of RNA-seq data was conducted in collaboration with Anton Enright (EBI-EMBL). The expression of the protein coding host genes 2.5 Mb either side of the HPV16 integration site were compared with the control 6-clone average to determine whether host gene expression changed as a result of integration. In each of the W12 clones analysed (G2, D2, H, F and A5) significant changes to protein coding host gene expression — both over and under expressed — were seen across the entire 5 Mb region (Figure 5.11 D (W12 G2), 5.12 D (W12 D2), 5.13 D (W12 H), 5.14 D (W12 F) and 5.15 D (W12 A5)). Most notably, where HPV16 had integrated within a host gene, the expression of that gene was consistently upregulated; the change for TENM2 expression in clone D2 was 4.79-fold greater than the 6-clone average, MAPK10 expression was increased by 4.47-fold and RASSF6 increased by 1.62- and 1.64-fold in clone A5 and F, respectively. In addition, the HPV16-host 3D interaction to the first intron of ARL15 in clone G2 led to increased expression of the gene with a 0.30-fold increase compared with the 6-clone average ($p<0.05$) (Figure 5.11 D).

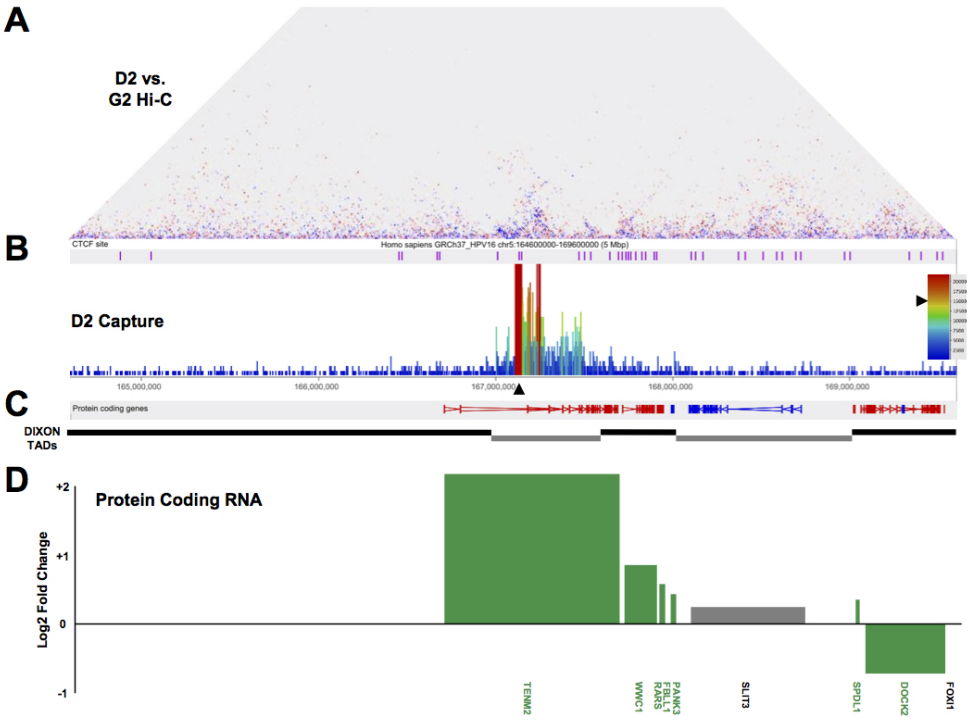

Figure 5.12: Changes in host genome architecture and gene expression as a result of HPV16 integration in W12 clone D2. A) Comparative Hi-C map of D2 vs. G2 control. Blue dots represent host-host interaction sites that occur less frequently in D2, while red dots represent interactions that are stronger in D2 compared with G2. B) W12 D2 SCRiBL Hi-C data showing 3D interactions between the integrated virus and the host, CTCF sites (purple) are aligned in the top track. The scale bar indicates the normalised read count and a black arrowhead indicates the HPV16 integration site. C) Aligned protein coding track (genes coloured according to their orientation: red = 5' to 3', blue = 3' to 5') above an alignment of TAD boundaries from the publically available IMR90 (human lung fibroblast) cell line (Dixon *et al.* 2012). D) Chart indicating the relative expression of W12 D2 protein-coding genes 2.5 Mb either side of the HPV16 integration site. The chart indicates the log fold-change of host gene expression in the clone of interest compared with a 6-clone integrant average. Significant changes in host gene expression are indicated by genes coloured green, non-significant expression changes are coloured grey (p<0.05, negative binomial Wald test).

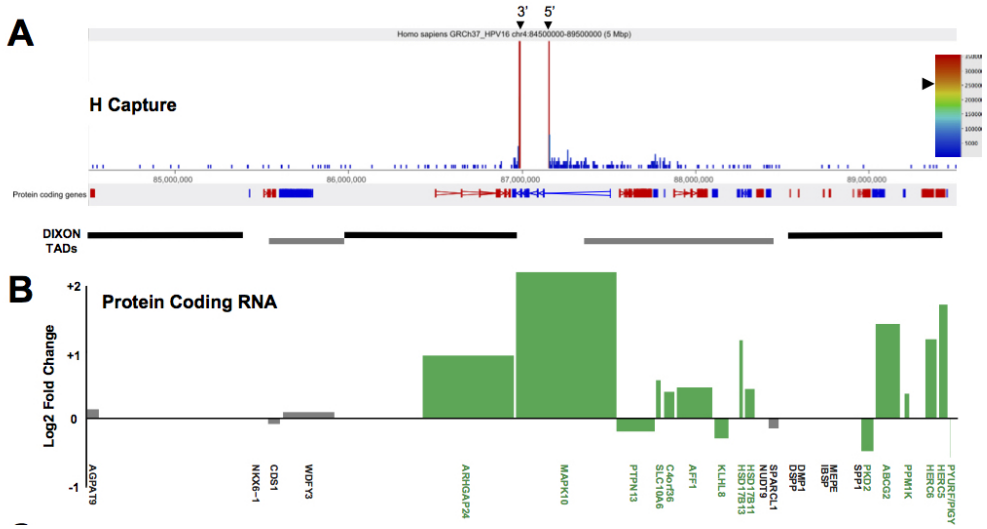

Figure 5.13: Host gene expression changes across the 5 Mb HPV16 integration loci in clones H. A) SCRiBL-Hi-C data track with aligned protein-coding genes shown beneath; genes coloured according to their orientation: red = 5' to 3', blue = 3' to 5'. TAD boundaries from the publically available IMR90 (human lung fibroblast) cell line (Dixon *et al.* 2012) are aligned beneath. B) Chart indicating the relative expression of W12 H protein-coding genes 2.5 Mb either side of the HPV16 integration site. The chart indicates the log fold-change of host gene expression in the clone of interest compared with a 6-clone integrant average. Significant changes in host gene expression are indicated by genes coloured green, non-significant expression changes are coloured grey (p<0.05, negative binomial Wald test). For each panel the scale bar indicates the normalised read count.

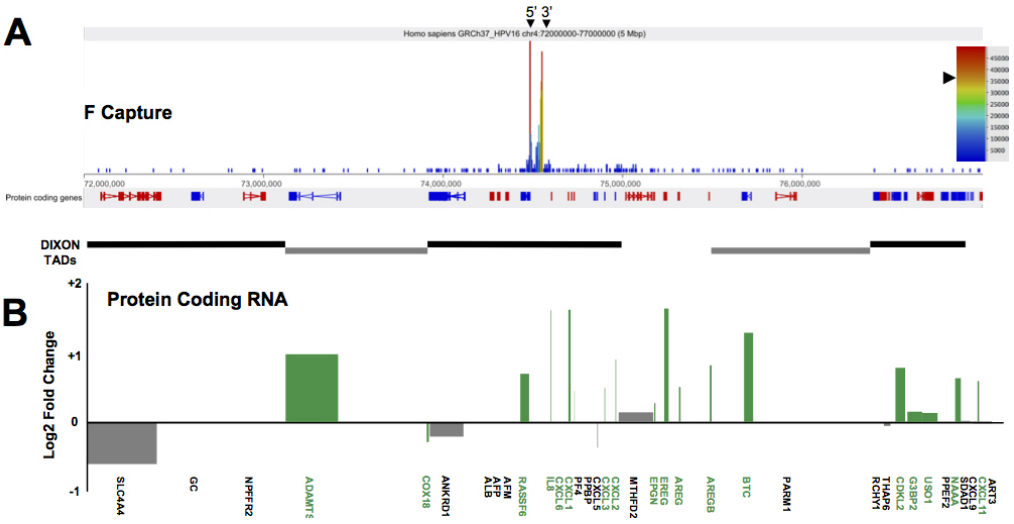

Figure 5.14: Host gene expression changes across the 5 Mb HPV16 integration loci in clones F. A) SCRiBL-Hi-C data track with aligned protein-coding genes shown beneath; genes coloured according to their orientation: red = 5' to 3', blue = 3' to 5'. TAD boundaries from the publically available IMR90 (human lung fibroblast) cell line (Dixon *et al.* 2012) are aligned beneath. B) Chart indicating the relative expression of W12 F protein-coding genes 2.5 Mb either side of the HPV16 integration site. The chart indicates the log fold-change of host gene expression in the clone of interest compared with a 6-clone integrant average. Significant changes in host gene expression are indicated by genes coloured green, non-significant expression changes are coloured grey (p<0.05, negative binomial Wald test). For each panel the scale bar indicates the normalised read count.

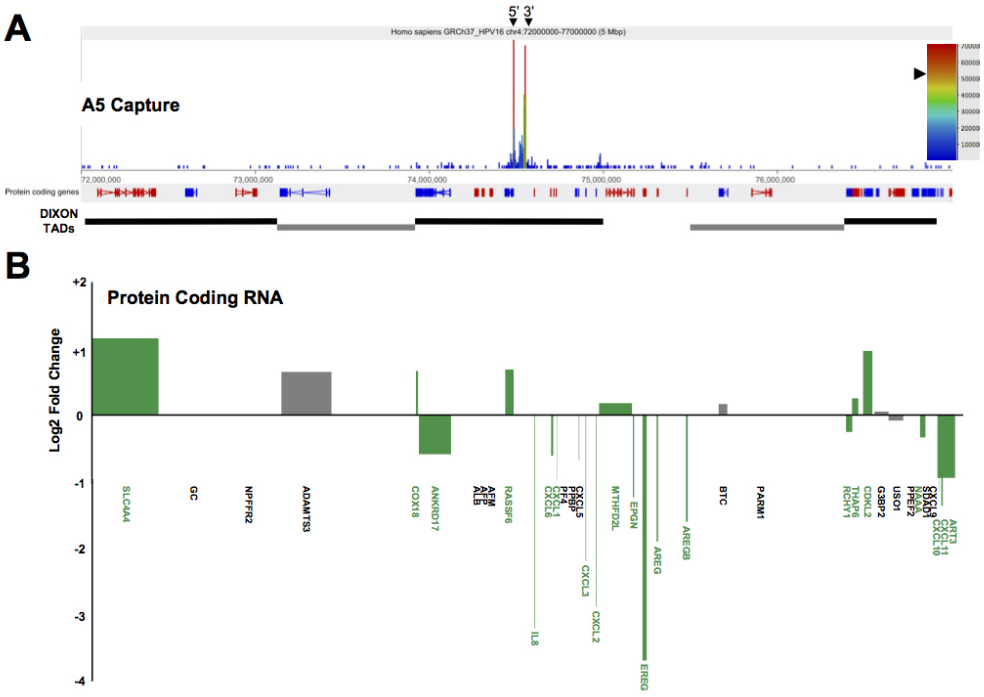

Figure 5.15: Host gene expression changes across the 5 Mb HPV16 integration loci in clones A5. A) SCRiBL-Hi-C data track with aligned protein-coding genes shown beneath; genes coloured according to their orientation: red = 5' to 3', blue = 3' to 5'. TAD boundaries from the publically available IMR90 (human lung fibroblast) cell line (Dixon *et al.* 2012) are aligned beneath. B) Chart indicating the relative expression of W12 A5 protein-coding genes 2.5 Mb either side of the HPV16 integration site. The chart indicates the log fold-change of host gene expression in the clone of interest compared with a 6-clone integrant average. Significant changes in host gene expression are indicated by genes coloured green, non-significant expression changes are coloured grey (p<0.05, negative binomial Wald test). For each panel the scale bar indicates the normalised read count.

To further investigate the effect of HPV16 integration on host gene expression, the variance in gene expression in the genomic regions adjacent to the HPV16 integration site was compared with that of the whole chromosome. Analysis of the gene expression variance was carried out in collaboration with Anton Enright (EBI-EMBL) and was performed using the RNA-seq datasets from W12 clones G2, D2, H, F, A5, B and R2. Host genes — including both protein-coding and non-coding genes — either side of the HPV16 integration site were grouped into bins, each containing five genes. The range and variance of gene expression of each bin was plotted against the mean level of gene expression across the whole chromosome indicating that across the W12 clones expression of genes in this regions was highly variable (Figure 5.16 A, C, E, G and I). In each of the clones analysed (G2, D2, H, F and A5), the variance in gene expression of multiple bins within genomic regions adjacent to the HPV16 integration site were highly significant ($p<0.05$ and $p<0.001$), indicating that integration of HPV16 has a direct influence on host gene expression at and around the integration site (Figure 5.16 B, D, F, H and J). Notably, significant changes to the variance in host gene expression were felt within bins directly at the site of virus integration (W12 A5: Fig. 5.16 J) or within bins very close to this region (≤6 bins).

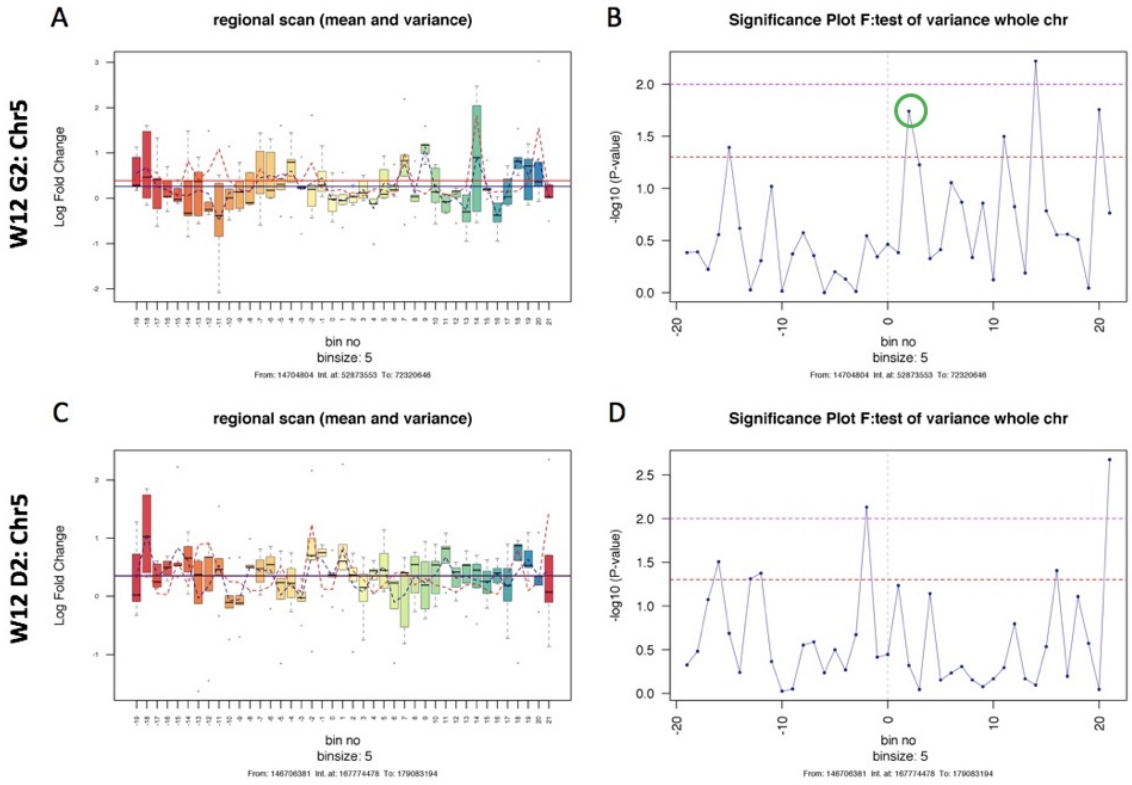

Figure 5.16: Variance in host gene expression across the host genomic region containing the HPV16 integration site in W12 clones G2, D2. Each left panel indicates the range and variance of host gene expression in W12 integrant clones [A) W12 G2, C) W12 D2, E) W12 H, G) W12 F, I) W12 A5], focussing on 100 genes either side of the HPV16 integration site. For each clone, gene expression levels were compared with those in six other clones, based on the full dataset from clones G2, D2, H, F, A5, B and R2. In each panel, the HPV16 integration site is centred on bin 0. Each bin contains five genes, with no overlap between bins. The box and whisker plots illustrate the range of gene expression levels within each bin, with the bar indicating median values, the box the IQR and the whiskers the range. The mean gene expression across the whole chromosome is indicated by the solid blue line, while the mean level of gene expression across individual bins is shown by the dotted blue line. The mean variance of gene expression across the whole chromosome is indicated by the solid red line, while the mean level of gene expression across individual bins is shown by a dotted red line. Each right hand panel shows the significance of the variance in gene expression within each bin. Each point represents a five-gene bin, corresponding to those in the left-hand panels. The horizontal lines indicate the significance of the variance in each bin, compared with the variance in gene expression across the whole chromosome (above the dashed red line, $p<0.05$; above the dashed pink line, $p<0.01$).

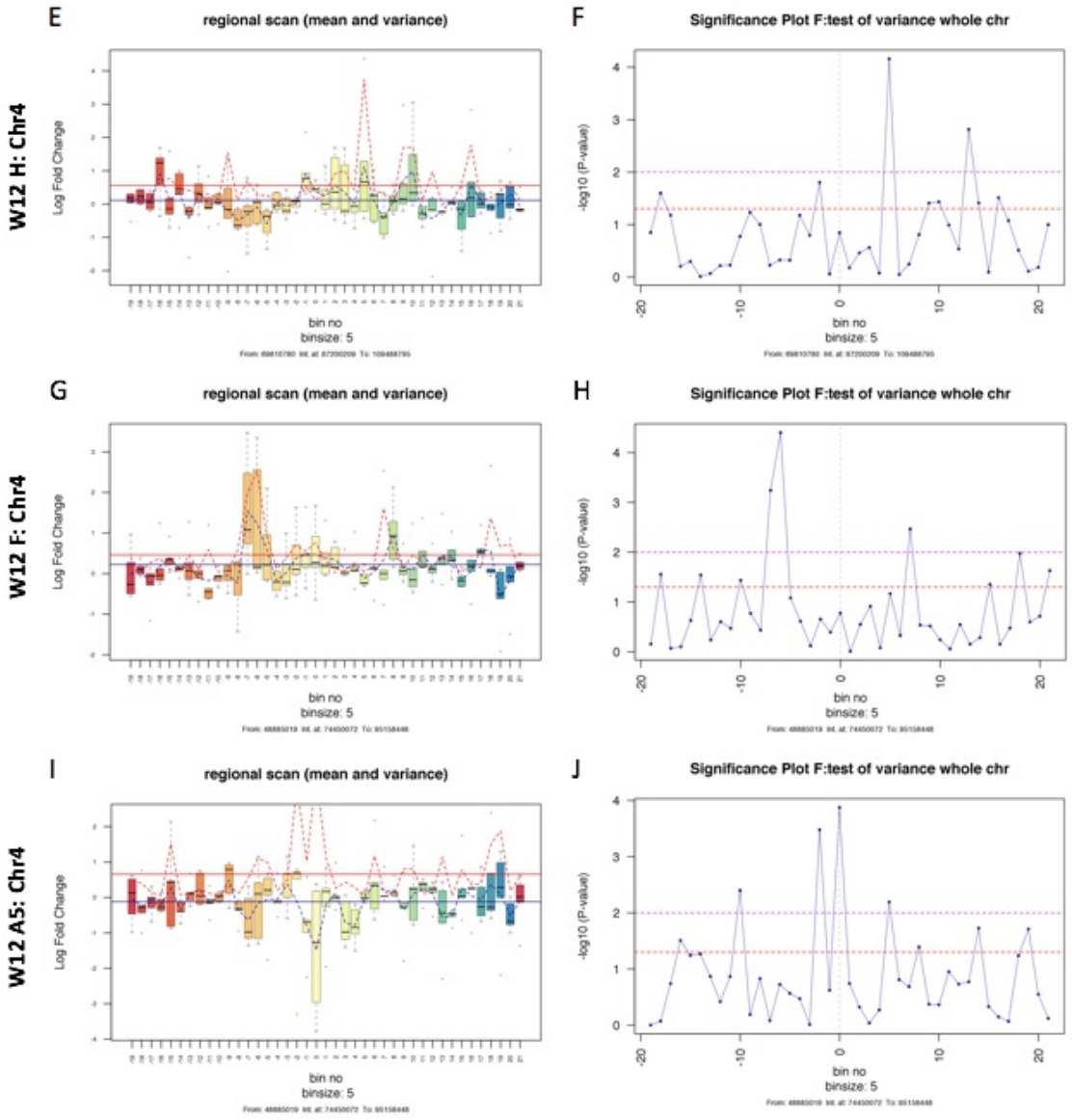

Figure 5.16: Continued. Variance in host gene expression across the host genomic region containing the HPV16 integration site. E and F) W12 H, G and H) W12 F and I and J) W12 A5.

5.2.6 HPV16 integration results in virus-host fusion transcripts

Evaluation of RNA-seq data aligned to the human Hg19 reference genome for clone G2 revealed the presence of a number of virus-host fusion transcripts that originated from the virus early promoter. The fusion transcripts were generated by splicing from either the 225 or 879 splice donor sites in the HPV16 genome into either the 52,664,903 or 52,665,662 cryptic splice acceptor sites in the host; this occurs as a result of the loss of splice acceptor sites in the virus due to genome truncation during the integration process. In addition, transcription across the virus-host breakpoint junction itself was identified. All virus-host transcripts are summarised in Figure 5.17 A and C. To test the abundance of each fusion transcript in clone G2, RT-qPCR primer pairs were designed according to the Hg19 coordinates produced by RNA-seq. In clone G2 the 225-52,665,662 fusion transcript was most abundant; the levels of the breakpoint and 879-52,664,903 fusion transcripts were similar; and the relative abundance of the 879-52,664,903 transcript was extremely low (Figure 5.17 B). To determine the specificity of the fusion-transcripts, the same primer pairs were tested in clone A5 — a W12 integrant with a different site of HPV16 integration — and the episomal W12par1 cell line; in both control cell lines G2 fusion transcripts were not identified. As a positive control, the level of E6/E7 (E6all) mRNA, which includes all transcripts produced from the virus early promoter, was tested in each of W12 cell lines.

A

HPV coordinate	Host coordinate	RNA
2761	52655**799**	Host-virus junction
879	52664**903**	Splice product
225	52665**662**	Splice product
879	52665**662**	Splice product

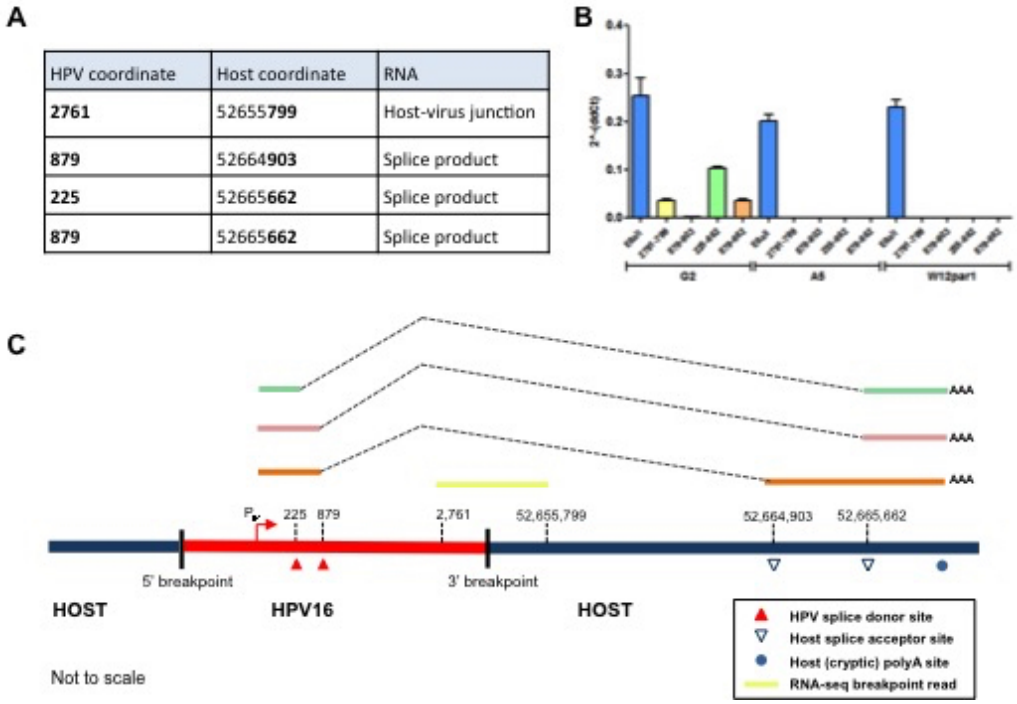

Figure 5.17: Differential splicing of virus-host fusion transcripts in W12 clone G2. A) Table summarising the the fusion transcripts between HPV16 and the host in clone G2 identified by RNA-sequencing. B) The abundance of each virus-host splice transcript analysed by RT-qPCR in clone G2, clone A5 and W12parI (episomal). Four mRNA transcripts were identified: RNA-seq breakpoint read (yellow); one read from the virus splice donor site at 225 bp to host acceptor site at 52,665,662 (green); and two reads from the virus splice donor site at 879 bp to host acceptor sites at 52,664,903 (pink) and 52,665,662 (orange). The abundance of each transcript was compared to a positive 'E6all' control (blue), which captures all viral transcripts from the early promoter. C) Schematic of the virus-host fusion transcripts illustrating the virus splice donor sites (red triangles) and host acceptor sites (inverted blue triangles) in integrated W12 clone G2. The virus genome and host DNA are represented by red and blue lines, respectively and the virus-host breakpoints indicated with a black line. Spliced transcripts are coloured to match the RT-qPCR analysis in B. Diagram to scale.

5.3 Discussion

Using a panel of W12 integrant clones, the results presented in this chapter illustrate that the HPV16 genome integrates into distinct regions of the host resulting in minor changes to host gene expression. Moreover, 3D interactions between the integrated HPV16 genome and the host genome have been identified. These data indicate consequences of HPV16 integration that are a typical feature of all HPV integrants and are not restricted to cells with a selective growth advantage.

In the first part of this chapter, the identification of the 5' and 3' virus-host breakpoints of five W12 clones (F, A5, D2, H and G2) were described at nucleotide resolution. It is important to note that the HPV16 integration sites in four out of five clones identified in this study differ from those that have been previously published in Dall *et al.*, 2008[104]; the discrepancies are summarised in Table 5.1.

Table 5.1: W12 integrant clone integration sites.

Clone	Dall *et al.* integration site	Coordinates	Technique	Current integration site identification	Coordinates	Technique
F	4q13.3 & 8q24.21	5': 746,995,017 3': 128,476,710	RS-PCR	4q13.3	5': 74,549,681 3': 74,480,662	NGS
A5	8p11.21	3': 41,625,232	RS-PCR	4q13.3	5': 74,549,681 3': 74,480,662	NGS
D2	18q21.2	3': 52,165,073	RS-PCR	5q34	5': 167,112,984 3': 167,141,612	NGS
H	4q21.23	3': 87,202,219	RS-PCR	4q21.3	5': 87,153,458 3': 86,983,196	NGS
G2	21q22.1	3': 29,577,377	APOT	5q11.2	5': 52,681,626 3': 52,655,805	NGS

RS-PCR = restriction site polymerase chain reaction; APOT = Amplification of Papillomavirus Oncogene Transcripts; NGS = next generation sequencing.

The techniques used to elucidate the virus-host breakpoints in 2008, namely Restriction Site PCR (RS-PCR) and Amplification of Papillomavirus Oncogene Transcripts (APOT), were the most appropriate to use at the time; however, they are less sensitive and less accurate than next generation sequencing employed in this investigation. Additionally, and in contrast to previous findings, the results from this study revealed that the HPV16 integration site in W12 clones F and A5 are identical. Although both clones were isolated from the same mixed population of episomal W12 cells (W12Ser2 p12) (Figure 1.6 C), it is likely that clone A5 is a precursor

of clone F. Differences between the two clones are identified following continuous culture; after just twelve passes, clone F acquires additional genomic imbalances and has increased levels of HPV16 oncogene E6 and E7 expression when compared with clone A5 (unpublished data Cinzia Scarpini/Mark Pett). Despite having the same integration site, F and A5 have been treated as distinct clones in the host gene expression analysis of this investigation due to their phenotypic differences, including the rate of cell growth[183] and distinct patterns of host gene expression detected by subsequent RNA-seq analysis. Furthermore, the F/A5 5' and 3' virus-host junctions were found to be present in the W12 parental (episomal) cell line (Figure 5.2 D and E middle and right panels); this suggests that that across a mixed population, a small percentage of W12parI cells already harboured integrants.

Analysis of the HPV16 integration sites in the W12 clones showed that in the majority of cases the virus integrates within the coding region of a host gene; RASSF6 (Ras association domain family member 6; W12 F/A5), MAPK10 (mitogen-activated protein kinase 10; W12 H) and TENM2 (Teneurin transmembrane protein 2; W12 D2). Previous studies evaluating the HPV integration sites of cervical cancer samples have also indicated the region directly upstream of RASSF6[287] and MAPK10[101] as sites of integration. This is consistent with data that illustrates that HPV16 integration sites are located in host genes significantly more often than is expected; additionally in cases where the virus has integrated into an intergenic region of the host, the virus-host breakpoints are significantly closer to genes than predicted by chance[284] — the target region must be a transcriptionally competent area of host chromatin that can support viral oncogene expression. Preferential integration into gene-rich areas of the host genome is complemented by evidence that demonstrates in cervical SCC HPV16 preferentially integrates into regions of open and active chromatin, determined by the associated of integration sites with regions of DNaseI hypersensitivity and H3K4me3 methylation, respectively[288, 289]. The alignments of hallmark PTMs from the NHEK cell line across a window 2.5 Mb either side of the integration site of each W12 clone (Figure 5.5) correlates with these previous findings, indicating that this occurs regardless of subsequent cell selection. ChIP-seq data from NHEKs was determined the most appropriate publically available

comparator for the W12 clones as epidermal keratinocytes, being a target of HPV infection, were the only non-infected keratinocyte cell line available from ENCODE.

Evaluation of the integrated HPV16 genome revealed that the E2 ORF is most commonly disrupted as a result of linearisation of the virus episome. Although integration does not always lead to a large deletion of the HPV16 genome — deletion ranged from 36 bp (clone F/A5) to 2,131 bp (clone H) — disruption of the E2 ORF is sufficient for the protein regulator of the virus early promoter not to be transcribed; indeed, in clones G2, F and A5 a portion of the E2 ORF is placed upstream of the virus early promoter. The integration events seen in the W12 clones are representative of the earliest model of HPV integration that promotes oncogenesis by disrupting the E2 gene, alleviating E2 transcriptional repression of the HPV promoter and thus driving oncogene expression; however, as will be discussed, these integration events may also promote oncogenesis in a number of different ways including the formation of hybrid (viral-host) transcripts that are more stable than viral E6/E7 transcripts as well as the disruption of cellular genes and their flanking sequences, directly altering their expression and the expression of genes adjacent to the integration site[290].

Data from the capture-seq experiment and subsequent breakpoint verification (Figure 5.2) revealed that, regardless of virus genome copy number, only one 5' and 3' virus-host breakpoint exists in each clone. The observation of fewer breakpoints than the viral copy number has previously been reported and led the authors to hypothesise that the discrepancy is due to the amplification of viral integrants and flanking genomic sequences leading to redundant, identical breakpoints[176]. By evaluating the positions of the 5' and 3' breakpoints relative to the host genome it was deduced that there are two distinct mechanisms in which HPV16 integrates into the host genome in the W12 clones. Clone H is an example of direct integration of a single virus genome that results in a deletion of the human genome. In this model, the virus episome is linearised by two double strand breaks and inserts directly in the host (Figure 5.18 A). In W12 clones F and A5, a single copy of the virus integrates into the host genome via a looping mechanism. Looping results in the duplication of a portion of the host genome adjacent to the integrated virus, and the 5' and 3' virus-host junctions appear reversed when aligned to the human genome (Figure

5.18 B). Finally, clones G2 and D2 are examples of when multiple copies (n) of the HPV16 genome integrate via a looping mechanism in to the host, which results in amplification (n+1) of the adjacent region of the host (Figure 5.18 C). A summary of the virus and host genomes in each of the W12 clones is shown in Figure 5.19.

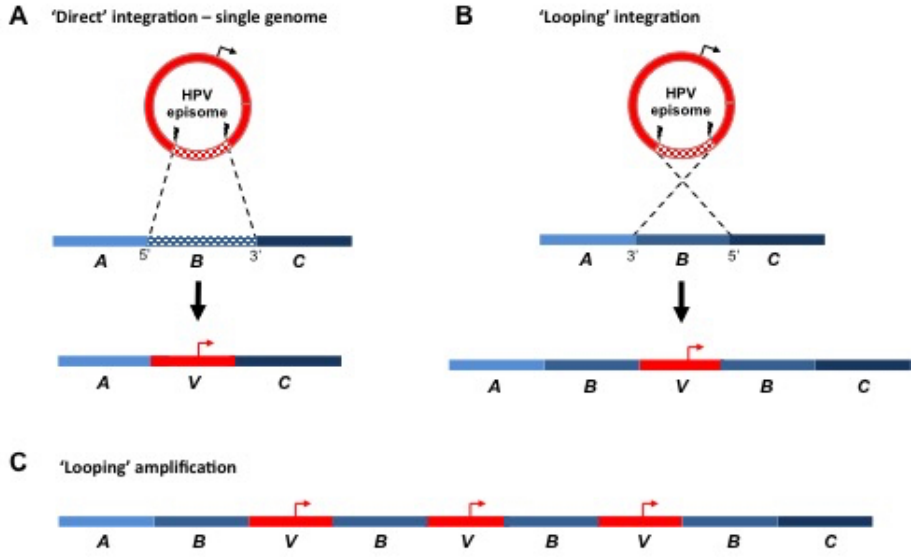

Figure 5.18: Mechanisms of HPV16 integration in the W12 integrant clones. A) Schematic illustrating the direct integration of the HPV16 genome (red) into the host (blue). Chequered regions indicate DNA that is deleted and/or broken during the integration process. B) Schematic illustrating a looping mechanism of integration of the HPV16 genome (red) into the host (blue). Chequered regions indicate DNA that is deleted in the integration process. HPV integration results in duplication of the host region labelled 'B', which flanks in the integrated virus on both 5' and 3' virus-host junctions. C) HPV16 integration via the looping mechanism can result in both the amplification of the virus genome and local regions of the host DNA. Schematic illustrates the fusion virus-host DNA linear organisation after insertion of three copies of the HPV16 genome. In all panels, different regions of the host have been labelled alphabetically and coloured in different shades of blue.

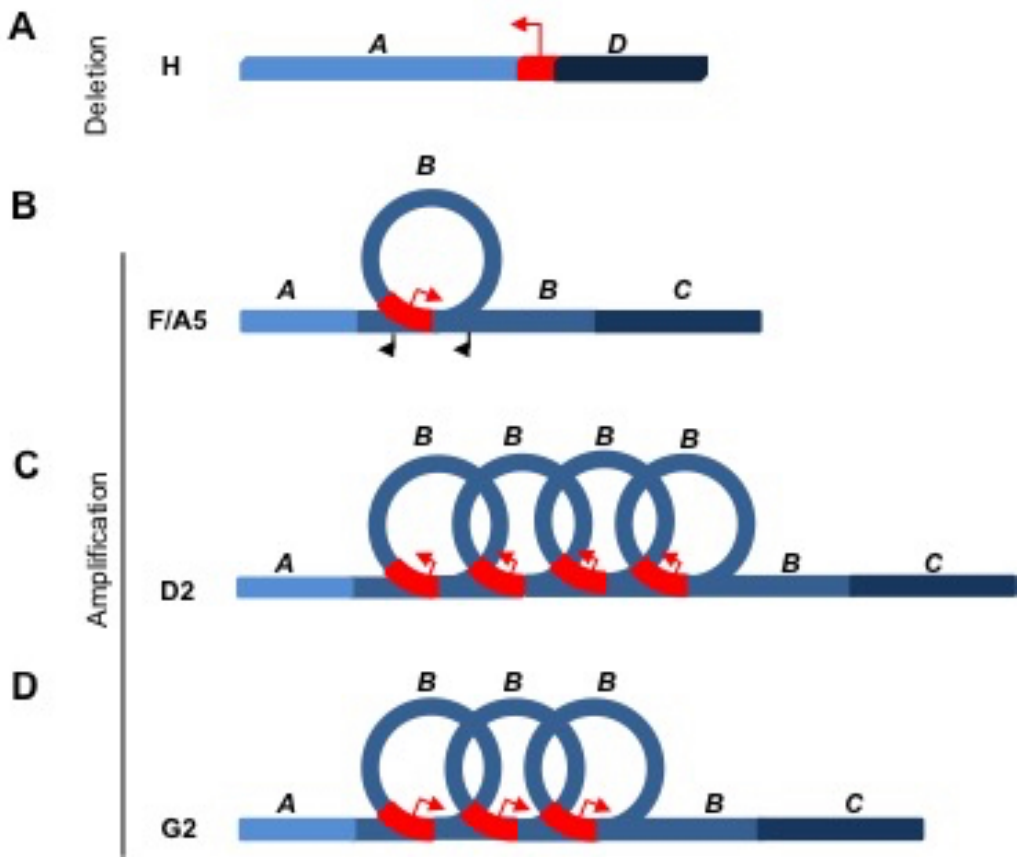

Figure 5.19: Summary of HPV16 integration in the W12 integrant clones. A) One copy of the HPV16 genome directly integrates into clone H resulting in a large deletion of host DNA. B) A single copy of HPV16 integrates via the looping mechanism in clone F/A5, which causes a duplication of the host region 'B'. C) Four copies of the HPV16 genome are integrated in clone D2 resulting in the amplification of host region 'B'. D) Three copies of the HPV16 genome are integrated in clone G2 resulting in the amplification of host region 'B'. In all panels, the HPV16 genome is coloured red and different regions of the host have been labelled alphabetically and coloured in different shades of blue. A red arrow denotes the virus early promoter; host gene promoters are illustrated with a black arrow.

Analysis of the nucleotide sequences of both the virus and the host at each breakpoint in the W12 clones adds to the existing body of evidence that suggests HPV integration likely occurs through microhomology-mediated repair mechanisms[291, 282, 87]. Interestingly, although a number of breakpoints have at least five out of ten homologous nucleotides at the breakpoint (G2: 5', D2: 5' and F: 3') they were not called as significant despite other breakpoints with the same (G2: 3') or fewer (H: 5' and F: 5') homologous nucleotides having a p-value <0.05 (Figure 5.4). This is a result of the analysis that compares the microhomology of the ten nucleotides directly adjacent to the breakpoint to that of the immediate flanking region (1,000 nt) rather than to randomly selected parts of the genome; this analysis includes corrections for local biases of AT rich regions, a consideration that has not been made in current publications that comment on microhomology at the HPV-host integration breakpoints.

The second part of this chapter focussed on identifying 3D interactions between the integrated virus and host genomes. A technical factor that affects the appearance of individual Circos plots and further analysis of the SCRiBL-Hi-C data is the number of sequencing reads that have at least one end that maps to the HPV16 genome. The number of captured reads varied dramatically between W12 clones, which is reflective of the integrated viral genome copy, and also between replicates of the same clone indicating the importance of performing biological replicates. Combining replicate datasets resulted in removing library complexity; as such, the biological replicate of each clone with the greatest number of virus-host reads have been analysed in this thesis. Quality control checks throughout library preparation, and comparing the SCRiBL-Hi-C data for both replicates, indicate that the results obtained are very similar; however, the replicate with greater read depth enables more information and robust conclusions to be drawn. Moreover, for W12 clones G2 and D2 where both the Hi-C and SCRiBL-Hi-C datasets were obtained, the enrichment of HPV-containing reads was vast and ranged from 170-fold (W12 D2 rep II) to 320-fold (W12 G2 repI) (data not shown).

In W12 clones containing more than one copy of the virus genome, both short- (defined as <500 kbp) and long-range (>500 kbp) interactions between the integrated virus and regions of the host were identified. The significance of distal enhancer-

promoter contacts via chromatin looping has been demonstrated by the generation of artificially forced loops, which were found sufficient to activate gene expression highly[292]. The 3D interactions between HPV16 and the host genome were commonly associated with CTCF binding sites in the host. CTCF is a zinc-finger protein that binds to an insulator region in genomic DNA and plays a fundamental role in transcriptional regulation and controlling higher order chromatin structure; furthermore, CTCF proteins dimerise when bound to different DNA sequences, mediating long-range chromatin looping[293]. It has previously been shown that CTCF is recruited to the CTCF binding sites of the HPV18 genome[294]. HPV integration may therefore result in the insertion of an ectopic CTCF-binding site into the human genome, which is able to form interacting loops with pre-existing host CTCF binding sites in a similar manner to that observed as a result of HTLV-1 integration[295]. In W12 clone G2 all virus-host interactions occurred within a single chromatin interaction domain (TAD) with CTCF binding sites marking the TAD boundaries (IMR90, Dixon *et al.*, 2012[252]) (Figure 5.11); these data are in agreement similar observations made at the colorectal cancer risk loci 8q24.21[256]. It is likely that CTCF binds to the host genome at these sites and acts as an insulator preventing more distant virus-host interactions. Additionally, in W12 clone H HPV16 integrates into the host genome just outside a TAD boundary (Figure 5.13). In this instance integration resulted in a substantial deletion of the host genome. It is possible that the positioning of the TAD and associated boundary elements prevented an even greater deletion.

Integration in clone G2 results in the formation of a new 'viro-loop' — HPV16 genome to ARL15 intron 1. The interacting loop between the integrated HPV16 genome in clone G2 and the host genome within the first intron on gene ARL15 was shown to be a 'relevant interaction', which is defined as an interaction that co-localises at a significantly higher frequency that random control probes at a similar distance[256]; verification of this virus-host interaction by fluorescent in situ hybridisation (FISH) demonstrated that the HPV16-host long-range interaction brings the two genomic regions into closer physical proximity.

Analysis of RNA-seq data indicates that HPV16 integration results in disruption of host gene expression at least 2.5 Mb either side of the integration site; both in-

creases and decreases to host gene expression as a result of HPV integration have previously been reported but are limited to <1.8 Mb across the integration window[87]. Interestingly, although the vast majority of 3D interactions were restricted to within a TAD, changes to gene expression, as a result of HPV16 integration, were observed beyond TAD boundaries. This may be a result of dramatic increases in viral oncogene E6 and E7 expression, which, as previously mentioned, has far-reaching consequences and affects many cellular processes including the modulation of host gene expression[296]; E6/E7-mediated changes to a plethora of host transcription factors is also a likely reason for changes to host gene expression across a broad genomic locus. More recently, it has been shown that gene activation can occur via genetic alterations that disrupt insulated neighbourhoods (TADs) as a consequence of aberrant activation by enhancers that are normally located outside of the TAD[297, 298]. Therefore, it is possible that changes to the genomic sequence as a result of HPV16 integration, including deletion and focal amplification, could disrupt pre-existing TAD boundaries causing changes to host gene expression.

In cases where HPV16 integrated within the coding region of a host gene, the expression of that gene was consistently upregulated when compared with a 6-clone average; this indicates that the introduction of an additional virus promoter, and its associated regulatory/enhancer region (LCR), results in a greater level of transcription. Amplified gene expression may be caused by increased transcription factor-mediated RNAPII recruitment to the extra target sequence i.e. the p97 promoter[163]. It has previously been shown that HPV-host fusion transcripts containing sequences of known cellular genes that have both the viral and cellular sequences in sense orientation — as they are in W12 clone H — the viral sequence was spliced to a cellular exon sequence[299]; therefore, it is hypothesised that further analysis of the RNA-seq data at the exon level may indicate increased expression of specific exons downstream of the virus promoter illustrating that the virus promoter is able to contribute to host gene expression.

Data presented in this chapter also demonstrate how virus-host 3D interactions can affect host gene expression; the interaction between integrated HPV16 and ARL15 in clone G2 resulted in a significant increase in ARL15 expression (Figure

5.11 D). As previously discussed, this interaction is likely driven by virus and host CTCF dimerisation leading to changes in the nuclear architecture of the host in this region; it is hypothesised that in this situation the formation of a virus-host interacting region and the introduction of the virus promoter is sufficient to increase the transcription efficiency of host gene ARL15 as a result of increasing the local concentration of gene promoters[298]. These data indicate that influence of HPV integration can be exerted over greater distances by forming larger virus-host interactions than previously described[285].

Directly correlating HPV16 integration in a particular W12 clone with changes to host gene expression by comparing the expression of host genes at and around the integration site compared with random regions of the genome is an inadequate method of comparison due to natural fluctuations in gene expression due to copy number variations (CNVs) as well as chromatin structure affecting TF accessibility etc. In addition, analysis of the RNA-seq data illustrated that when comparing the W12 clones (A5, B, D2, G2, F, H, R) changes to gene expression occurred across the whole genome. To address this issue we analysed the variance of gene expression across the W12 clones and compared this to the chromosome in which the virus had integrated. This analysis indicated that although significant changes were found across the whole region (100 genes either side of the integration site), they were predominantly found at, or close to, the site of HPV integration. These findings are in keeping with the observation made by Ojesina *et al.* that gene expression levels at sites of HPV integration were significantly higher in tumours with HPV integration compared with the expression levels of the same genes across other tumours without integration at that site[97]. Interestingly, however, HPV16 integration into the host genome in W12 clone A5 resulted in the down regulation of the majority of host genes near the integration site (Figure 5.15). A gene of interest in this region is IL-8, a chemotactic factor for neutrophils and T-lymphocytes. IL-8 has been shown to be downregulated as a result of HPV16 E6 and E7 onocoprotein expression[300]. The downregulation of IL-8 expression may contribute to the ability of infected cells to avoid the host immune response[67].

Additionally, the variance in gene expression correlated with the gene expression

data indicating that in clone G2 the bin containing ARL15 (bin 3) — the protein-coding gene to which integrated HPV interacts with in 3D — was significant when compared to the other six W12 clones (Figure 5.16 B). Additionally, the analysis of the 3D interactions between the virus and the host illustrated that virus-host loops and resultant changes to host architecture occurred within a TAD (W12 D2 Figure 5.8 & 5.12, W12 G2 Figure 5.6 & 5.11); however, changes to host gene expression and indeed the variance in gene expression extend beyond these architectural boundaries perhaps as a consequence of broader epigenetic changes.

Finally, analysis of the RNA-seq data highlighted the presence of virus-host transcripts in G2 (Figure 5.18). Whilst it is known that splicing occurs from splice donor sites in the virus into slice acceptor sites in the host, these genomic positions occur within the amplified virus-host region previously described. As such, it is not possible to determine which of the integrated virus genomes and/or from how many genomes splicing occurs. Furthermore, virus-host fusion transcripts have also been detected originating from the splice donor site 880 within the virus E1 ORF[299]; this site is present in the integrated HPV genome in clone G2 and therefore represents an additional site of potential virus-host fusion transcript generation.

Overall, results presented in this chapter provide evidence that modifications to the host genome as a result of HPV integration that are present in advanced SCCs also occur in pre-malignant integrant cells derived in the absence of selective pressure and are therefore characteristic of all HPV integration events. Further work is needed to establish whether long-range virus-host interactions contribute to the growth advantage and selection of particular cells across the mixed population of a polyclonal SIL.

Chapter 6

Concluding discussion and future work

Cervical malignancy is the fourth most common cause of cancer-related deaths in women worldwide[1]. Despite recent advances in the prevention methods for cervical cancer, including implementation of population-wide cervical cytology screening tests and introduction of HPV vaccination, approximately 300,000 women still die as a result of cervical cancer every year[1]. As a result, there remains considerable interest in developing novel targeted therapeutics based on improved understanding of the biology of advanced cervical cancers. It is known that infection with HPV is a necessary cause of cervical cancer[15]. In addition, integration of HRHPV types into the host genome and subsequent deregulated expression of HPV oncogenes E6 and E7 disrupts cell cycle control resulting in genetic instability and cell transformation. As such, integration is known to be the major risk factor associated with disease progression[27]. However, relatively little is known about how particular cells containing integrated HPV gain a growth advantage and are selected over other cells with HPV integrated elsewhere in the genome. This thesis sought to determine epigenetic mechanisms by which the transcription of integrated HPV16 oncogenes as well as host genes are deregulated during the early stages of carcinogenesis and to elucidate differences between cells that have high levels of oncogene expression compared with those that express the virus genome at low levels; it is important to understand these processes as they may indicate new targets for silencing transcriptionally active HRHPV integrants.

The work described and discussed in Chapter 3 demonstrated that the level of HPV16 oncogene expression per integrated virus template and the selection of individual cells during cervical carcinogenesis is determined through multiple layers of epigenetic regulation of the integrated HRHPV genome. Using a panel of five W12 clones with significantly different E6/E7 expression per template and HPV16 genome copy number less than four, it was shown that cells with higher levels of virus expression per template were associated with increased levels of host post-translational modification (PTM) hallmarks of transcriptionally active chromatin and reduced levels of repressive marks. Additionally, it was shown that there was a greater abundance of the active/elongating form of the RNA polymerase-II en-

zyme (RNAPII-Ser2P) together with the components of the positive transcription elongation factor-b (P-TEFb), CDK9 and cyclin T1, responsible for Ser2 phosphorylation. Epigenetic regulation of the genome is highly complex; multiple mechanisms intertwine and impact on one another, which facilitates a robust and multifaceted transcriptional response. For example, in addition to relaxing chromatin structure permitting the binding of transcription factors (TFs) at gene promoters, histone acetylation — namely H3K27ac — plays an additional role in the regulation of RNAPII activation[301]. The combinatorial regulation of gene expression by PTMs and RNAPII function is reflected in the W12 clone system, as cells with high levels of HPV16 E6/E7 expression show significantly greater association of H3K27ac and RNAPII-Ser2P with the long control region (LCR), promoter (p97) and viral oncogenes of the integrated virus genome.

Epigenetic regulation of gene expression is a dynamic and reversible process; as such, there is a growing emphasis on using epigenetic therapies to reprogram neoplastic cells toward a normal state[302]. Epigenetic therapies currently in the clinic and at clinical trial have been designed to inhibit enzymes responsible for the writing and/or removal of specific marks that determine chromatin structure, namely DNA methyltransferases (DNMTs) and histone deacetylases (HDACs)[302]. My data demonstrate that enzymes responsible for the acetylation of histone tails (HATs: p300 and TIP60) as well as the transition of RNAPII from poised to actively elongating (CDK9) were functionally significant in the *in vitro* W12 cell system. Cells with higher HPV16 expression per template showed greater sensitivity to depletion and/or inhibition of HATs and CDK9, as well as reduced sensitivity to histone deacetylase inhibition; these data indicate that depending on acetylation levels and the cellular environment, the levels of transcription of the integrated HPV16 genome may be modulated. Indeed, the development of CDK9 inhibitors is currently an active area of investigation; a number of small molecule inhibitors of CDK9 have demonstrated potent anticancer activity against a number of different cell lines including cancer cells derived from the cervix[164].

Work presented in this thesis focussed on the impact of PTMs and associated writer enzymes and the role of RNAPII on the levels of transcription from the virus

promoter; however, further work carried out on the W12 clones (F, A5, D2, H and G2) revealed that other epigenetic mechanisms also contribute to the levels of integrated HPV16 expression. To further assess the level of chromatin accessibility, the positions of nucleosomes across the HPV16 long control region (LCR) and oncogenes E6 and E7 were assessed; nucleosome occupancy and positioning is dynamic and has a critical impact on gene expression and regulation[303]. Across the W12 clones the positions of nucleosomes were similar; however, clones with high expression per template showed greater amount of exogenously applied CpG methylation at the early promoter (p97) and directly after the transcription start site. This indicated a lower average occupancy of the nucleosomes in cells with high expression per template and therefore greater chromatin accessibility in this region[203].

Future work should include determining whether the epigenetic features of integrated HPV16 are acquired from the host genome at the site of integration or, conversely, whether chromatin is modified at the host sites as a result of HPV16 integration. Recent investigations into the virus and host epigenetic landscape following HPV16 integration in head and neck cancers (HNSCC) provides evidence suggesting that the epigenetic status of the virus is determined by the local chromatin environment of the host; levels of DNA methylation associated with the HPV16 genome were significantly altered as a result of integration and were reflective of the methylation status of the flanking host genome[304].

In addition to analysing epigenetic modifications made to the integrated virus chromatin that control the level of transcription from the virus promoter, work in this thesis aimed to evaluate the organisation of chromatin in three dimensions at the HPV integration locus to further elucidate mechanisms of virus and host gene regulation. The work presented in Chapter 4 describes how chromatin conformation capture techniques — SCRiBL Hi-C — can be adapted to generate a novel method to elucidate three-dimensional (3D) interactions between the integrated HPV16 genome and the host. Additionally, a capture sequencing experiment was designed and performed concurrently in order to accurately map the virus-host breakpoints in the W12 clones analysed.

Following next generation sequencing of W12 clone SCRiBL Hi-C and capture-seq DNA libraries, multiple analyses were conducted; these included virus-host breakpoint identification and characterisation of DNA sequence at HPV16 integration sites, the presence of virus-host interactions, nuclear architecture of the host and the effect of HPV16 integration on host gene expression — these data are presented and discussed in Chapter 5. Evaluation of the HPV16 integration sites in the W12 clones showed that the virus frequently integrates within the coding region of a host gene, consistent with data that indicates HPV16 integration sites in SCCs occur within host genes significantly more frequently than is expected[284]. Although HPV integration sites in SCCs have been found distributed widely across the human genome[282], studies have found that HPV commonly integrates into genomic 'hotspots' a number of which are at, or in close proximity to, common fragile sites (CFSs); a finding that has been replicated with oropharyngeal SCC tumour samples[305]. Additionally, data presented in this thesis are consistent with evidence from SCCs that demonstrates HPV16 preferentially integrated into regions of open and active chromatin[288, 289] and in regions that exhibit microhomology between the virus and host DNA sequences[291, 282, 87, 283]. Combined, these factors suggest that HPV genome integration is non-random[96, 88]. It has been hypothesised that binding of the episomal HPV genome to active gene promoters may be one method in which the virus ensures the viral genome is retained in transcriptionally active regions of the nucleus[306]. In particular, it has been shown that the episomal HPV genome is tethered to host chromatin at regions associated with known fragile sites[307]; tethering to these regions, via interaction with the host BRD4 protein, greatly increases the chances of integration at that site and is also a mechanism for which non-random areas of the genome are targeted for retrotransposon and retrovirus integration[308].

Analysis of the HPV integration sites across the five W12 clones showed that HPV16 integrates into the host genome via two distinct mechanisms. The first, and most frequent integration mechanism, resulted in looping of the virus genome and amplification of a portion of the host genome adjacent to the integrated virus. Integration via looping meant that in cells with virus copy number greater than one, only one 5' and 3' virus-host breakpoint exist. Both the looping mechanism of in-

tegration and the presence of redundant, identical breakpoints have previously been observed in advanced HPV-positive cancer cell lines and provide evidence that HPV directly promotes genomic instability[176]. Here my work demonstrates that a similar method of integration occurs in all cells regardless of whether they are selected during carcinogenesis. The second type of integration mechanism saw the HPV16 genome integrated directly into a host gene, MAPK10, resulting in a large deletion of host DNA; additionally, in clone H the virus and host sequences are both in the same orientation raising the possibility of coding virus-host fusion transcripts. These findings warrant further investigation; western blot analysis using a polyclonal antibody for MAPK10 would detect if, as a result of truncation, mRNAs of different lengths are successfully transcribed into proteins of different molecular weight. Additionally, protein expression analysis of host genes into which HPV16 has integrated could be used to determine whether the abundance of protein is reflective of the transcript levels determined by RNA-seq. Furthermore, in-depth analysis of the RNA-seq data at the gene exon level will provide a greater insight into the direct effect of HPV integration on the expression of host genes into which the virus has integrated.

Next generation sequencing (NGS) is far superior to earlier techniques employed to identify sites of HPV integration. In this investigation the virus-host breakpoints in the W12 clones differed to those previously identified[104]; in particular, it was found that clones F and A5 have identical sites of integration. In light of this, it is important to view recent meta-analyses of HPV integration sites with uncertainty, as it is likely that a reasonable number of integration sites included in the analysis are incorrect having been previously identified using inadequate technologies[284, 289]. In the future it would be prudent to perform a similar capture sequencing reaction using similar methodology employed in this thesis to accurately identify virus-host breakpoints across a large number of W12 integrant clones to fully understand and exploit the unique model system.

Initial investigations into the potential role of host gene expression on cell selection included analysis of RNA-seq data for each of the W12 clones. Analysis in this thesis sought to determine the effect of HPV16 integration on gene expression locally, therefore resultant changes to host gene expression 2.5 Mb either side of the

integration site were determined. These data indicated that HPV16 integration results in significant disruption of host gene expression across the entire 5 Mb window. In addition, in cases where HPV16 integrated within the coding region of a gene, the expression of that gene was significantly upregulated. These data are consistent with studies that have demonstrated host gene expression in SCCs is modulated as a result of HPV16 integration[87, 296, 96].

To determine whether the disruption of host genes contribute to early cell selection events, *in vitro* gene function experiments should be carried out comparing W12 clones with similar levels of HPV16 E6/E7 oncoproteins but significantly different cell growth rates[183]. Employing the recently emerged CRISPR-Cas9 (clustered, regularly interspaced, short palindromic repeats (CRISPR)-CRISPR-associated protein 9) genome editing technologies would enable efficient and precise genetic manipulation of genomes. Specifically, a CRISPR-Cas9 system could be designed and used to knock out amplified or over expressed host genes at the site of HPV16 integration or genes which interact with the virus in 3D, identified by SCRiBL Hi-C. Based on data from this thesis, upregulated host gene ARL15 (ADP Ribosylation Factor Like GTPase 15), which contacts the integrated virus in 3D in clone G2, represents a good candidate to perform functional experiments. Phenotypic effects of host gene modulation should be quantified and could include proliferation rates, cell cycle assays (both via FACS), clonogenicity, migration and invasion. Alternatively the CRISPR/Cas9 system could be altered by redesigning the guide RNA (gRNA) to target the HPV16 genome; gene expression analysis following resultant knockout of the HPV16 DNA fragment would determine whether removal of the virus promoter/enhancer from the genomic region affects host gene expression. Recently, similar methodology was used to demonstrate that removal of the HPV18 genome reduced *MYC* gene expression by approximately 30%[309].

Future study of the effect of HPV and host gene expression on cell selection could also involve epigenome editing using nuclease-null or 'dead' Cas9 (dCas9) CRIPSR-Cas9 technologies. dCas9 is generated by introducing point mutations in the nuclease domains of Cas9 and its use permits targeting specific DNA loci without cleavage[310]. dCas9 can subsequently be fused to an epigenetic effector; previous studies have

shown that dCas9 tethered to a HAT results in transcriptional activation[311] whilst gene silencing has been caused as a result of dCas9 fused to the Krupel-associated box (KRAB) repressor[312]. To induce gene specific silencing, promoter sites could be artificially supressed by direct targeting of dCas9 fused to the catalytic domain of a DNA methyltransferase (e.g. DNMT3A)[313, 314]. Again, both the HPV16 genome — specifically the viral promoter — as well as the promoters of interacting host genes identified by SCRiBL Hi-C, i.e. ARL15, could be targeted using this technology.

Following the discovery of a long-range interaction (508 kb) between integrated HPV18 and the *MYC* gene locus and associated increases in *MYC* gene expression in the advanced carcinoma HeLa cell line[285], this thesis sought to determine whether such long-range genomic interactions are common in SCC cells and, additionally, whether virus-host interactions exist prior to cell selection in premalignant lesions. Analysis of the SCRiBL Hi-C libraries revealed that W12 clones that contain more than one copy of the virus genome formed both *cis* short- and long-range HPV16-host interactions; the significant virus-host long-range (~1Mb) interaction with ARL15 in clone G2 was particularly informative, indicating not only that virus-host interactions are a feature of HPV integration, but also that 3D interactions from the virus can significantly modulate host gene expression. This long-range interaction may be mediated by CTCF dimerisation and results in changes to the nuclear architecture of the host in this region. The association between genome topology and transcriptional activity has been widely reported, with the disruption of topologically-associating domains (TADs) linked to ectopic and/or aberrant enhancer expression causing disease[315]. It would be interesting to determine whether knockout of either of the CTCF binding sites facilitating the virus-host interaction — either the ectopic putative site within HPV16 genome or the individual CTCF site within ARL15 — using CRISPR/Cas9 technology would result in decreased expression of ARL15 and/or whether the virus genome forms different long-range interactions within the TAD, as has been demonstrated[315].

In the future, the same SCRiBL Hi-C analysis should be performed in the outgrowth clone W12Ser2p31, in which the HPV genome integrates into the 8q24.21

genomic locus. It would be particularly interesting to see whether, as in HeLa cells, integrated HPV16 forms a long-range interaction to the proto-oncogene *MYC*; evaluating the 3D interactions of the outgrowth clone may indicate whether this plays a role in the selection of individual integrant clone following long-term culture.

In depth analysis of the 3D structure of the nucleus and how chromatin organisation affects gene regulation is a rapidly developing area of research, particularly in elucidating structural changes and disruption that result in disease. Despite the recent developments of 3C technologies in the last fifteen years — 3C[236], 4C[242], 5C[244], Hi-C[252, 246, 251] and ChIA-PET[316] — alternative methods are already being generated to improve existing methods used to evaluate nuclear architecture. 3C-based approaches have technical limitations; methods are reliant on restriction enzyme digest and subsequent ligation to capture interactions between DNA segments, as such, biases in GC content, protein occupancy and restriction site density occur. One such method redesigned the Capture-C protocol using DNA rather than RNA biotinylated oligonucleotides; this has led to increased sensitivity, identifying both weak *cis* and *trans* interactions, and superior efficiency — multiple independent 3C libraries from different samples can be processed in a single reaction, minimizing experimental variation and allowing for precise comparison of chromosome conformations in different cell types[317]. Even more recently, a different approach to deciphering nuclear architecture is genome architecture mapping (GAM); GAM is the first genome-wide method for capturing 3D proximities between any number of genomic loci without ligation and measures 3D distances by combining ultrathin cryosectioning with laser microdissection and DNA sequencing[318]. Moreover, to fully appreciate the dynamics of nuclear organisation, it will be important to combine population-based models with data on individual cells at high resolution; this will require a multidisciplinary approach bringing together genomics, biophysics and imaging. The 4D Nucleome Program (NIH) is one initiative that seeks to combine expertise in the aforementioned fields to understand: the principles underlying nuclear organization in space and time; the role nuclear organization plays in gene expression and cellular function; and how changes in nuclear organization affect normal development and differenti-

ation as well as various diseases. In the future it would be beneficial to review and possibly implement some of the emerging technologies being used to evaluate the nuclear architecture of cells containing integrated HPV16 at even greater resolution.

The findings presented in this doctoral thesis further our understanding of the impact of HPV integration during early cervical carcinogenesis. To date, the majority of studies have focussed on the biological properties of advanced SCCs — the endpoint of the clonal selection process — and have not addressed the dynamic changes that underpin progression from pre-malignancies to carcinomas. Results presented here illustrate that many of the virus and host alterations seen in advanced SCCs brought about as a consequence of HPV integration are characteristic of all integration events and are present in cells regardless of whether they are ultimately selected during carcinogenesis. This thesis forms the foundation for future work that may elucidate potential therapeutic targets for epigenetic therapies in cervical SCCs containing integrated HRHPV.

Chapter 7

Appendix 1

Key:
NNNNNN: Nucleotide base change
NNNNNN: Nucleotide deletion in ~1/7 copies of viral genome
NNNNNN: Nucleotide deletion
~~NNNNNN~~: Unread sequence

```
W12 E   ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAGCTACACCCACC
Par 1   ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAA~~GCTACACCCACC
Par 2   ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAA~~AACGAAAAGCTACACCCACC
W12 F   ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAGCTACAC~~CCACC
W12 A5  ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGA~~AAAGCTACACCCACC
W12 D2  ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAGCTACACCCACC
W12 H   ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAGCTACA~~CCCACC
W12 J3  ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAG~~CTACACCCACC
W12 R2  ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAGCTACACCCACC
W12 Q   ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAGCTACA~~CCCACC
W12 H2  ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAGC~~TACACCCACC
W12 G2  ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGA~~AAAGCTACACCCACC
W12 S2  ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAGCTACA~~CCCACC
W12 3   ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAGCTACA~~CCCACC
W12 E3  ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAGCTACACCCACC~~
SiHa    ~~ACAAGCAGGATTGAAGGCCAAACCAAAATTTACATTAGGAAAACGAAAAGCTACACCCA~~CC

W12 E   ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
Par 1   ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
Par 2   ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 F   ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 A5  ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 D2  ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 H   ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 J3  ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 R2  ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 Q   ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 H2  ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 G2  ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 S2  ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 3   ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
W12 E3  ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT
SiHa    ACCTCATCTACCTCTACAACTGCTAAACGCAAAAAACGTAAGCTGTAAGTATTGTATGTAT

W12 E   GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
Par 1   GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
Par 2   GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 F   GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 A5  GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 D2  GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 H   GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 J3  GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 R2  GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 Q   GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 H2  GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 G2  GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 S2  GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 3   GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
W12 E3  GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
SiHa    GTTGAATTAGTGTTGTTTGTTGTTTATATGTTTGTATGTGCTTGTATGTGCTTGTAAATAT
```

```
W12 E   TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
Par 1   TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
Par 2   TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 F   TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 A5  TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 D2  TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 H   TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 J3  TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 R2  TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 Q   TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 H2  TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 G2  TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 S2  TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 3   TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
W12 E3  TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG
SiHa    TAAGTTGTATGTGTGTTTGTATGTATGGTATAATAAACACGTGTGTATGTGTTTTTAAATG

W12 E   CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
Par 1   CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
Par 2   CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 F   CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 A5  CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 D2  CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 H   CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 J3  CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 R2  CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 Q   CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 H2  CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 G2  CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 S2  CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 3   CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
W12 E3  CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG
SiHa    CTTGTGTAACTATTGTGTCATGCAACATAAATAAACTTATTGTTTCAACACCTACTAATTG

W12 E   TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
Par 1   TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
Par 2   TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 F   TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 A5  TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 D2  TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 H   TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 J3  TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 R2  TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 Q   TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 H2  TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 G2  TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 S2  TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 3   TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
W12 E3  TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
SiHa    TGTTGTGGTTATTCATTGTATATAAACTATATTTGCTACATCCTGTTTTTGTTTTATATAT
```

```
W12 E   ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
Par 1   ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
Par 2   ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 F   ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 A5  ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 D2  ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 H   ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 J3  ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 R2  ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 Q   ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 H2  ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 G2  ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 S2  ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 3   ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
W12 E3  ACTAAATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT
SiHa    ACTATATTTTGTAGCGCCAGCGGCCATTTTGTAGCTTCAACCGAATTCGGTTGCATGCTTT

W12 E   TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
Par 1   TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
Par 2   TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 F   TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 A5  TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 D2  TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 H   TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 J3  TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 R2  TTGGCACAAAATGTGTTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 Q   TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 H2  TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 G2  TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 S2  TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 3   TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
W12 E3  TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC
SiHa    TTGGCACAAAATGTGTTTTTTTAAATAGTTCTATGTCAGCAACTATAGTTTAAACTTGTAC

W12 E   GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
Par 1   GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
Par 2   GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 F   GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 A5  GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 D2  GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 H   GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 J3  GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 R2  GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 Q   GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 H2  GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 G2  GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 S2  GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 3   GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
W12 E3  GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
SiHa    GTTTCCTGCTTGCCATGCGTGCCAAATCCCTGTTTTCCTGACCTGCACTGCTTGCCAACCA
```

```
W12 E   TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
Par 1   TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
Par 2   TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 F   TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 A5  TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 D2  TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 H   TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 J3  TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 R2  TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 Q   TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 H2  TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 G2  TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 S2  TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 3   TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
W12 E3  TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT
SiHa    TTCCATTGTTTTTTACACTGCACTATGTGCAACTACTGAATCACTATGTACATTGTGTCAT

W12 E   ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
Par 1   ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
Par 2   ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 F   ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 A5  ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 D2  ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 H   ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 J3  ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 R2  ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 Q   ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 H2  ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 G2  ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 S2  ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 3   ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
W12 E3  ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT
SiHa    ATAAAATAAATCACTATGCGCCAACGCCTTACATACCGCTGTTAGGCACATATTTTTGGCT

W12 E   TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
Par 1   TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
Par 2   TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 F   TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 A5  TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 D2  TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 H   TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 J3  TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 R2  TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 Q   TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 H2  TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 G2  TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 S2  TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 3   TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
W12 E3  TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
SiHa    TGTTTTAACTAACCTAATTGCATATTTGGCATAAGGTTTAAACTTCTAAGGCCAACTAAAT
```

```
W12 E   GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
Par 1   GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
Par 2   GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 F   GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 A5  GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 D2  GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 H   GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 J3  GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 R2  GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 Q   GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 H2  GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 G2  GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 S2  GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 3   GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
W12 E3  GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC
SiHa    GTCACCCTAGTTCATACATGAACTGTGTAAAGGTTAGTCATACATTGTTCATTTGTAAAAC

W12 E   TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
Par 1   TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
Par 2   TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 F   TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 A5  TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 D2  TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 H   TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 J3  TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 R2  TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 Q   TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 H2  TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 G2  TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 S2  TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 3   TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
W12 E3  TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT
SiHa    TGCACATGGGTGTGTGCAAACCGTTTTGGGTTACACATTTACAAGCAACTTATATAATAAT

W12 E   ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
Par 1   ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
Par 2   ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 F   ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 A5  ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 D2  ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 H   ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 J3  ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 R2  ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 Q   ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 H2  ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 G2  ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 S2  ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 3   ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
W12 E3  ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
SiHa    ACTAAACTACAATAATTCATGTATAAAACTAAGGGCGTAACCGAAATCGGTTGAACCGAAA
```

```
W12 E   CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
Par 1   CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
Par 2   CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 F   CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 A5  CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 D2  CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 H   CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 J3  CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 R2  CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 Q   CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 H2  CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 G2  CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 S2  CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 3   CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
W12 E3  CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC
SiHa    CCGGTTAGTATAAAAGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCC

W12 E   ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
Par 1   ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
Par 2   ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 F   ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 A5  ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 D2  ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 H   ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 J3  ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 R2  ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 Q   ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 H2  ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 G2  ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 S2  ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 3   ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
W12 E3  ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT
SiHa    ACAGGAGCGACCCAGAAAGTTACCACAGTTATGCACAGAGCTGCAAACAACTATACATGAT

W12 E     ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
Par 1     ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
Par 2     ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 F     ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 A5    ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 D2    ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 H     ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 J3    ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 R2    ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 Q     ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 H2    ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 G2    ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 S2    ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 3     ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
W12 E3    ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
SiHa      ATAATATTAGAATGTGTGTACTGCAAGCAACAGTTACTGCGACGT
```

Chapter 8

Appendix 2

OPEN

Oncogene (2016), 1–14

www.nature.com/onc

ORIGINAL ARTICLE

HPV16 oncogene expression levels during early cervical carcinogenesis are determined by the balance of epigenetic chromatin modifications at the integrated virus genome

IJ Groves, ELA Knight, QY Ang, CG Scarpini and N Coleman

In cervical squamous cell carcinomas, high-risk human papillomavirus (HRHPV) DNA is usually integrated into host chromosomes. Multiple integration events are thought to be present within the cells of a polyclonal premalignant lesion and the features that underpin clonal selection of one particular integrant remain poorly understood. We previously used the W12 model system to generate a panel of cervical keratinocyte clones, derived from cells of a low-grade premalignant lesion naturally infected with the major HRHPV type, HPV16. The cells were isolated regardless of their selective advantage and differed only by the site of HPV16 integration into the host genome. We used this resource to test the hypothesis that levels of HPV16 E6/E7 oncogene expression in premalignant cells are regulated epigenetically. We performed a comprehensive analysis of the epigenetic landscape of the integrated HPV16 DNA in selected clones, in which levels of virus oncogene expression per DNA template varied ~6.6-fold. Across the cells examined, higher levels of virus expression per template were associated with more open chromatin at the HPV16 long control region, together with greater loading of chromatin remodelling enzymes and lower nucleosome occupancy. There were higher levels of histone post-translational modification hallmarks of transcriptionally active chromatin and lower levels of repressive hallmarks. There was greater abundance of the active/elongating form of the RNA polymerase-II enzyme (RNAPII-Ser2P), together with CDK9, the component of positive transcription elongation factor b complex responsible for Ser2 phosphorylation. The changes observed were functionally significant, as cells with higher HPV16 expression per template showed greater sensitivity to depletion and/or inhibition of histone acetyltransferases and CDK9 and less sensitivity to histone deacetylase inhibition. We conclude that virus gene expression per template following HPV16 integration is determined through multiple layers of epigenetic regulation, which are likely to contribute to selection of individual cells during cervical carcinogenesis.

Oncogene advance online publication, 15 February 2016; doi:10.1038/onc.2016.8

INTRODUCTION

Infection with high-risk human papillomavirus (HRHPV) is responsible for over 600 000 new cancers per annum, including over 500 000 carcinomas of the cervix.[1] The majority of cervical malignancies are squamous cell carcinomas (SCCs), which arise from a mixed population of HRHPV-infected cells by clonal selection of cells with the greatest competitive growth advantage.[2,3] In ~85% of cervical SCCs the selected cells contain HRHPV DNA that is integrated into host chromosomes. In the remaining ~15% of cases the virus genome remains in the extrachromosomal (episomal) state, as is also seen in the normal virus lifecycle.[4–6]

In the squamous epithelial lesions that result from productive HRHPV infections, there are ~100 virus episome copies in each basal layer cell.[7,8] In the lower cell layers, the necessary expression of the HRHPV early genes E6 and E7 occurs through transcriptional initiation at the virus early promoter (p97 in the case of the major HRHPV, HPV16), while cell maturation is associated with activation of the virus late promoter (p670 for HPV16) and expression of late virus genes. These events are linked to changes in transcription factor binding and altered chromatin structure, based on histone post-translational modifications (PTMs) at nucleosomes associated with the HRHPV genome.[3,9–13]

Integration of HRHPV genomes is thought to occur in premalignant squamous intraepithelial lesions (SILs). The probability of integration increases with time[14] and multiple integration events are thought to be present across the cells of a polyclonal SIL. However, relatively little is known about how particular cells containing integrated HPV gain a growth advantage over other cells with HPV integrated elsewhere in the genome. Notably, the significance of virus transcriptional deregulation in individual integrants during these early events in cervical carcinogenesis is poorly understood. Most studies to date have concerned the end point of the clonal selection process, by focusing on the virus integrants seen in the SCC cells themselves, and have not addressed the dynamic changes that underpin progression from SILs to carcinomas. It is difficult to study such processes by cross-sectional analysis of clinical samples, as the key events that precede clonal selection early in cervical carcinogenesis occur in the basal epithelial cells of low-grade SILs (LSILs),[4,15] which would need to be isolated by tissue micro-dissection. A more informative approach has been to study experimental *in vitro* models, including W12.

The W12 system was developed from a polyclonal culture of cervical squamous cells (keratinocytes) naturally infected with HPV16, which were derived by explant culture of a cervical LSIL.[7]

Department of Pathology, University of Cambridge, Cambridge, UK. Correspondence: Professor N Coleman, Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge, Cambs CB2 1QP, UK.
E-mail: nc109@cam.ac.uk
Received 9 September 2015; revised 14 December 2015; accepted 29 December 2015

At early passages, these 'parental' W12 cells are phenotypically and genetically stable. They allow maintenance of HPV16 episomes at ~100 copies per cell and recapitulate an LSIL in three-dimensional organotypic culture. Following long-term culture of W12, however, the cells lose these properties and closely mirror the virus and host events associated with cervical carcinogenesis *in vivo*, with phenotypic progression of the reformed epithelia to high-grade SIL and then SCC.[4] These events may be associated with deregulation of episome numbers and transcriptional control (e.g. W12 series 4 and W12E cells).[2,6] More typically, however, there is a change in the virus physical state from episomal to integrated, due to loss of trans-repressive episomes and emergence of a clonal population containing the HPV16 integration event that confers the greatest growth advantage.[16] In different W12 series, different integration sites are seen in the selected cells.[16]

We previously used limiting dilution cloning of polyclonal parental W12 cells at early passage, to sample the range of integration events that exists prior to episome clearance and integrant emergence.[14] The cells were selected under non-competitive conditions, allowing isolation of clones regardless of whether they had a selective advantage in mixed cell populations. By this method, we derived a series of clones from an identical genetic background that differed only by the site of HPV16 integration into the host genome. The large majority of clones showed no evidence of full-length HPV16 concatemerisation[17] and were therefore so called type I integrants.[2] Several clones contained multiple copies of the E6/E7 oncogenes, consistent with local DNA rearrangements following integration. At the early passages examined (i.e. prior to clonal evolution events), all clones recapitulated a premalignant SIL phenotype in organotypic tissue culture, with no evidence of invasiveness.[17]

The W12 clones therefore represent a unique system to examine the host and virus factors that determine selection of a particular HPV16 integrant from the range that exists in a typical polyclonal population of premalignant cervical keratinocytes. Across 17 representative clones analysed, levels of HPV16 E6 and E7 transcripts per cell varied by ~6-fold and correlated closely. Only seven of the clones analysed (41%) showed significantly greater expression of HPV16 E6 and E7 than the episome-containing LSIL-like cells from which they were derived, indicating that HPV integration *per se* does not necessarily lead to increased levels of virus oncogenes per cell.[17] Interestingly, levels of E6/E7 transcript per DNA template across the clones varied by ~16-fold.[17]

In the present study, we used the W12 clones to investigate how different HPV16 integration events in basal-type premalignant cervical keratinocytes lead to different levels of virus oncogene expression. In order to provide a tractable system for our experiments, we chose cells without full-length HPV16 concatemerisation and with four or less copies of integrated virus DNA per cell. Of the five such clones available, two (F and A5) showed high levels of E6/E7 expression per template, two (D2 and H) showed medium levels and one (G2) showed low levels, with ~6.6-fold variation in expression levels across the five clones (Table 1). In our previous preliminary analysis of a restricted sequence of the HPV16 genome in the five clones,[17] we found that levels of HPV16 expression per template were associated with different distributions of a selected small number of histone PTMs.[17,18] We therefore hypothesised that variation in levels of expression per DNA template following HPV16 integration were due to epigenetic differences in the virus chromatin. We used the W12 clones to undertake a detailed and extensive analysis of the epigenetic landscape on the integrated HPV16 genome, focussing on the relationships between virus oncogene expression per template and chromatin accessibility, histone PTMs and activity of RNA polymerase-II (RNAPII).

RESULTS

HPV16 oncogene expression per template associates with accessibility of virus chromatin

No mutations were seen in any of the five W12 clones following PCR amplification and sequencing of the HPV16 long control region (LCR) (data not shown). By formaldehyde-assisted isolation of regulatory elements, enrichment of open chromatin (i.e. with lower nucleosome occupancy) at the HPV16 LCR and early promoter was greatest in cells with high levels of virus gene expression per template (F and A5) and showed progressive reductions through cells with medium expression per template (D2 and H) to cells with low expression per template (G2) (Figure 1a). The positions of nucleosomes, as indicated by nucleosome occupancy and methylome sequencing, were similar across all clones and usually spaced 150–200 bp apart. These positions were indicated by low levels of exogenously applied GpC methylation. However, clones with high expression per template showed greater amounts of exogenously applied GpC methylation at the early promoter and directly after the transcription start site (Figure 1b), indicating lower average occupancy of the nucleosomes and therefore greater chromatin accessibility in this region. Cells with higher virus expression levels per template also showed a greater abundance of the ATP-dependent chromatin remodelling enzymes BRG1 and INI1 across the virus genome (Figures 1c and d), in keeping with greater openness/accessibility of the HPV16 chromatin in these cells. Both enzymes were most abundant over the virus early region, including the early and late promoters. There was a striking peak

Table 1. Details of the W12 clones studied

Clone	*Integration site*	*Ploidy*	*HPV16 gene copy number*					*HPV16 E6/E7 expression per template*	*Expression per template category*
			E6	*E7*	*Mean E6/E7*	*E2-5′*	*E2-3′*		
F	4q13.3	2N	1	1	1	1	1	248.6 (±31.8)	HIGH
A5	8p11.21	2N	1	1	1	1	1	215.6 (±14.9)	
D2	18q21.2	2N	3	4	4	0	3	118.5 (±12.0)	MEDIUM
H	4q21.23	2N	1	1	1	0	1	100.1 (±12.4)	
G2	21q22.1	2N	3	3	3	3	0	37.5 (±4.2)	LOW

All virus gene copy numbers were adjusted for cell ploidy and rounded to the nearest whole number. Levels of HPV16 E6 and E7 transcripts per template were referenced individually to low passage episome-containing W12 cells (W12 Series6 p11) and mean values (±s.e.m.) were determined from three biological replicates. Clone G2 showed three different virus–host junction transcripts by RNA-sequencing and clone D2 showed four different virus–host junction transcripts (data not shown). All clones tested (F, A5, D2 and G2) reformed an LSIL in organotypic tissue culture.

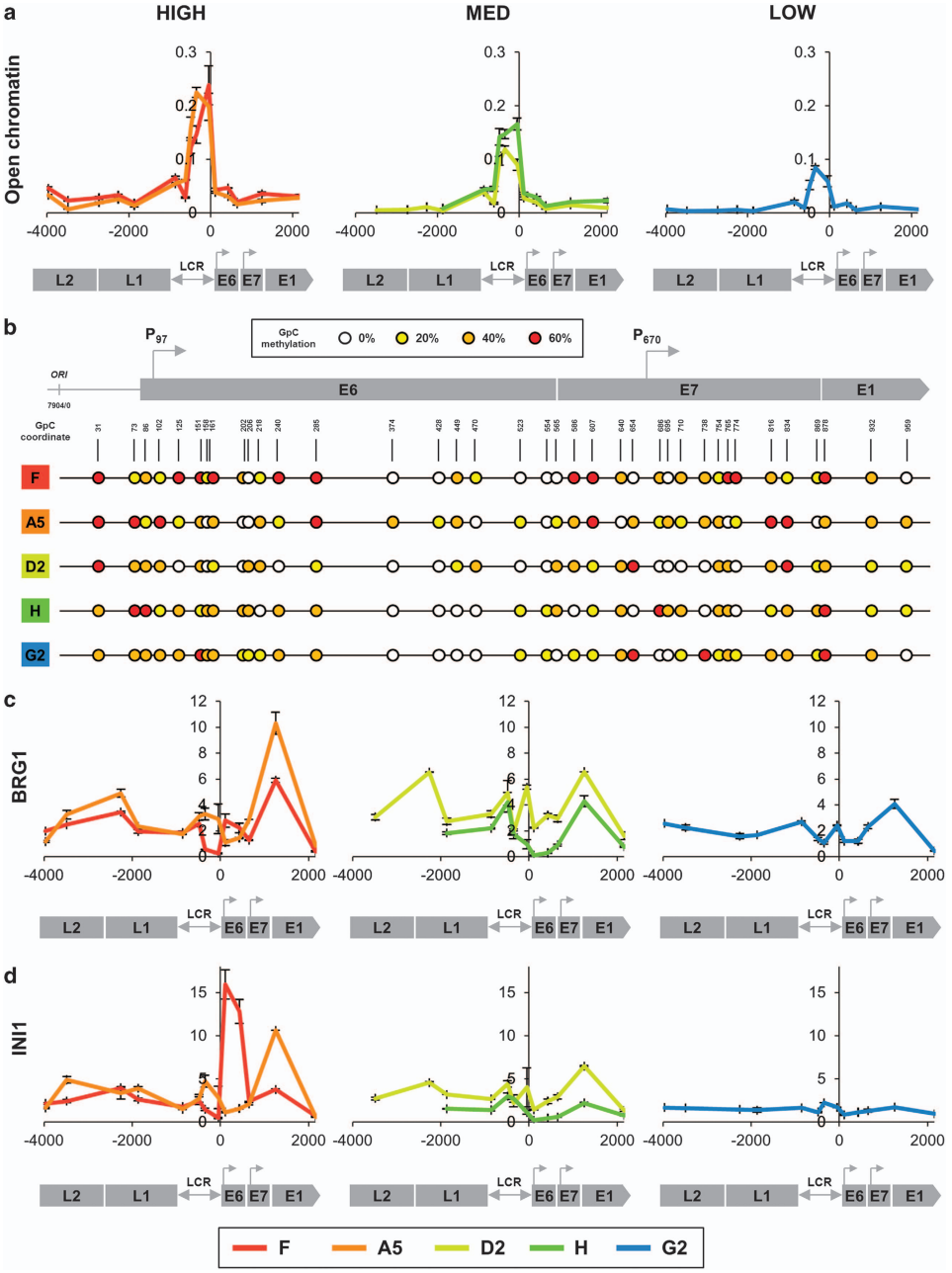

Figure 1. Levels of HPV16 transcription per template associate with virus genome accessibility. (**a**) In each graph the *y*-axis shows fold enrichment of open chromatin across the HPV16 genome, as determined by FAIRE using three biological replicates. Values were normalised to the efficiency of enrichment, as determined by the ratio of *GAPDH* promoter to *GAPDH* open reading frame qPCR. The *x*-axis and underlying schematic show the region of the HPV16 genome analysed. The panels show data for the clones in which transcription levels per template were high, medium (MED) or low. (**b**) Virus genome occupancy by nucleosomes or other DNA-binding proteins, as determined by NOMe sequencing using four biological replicates. Regions with a lower rate of occupancy are indicated by higher levels of exogenously applied GpC methylation. The degree of GpC methylation is shown as a heat map (see key), with circles at individual nucleotide positions. (**c**, **d**) Association of chromatin remodelling enzymes BRG1 and INI1 with the integrated HPV16 genome across the cell lines. The *y*-axis shows relative levels of enrichment of BRG1 (**c**) and INI1 (**d**), derived from three biological replicates in each case and normalised to host control target regions (see Supplementary Table S1). The *x*-axis and underlying schematic show the region of the HPV16 genome analysed. In all panels, data for each of the five clones are colour coded according to the key at the foot of the figure. This code is maintained in all subsequent figures. In all panels, bars = mean ± s.e.m. Abbreviation: FAIRE, formaldehyde-assisted isolation of regulatory element; NOMe, nucleosome occupancy and methylation.

of INI1 abundance at the early promoter and transcription start site in clone F.

High HPV16 expression requires activating chromatin marks

We next quantified levels of histone PTMs on the integrated virus chromatin. Higher virus expression per template was associated with greater levels of histone 3 lysine 4 tri-methylation (H3K4me3), a hallmark of transcriptional activity. We extended our previous observations[17] by showing that this mark was only present downstream of the HPV16 early promoter and was absent from the late region (Figure 2a). Key enzymatic writers of the mark, SETD1A and MLL1, were also more abundant at the virus genome in cells with higher expression per template, with consistent enrichment at the LCR and early promoter (Figures 2b and c). There was more variable enrichment over the late and early regions, with higher levels of SETD1A in clones A5 and D2 and MLL1 in A5. The cells with high expression per template also showed enrichment of histone PTMs associated with gene enhancer/promoter regions, with strong enrichment of H3K4me1 across the entire virus genome, including the late genes (Figure 2d), and greater abundance of H3K27ac, predominantly at the LCR and early genes (Figure 2e).

Conversely, lower levels of expression per template were associated with higher levels of repressive histone PTMs, namely di-methylation of histone 3 lysines 9 and 27 (H3K9me2 and H3K27me2) (Supplementary Figures S1A and C). However, there was very little enrichment of tri-methylated forms of these histones (H3K9me3 and H3K27me3) at any of the integrated HPV16 genomes (Supplementary Figures S1B and D). The cells with lower expression also showed higher levels of endogenous CpG DNA methylation across the HPV16 genomic region analysed (nt 6731 to 1287) (Figure 3a). There was a prominent peak of DNA methylation at the LCR in G2, the only clone showing low expression per template (Figure 3b). Levels of methylation at *L1* were variable, including between clones with similar levels of virus gene expression per template (Figure 3b).

HPV16 transcription per template associates with histone methylation modifying enzymes

Higher virus expression per template was associated with higher levels of general histone 3 acetylation (H3ac) (Figure 4a), together with greater abundance of the histone acetyltransferases (HATs) p300 and TIP60, across the entire HPV16 genome (Figures 4b and c). The abundance of these enzymes at the HPV16 genome showed no relation to total levels in the cells, indicating specific loading onto the virus chromatin (Supplementary Figure S2). High levels of p300 across the HPV16 genome were associated with high overall abundance of cJun (Supplementary Figure S3A), which can act as a p300 recruiter protein.[19] However, there was no close association between levels of p300 and cJun at individual sites on the virus genome. Levels of TIP60 were not associated with those of its potential recruiter protein YY1 (Supplementary Figure S3B), but did associate closely with levels of H3K4me1 (Figure 2d).

We tested the functional significance of HAT recruitment in determining levels of HPV16 transcript expression. We inhibited p300 or TIP60 in the cells with the highest and lowest levels of virus early gene expression per template (clones F and G2, respectively) (Figures 4d–k). We did not examine post-transcriptional effects on HPV16 oncoprotein levels in these experiments. We observed significantly greater reductions in E6/E7 transcript levels in clone F vs G2 when p300 was depleted using siRNA (Figure 4h) or specifically inhibited using C646 (Figure 4j) and when TIP60 was depleted using siRNAs (Figure 4i) or specifically inhibited using MG149 (Figure 4k).

Mirroring these observations with HATs, cells with lower virus transcript levels per template showed higher abundance of histone deacetylase 1 (HDAC1) (Figure 5a). In the absence of specific siRNAs targeting HDAC1, we used the class I/II HDAC-specific small-molecule inhibitor Trichostatin-A. After 16 h of treatment, this produced significantly greater increases in HPV16 E6/E7 transcript levels in clone G2 than in clone F (Figure 5b).

Transcript levels per template associate with active RNAPII, determined by P-TEFb (CDK9)

There were no differences across the clones in the overall amounts of RNAPII associated with the HPV16 genome (Figure 6a). However, cells with lower virus expression per template showed higher amounts of the poised/paused or stalled form of RNAPII, Ser5P, across the early genes (Figure 6b). Conversely, cells with higher expression per template showed greater amounts of the active/elongating form of RNAPII, Ser2P, across the virus LCR and early genes (Figure 6c), together with higher levels of histone 3 lysine 36 tri-methylation (H3K36me3), a histone PTM associated with transcriptional elongation (Figure 6d). There were also higher levels of the positive transcription elongation factor b (P-TEFb) complex kinase CDK9 (Figure 7a), which is responsible for phosphorylation of the RNAPII C-terminal domain at Ser2. Depletion of CDK9 (Figures 7b and c) produced significantly greater reductions in E6/E7 transcript levels in clone F (higher expression per template) than in clone G2 (lower expression per template) (Figure 7d) (F vs G2 $P < 0.001$).

We next investigated the consequences of inhibiting CDK9 function in high expressing clone F cells. As the chromatin yield from siRNA experiments was too low, we used the small-molecule inhibitor Flavopiridol, which caused 87% reduction in E6/E7 transcript levels (Supplementary Figure S4A). Similar effects were also seen using other small inhibitors with predominant specificity for CDK9 (Supplementary Figure S4A). While Flavopiridol produced no change in overall levels of CDK9 recruitment at the HPV16 genome (Figure 8a), there were reduced levels of total RNAPII, particularly downstream of the virus early promoter (Figure 8b). There was also less elongating RNAPII-Ser2P downstream of the transcription start site, with evidence of redistribution to the LCR/early promoter region (Figure 8c). In addition, the LCR and early genes showed striking decreases in the histone PTM mark of transcriptional activation, H3K4me3 (Figure 8d), mirrored by increases in the mark of constitutive heterochromatin and transcriptional repression, H3K9me2 (Figure 8e).

Similar observations to those made in clone F were seen in the cervical SCC cell line SiHa. Indeed, the CDK9 inhibitors (including Flavopiridol) produced greater reductions in E6/E7 expression in SiHa than in clone F (95%) (Supplementary Figure S4A), while Flavopiridol led to more pronounced shifts in epigenetic marks and reduced RNAPII-Ser2P levels (Figures 8f–j). Transcript levels in SiHa reduced by $>80\%$ over the first 8 h of Flavopiridol treatment (Supplementary Figure S4B), consistent with profound transcriptional shut off of the integrated HPV16 DNA. These changes were associated with complete inhibition of cell growth (Supplementary Figure S4C).

DISCUSSION

The W12 cell clones represent a unique resource that has enabled us to study the factors associated with the large differences in virus oncogene expression per template observed following natural HPV16 integration events in premalignant basal cervical keratinocytes. We focussed on five clones from the same genetic background, in which HPV16 was integrated at low copy number without full-length virus concatemers. The cells were studied at a very early stage after cloning, when levels of E6/E7 varied by ~6.6-fold but the cells had not shown the effects of HPV16 oncoprotein-driven genomic instability and still recapitulated an LSIL phenotype in organotypic tissue culture. Our findings indicate that levels of HPV16 expression following integration are determined through multiple layers of epigenetic regulation.

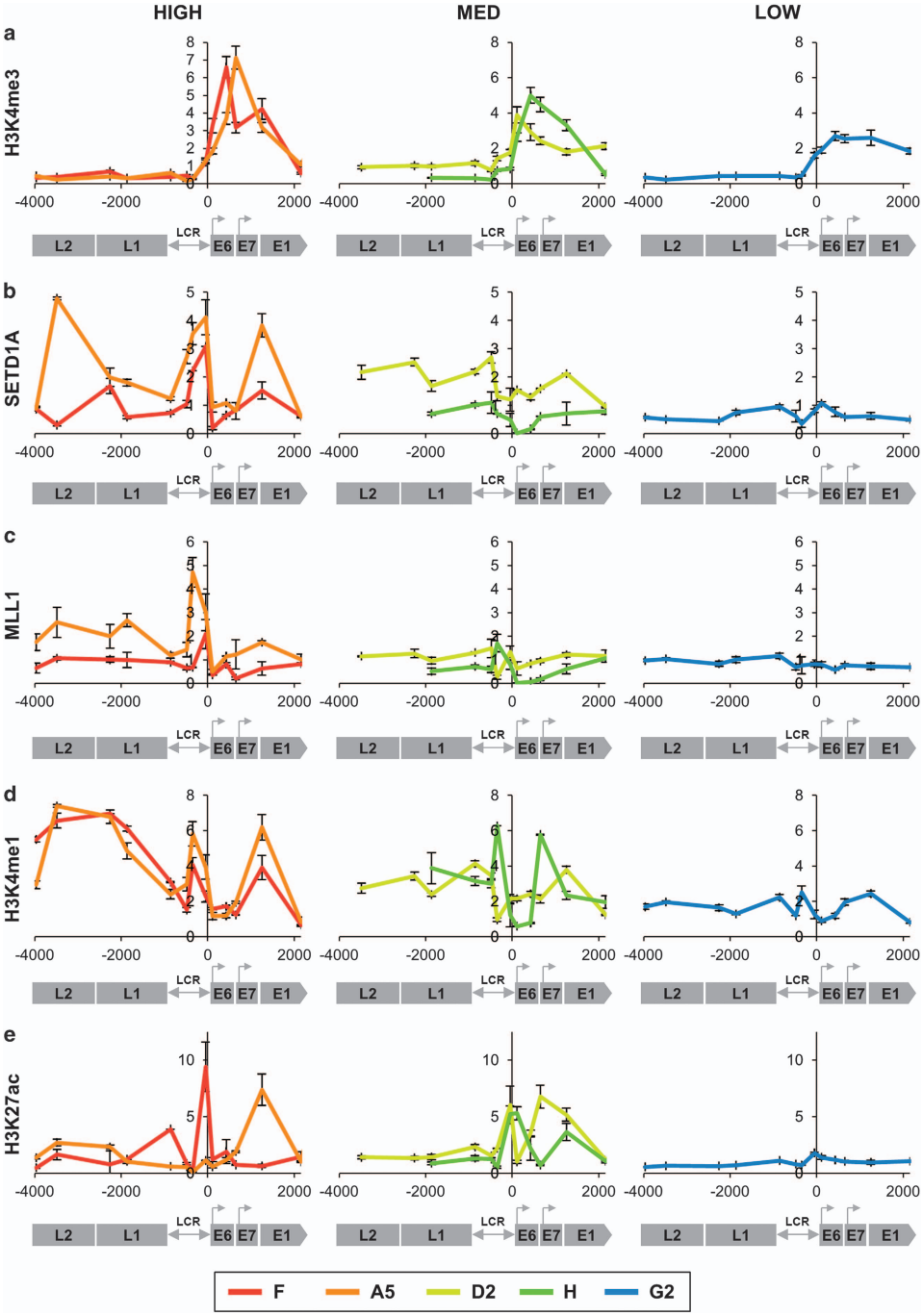

Figure 2. Associations with active histone PTMs and modifying enzymes. Levels of association of the H3K4me3 histone PTM (derived from four biological replicates) (**a**) and the associated histone-modifying enzymes SETD1A (three replicates) (**b**) and MLL1 (four replicates) (**c**); as well as the transcriptional enhancer marks H3K4me1 (three replicates) (**d**) and H3K27ac (two replicates) (**e**). In each graph, the *y*-axis shows the relative levels of enrichment, normalised to host control target regions (see Supplementary Table S1). The *x*-axis and underlying schematic show the region of the HPV16 genome analysed. In all panels, data are colour coded according to the key at the foot of the figure. Bars = mean ± s.e.m.

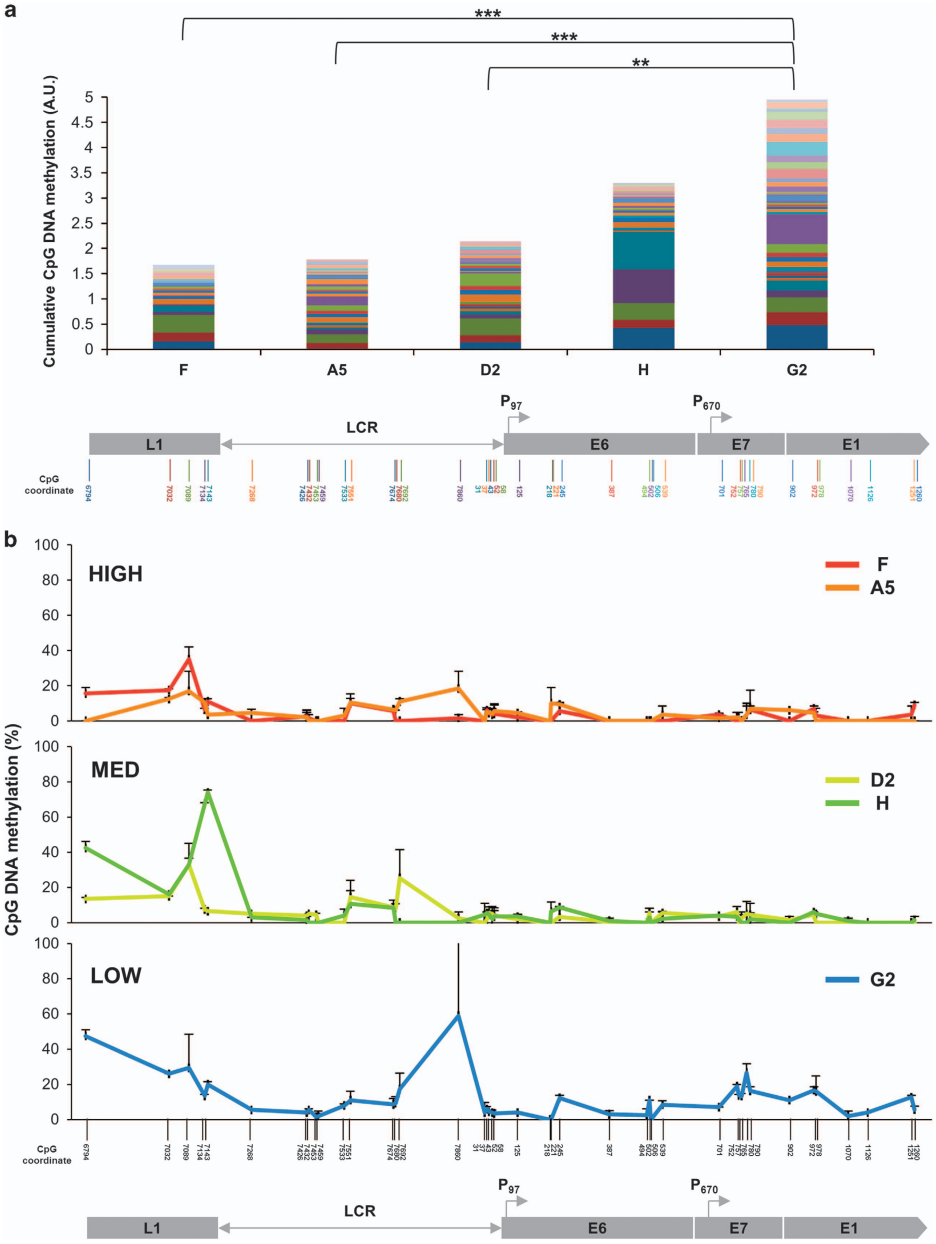

Figure 3. Associations with CpG DNA methylation. (**a**) Cumulative levels of endogenous CpG DNA methylation across the integrated HPV16 genomes, derived from three biological replicates. The coloured bars in each stack correspond to individual CpG sites. The order of the bars in each stack (from bottom to top) corresponds to the order of the CpG coordinates (from left to right) in the genome map at the base of the panel. *P*-values (Student's *t*-test): $^{**}P < 0.01$, $^{***}P < 0.001$. (**b**) Percentage of endogenous DNA methylation at CpG dinucleotides across the HPV16 genome (*y*-axis). The *x*-axis and underlying schematic show the region of the HPV16 genome analysed. Bars = mean+s.e.m.

Our initial data showed that high virus expression per template was associated with open chromatin at the HPV16 LCR, together with greater loading of chromatin remodelling enzymes and less nucleosome occupancy across the HPV16 early promoter and the oncogenes E6/E7. Together, these changes would be expected to increase template accessibility for the cellular transcriptional machinery, enabling transcriptional activation and RNAPII elongation. The reasons for the relative abundance of BRG1 and INI1 over the virus early region are not certain but may be related to the ability of these enzymes to orchestrate long-range interactions between promoter-enhancer regions.[20,21]

Levels of HPV16 expression per template were positively associated with higher abundance of histone PTMs that marked transcriptionally active chromatin, together with the cognate

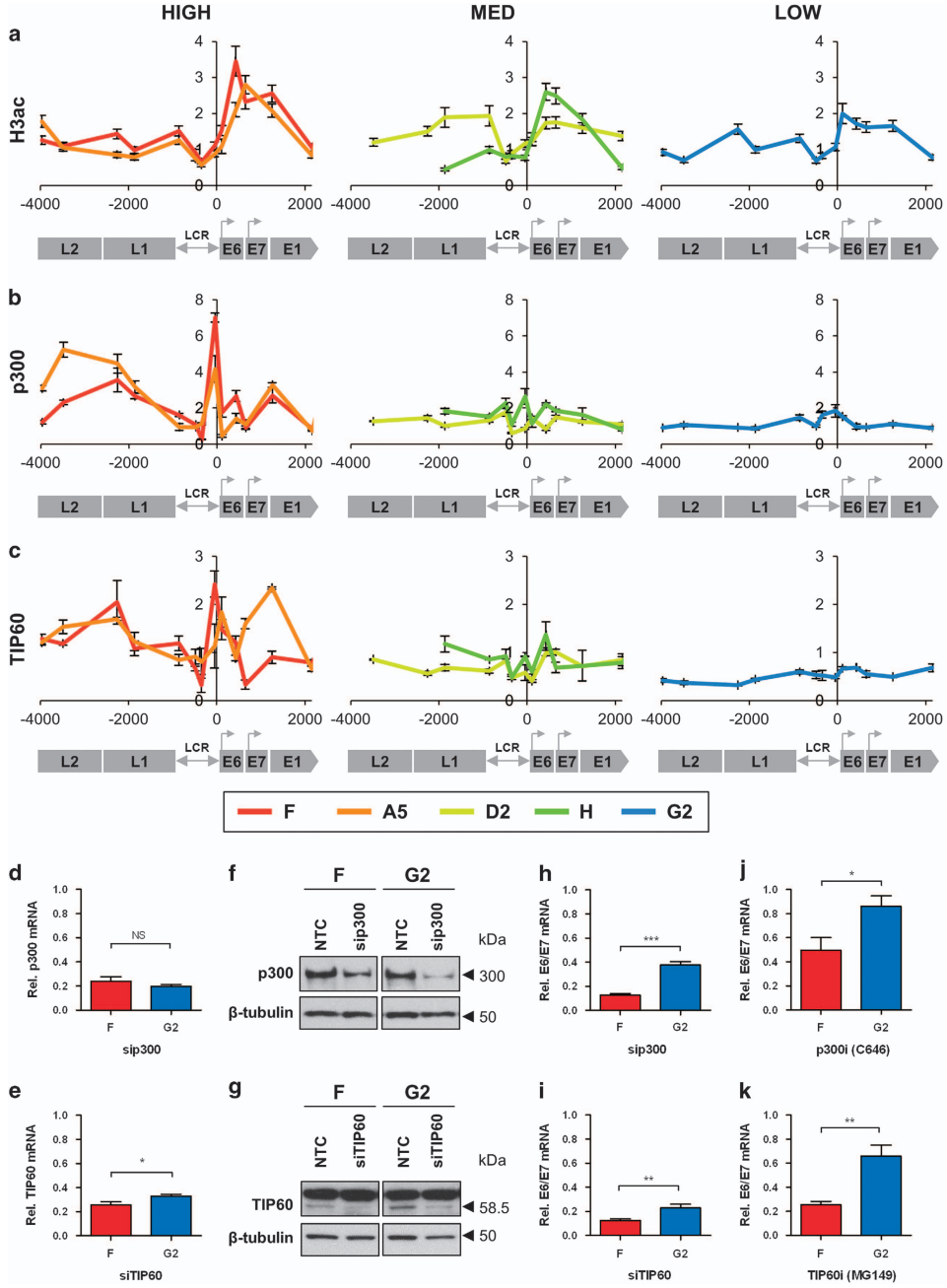

Figure 4. Associations with histone acetylation and HAT abundance/activity. (**a–c**) Levels of the H3ac histone PTM (derived from three biological replicates) (**a**) and the associated HAT enzymes p300 (three replicates) (**b**) and TIP60 (three replicates) (**c**). In each graph, the *y*-axis shows the relative levels of enrichment, normalised to host control target regions (see Supplementary Table S1). The *x*-axis and underlying schematic show the region of the HPV16 genome analysed. In all panels, data are colour coded according to the key beneath panel c. (**d–k**) Depletion/inhibition in clones F and G2 of HAT enzymes p300 (upper row) and TIP60 (lower row). The panels show levels of depletion of target mRNAs (**d, e**), target protein (**f, g**) and HPV16 E6/E7 transcripts (**h, i**) in siRNA-treated vs NTC-treated cells, together with HPV16 E6/E7 transcript levels in cells treated with specific small-molecule inhibitors, vs cells treated with vehicle only (**j, k**). All data for p300 were derived from four biological replicates and all data for TIP60 from six biological replicates. Each western blot used protein samples from all replicate experiments combined. Bars = mean ± s.e.m. *P*-values (Student's *t*-test): $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, NS = not significant. Abbreviation: NTC, non-targeting control.

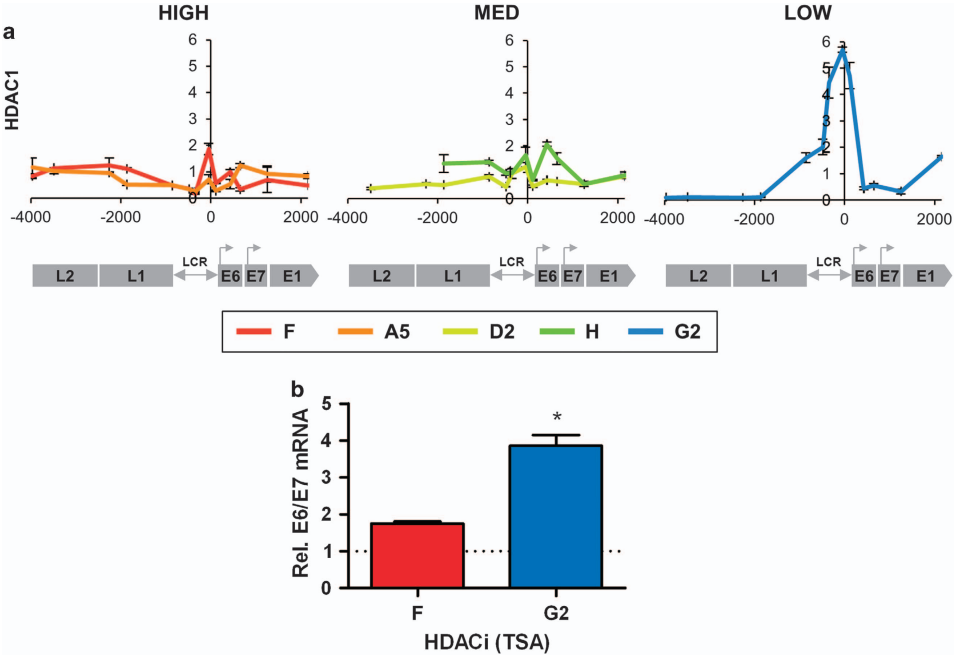

Figure 5. Associations with HDAC abundance/activity. (**a**) Levels of association of HDAC1 enzyme. The *y*-axis shows the relative levels of enrichment, derived from two biological replicates and normalised to host control target regions (see Supplementary Table S1). The *x*-axis and underlying schematic show the region of the HPV16 genome analysed. (**b**) Changes in HPV16 E6/E7 transcript levels following type I/type II HDAC inhibition with TSA in clones F and G2, derived from three biological replicates. Bars = mean ± s.e.m. *P*-values (Student's *t*-test): $^{*}P < 0.05$.

writer enzymes. The presence of the transcriptional activation mark H3K4me3 was associated with consistent enrichment for the H3K4 methylases SETD1A and MLL1 at the LCR, where the enzymes would be recruited to the activating RNAPII complex. This observation is paralleled by evidence that a specific isoform of MLL5 (MLL5β) is recruited via a distal AP1 site at the HPV18 LCR and is necessary for virus oncogene expression.[22] The reasons for the different distributions of H3K4me3 and H4K4me1 are unclear and may be related to the relative distribution or balance of the H3K4 methylases and their cofactors.[23]

Expression levels per template were negatively associated with repressive heterochromatin marks and with overall levels of endogenous CpG DNA methylation. In the type I HPV16 integrants studied here, there was no clear relationship between virus expression per template and *L1* methylation. At present, there is considerable interest in using HRHPV methylation as a clinical diagnostic test, for example to triage cytology samples.[24] Our data indicate a need for further investigations of the associations between HRHPV *L1* methylation and virus parameters (e.g. physical state, presence or absence of full-length concatemers, levels of early gene expression per template), in order to understand better the potentially complex relationship between *L1* methylation and cervical neoplastic progression.

Interestingly, the repressive heterochromatin marks H3K9me2 and H3K27me2 were present at much greater overall abundance than the equivalent tri-methyl marks H3K9me3 and H3K27me3. Previous work has shown a global reduction in H3K27me3 in HRHPV-infected cells, caused by virus-driven upregulation of H3K27 demethylases KDM6A and KDM6B, and inhibition of the polycomb repressive complex 2, the writer of the H3K27me3 mark.[25,26] The absence of these heterochromatic tri-methyl marks is also consistent with chromatin immunoprecipitation sequencing data from the HPV18-positive cervical adenocarcinoma cell line HeLa[27] and analyses of undifferentiated and differentiated squamous epithelial cells containing HPV31 episomal genomes.[11] Indeed, for naturally occurring HRHPV integrants (as opposed to those generated experimentally) significant levels of heterochromatic marks, including H4K20me3, have only been reported in CaSki cervical SCC cells, in which there is an unusually high number of integrated HPV16 genomes (~600 copies).[28]

Virus expression per template was associated with histone acetylation at the integrated HPV16 genomes, consistent with our previous findings in episome-associated cervical carcinogenesis[6] and with observations using genetically modified HPV16 templates.[29] Histone acetylation associated positively with levels of both HATs examined, p300 and TIP60. High levels of p300 were associated with greater overall abundance of cJun, a potential component of the AP1 complex, which is a possible mechanism of p300 recruitment. While it has previously been shown that p300 can activate HPV gene expression,[19,30,31] our data demonstrate a functional, dose-dependent relationship between levels of p300 and HPV16 gene expression, as cells with high virus expression per template showed significantly greater sensitivity to p300 depletion or inhibition than those with low expression per template. Similar observations were made when inhibiting TIP60. Interestingly, there is evidence that TIP60 is a transcriptional repressor at the HPV18 early enhancer/promoter and can be targeted for degradation by the HPV18 E6 protein.[31,32] Inhibition using MG149 also indicated an activating role for TIP60 in episome-containing parental W12 cells (data not shown), despite the presence of E2 protein, which has been shown to organise TIP60-mediated repression of the HPV18 LCR.[33] Therefore, the function of TIP60 at HPV16 genomes is not obviously dictated by template structure.

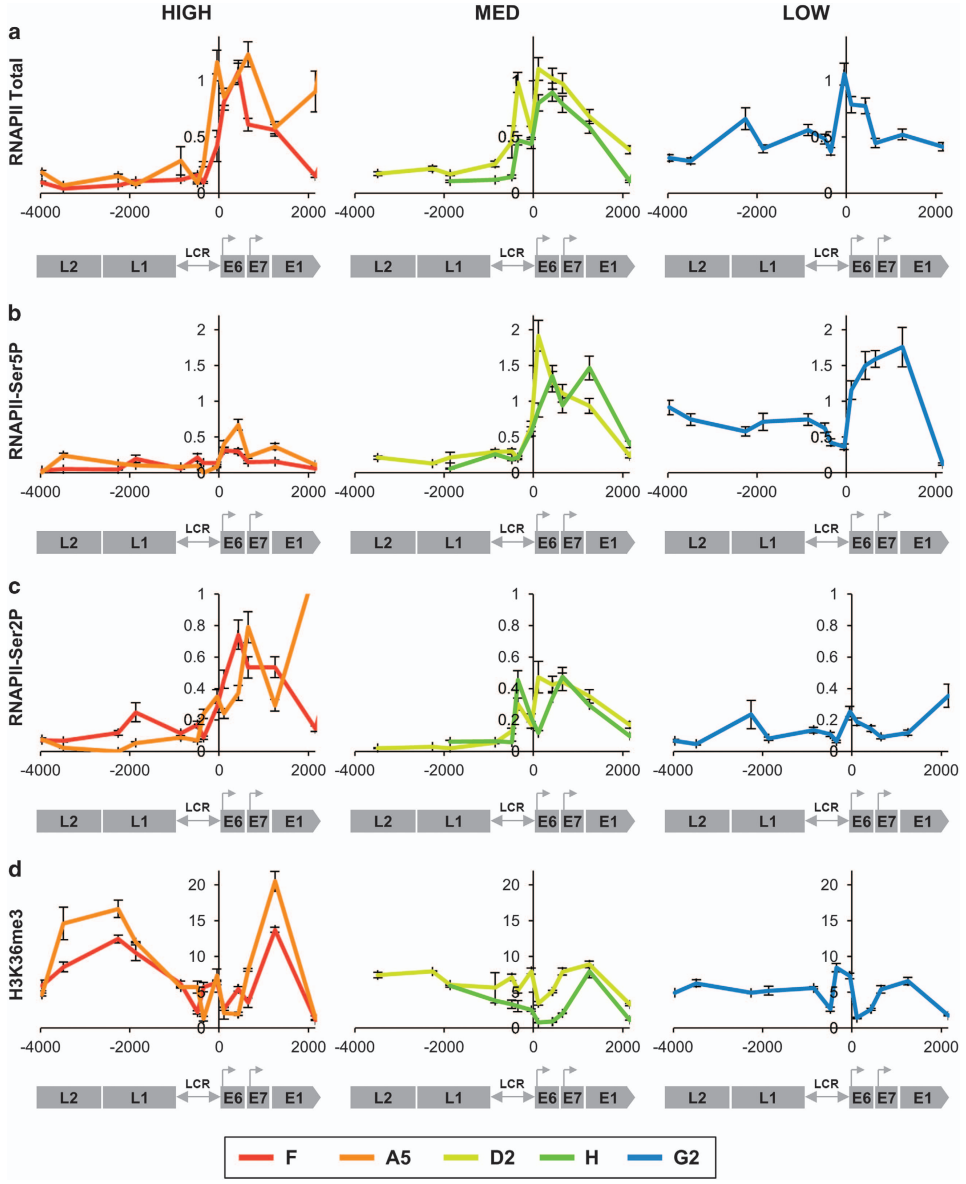

Figure 6. Associations with RNAPII and H3K36me3. Levels of association of total RNAPII (derived from three biological replicates) (**a**), RNAPII-Ser5P (poised/paused) (three replicates) (**b**), RNAPII-Ser2P (active/elongating) (three replicates) (**c**) and H3K36me3 (two replicates) (**d**). In each graph, the *y*-axis shows the relative levels of enrichment, normalised to host control target regions (see Supplementary Table S1). The *x*-axis and underlying schematic show the region of the HPV16 genome analysed. In all panels, data are colour coded according to the key at the foot of the figure. Bars = mean ± s.e.m.

The reasons for the disparate observations concerning TIP60 function are unclear. The mechanism of TIP60 recruitment may be relevant, as we observed no overall association between levels of TIP60 and YY1 in W12 cells, whereas YY1 was found at the integrated HPV18 genome in HeLa cells, where TIP60 is repressive.[31] In the absence of YY1, TIP60 can be recruited to chromatin via activated RNAPII-Ser2P itself[34] and by various other transcription factors including E2F1, MYC, MAX and MXI1, all of which have been found at the HPV18 LCR.[27] Indeed, increased TIP60 recruitment to the hTERT promoter, likely through MYC interaction, was seen in human foreskin keratinocytes expressing HPV16 E6 protein.[35] TIP60 has also been shown to interact directly with chromatin through its chromodomain. This can occur via the repressive mark H3K9me3 at DNA double-strand breaks[36] but also via the active marks H3K4me3 (enabling TIP60 to act as a histone code reader/translator)[37] and H3K4me1.[38] The latter, when combined with H3K27ac, is an indicator of active enhancers.[39] In the W12 cells with high virus expression per template, these marks

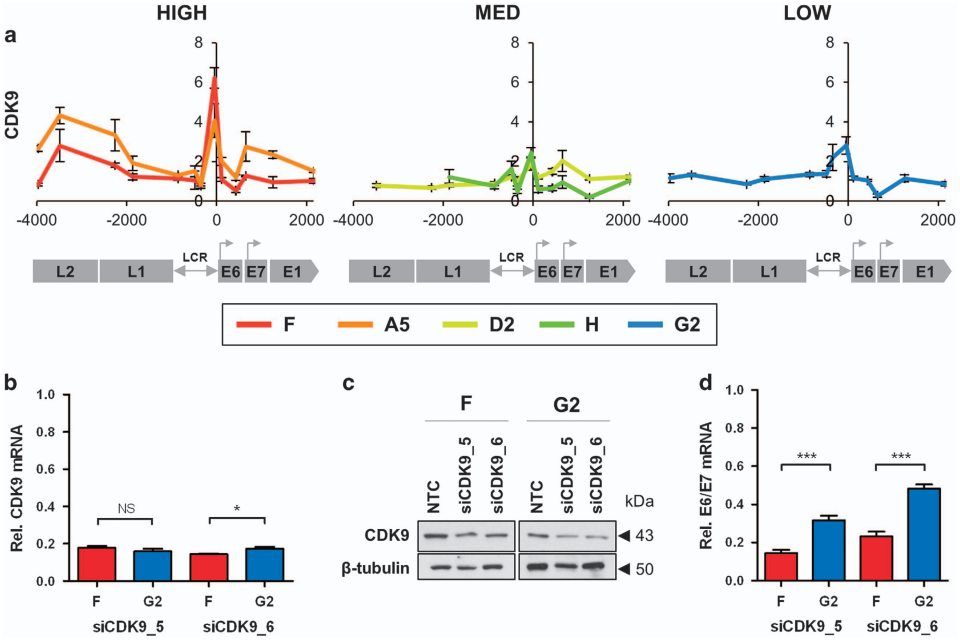

Figure 7. Associations with CDK9 abundance/activity. (**a**) Levels of association of CDK9. The *y*-axis shows the relative levels of enrichment, derived from three biological replicates and normalised to host control target regions (see Supplementary Table S1). The *x*-axis and underlying schematic show the region of the HPV16 genome analysed. (**b–d**) Depletion of CDK9 using siRNAs, showing levels of target mRNA (**b**) and protein (**c**), together with changes in HPV16 E6/E7 transcript levels (**d**), in siRNA-treated vs NTC-treated cells. All data were derived from two biological replicates. The western blot used protein samples from both replicates combined. Bars = mean ± s.e.m. *P*-values (Student's *t*-test): $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, NS = not significant. Abbreviation: NTC, non-targeting control.

were present, together with p300, at the integrated HPV16 LCR, which therefore appears to be acting as a canonical enhancer of transcription. Interestingly, such marks were also present over the virus late gene region, which, when out of the context of the episomal genome, may augment integrated HPV16 gene expression.

The differences in HAT recruitment between the clones were mirrored by differences in HDAC1 abundance at the HPV16 genome. HDAC1 levels were greater in cells with less virus gene expression per template, which showed significantly greater increases in transcript levels following HDAC inhibition. However, HDAC1 was detectable at the virus genome in all clones and all showed increased gene expression levels following HDAC inhibition with TSA over a relatively long duration of 16 h. These observations are consistent with data describing the necessity for HDAC presence at gene promoter regions, in order to allow resetting of histone acetylation during the dynamic turnover of these marks that accompanies RNAPII progression.[34,40]

While virus expression per template showed no association with overall levels of RNAPII at the HPV16 genome, there was an association with levels of the active/elongating form of the enzyme (RNAPII-Ser2P), together with those of CDK9, the component of P-TEFb responsible for phosphorylating Ser2 of the RNAPII C-terminal domain. The CDK9 enzyme was functionally significant, as evidenced by a greater sensitivity to depletion in cells with higher HPV16 gene expression per template. We also observed striking changes in the distribution of RNAPII and chromatin marks following treatment with Flavopiridol. While this small molecule can inhibit multiple CDKs and affect cell cycle progression, its major mode of action is considered to be inhibition of CDK9.[41,42] The importance of P-TEFb/CDK9 in transcription of integrated HPV16 supports observations for other viruses. For example, CDK9 is necessary to relieve RNAPII pausing at the Epstein-Barr virus C promoter and drive transcription of polycistronic virus mRNAs,[43] while P-TEFb is required for Tat-driven transcriptional elongation at the human immunodeficiency virus (HIV) long terminal repeat.[44]

Together, our data are consistent with the model shown in Figure 9. Integrated HPV16 templates showing higher levels of oncogene expression are associated with more accessible DNA, via the action of chromatin remodellers. This accessibility leads to the recruitment of activating histone-modifying enzymes, either directly or via transcription factors. In turn, these enzymes methylate and acetylate histone tails, so that the recruitment and activation of RNAPII can occur through activating complexes such as P-TEFb. While the integrated templates with lower expression levels are still able to activate RNAPII, there is a shift in the balance of activating and repressing enzymes that affects gene expression levels. In future work, it will be important to study the mechanisms by which initial virus template accessibility is determined, including whether HPV16 acquires the features of the host chromatin at integration sites. The W12 system will allow detailed dissection of the relative roles of virus factors, such as those described here, and host genes in providing individual cells with a selective advantage during the early stages of cervical neoplastic progression.

MATERIALS AND METHODS

Cell culture

Previous publications have given detailed descriptions of the W12 system,[6,16,45] including generation of the W12 cell clones.[14,17] The five clones selected for further investigation (Table 1) were episome-free and did not express the HPV16 transcriptional regulator E2.[17] All W12 cells were

grown in monolayer culture, as described,[46] in order to restrict cell differentiation and maintain the phenotype of the basal epithelial cell layer, the key site of HRHPV transcriptional deregulation in cervical carcinogenesis.[4,15] Cells were analysed at the lowest available passage (p) after cloning (typically p3 to p8), in order to minimise any effects of genomic instability caused by deregulated HPV16 oncogene expression. We also used the HPV16-positive cervical SCC cell line SiHa,[47] which contains ~2 integrated virus copies and was grown as described.[48]

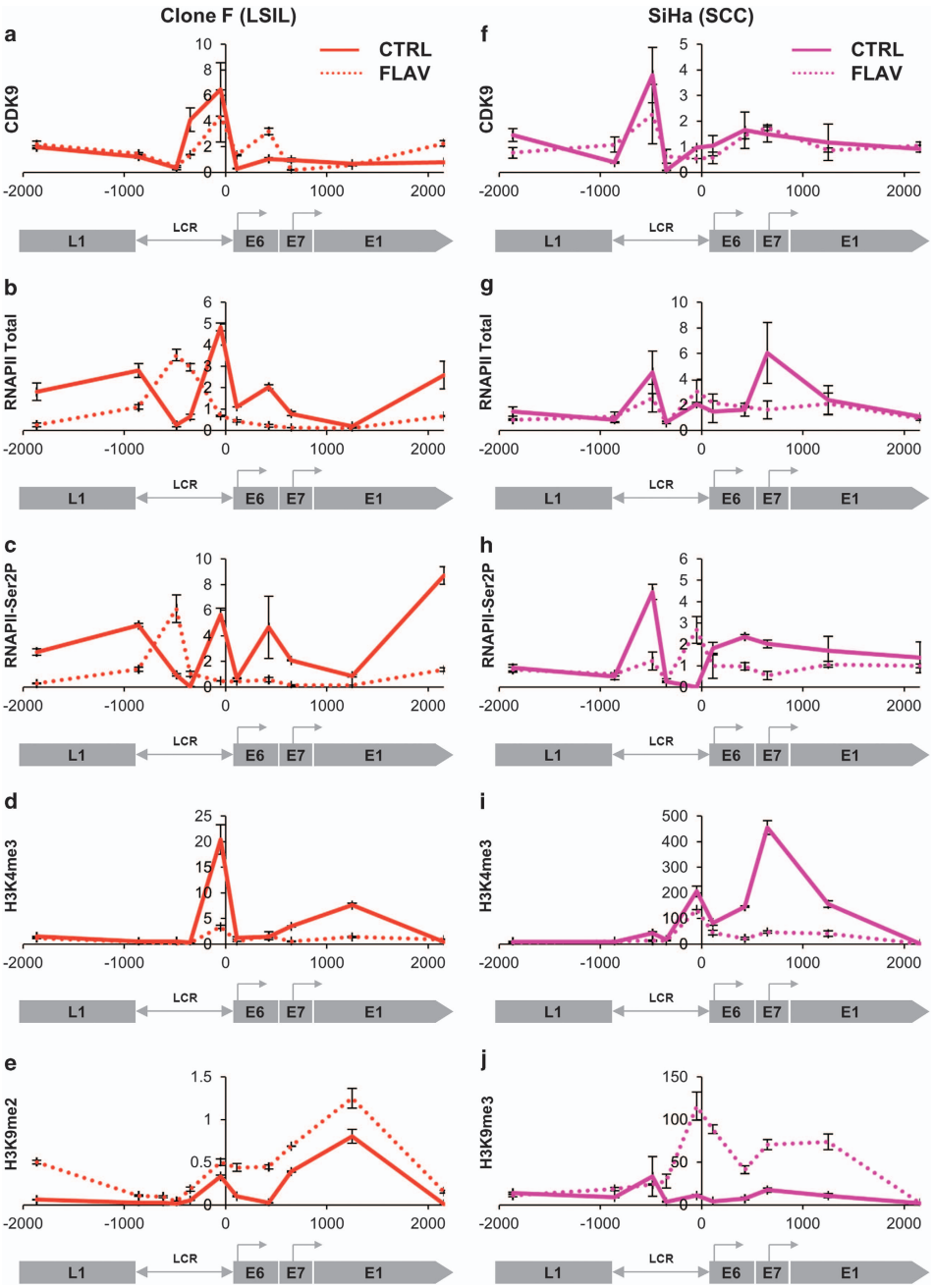

Figure 8. Effects of CDK9 inhibition. Effects of Flavopiridol in clone F (LSIL phenotype) and SiHa (SCC phenotype). Rows show levels of CDK9 (**a, f**), total RNAPII (**b, g**), RNAPII-Ser2P (active/elongating) (**c, h**), H3K4me3 (active) (**d, i**) and H3K9me2/3 (repressed) (**e, j**). In each graph, the *y*-axis shows the relative levels of enrichment, derived from two biological replicates and normalised to host control target regions (see Supplementary Table S1). The *x*-axis and underlying schematic show the region of the HPV16 genome analysed. Solid lines = control-treated cells; dotted lines = Flavopiridol-treated cells. Bars = mean ± s.e.m.

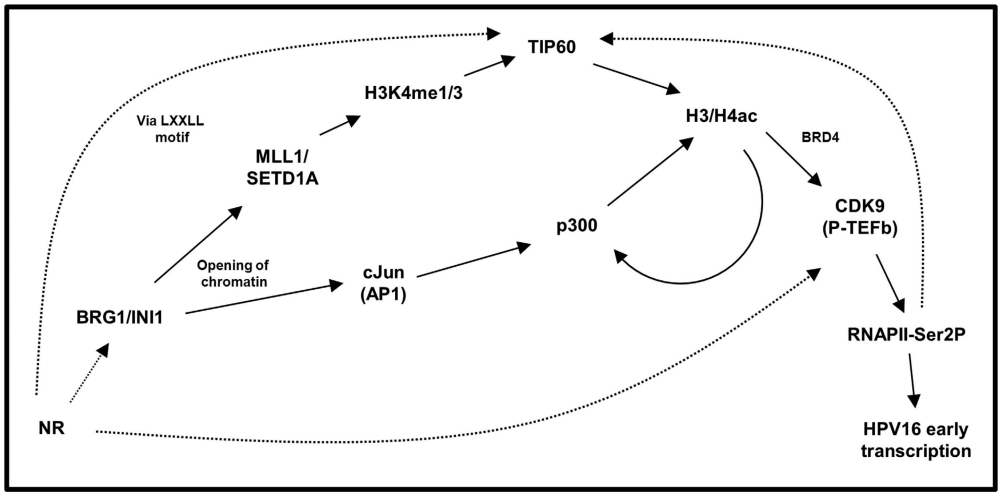

Figure 9. Working model of the multilayered epigenetic changes that enable high levels of virus gene expression per template following HPV16 integration. Recruitment of chromatin remodelling enzymes (BRG1/INI1) to the HPV16 genome, possibly through host steroid hormone nuclear receptors (NR), allows greater accessibility to the virus chromatin of transcription factors (e.g. cJun/AP1) and histone-modifying enzymes, including MLL1 and SETD1A, which can write the H3K4me1/3 marks. Recruitment of HATs can occur though interactions with transcription factors (e.g. p300) and histone PTMs, which may allow TIP60 recruitment through its chromodomain. Once acetylation of histones has occurred, recruitment of CDK9 (the enzymatic component of P-TEFb) is able to activate RNAPII through phosphorylation of Ser2 at the C-terminal domain, leading to stimulation of transcription from the HPV16 early promoter.

Treatment with small-molecule inhibitors

Cells were treated for 16 h with medium supplemented with small-molecule inhibitors, using the highest doses that did not produce cell death over the timecourse of the experiments. The small-molecule inhibitors used were: p300 inhibitor, C646 (SML0002; Sigma-Aldrich, Dorset, UK; 25 μM); TIP60 inhibitor, MG149 (Axon 1785; Axon Medchem, Groningen, Netherlands; 150 μM); HDAC inhibitor, Trichostatin-A (T1952; Sigma-Aldrich; 400 nM); or CDK9 inhibitors, Flavopiridol (F3055; Sigma-Aldrich; 150 nM), Roscovitine (C3249; Sigma-Aldrich; 20 μM) or DRB (D1916; Sigma-Aldrich; 50 μM). For analysis of cell growth, cells were seeded at 5×10^4 per well and treated with Flavopiridol after 24 h. Total live cell counts were carried out every 24 h over 5 days, using Trypan blue staining. In all experiments, negative control cells were treated with equivalent volumes of DMSO vehicle (vol/vol).

Gene depletion

Each target gene was depleted using human Flexitube siRNAs (Qiagen, Crawley, UK): *CDK9* (CDK9_5 SI00605066; CDK9_6 SI00605073); *p300* (EP300_7 SI02626267); *TIP60* (KAT5_2 SI05120304); non-targeting control (AllStars Negative Control siRNA, 1027280). All siRNAs were used at 10 nM, with cells being transfected at 20–30% confluence using Lipofectamine RNAiMAX (Invitrogen, Paisley, UK) as described.[49,50]

Quantification of host proteins and HPV16 transcripts

Quantitative western immunoblotting was carried out as described,[6,17,51] using the primary antibodies listed in Supplementary Table S1. Protein concentrations were compared with those of the β-tubulin loading control (Abcam, Cambridge, UK; 6ng/ml), using ImageJ software. Levels of HPV16 E6 and E7 transcripts were measured using SYBRGreen quantitative reverse transcription–PCR (qRT–PCR), as described.[17] Primers and conditions are given in Supplementary Table S2. Relative transcript levels were determined using the Pfaffl equation,[52] normalised to the mean of four housekeeping genes[53] and residual levels of the target protein, then referenced to control samples.

Chromatin immunoprecipitation

Chromatin immunoprecipitation was performed as described,[6,17] using chromatin immunoprecipitation-validated primary antibodies and appropriate serum/IgG negative controls (Supplementary Table S1). In contrast to our previous assessment of a relatively limited region of the HPV16 genome, we analysed 6094 nucleotides (nt) of HPV sequence, from the *L2* gene, through the LCR, to the *E1* gene (nt 3936 to 2158). This genomic region was present in all five clones, with the exception of nt 3936 to 6039 in clone H and nt 3936 to 4419 in clone D2. Primers and conditions used for qPCR are given in Supplementary Table S3. Efficiency of immunoprecipitation of each target was normalised using control region qPCR primers (Supplementary Table S4).

Formaldehyde-assisted isolation of regulatory elements and nucleosome occupancy and methylome sequencing

Formaldehyde-assisted isolation of regulatory element was carried out as described.[54] Quantification of HPV16 DNA sequences was carried out by qPCR and normalised to the efficiency of enrichment, as determined by the ratio of *GAPDH* promoter (open)[55] to *GAPDH* open reading frame (closed).[43] Primers and conditions for qPCR were those in Supplementary Tables S3 and S4. The occupancy of nucleosomes or other DNA-binding proteins between the HPV16 early promoter and *E1* gene (nt 7902 to 1012) was assessed by nucleosome occupancy and methylome sequencing (Active Motif, La Hulpe, Belgium), which measures the distribution of exogenous GpC DNA methylation.[56] Samples were amplified in duplicate using PCR primers designed to exclude either GpC or CpG dinucleotides, in order to eliminate amplification bias (Supplementary Table S5). PCR products were Sanger sequenced, using 5′- and 3′-end primers to confirm reads from each end of the product. Each analysis was carried out in duplicate and the degree of cytosine methylation for each nucleotide position averaged across replicates. Percentage GpC methylation was scored in 20% intervals, from which a heatmap was generated.

HPV16 DNA methylation

Five hundred nanograms of genomic DNA were bisulphite-converted using the EpiTect Bisulfite Kit (59104; Qiagen), then desulphonated, washed and eluted in 40 μl of buffer. PCR amplification of HPV16 sequences was carried out using Immolase (Bioline, London, UK) and the primers listed in Supplementary Table S6. *LINE1* amplification was also carried out as a methylation-positive conversion control. Sequencing primers were designed using PyroQ software (Pyromark MD, Qiagen) and analysis performed on a Pyromark MD pyrosequencer, using standard protocols and controls. For each cell line, assays were performed in duplicate on a minimum of three independently prepared bisulphite-converted DNA samples.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by Cancer Research UK (Programme Grant A13080); the Medical Research Council; The Pathological Society of Great Britain and Ireland (ELAK); and the Agency for Science, Technology and Research, Singapore (QYA). Funding for open access charge: Cancer Research UK and the Medical Research Council.

REFERENCES

1 Forman D, de Martel C, Lacey CJ, Soerjomataram I, Lortet-Tieulent J, Bruni L *et al.* Global burden of human papillomavirus and related diseases. *Vaccine* 2012; **30**: F12–F23.
2 Jeon S, Allen-Hoffmann BL, Lambert PF. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *J Virol* 1995; **69**: 2989–2997.
3 Groves IJ, Coleman N. Pathogenesis of human papillomavirus-associated mucosal disease. *J Pathol* 2015; **235**: 527–538.
4 Pett M, Coleman N. Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *J Pathol* 2007; **212**: 356–367.
5 Wentzensen N, Vinokurova S, von Knebel Doeberitz M. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res* 2004; **64**: 3878–3884.
6 Gray E, Pett MR, Ward D, Winder DM, Stanley MA, Roberts I *et al.* In vitro progression of human papillomavirus 16 episome-associated cervical neoplasia displays fundamental similarities to integrant-associated carcinogenesis. *Cancer Res* 2010; **70**: 4081–4091.
7 Stanley MA, Browne HM, Appleby M, Minson AC. Properties of a non-tumorigenic human cervical keratinocyte cell line. *Int J Cancer* 1989; **43**: 672–676.
8 Bedell MA, Hudson JB, Golub TR, Turyk ME, Hosken M, Wilbanks GD *et al.* Amplification of human papillomavirus genomes in vitro is dependent on epithelial differentiation. *J Virol* 1991; **65**: 2254–2260.
9 Stunkel W, Bernard HU. The chromatin structure of the long control region of human papillomavirus type 16 represses viral oncoprotein expression. *J Virol* 1999; **73**: 1918–1930.
10 Bernard HU. Regulatory elements in the viral genome. *Virology* 2013; **445**: 197–204.
11 Wooldridge TR, Laimins LA. Regulation of human papillomavirus type 31 gene expression during the differentiation-dependent life cycle through histone modifications and transcription factor binding. *Virology* 2008; **374**: 371–380.
12 Carson A, Khan SA. Characterization of transcription factor binding to human papillomavirus type 16 DNA during cellular differentiation. *J Virol* 2006; **80**: 4356–4362.
13 del Mar Pena LM, Laimins LA. Differentiation-dependent chromatin rearrangement coincides with activation of human papillomavirus type 31 late gene expression. *J Virol* 2001; **75**: 10005–10013.
14 Dall KL, Scarpini CG, Roberts I, Winder DM, Stanley MA, Muralidhar B *et al.* Characterization of naturally occurring HPV16 integration sites isolated from cervical keratinocytes under noncompetitive conditions. *Cancer Res* 2008; **68**: 8249–8259.
15 Stoler MH, Rhodes CR, Whitbeck A, Wolinsky SM, Chow LT, Broker TR. Human papillomavirus type 16 and 18 gene expression in cervical neoplasias. *Hum Pathol* 1992; **23**: 117–128.
16 Pett MR, Alazawi WO, Roberts I, Dowen S, Smith DI, Stanley MA *et al.* Acquisition of high-level chromosomal instability is associated with integration of human papillomavirus type 16 in cervical keratinocytes. *Cancer Res* 2004; **64**: 1359–1368.
17 Scarpini CG, Groves IJ, Pett MR, Ward D, Coleman N. Virus transcript levels and cell growth rates after naturally occurring HPV16 integration events in basal cervical keratinocytes. *J Pathol* 2014; **233**: 281–293.
18 Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res* 2011; **21**: 381–395.
19 Wang WM, Wu SY, Lee AY, Chiang CM. Binding site specificity and factor redundancy in activator protein-1-driven human papillomavirus chromatin-dependent transcription. *J Biol Chem* 2011; **286**: 40974–40986.
20 Euskirchen GM, Auerbach RK, Davidov E, Gianoulis TA, Zhong G, Rozowsky J *et al.* Diverse roles and interactions of the SWI/SNF chromatin remodeling complex revealed using global approaches. *PLoS Genet* 2011; **7**: e1002008.
21 Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 2012; **148**: 84–98.
22 Yew CW, Lee P, Chan WK, Lim VK, Tay SK, Tan TM *et al.* A novel MLL5 isoform that is essential to activate E6 and E7 transcription in HPV16/18-associated cervical cancers. *Cancer Res* 2011; **71**: 6696–6707.
23 Dou Y, Milne TA, Ruthenburg AJ, Lee S, Lee JW, Verdine GL *et al.* Regulation of MLL1 H3K4 methyltransferase activity by its core components. *Nat Struct Mol Biol* 2006; **13**: 713–719.
24 Wentzensen N, Sun C, Ghosh A, Kinney W, Mirabello L, Wacholder S *et al.* Methylation of HPV18, HPV31, and HPV45 genomes and cervical intraepithelial neoplasia grade 3. *J Natl Cancer Inst* 2012; **104**: 1738–1749.
25 McLaughlin-Drubin ME, Crum CP, Munger K. Human papillomavirus E7 oncoprotein induces KDM6A and KDM6B histone demethylase expression and causes epigenetic reprogramming. *Proc Natl Acad Sci USA* 2011; **108**: 2130–2135.
26 Hyland PL, McDade SS, McCloskey R, Dickson GJ, Arthur K, McCance DJ *et al.* Evidence for alteration of EZH2, BMI1, and KDM6A and epigenetic reprogramming in human papillomavirus type 16 E6/E7-expressing keratinocytes. *J Virol* 2011; **85**: 10999–11006.
27 Johannsen E, Lambert PF. Epigenetics of human papillomaviruses. *Virology* 2013; **445**: 205–212.
28 De-Castro Arce J, Gockel-Krzikalla E, Rosl F. Silencing of multi-copy HPV16 by viral self-methylation and chromatin occlusion: a model for epigenetic virus-host interaction. *Hum Mol Genet* 2012; **21**: 1693–1705.
29 Johansson C, Jamal Fattah T, Yu H, Nygren J, Mossberg AK, Schwartz S. Acetylation of intragenic histones on HPV16 correlates with enhanced HPV16 gene expression. *Virology* 2015; **482**: 244–259.
30 Kruppel U, Muller-Schiffmann A, Baldus SE, Smola-Hess S, Steger G. E2 and the co-activator p300 can cooperate in activation of the human papillomavirus type 16 early promoter. *Virology* 2008; **377**: 151–159.
31 He H, Luo Y. Brg1 regulates the transcription of human papillomavirus type 18 E6 and E7 genes. *Cell Cycle* 2012; **11**: 617–627.
32 Jha S, Vande Pol S, Banerjee NS, Dutta AB, Chow LT, Dutta A. Destabilization of TIP60 by human papillomavirus E6 results in attenuation of TIP60-dependent transcriptional regulation and apoptotic pathway. *Mol Cell* 2010; **38**: 700–711.
33 Smith JA, Haberstroh FS, White EA, Livingston DM, DeCaprio JA, Howley PM. SMCX and components of the TIP60 complex contribute to E2 regulation of the HPV E6/E7 promoter. *Virology* 2014; **468-470**: 311–321.
34 Wang Z, Zang C, Cui K, Schones DE, Barski A, Peng W *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* 2009; **138**: 1019–1031.
35 Xu M, Katzenellenbogen RA, Grandori C, Galloway DA. An unbiased in vivo screen reveals multiple transcription factors that control HPV E6-regulated hTERT in keratinocytes. *Virology* 2013; **446**: 17–24.
36 Sun Y, Jiang X, Xu Y, Ayrapetov MK, Moreau LA, Whetstine JR *et al.* Histone H3 methylation links DNA damage detection to activation of the tumour suppressor Tip60. *Nat Cell Biol* 2009; **11**: 1376–1382.
37 Kim CH, Kim JW, Jang SM, An JH, Seo SB, Choi KH. The chromodomain-containing histone acetyltransferase TIP60 acts as a code reader, recognizing the epigenetic codes for initiating transcription. *Biosci Biotechnol Biochem* 2015; **79**: 532–538.
38 Jeong KW, Kim K, Situ AJ, Ulmer TS, An W, Stallcup MR. Recognition of enhancer element-specific histone methylation by TIP60 in transcriptional activation. *Nat Struct Mol Biol* 2011; **18**: 1358–1365.
39 Bogdanovic O, Fernandez-Minan A, Tena JJ, de la Calle-Mustienes E, Hidalgo C, van Kruysbergen I *et al.* Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res* 2012; **22**: 2043–2053.
40 Hazzalin CA, Mahadevan LC. Dynamic acetylation of all lysine 4-methylated histone H3 in the mouse nucleus: analysis at c-fos and c-jun. *PLoS Biol* 2005; **3**: e393.
41 Wang S, Fischer PM. Cyclin-dependent kinase 9: a key transcriptional regulator and potential drug target in oncology, virology and cardiology. *Trends Pharmacol Sci* 2008; **29**: 302–313.
42 Chao SH, Fujinaga K, Marion JE, Taube R, Sausville EA, Senderowicz AM *et al.* Flavopiridol inhibits P-TEFb and blocks HIV-1 replication. *J Biol Chem* 2000; **275**: 28345–28348.
43 Palermo RD, Webb HM, Gunnell A, West MJ. Regulation of transcription by the Epstein-Barr virus nuclear antigen EBNA 2. *Biochem Soc Trans* 2008; **36**: 625–628.
44 Zhu Y, Pe'ery T, Peng J, Ramanathan Y, Marshall N, Marshall T *et al.* Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. *Genes Dev* 1997; **11**: 2622–2632.
45 Hanning JE, Saini HK, Murray MJ, Caffarel MM, van Dongen S, Ward D *et al.* Depletion of HPV16 early genes induces autophagy and senescence in a cervical carcinogenesis model, regardless of viral physical state. *J Pathol* 2013; **231**: 354–366.

46 Coleman N, Greenfield IM, Hare J, Kruger-Gray H, Chain BM, Stanley MA. Characterization and functional analysis of the expression of intercellular adhesion molecule-1 in human papillomavirus-related disease of cervical keratinocytes. *Am J Pathol* 1993; **143**: 355–367.
47 Friedl F, Kimura I, Osato T, Ito Y. Studies on a new human cell line (SiHa) derived from carcinoma of uterus. I. Its establishment and morphology. *Proc Soc Exp Biol Med* 1970; **135**: 543–545.
48 Coleman N, Stanley MA. Expression of the myelomonocytic antigens CD36 and L1 by keratinocytes in squamous intraepithelial lesions of the cervix. *Hum Pathol* 1994; **25**: 73–79.
49 Hanning JE, Saini HK, Murray MJ, van Dongen S, Davis MP, Barker EM *et al*. Lack of correlation between predicted and actual off-target effects of short-interfering RNAs targeting the human papillomavirus type 16 E7 oncogene. *Br J Cancer* 2013; **108**: 450–460.
50 Hanning JE, Groves IJ, Pett MR, Coleman N. Depletion of polycistronic transcripts using short interfering RNAs: cDNA synthesis method affects levels of non-targeted genes determined by quantitative PCR. *Virol J* 2013; **10**: 159.
51 Herdman MT, Pett MR, Roberts I, Alazawi WO, Teschendorff AE, Zhang XY *et al*. Interferon-beta treatment of cervical keratinocytes naturally infected with human papillomavirus 16 episomes promotes rapid reduction in episome numbers and emergence of latent integrants. *Carcinogenesis* 2006; **27**: 2341–2353.
52 Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 2001; **29**: e45.
53 Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A *et al*. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002; **3**: RESEARCH0034.
54 Simon JM, Giresi PG, Davis IJ, Lieb JD. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc* 2012; **7**: 256–267.
55 Groves IJ, Reeves MB, Sinclair JH. Lytic infection of permissive cells with human cytomegalovirus is regulated by an intrinsic 'pre-immediate-early' repression of viral gene expression mediated by histone post-translational modification. *J Gen Virol* 2009; **90**: 2364–2374.
56 Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* 2012; **22**: 2497–2506.

Supplementary Information accompanies this paper on the Oncogene website (http://www.nature.com/onc)

Bibliography

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," *International Journal of Cancer*, vol. 136, pp. E359–E386, mar 2015.

[2] I. J. Groves and N. Coleman, "Pathogenesis of human papillomavirus-associated mucosal disease," *Journal of Pathology*, vol. 235, no. 4, pp. 527–538, 2015.

[3] A. Molijn, D. Jenkins, W. Chen, X. Zhang, E. Pirog, W. Enqi, B. Liu, J. Schmidt, J. Cui, Y. Qiao, and W. Quint, "The complex relationship between human papillomavirus and cervical adenocarcinoma," *International Journal of Cancer*, vol. 138, pp. 409–416, jan 2016.

[4] G. N. Papanicolaou and H. F. Traut, "The diagnostic value of vaginal smears in carcinoma of the uterus. 1941.," *Archives of Pathology & Laboratory Medicine*, vol. 121, pp. 211–224, mar 1997.

[5] C. E. Anderson, A. J. Lee, K. M. McLaren, S. Cairns, C. Cowen, F. McQueen, N. J. Mayer, and H. M. Kamel, "Level of agreement and biopsy correlation using two- and three-tier systems to grade cervical dyskaryosis," *Cytopathology*, vol. 15, pp. 256–262, oct 2004.

[6] M. Motamedi, G. Böhmer, H. H. Neumann, and R. von Wasielewski, "CIN III lesions and regression: retrospective analysis of 635 cases," *BMC Infectious Diseases*, vol. 15, no. 1, p. 541, 2015.

[7] R. M. Richart, "Cervical intraepithelial neoplasia.," *Pathology annual*, vol. 8, no. 21, pp. 301–328, 1973.

[8] D. Soloman, "The 1988 Bethesda system for reporting cervical/vaginal cytologic diagnoses.," *Diagnostic Cytopathology*, vol. 5, pp. 331–334, jul 1989.

[9] E. S. Cibas, "Cervical and vaginal cytology," in *Cytology*, pp. 1–36, Elsevier, 2002.

[10] K. Matsumoto, A. Oki, R. Furuta, H. Maeda, T. Yasugi, N. Takatsuka, A. Mitsuhashi, T. Fujii, Y. Hirai, T. Iwasaka, N. Yaegashi, Y. Watanabe, Y. Nagai, T. Kitagawa, and H. Yoshikawa, "Predicting the progression of cervical precursor lesions by human papillomavirus genotyping: A prospective cohort study," *International Journal of Cancer*, vol. 128, no. 12, pp. 2898–2910, 2011.

[11] C. L. Trimble, S. Piantadosi, P. Gravitt, B. Ronnett, E. Pizer, A. Elko, B. Wilgus, W. Yutzy, R. Daniel, and K. Shah, "Spontaneous Regression of High-Grade Cervical Dysplasia: Effects of Human Papillomavirus Type and HLA Phenotype," *Clinical Cancer Research*, vol. 11, no. 13, pp. 4717–4723, 2005.

[12] A. Bermudez, N. Bhatla, and E. Leung, "Cancer of the cervix uteri," *International Journal of Gynecology and Obstetrics*, vol. 131, pp. S88–S95, oct 2015.

[13] H. Zur Hausen, "Viruses in human cancers," *European Journal of Cancer*, vol. 35, pp. 1878–1885, nov 1999.

[14] P. F. Lambert and B. Sugden, "Viruses and Human Cancer," *Abeloff's Clinical Oncology*, vol. 193, pp. 115–122, 2014.

[15] M. Dürst, L. Gissmann, H. Ikenberg, and H. zur Hausen, "A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 80, pp. 3812–3815, jun 1983.

[16] J. M. Walboomers, M. V. Jacobs, M. M. Manos, F. X. Bosch, J. A. Kummer, K. V. Shah, P. J. Snijders, J. Peto, C. J. Meijer, and N. Muñoz, "Human papillomavirus is a necessary cause of invasive cervical cancer worldwide," *Journal of Pathology*, vol. 189, pp. 12–19, sep 1999.

[17] M. Plummer, C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi, "Global burden of cancers attributable to infections in 2012: a synthetic analysis," *The Lancet Global Health*, vol. 4, pp. e609–e616, sep 2016.

[18] E. M. Burd, "Human papillomavirus and cervical cancer," *Clinical Microbiology Reviews*, vol. 16, pp. 1–17, jan 2003.

[19] J. Doorbar, W. Quint, L. Banks, I. G. Bravo, M. Stoler, T. R. Broker, and M. A. Stanley, "The biology and life-cycle of human papillomaviruses," *Vaccine*, vol. 30, pp. F55–70, nov 2012.

[20] E. F. Dunne and L. E. Markowitz, "Genital Human Papillomavirus Infection," *Source Clinical Infectious Diseases*, vol. 43, pp. 624–629, sep 2006.

[21] M. Scheurer, G. Tortolero-Luna, and K. Adler-Storthz, "Human papillomavirus infection: biology, epidemiology, and prevention," *International Journal of Gynecological Cancer*, vol. 15, pp. 727–746, sep 2005.

[22] Z.-M. Zheng and C. C. Baker, "Papillomavirus genome structure, expression, and post-transcriptional regulation.," *Frontiers in bioscience : a journal and virtual library*, vol. 11, pp. 2286–302, sep 2006.

[23] N. Shulzhenko, H. Lyng, G. F. Sanson, and A. Morgun, "Ménage à trois: An evolutionary interplay between human papillomavirus, a tumor, and a woman," *Trends in Microbiology*, vol. 22, pp. 345–353, jun 2014.

[24] S. D. Datta and M. Saraiya, "Cervical cancer screening among women who attend sexually transmitted diseases (STD) clinics: Background paper for 2010 STD treatment guidelines," *Clinical Infectious Diseases*, vol. 53, pp. S153–S159, dec 2011.

[25] Y. Chen, V. Williams, M. Filippova, V. Filippov, and P. Duerksen-Hughes, "Viral carcinogenesis: Factors inducing DNA damage and virus integration," *Cancers*, vol. 6, no. 4, pp. 2155–2186, 2014.

[26] M. Stanley, "Immunobiology of HPV and HPV vaccines," *Gynecologic Oncology*, vol. 109, no. 2 SUPPL., pp. S15–S21, 2008.

[27] M. Pett and N. Coleman, "Integration of high-risk human papillomavirus: A key event in cervical carcinogenesis?," *Journal of Pathology*, vol. 212, pp. 356–367, aug 2007.

[28] A. P. Vizcaino, V. Moreno, F. X. Bosch, N. Munoz, X. M. Barros-Dios, J. Borras, and D. M. Parkin, "International trends in incidence of cervical cancer: II.

Squamous-cell carcinoma," *International Journal of Cancer*, vol. 86, pp. 429–435, may 2000.

[29] K. Nanda, D. C. McCrory, E. R. Myers, L. A. Bastian, V. Hasselblad, J. D. Hickey, and D. B. Matchar, "Accuracy of the papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: A systematic review," *Annals of Internal Medicine*, vol. 132, pp. 810–819, may 2000.

[30] NICE, "Guidance on the use of liquid-based cytology for cervical screening," *NICE technology appraisal guidance 69*, no. October, 2003.

[31] M. Arbyn, G. Ronco, A. Anttila, C. J. Meijer, M. Poljak, G. Ogilvie, G. Koliopoulos, P. Naucler, R. Sankaranarayanan, and J. Peto, "Evidence Regarding Human Papillomavirus Testing in Secondary Prevention of Cervical Cancer," *Vaccine*, vol. 30, pp. F88–F99, 2012.

[32] A. Subramaniam, J. M. Fauci, K. E. Schneider, J. M. Whitworth, B. K. Erickson, K. Kim, and W. K. Huh, "Invasive Cervical Cancer and Screening," *Journal of Lower Genital Tract Disease*, vol. 15, pp. 110–113, apr 2011.

[33] N. Wentzensen and M. Arbyn, "HPV-based cervical cancer screening- facts, fiction, and misperceptions," *Preventive Medicine*, vol. 98, pp. 33–35, may 2017.

[34] N. Kash, M. Lee, R. Kollipara, C. Downing, J. Guidry, and S. Tyring, "Safety and Efficacy Data on Vaccines and Immunization to Human Papillomavirus," *Journal of Clinical Medicine*, vol. 4, pp. 614–633, apr 2015.

[35] J. T. Bryan, B. Buckland, J. Hammond, and K. U. Jansen, "Prevention of cervical cancer: Journey to develop the first human papillomavirus virus-like particle vaccine and the next generation vaccine," *Current Opinion in Chemical Biology*, vol. 32, pp. 34–47, jun 2016.

[36] L. Zhai and E. Tumban, "Gardasil-9: A global survey of projected efficacy," *Antiviral Research*, vol. 130, pp. 101–109, jun 2016.

[37] A. Yang, E. Farmer, T. C. Wu, and C.-F. Hung, "Perspectives for therapeutic HPV vaccine development," *Journal of Biomedical Science*, vol. 23, p. 75, dec 2016.

[38] A. Yang, J. Jeang, K. Cheng, T. Cheng, B. Yang, T.-C. Wu, and C.-F. Hung, "Current state in the development of candidate therapeutic HPV vaccines.," *Expert Review of Vaccines*, vol. 0584, pp. 1–19, aug 2016.

[39] P. Vici, L. Pizzuti, L. Mariani, G. Zampa, D. Santini, L. Di Lauro, T. Gamucci, C. Natoli, P. Marchetti, M. Barba, M. Maugeri-Saccà, D. Sergi, F. Tomao, E. Vizza, S. Di Filippo, F. Paolini, G. Curzio, G. Corrado, A. Michelotti, G. Sanguineti, A. Giordano, R. De Maria, and A. Venuti, "Targeting immune response with therapeutic vaccines in premalignant lesions and cervical cancer: hope or reality from clinical studies," *Expert Review of Vaccines*, vol. 15, pp. 1327–1336, oct 2016.

[40] C. K. Howlader N, Noone AM, Krapcho M, Miller D, Bishop K, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, "SEER Cancer Statistics Review," 2013.

[41] A. Dueñas-González, M. Lizano, M. Candelaria, L. Cetina, C. Arce, and E. Cervera, "Epigenetics of cervical cancer. An overview and therapeutic perspectives.," *Molecular Cancer*, vol. 4, no. 1, p. 38, 2005.

[42] N. Bizzarri, V. Ghirardi, F. Alessandri, P. L. Venturini, M. Valenzano Menada, S. Rundle, U. Leone Roberti Maggiore, and S. Ferrero, "Bevacizumab for the treatment of cervical cancer," *Expert Opinion on Biological Therapy*, vol. 16, pp. 407–419, mar 2016.

[43] S. V. Graham and A. A. A. Faizo, "Control of human papillomavirus gene expression by alternative splicing," *Virus Research*, vol. 231, pp. 83–95, 2017.

[44] J. Doorbar, "The papillomavirus life cycle," *Journal of Clinical Virology*, vol. 32, pp. 7–15, mar 2005.

[45] Y. Xue, S. Bellanger, W. Zhang, D. Lim, J. Low, D. Lunny, and F. Thierry, "HPV16 E2 is an immediate early marker of viral infection, preceding E7 expression in precursor structures of cervical carcinoma," *Cancer Research*, vol. 70, no. 13, pp. 5316–5325, 2010.

[46] S. V. Graham, "Human papillomavirus: gene expression, regulation and prospects for novel diagnostic methods and antiviral therapies.," *Future Microbiology*, vol. 5, pp. 1493–506, oct 2010.

[47] J. M. Bodily, K. P. Mehta, and L. A. Laimins, "Human papillomavirus E7 enhances hypoxia-inducible factor 1-mediated transcription by inhibiting binding of histone deacetylases," *Cancer Research*, vol. 71, pp. 1187–1195, feb 2011.

[48] S. V. Graham, "Papillomavirus 3' UTR regulatory element," *Frontiers in Bioscience*, vol. 13, pp. 5646–63, 2008.

[49] A. J. Levine, "The common mechanisms of transformation by the small DNA tumor viruses: The inactivation of tumor suppressor gene products: p53," *Virology*, vol. 384, pp. 285–293, feb 2009.

[50] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, pp. 57–70, jan 2000.

[51] H. L. Howie, R. A. Katzenellenbogen, and D. A. Galloway, "Papillomavirus E6 proteins," *Virology*, vol. 384, no. 2, pp. 324–334, 2009.

[52] 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*, vol. 75, pp. 495–505, nov 1993.

[53] M. C. Thomas and C. M. Chiang, "E6 oncoprotein represses p53-dependent gene activation via inhibition of protein acetylation independently of inducing p53 degradation," *Molecular Cell*, vol. 17, pp. 251–264, jan 2005.

[54] X. Xie, L. Piao, B. N. Bullock, A. Smith, T. Su, M. Zhang, T. N. Teknos, P. S. Arora, and Q. Pan, "Targeting HPV16 E6-p300 interaction reactivates

p53 and inhibits the tumorigenicity of HPV-positive head and neck squamous cell carcinoma," *Oncogene*, vol. 33, pp. 1037–1046, feb 2014.

[55] P. Gariglio and J. Organista Nava, "Role of HR-HPVs E6 and E7 Oncoproteins in Cervical Carcinogenesis," *Journal of Molecular and Genetic Medicine*, vol. 10, no. 2, pp. 1–11, 2016.

[56] W. C. Hahn, "Immortalization and transformation of human cells.," *Molecules and Cells*, vol. 13, no. 3, pp. 351–361, 2002.

[57] L. Gewin, H. Myers, T. Kiyono, and D. A. Galloway, "Identification of a novel telomerase repressor that interacts with the human papillomavirus type-16 E6/E6-AP complex," *Genes and Development*, vol. 18, pp. 2269–2282, sep 2004.

[58] T. Yugawa and T. Kiyono, "Molecular mechanisms of cervical carcinogenesis by high-risk human papillomaviruses: Novel functions of E6 and E7 oncoproteins," *Reviews in Medical Virology*, vol. 19, no. 2, pp. 97–113, 2009.

[59] M. E. McLaughlin-Drubin and K. Münger, "The human papillomavirus E7 oncoprotein," *Virology*, vol. 384, pp. 335–344, feb 2009.

[60] S. Duensing and K. Münger, "Human papillomaviruses and centrosome duplication errors: modeling the origins of genomic instability," *Oncogene*, vol. 21, pp. 6241–6248, sep 2002.

[61] E. S. Hickman, M. C. Moroni, and K. Helin, "The role of p53 and pRB in apoptosis and cancer," *Current Opinion in Genetics and Development*, vol. 12, pp. 60–66, feb 2002.

[62] C. A. Moody and L. A. Laimins, "Human papillomavirus oncoproteins: pathways to transformation," *Nature Reviews Cancer*, vol. 10, pp. 550–560, aug 2010.

[63] H. U. Bernard, "Regulatory elements in the viral genome," *Virology*, vol. 445, pp. 197–204, oct 2013.

[64] H. U. Bernard, "Gene expression of genital human papillomaviruses and considerations on potential antiviral approaches," *Antiviral Therapy*, vol. 7, pp. 219–237, dec 2002.

[65] W. Stünkel and H. U. Bernard, "The chromatin structure of the long control region of human papillomavirus type 16 represses viral oncoprotein expression.," *Journal of Virology*, vol. 73, no. 3, pp. 1918–30, 1999.

[66] S. V. Graham, "Human Papillomavirus E2 Protein: Linking Replication, Transcription, and RNA Processing," *Journal of Virology*, vol. 90, no. 19, pp. 8384–8388, 2016.

[67] J. Durzynska, K. Lesniewicz, and E. Poreba, "Human papillomaviruses in epigenetic regulations," *Mutation Research*, vol. 772, pp. 36–50, 2017.

[68] A. A. McBride, "The Papillomavirus E2 Proteins," *Virology*, vol. 445, no. 0, pp. 57–79, 2013.

[69] T. Iftner, J. Haedicke-Jarboui, S. Y. Wu, and C. M. Chiang, "Involvement of Brd4 in different steps of the papillomavirus life cycle," *Virus Research*, vol. 231, pp. 76–82, 2017.

[70] F. Thierry and M. Yaniv, "The BPV1-E2 trans-acting protein can be either an activator or a repressor of the HPV18 regulatory region.," *The EMBO journal*, vol. 6, no. 11, pp. 3391–7, 1987.

[71] E. E. Hernandez-Ramon, J. E. Burns, W. Zhang, H. F. Walker, S. Allen, A. A. Antson, and N. J. Maitland, "Dimerization of the Human Papillomavirus Type 16 E2 N Terminus Results in DNA Looping within the Upstream Regulatory Region," *Journal of Virology*, vol. 82, no. 10, pp. 4853–4861, 2008.

[72] R. Li, J. D. Knight, S. P. Jackson, R. Tjian, and M. R. Botchan, "Direct interaction between Sp1 and the BPV enhancer E2 protein mediates synergistic activation of transcription," *Cell*, vol. 65, no. 3, pp. 493–505, 1991.

[73] C. Helfer, J. Yan, and J. You, "The Cellular Bromodomain Protein Brd4 has Multiple Functions in E2-Mediated Papillomavirus Transcription Activation," *Viruses*, vol. 6, no. 8, pp. 3228–3249, 2014.

[74] M.-r. Schweiger, J. You, and P. M. Howley, "Bromodomain Protein 4 Mediates the Papillomavirus E2 Transcriptional Activation Function Bromodomain Protein 4 Mediates the Papillomavirus E2 Transcriptional Activation Function," *Journal of Virology*, vol. 80, no. 9, pp. 4276–4285, 2006.

[75] M. G. McPhillips, J. G. Oliveira, J. E. Spindler, R. Mitra, and A. A. McBride, "Brd4 is required for E2-mediated transcriptional activation but not genome partitioning of all papillomaviruses.," *Journal of Virology*, vol. 80, no. 19, pp. 9530–43, 2006.

[76] S. Y. Wu and C. M. Chiang, "The double bromodomain-containing chromatin adaptor Brd4 and transcriptional regulation," *Journal of Biological Chemistry*, vol. 282, no. 18, pp. 13141–13145, 2007.

[77] C. a. J. Horvath, G. a. V. Boulet, V. M. Renoux, P. O. Delvenne, and J.-P. J. Bogers, "Mechanisms of cell entry by human papillomaviruses: an overview.," *Virology Journal*, vol. 7, p. 11, 2010.

[78] J. Broniarczyk, D. Pim, P. Massimi, M. Bergant, A. Goździcka-Józefiak, C. Crump, and L. Banks, "The VPS4 component of the ESCRT machinery plays an essential role in HPV infectious entry and capsid disassembly," *Scientific Reports*, vol. 7, no. March, p. 45159, 2017.

[79] S. DiGiuseppe, W. Luszczek, T. R. Keiffer, M. Bienkowska-Haba, L. G. M. Guion, and M. J. Sapp, "Incoming human papillomavirus type 16 genome resides in a vesicular compartment throughout mitosis," *PNAS*, vol. 113, no. 22, pp. 6289–6294, 2016.

[80] S. DiGiuseppe, M. Bienkowska-Haba, and M. Sapp, "Human papillomavirus entry: hiding in a bubble," *Journal of Virology*, vol. 90, no. 18, pp. JVI.01065–16, 2016.

[81] K. L. Conger, J.-s. Liu, S.-r. Kuo, L. T. Chow, and T. S. Wang, "Human Papillomavirus DNA Replication," *The Journal of Biological Chemistry*, vol. 274, pp. 2696–2705, jan 1999.

[82] E. J. Crosbie, M. H. Einstein, S. Franceschi, and H. C. Kitchener, "Human papillomavirus and cervical cancer," *The Lancet*, vol. 382, pp. 889–899, sep 2013.

[83] M. A. Stanley, M. R. Pett, and N. Coleman, "HPV: from infection to cancer.," *Biochemical Society Transactions*, vol. 35, pp. 1456–60, dec 2007.

[84] M. Stanley, D. R. Lowy, and I. Frazer, "Chapter 12: Prophylactic HPV vaccines: Underlying mechanisms," *Vaccine*, vol. 24, pp. S106–S113, aug 2006.

[85] M. Schiffman, P. E. Castle, J. Jeronimo, A. C. Rodriguez, and S. Wacholder, "Human papillomavirus and cervical cancer," *Lancet*, vol. 370, no. 9590, pp. 890–907, 2007.

[86] N. Wentzensen, S. Vinokurova, and M. Von Knebel Doeberitz, "Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract," *Cancer Research*, vol. 64, no. 11, pp. 3878–3884, 2004.

[87] Z. Hu, D. Zhu, W. Wang, W. Li, W. Jia, X. Zeng, W. Ding, L. Yu, X. Wang, and L. Wang, "Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism," *Nature Genetics*, vol. 47, no. 2, pp. 158–163, 2015.

[88] J.-W. Zhao, F. Fang, Y. Guo, T.-L. Zhu, Y.-Y. Yu, F.-F. Kong, L.-F. Han, D.-S. Chen, and F. Li, "HPV16 integration probably contributes to cervical oncogenesis through interrupting tumor suppressor genes and inducing chromosome instability," *Journal of Experimental & Clinical Cancer Research*, vol. 35, no. 1, p. 180, 2016.

[89] S. Jeon, B. L. Allen-Hoffmann, and P. F. Lambert, "Integration of human pa-

pillomavirus type 16 into the human genome correlates with a selective growth advantage of cells.," *Journal of Virology*, vol. 69, no. 5, pp. 2989–97, 1995.

[90] M. R. Pett, W. O. Alazawi, I. Roberts, S. Dowen, D. I. Smith, M. A. Stanley, and N. Coleman, "Acquisition of High-Level Chromosomal Instability Is Associated with Integration of Human Papillomavirus Type 16 in Cervical Keratinocytes," *Cancer Research*, vol. 64, no. 4, pp. 1359–1368, 2004.

[91] Y. Ueda, T. Enomoto, T. Miyatake, K. Ozaki, T. Yoshizaki, H. Kanao, Y. Ueno, R. Nakashima, K. R. Shroyer, and Y. Murata, "Monoclonal expansion with integration of high-risk type human papillomaviruses is an initial step for cervical carcinogenesis: association of clonal status and human papillomavirus infection with clinical outcome in cervical intraepithelial neoplasia," *Laboratory Investigation*, vol. 83, pp. 1517–1527, oct 2003.

[92] B. A. V. Tine, J. C. Kappes, N. S. Banerjee, J. Knops, L. Lai, R. D. M. Steenbergen, C. L. J. M. Meijer, P. J. F. Snijders, P. Chatis, T. R. Broker, P. T. Moen, and L. T. Chow, "Clonal Selection for Transcriptionally Active Viral Oncogenes during Progression to Cancer," *Journal of Virology*, vol. 78, pp. 11172–11186, oct 2004.

[93] C. Ziegert, N. Wentzensen, S. Vinokurova, F. Kisseljov, J. Einenkel, M. Hoeckel, and M. von Knebel Doeberitz, "A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques," *Oncogene*, vol. 22, pp. 3977–3984, jun 2003.

[94] E. C. Thorland, S. L. Myers, D. H. Persing, G. Sarkar, R. M. McGovern, B. S. Gostout, and D. I. Smith, "Human papillomavirus type 16 integrations in cervical tumors frequently occur in common fragile sites," *Cancer Research*, vol. 60, pp. 5916–5921, nov 2000.

[95] E. C. Thorland, S. L. Myers, B. S. Gostout, and D. I. Smith, "Common fragile sites are preferential targets for HPV16 integrations in cervical tumors," *Oncogene*, vol. 22, no. 8, pp. 1225–1237, 2003.

[96] M. Schmitz, C. Driesch, L. Jansen, I. B. Runnebaum, and M. Dürst, "Non-random integration of the HPV genome in cervical cancer," *PLoS ONE*, vol. 7, p. e39632, jun 2012.

[97] A. I. Ojesina, L. Lichtenstein, S. S. Freeman, C. S. Pedamallu, I. Imaz-Rosshandler, T. J. Pugh, A. D. Cherniack, L. Ambrogio, K. Cibulskis, and B. Bertelsen, "Landscape of genomic alterations in cervical carcinomas," *Nature*, vol. 506, no. 7488, pp. 371–375, 2013.

[98] R. D. Burk, Z. Chen, C. Saller, K. Tarvin, A. L. Carvalho, C. Scapulatempo-Neto, H. C. Silveira, J. H. Fregnani, C. J. Creighton, M. L. Anderson, and P. Castro, "Integrated genomic and molecular characterization of cervical cancer," *Nature*, vol. 543, no. 7645, pp. 378–384, 2017.

[99] D. M. Winder, M. R. Pett, N. Foster, M. K. K. Shivji, M. T. Herdman, M. A. Stanley, A. R. Venkitaraman, and N. Coleman, "An increase in DNA double-strand breaks, induced by Ku70 depletion, is associated with human papillomavirus 16 episome loss and de novo viral integration events," *Journal of Pathology*, vol. 213, pp. 27–34, sep 2007.

[100] M. Parfenov, C. S. Pedamallu, N. Gehlenborg, S. S. Freeman, L. Danilova, C. A. Bristow, S. Lee, A. G. Hadjipanayis, E. V. Ivanova, and M. D. Wilkerson, "Characterization of HPV and host genome interactions in primary head and neck cancers," *Proceedings of the National Academy of Sciences*, vol. 111, no. 43, pp. 15544–15549, 2014.

[101] A. Holmes, S. Lameiras, E. Jeannot, Y. Marie, L. Castera, X. Sastre-Garau, and A. Nicolas, "Mechanistic signatures of HPV insertions in cervical carcinomas," *npj Genomic Medicine*, vol. 1, no. 1, p. 16004, 2016.

[102] M. A. Stanley, H. M. Browne, M. Appleby, and A. C. Minson, "Properties of a nontumorigenic human cervical keratinocyte cell line," *International Journal of Cancer*, vol. 43, pp. 672–676, apr 1989.

[103] W. Alazawi, M. Pett, B. Arch, L. Scott, T. Freeman, M. A. Stanley, and

N. Coleman, "Changes in cervical keratinocyte gene expression associated with integration of human papillomavirus 16," *Cancer Research*, vol. 62, pp. 6959–6965, dec 2002.

[104] K. L. Dall, C. G. Scarpini, I. Roberts, D. M. Winder, M. A. Stanley, B. Muralidhar, M. T. Herdman, M. R. Pett, and N. Coleman, "Characterization of naturally occurring HPV16 integration sites isolated from cervical keratinocytes under noncompetitive conditions," *Cancer Research*, vol. 68, no. 20, pp. 8249–8259, 2008.

[105] A. Feinberg, B. Vogelstein, M. Droller, S. Baylin, and B. Nelkin, "Mutation affecting the 12th amino acid of the c-Ha-ras oncogene product occurs infrequently in human cancer," *Science*, vol. 220, no. 4602, pp. 1175–1177, 1983.

[106] K. W. Kinzler and B. Vogelstein, "Lessons from hereditary colorectal cancer," *Cell*, vol. 87, no. 2, pp. 159–170, 1996.

[107] C. H. Waddington, "The epigenotype. 1942.," *International Journal of Epidemiology*, vol. 41, no. 1, pp. 10–13, 2012.

[108] A. G. Muntean and J. L. Hess, "Epigenetic Dysregulation in Cancer," *The American Journal of Pathology*, vol. 175, no. 4, pp. 1353–1361, 2009.

[109] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[110] J. Sandoval and M. Esteller, "Cancer epigenomics: Beyond genomics," *Current Opinion in Genetics and Development*, vol. 22, no. 1, pp. 50–55, 2012.

[111] A. P. Feinberg and B. Vogelstein, "Hypomethylation distinguishes genes of some human cancers from their normal counterparts," *Nature*, vol. 301, pp. 89–92, jan 1983.

[112] D. B. Seligson, S. Horvath, M. A. McBrian, V. Mah, H. Yu, S. Tze, Q. Wang, D. Chia, L. Goodglick, and S. K. Kurdistani, "Global Levels of Histone Modifications Predict Prognosis in Different Cancers," *The American Journal of Pathology*, vol. 174, pp. 1619–1628, may 2009.

[113] S. Sharma, T. K. Kelly, and P. A. Jones, "Epigenetics in cancer," *Carcinogenesis*, vol. 31, no. 1, pp. 27–36, 2009.

[114] M. A. Dawson and T. Kouzarides, "Cancer epigenetics: From mechanism to therapy," *Cell*, vol. 150, no. 1, pp. 12–27, 2012.

[115] R. Holliday and J. Pugh, "DNA modification mechanisms and gene activity during development.," *Science*, vol. 187, no. 4173, pp. 226–232, 1975.

[116] B. E. Bernstein, A. Meissner, and E. S. Lander, "The Mammalian Epigenome," *Cell*, vol. 128, no. 4, pp. 669–681, 2007.

[117] M. Esteller, "Epigenetics provides a new generation of oncogenes and tumour-suppressor genes.," *British journal of cancer*, vol. 96 Suppl, no. 2, pp. R26–R30, 2007.

[118] A. Bird, "The essentials of DNA methylation," *Cell*, vol. 70, no. 1, pp. 5–8, 1992.

[119] Toraño, "Epigenetics in Cancer," *Molecular Origins of Cancer*, vol. 358, pp. 1–12, 2008.

[120] A. Eden, "Chromosomal Instability and Tumors Promoted by DNA Hypomethylation," *Science*, vol. 300, pp. 455–455, apr 2003.

[121] M. F. Fraga, M. Herranz, J. Espada, E. Ballestar, M. F. Paz, S. Ropero, E. Erkek, O. Bozdogan, H. Peinado, A. Niveleau, J. H. Mao, A. Balmain, A. Cano, and M. Esteller, "A mouse skin multistage carcinogenesis model reflects the aberrant DNA methylation patterns of human tumors," *Cancer Res*, vol. 64, no. 16, pp. 5527–5534, 2004.

[122] Z.-M. Zheng and X. Wang, "Regulation of cellular miRNA expression by human papillomaviruses," *Biochimica et Biophysica Acta*, vol. 1809, pp. 668–677, 2011.

[123] G. Calin and C. Croce, "MicroRNA signatures in human cancers.," *Nature Reviews Cancer*, vol. 6, no. 11, pp. 855–866, 2006.

[124] V. W. Zhou, A. Goren, and B. E. Bernstein, "Charting histone modifications and the functional organization of mammalian genomes," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 7–18, 2011.

[125] T. J. Richmond and C. A. Davey, "The structure of DNA in the nucleosome core," *Nature*, vol. 423, pp. 145–150, may 2003.

[126] A. J. Bannister and T. Kouzarides, "Regulation of chromatin by histone modifications," *Cell Research*, vol. 21, no. 3, pp. 381–395, 2011.

[127] K. J. Falkenberg and R. W. Johnstone, "Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders," *Nature Reviews Drug Discovery*, vol. 13, pp. 673–691, sep 2014.

[128] T. Kouzarides, "Chromatin Modifications and Their Function," *Cell*, vol. 128, no. 4, pp. 693–705, 2007.

[129] C. H. Arrowsmith, C. Bountra, P. V. Fish, K. Lee, and M. Schapira, "Epigenetic protein families: a new frontier for drug discovery," *Nature Reviews Drug Discovery*, vol. 11, pp. 384–400, apr 2012.

[130] V. G. Allfrey, R. Faulkner, and A. E. Mirsky, "Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis," *Proceedings of the National Academy of Sciences*, vol. 51, pp. 786–794, may 1964.

[131] A. H. Hassan, K. E. Neely, and J. L. Workman, "Histone acetyltransferase complexes stabilize SWI/SNF binding to promoter nucleosomes," *Cell*, vol. 104, pp. 817–827, mar 2001.

[132] S. K. Kurdistani and M. Grunstein, "Histone acetylation and deacetylation in yeast," *Nature Reviews Molecular Cell Biology*, vol. 4, pp. 276–284, nov 2003.

[133] C. K. Govind, F. Zhang, H. Qiu, K. Hofmeyer, and A. G. Hinnebusch, "Gcn5 Promotes Acetylation, Eviction, and Methylation of Nucleosomes in Transcribed Coding Regions," *Molecular Cell*, vol. 25, pp. 31–42, jan 2007.

[134] S. Mujtaba, L. Zeng, and M.-M. Zhou, "Structure and acetyl-lysine recognition of the bromodomain," *Oncogene*, vol. 26, pp. 5521–5527, aug 2007.

[135] D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao, "Dynamic Regulation of Nucleosome Positioning in the Human Genome," *Cell*, vol. 132, pp. 887–898, mar 2008.

[136] L. Ellis, P. W. Atadja, and R. W. Johnstone, "Epigenetics in cancer: Targeting chromatin modifications," *Molecular Cancer Therapeutics*, vol. 8, no. 6, pp. 1409–1420, 2009.

[137] Z. Wang, C. Zang, K. Cui, D. E. Schones, A. Barski, W. Peng, and K. Zhao, "Genome-wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes," *Cell*, vol. 138, no. 5, pp. 1019–1031, 2009.

[138] M. F. Fraga and M. Esteller, "Towards the human cancer epigenome: A first draft of histone modifications," *Cell Cycle*, vol. 4, pp. 1377–1381, oct 2005.

[139] D. B. Seligson, S. Horvath, T. Shi, H. Yu, S. Tze, M. Grunstein, and S. K. Kurdistani, "Global histone modification patterns predict risk of prostate cancer recurrence," *Nature*, vol. 435, pp. 1262–1266, jun 2005.

[140] R. Margueron, P. Trojer, and D. Reinberg, "The key to development: Interpreting the histone code?," *Current Opinion in Genetics and Development*, vol. 15, pp. 163–176, apr 2005.

[141] K. S. Champagne and T. G. Kutateladze, "Structural insight into histone recognition by the ING PHD fingers.," *Current Drug Targets*, vol. 10, pp. 432–441, may 2009.

[142] J. Kim, J. Daniel, A. Espejo, A. Lake, M. Krishna, L. Xia, Y. Zhang, and M. T. Bedford, "Tudor, MBT and chromo domains gauge the degree of lysine methylation," *EMBO Reports*, vol. 7, pp. 397–403, apr 2006.

[143] N. Kitkumthorn, P. Yanatatsanajit, S. Kiatpongsan, C. Phokaew, S. Triratanachat, P. Trivijitsilp, W. Termrungruanglert, D. Tresukosol, S. Niruthis-

ard, and A. Mutirangura, "Cyclin A1 promoter hypermethylation in human papillomavirus-associated cervical cancer.," *BMC Cancer*, vol. 6, p. 55, 2006.

[144] J. de Wilde, J. M. Kooter, R. M. Overmeer, D. Claassen-Kramer, C. J. L. M. Meijer, P. J. F. Snijders, and R. D. M. Steenbergen, "hTERT promoter activity and CpG methylation in HPV-induced carcinogenesis.," *BMC Cancer*, vol. 10, p. 271, 2010.

[145] Q. Tao and K. D. Robertson, "Stealth technology: How Epstein-Barr virus utilizes DNA methylation to cloak itself from immune detection," *Clinical Immunology*, vol. 109, no. 1, pp. 53–63, 2003.

[146] M. Melar-New and L. A. Laimins, "Human papillomaviruses modulate expression of microRNA 203 upon epithelial differentiation to control levels of p63 proteins.," *Journal of Virology*, vol. 84, no. 10, pp. 5212–21, 2010.

[147] T. Yao and Z. Lin, "MiR-21 is involved in cervical squamous cell tumorigenesis and regulates CCL20.," *Biochimica et Biophysica Acta*, vol. 1822, no. 2, pp. 248–60, 2012.

[148] A. Jansma, M. Martinez-Yamout, R. Liao, P. Sun, H. Dyson, and P. Wright, "The high-risk HPV16 E7 oncoprotein mediates interaction between the transcriptional coactivator CBP and the retinoblastoma protein pRb," *Journal of Molecular Biology*, vol. 426, no. 24, pp. 4030–4048, 2014.

[149] N. Avvakumov, J. Torchia, and J. Mymryk, "Interaction of the HPV E7 proteins with the pCAF acetyltransferase.," *Oncogene*, vol. 22, no. 25, pp. 3833–3841, 2003.

[150] W. A. Burgers, L. Blanchon, S. Pradhan, Y. de Launoit, T. Kouzarides, and F. Fuks, "Viral oncoproteins target the DNA methyltransferases," *Oncogene*, vol. 26, pp. 1650–1655, mar 2007.

[151] M. E. McLaughlin-Drubin, C. P. Crum, and K. Münger, "Human papillomavirus E7 oncoprotein induces KDM6A and KDM6B histone demethylase

expression and causes epigenetic reprogramming," *Proceedings of the National Academy of Sciences*, vol. 108, no. 5, pp. 2130–2135, 2011.

[152] A. Brehm, S. J. Nielsen, E. A. Miska, D. J. McCance, J. L. Reid, A. J. Bannister, and T. Kouzarides, "The E7 oncoprotein associates with Mi2 and histone deacetylase activity to promote cell growth," *EMBO Journal*, vol. 18, pp. 2449–2458, may 1999.

[153] E. Ballestar, M. F. Paz, L. Valle, S. Wei, M. F. Fraga, J. Espada, J. C. Cigudosa, T. H. M. Huang, and M. Esteller, "Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer," *EMBO Journal*, vol. 22, no. 23, pp. 6335–6345, 2003.

[154] D. S. Wendler, "Problems with the consensus definition of the therapeutic misconception," *Journal of Clinical Ethics*, vol. 24, no. 4, pp. 387–394, 2013.

[155] F. Barlési, G. Giaccone, M. I. Gallegos-Ruiz, A. Loundou, S. W. Span, P. Lefesvre, F. A. E. Kruyt, and J. A. Rodriguez, "Global histone modifications predict prognosis of resected non-small-cell lung cancer," *Journal of Clinical Oncology*, vol. 25, pp. 4358–4364, oct 2007.

[156] F. Barbisan, R. Mazzucchelli, A. Santinelli, D. Stramazzotti, M. Scarpelli, A. Lopez-Beltran, L. Cheng, and R. Montironi, "Immunohistochemical evaluation of global DNA methylation and histone acetylation in papillary urothelial neoplasm of low malignant potential," *International Journal of Immunopathology and Pharmacology*, vol. 21, pp. 615–623, jul 2008.

[157] C. B. Yoo and P. A. Jones, "Epigenetic therapy of cancer: past, present and future," *Nature Reviews Drug Discovery*, vol. 5, pp. 37–50, jan 2006.

[158] E. Glass and P. H. Viale, "Histone deacetylase inhibitors: Novel agents in cancer treatment," *Clinical Journal of Oncology Nursing*, vol. 17, pp. 34–40, feb 2013.

[159] E. Kaminskas, A. Farrell, S. Abraham, A. Baird, L. S. Hsieh, S. L. Lee, J. K. Leighton, H. Patel, A. Rahman, R. Sridhara, Y. C. Wang, and R. Pazdur,

"Approval summary: Azacitidine for treatment of myelodysplastic syndrome subtypes," *Clinical Cancer Research*, vol. 11, pp. 3604–3608, may 2005.

[160] G. Lolli, "Binding to DNA of the RNA-polymerase II C-terminal domain allows discrimination between Cdk7 and Cdk9 phosphorylation," *Nucleic Acids Research*, vol. 37, pp. 1260–1268, mar 2009.

[161] K. Adelman and J. T. Lis, "Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans," *Nature Reviews Genetics*, vol. 13, pp. 720–731, sep 2012.

[162] S. Baumli, G. Lolli, E. D. Lowe, S. Troiani, L. Rusconi, A. N. Bullock, J. E. É. E. Debreczeni, S. Knapp, and L. N. Johnson, "The structure of P-TEFb (CDK9/cyclin T1), its complex with flavopiridol and regulation by phosphorylation.," *EMBO Journal*, vol. 27, pp. 1907–1918, jul 2008.

[163] I. Jonkers and J. T. Lis, "Getting up to speed with transcription elongation by RNA polymerase II," *Nature Reviews Molecular Cell Biology*, vol. 16, no. 3, pp. 167–177, 2015.

[164] F. Morales and A. Giordano, "Overview of CDK9 as a target in cancer research," *Cell Cycle*, vol. 15, no. 4, pp. 519–527, 2016.

[165] K. M. Harlen and L. S. Churchman, "The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain," *Nature Reviews Molecular Cell Biology*, vol. 18, no. 4, pp. 263–273, 2017.

[166] B. M. Peterlin and D. H. Price, "Controlling the Elongation Phase of Transcription with P-TEFb," *Molecular Cell*, vol. 23, pp. 297–305, aug 2006.

[167] B. J. Krueger, K. Varzavand, J. J. Cooper, and D. H. Price, "The mechanism of release of P-TEFb and HEXIM1 from the 7SK snRNP by viral and cellular activators includes a conformational change in 7SK," *PLoS ONE*, vol. 5, p. e12335, aug 2010.

[168] Z. Luo, C. Lin, and A. Shilatifard, "The super elongation complex (SEC) family in transcriptional control," *Nature Reviews Molecular Cell Biology*, vol. 13, pp. 543–547, aug 2012.

[169] P. B. Rahl, C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine, C. B. Burge, P. A. Sharp, and R. A. Young, "C-Myc regulates transcriptional pause release," *Cell*, vol. 141, pp. 432–445, apr 2010.

[170] M. C. Patel, M. Debrosse, M. Smith, A. Dey, W. Huynh, N. Sarai, T. D. Heightman, T. Tamura, and K. Ozato, "BRD4 Coordinates Recruitment of Pause Release Factor P-TEFb and the Pausing Complex NELF/DSIF To Regulate Transcription Elongation of Interferon-Stimulated Genes," *Molecular and Cellular Biology*, vol. 33, pp. 2497–2507, jun 2013.

[171] T. I. Lee and R. A. Young, "Transcriptional regulation and its misregulation in disease," *Cell*, vol. 152, no. 6, pp. 1237–1251, 2013.

[172] K. J. Moon, K. Mochizuki, M. Zhou, H. S. Jeong, J. N. Brady, and K. Ozato, "The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription," *Molecular Cell*, vol. 19, pp. 523–534, aug 2005.

[173] J. Shi and C. R. Vakoc, "The Mechanisms behind the Therapeutic Activity of BET Bromodomain Inhibition," *Molecular Cell*, vol. 54, no. 5, pp. 72–736, 2014.

[174] F. Neri, S. Rapelli, A. Krepelova, D. Incarnato, C. Parlato, G. Basile, M. Maldotti, F. Anselmi, and S. Oliviero, "Intragenic DNA methylation prevents spurious transcription initiation," *Nature*, vol. 543, pp. 72–77, feb 2017.

[175] S. B. Baylin and P. A. Jones, "A decade of exploring the cancer epigenome biological and translational implications," *Nature Reviews Cancer*, vol. 11, pp. 726–734, sep 2011.

[176] K. Akagi, J. Li, T. R. Broutian, H. Padilla-Nash, W. Xiao, B. Jiang, J. W. Rocco, T. N. Teknos, B. Kumar, D. Wangsa, D. He, T. Ried, D. E. Symer, and

M. L. Gillison, "Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability," *Genome Research*, vol. 24, no. 2, pp. 185–199, 2014.

[177] F. Friedl, I. Kimura, T. Osato, Y. Ito, M. Namiki, M. Inoue, S. Ratnam, F. Coutlee, E. Franco, and D. Wallwiener, "Studies on a New Human Cell Line (SiHa) Derived from Carcinoma of Uterus. I. Its Establishment and Morphology," *Experimental Biology and Medicine*, vol. 135, pp. 543–545, nov 1970.

[178] C. C. Baker, W. C. Phelps, V. Lindgren, M. J. Braun, M. A. Gonda, and P. M. Howley, "Structural and transcriptional analysis of human papillomavirus type 16 sequences in cervical carcinoma cell lines.," *Journal of Virology*, vol. 61, pp. 962–971, apr 1987.

[179] G. Tadaro and H. Green, "Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines.," *The Journal of cell biology*, vol. 17, pp. 299–313, may 1963.

[180] M. W. Pfaffl, "A new mathematical model for relative quantification in real-time RT-PCR," *Nucleic Acids Research*, vol. 29, no. 9, pp. 45e–45, 2001.

[181] M. Pfaffl, "Quantification strategies in real-time PCR Michael W . Pfaffl," *A-Z of Quantitative PCR*, pp. 87–112, 2004.

[182] D. J. Bolland, M. R. King, W. Reik, A. E. Corcoran, and C. Krueger, "Robust 3D DNA FISH Using Directly Labeled Probes," *Journal of Visualized Experiments*, no. 78, pp. 1–9, 2013.

[183] C. G. Scarpini, I. J. Groves, M. R. Pett, D. Ward, and N. Coleman, "Virus transcript levels and cell growth rates after naturally occurring HPV16 integration events in basal cervical keratinocytes," *Journal of Pathology*, vol. 233, no. 3, pp. 281–293, 2014.

[184] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch, "Histone H3K27ac separates active from

poised enhancers and predicts developmental state," *Proceedings of the National Academy of Sciences*, vol. 107, pp. 21931–21936, dec 2010.

[185] D. E. Sterner and S. L. Berger, "Acetylation of histones and transcription-related factors.," *Microbiology and molecular biology reviews : MMBR*, vol. 64, no. 2, pp. 435–459, 2000.

[186] V. Sapountzi, I. R. Logan, and C. N. Robson, "Cellular functions of TIP60," *International Journal of Biochemistry and Cell Biology*, vol. 38, no. 9, pp. 1496–1509, 2006.

[187] S. A. Gayther, S. J. Batley, L. Linger, A. Bannister, K. Thorpe, S.-F. Chin, Y. Daigo, P. Russell, A. Wilson, H. M. Sowter, J. D. Delhanty, B. A. Ponder, T. Kouzarides, and C. Caldas, "Mutations truncating the EP300 acetylase in human cancers," *Nature Genetics*, vol. 24, no. 3, pp. 300–303, 2000.

[188] N. G. Iyer, H. Özdag, and C. Caldas, "p300/CBP and cancer," *Oncogene*, vol. 23, pp. 4225–4231, may 2004.

[189] J. You and P. Jones, "Cancer Genetics and Epigenetics : Two Sides of the Same Coin ?," *Cancer Cell*, vol. 22, no. 1, pp. 9–20, 2012.

[190] G. Romano, "Deregulations in the cyclin-dependent kinase-9-related pathway in cancer: implications for drug discovery and development.," *ISRN Oncology*, vol. 2013, pp. 18–20, 2013.

[191] E. Straub, J. Fertey, M. Dreer, T. Iftner, and F. Stubenrauch, "Characterization of the Human Papillomavirus 16 E8 Promoter.," *Journal of Virology*, vol. 89, no. 14, pp. 7304–13, 2015.

[192] T. Van Den Bosch, A. Boichenko, N. G. Leus, M. E. Ourailidou, H. Wapenaar, D. Rotili, A. Mai, A. Imhof, R. Bischoff, H. J. Haisma, and F. J. Dekker, "The histone acetyltransferase p300 inhibitor C646 reduces pro-inflammatory gene expression and inhibits histone deacetylases," *Biochemical Pharmacology*, vol. 102, pp. 130–140, 2016.

[193] M. Ghizzoni, J. Wu, T. Gao, H. J. Haisma, F. J. Dekker, and Y. George Zheng, "6-alkylsalicylates are selective Tip60 inhibitors and target the acetyl-CoA binding site," *European Journal of Medicinal Chemistry*, vol. 47, no. 1, pp. 337–344, 2012.

[194] J. P. Mosmann, M. S. Monetti, M. C. Frutos, A. X. Kiguen, R. F. Venezuela, and C. G. Cuffini, "Mutation detection of E6 and LCR genes from HPV 16 associated with carcinogenesis," *Asian Pacific Journal of Cancer Prevention*, vol. 16, no. 3, pp. 1151–1157, 2015.

[195] I. Cornet, T. Gheit, S. Franceschi, J. Vignat, R. D. Burk, B. S. Sylla, M. Tommasino, and G. M. Clifford, "Human Papillomavirus Type 16 Genetic Variants: Phylogeny and Classification Based on E6 and LCR," *Journal of Virology*, vol. 86, no. 12, pp. 6855–6861, 2012.

[196] I. Cornet, T. Gheit, M. R. Iannacone, J. Vignat, B. S. Sylla, A. Del Mistro, S. Franceschi, M. Tommasino, and G. M. Clifford, "HPV16 genetic variation and the development of cervical cancer worldwide," *British Journal of Cancer*, vol. 108, no. 1, pp. 240–244, 2013.

[197] L. Ho, S. Y. Chan, V. Chow, T. Chong, S. K. Tay, L. L. Villa, and H. U. Bernard, "Sequence Variants of Human Papillomavirus Type-16 in Clinical-Samples Permit Verification and Extension of Epidemiologic Studies and Construction of a Phylogenetic Tree," *Journal of Clinical Microbiology*, vol. 29, no. 9, pp. 1765–1772, 1991.

[198] M. J. Lace, C. Isacson, J. R. Anson, A. T. Lörincz, S. P. Wilczynski, T. H. Haugen, and L. P. Turek, "Upstream regulatory region alterations found in human papillomavirus type 16 (HPV-16) isolates from cervical carcinomas increase transcription, ori function, and HPV immortalization capacity in culture.," *Journal of Virology*, vol. 83, no. 15, pp. 7457–66, 2009.

[199] C. Pientong, P. Wongwarissara, T. Ekalaksananan, P. Swangphon, P. Kleebkaow, B. Kongyingyoes, S. Siriaunkgul, K. Tungsinmunkong, and C. Suthip-

intawong, "Association of human papillomavirus type 16 long control region mutation and cervical cancer.," *Virology Journal*, vol. 10, no. 1, p. 30, 2013.

[200] Y. Liu, J. Z. Li, X. H. Yuan, K. Adler-Storthz, and Z. Chen, "An AP-1 binding site mutation in HPV-16 LCR enhances E6/E7 promoter activity in human oral epithelial cells," *Virus Genes*, vol. 24, no. 1, pp. 29–37, 2002.

[201] Y. Dou, T. A. Milne, A. J. Ruthenburg, S. Lee, J. W. Lee, G. L. Verdine, C. D. Allis, and R. G. Roeder, "Regulation of MLL1 H3K4 methyltransferase activity by its core components," *Nature Structural & Molecular Biology*, vol. 13, pp. 713–719, aug 2006.

[202] P. L. Hyland, S. S. McDade, R. McCloskey, G. J. Dickson, K. Arthur, D. J. McCance, and D. Patel, "Evidence for Alteration of EZH2, BMI1, and KDM6A and Epigenetic Reprogramming in Human Papillomavirus Type 16 E6/E7-Expressing Keratinocytes," *Journal of Virology*, vol. 85, no. 21, pp. 10999–11006, 2011.

[203] I. J. Groves, E. L. A. Knight, Q. Y. Ang, C. G. Scarpini, and N. Coleman, "HPV16 oncogene expression levels during early cervical carcinogenesis are determined by the balance of epigenetic chromatin modifications at the integrated virus genome," *Oncogene*, vol. 35, no. 36, pp. 4773–4786, 2016.

[204] R. J. Klose and A. P. Bird, "Genomic DNA methylation: The mark and its mediators," *Trends in Biochemical Sciences*, vol. 31, no. 2, pp. 89–97, 2006.

[205] U. Krüppel, A. Müller-Schiffmann, S. E. Baldus, S. Smola-Hess, and G. Steger, "E2 and the co-activator p300 can cooperate in activation of the human papillomavirus type 16 early promoter," *Virology*, vol. 377, no. 1, pp. 151–159, 2008.

[206] W. M. Wang, S. Y. Wu, A. Y. Lee, and C. M. Chiang, "Binding site specificity and factor redundancy in activator protein-1-driven human papillomavirus chromatin-dependent transcription," *Journal of Biological Chemistry*, vol. 286, pp. 40974–40986, nov 2011.

[207] K. W. Jeong, K. Kim, A. J. Situ, T. S. Ulmer, W. An, and M. R. Stallcup, "Recognition of enhancer elementspecific histone methylation by TIP60 in transcriptional activation," *Nature Structural & Molecular Biology*, vol. 18, no. 12, pp. 1358–1365, 2011.

[208] H. He and Y. Luo, "Brg1 regulates the transcription of human papillomavirus type 18 E6 and E7 genes," *Cell Cycle*, vol. 11, no. 3, pp. 617–627, 2012.

[209] H. J. Szerlong, J. E. Prenni, J. K. Nyborg, and J. C. Hansen, "Activator-dependent p300 acetylation of chromatin in vitro: Enhancement of transcription by disruption of repressive nucleosome-nucleosome interactions," *Journal of Biological Chemistry*, vol. 285, no. 42, pp. 31954–31964, 2010.

[210] A. Kimura and M. Horikoshi, "Tip60 acetylates six lysines of a specific class in core histones in vitro," *Genes to Cells*, vol. 3, no. 12, pp. 789–800, 1998.

[211] S. Jha, S. Vande Pol, N. S. Banerjee, A. B. Dutta, L. T. Chow, and A. Dutta, "Destabilization of TIP60 by Human Papillomavirus E6 Results in Attenuation of TIP60-Dependent Transcriptional Regulation and Apoptotic Pathway," *Molecular Cell*, vol. 38, no. 5, pp. 700–711, 2010.

[212] Y. Sun, X. Jiang, Y. Xu, M. K. Ayrapetov, L. A. Moreau, J. R. Whetstine, and B. D. Price, "Histone H3 methylation links DNA damage detection to activation of the tumour suppressor Tip60.," *Nature cell biology*, vol. 11, no. 11, pp. 1376–82, 2009.

[213] C. H. Kim, J. W. Kim, S. M. Jang, J. H. An, S. B. Seo, and K. H. Choi, "The chromodomain-containing histone acetyltransferase TIP60 acts as a code reader, recognizing the epigenetic codes for initiating transcription," *Bioscience, Biotechnology and Biochemistry*, vol. 79, no. 4, pp. 532–538, 2015.

[214] F. Itzen, A. K. Greifenberg, C. A. Bösken, and M. Geyer, "Brd4 activates P-TEFb for RNA polymerase II CTD phosphorylation," *Nucleic Acids Research*, vol. 42, no. 12, pp. 7577–7590, 2014.

[215] Y. Katan-Khaykovich and K. Struhl, "Dynamics of global histone acetylation and deacetylation in vivo: Rapid restoration of normal histone acetylation status upon removal of activators and repressors," *Genes and Development*, vol. 16, no. 6, pp. 743–752, 2002.

[216] C. A. Hazzalin and L. C. Mahadevan, "Dynamic acetylation of all lysine 4-methylated histone H3 in the mouse nucleus: Analysis at c-fos and c-jun," *PLoS Biology*, vol. 3, no. 12, pp. 1–16, 2005.

[217] A. Dean, "On a chromosome far, far away: LCRs and gene expression," *Trends in Genetics*, vol. 22, no. 1, pp. 38–45, 2006.

[218] R. Palermo, H. Webb, A. Gunnell, and M. West, "Regulation of transcription by the EpsteinBarr virus nuclear antigen EBNA 2," *Biochemical Society Transactions*, vol. 36, no. 4, pp. 625–628, 2008.

[219] Y. Zhu, T. Pe'ery, J. Peng, Y. Ramanathan, N. Marshall, T. Marshall, B. Amendt, M. B. Mathews, and D. H. Price, "Transcription elongation factor P-TEFb is required for HIV-1 Tat transactivation in vitro," *Genes and Development*, vol. 11, no. 20, pp. 2622–2632, 1997.

[220] M. Bulger and M. Groudine, "Functional and mechanistic diversity of distal transcription enhancers," *Cell*, vol. 144, no. 3, pp. 327–339, 2011.

[221] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker, "The long-range interaction landscape of gene promoters," *Nature*, vol. 489, no. 7414, pp. 109–113, 2012.

[222] W. A. Bickmore, "The Spatial Organization of the Human Genome," *Annual Review of Genomics and Human Genetics*, vol. 14, no. 1, pp. 67–84, 2013.

[223] E. Heitz, "Das Verhalten von Kern und Chloroplasten bei der Regeneration," *Zeitschrift fur Zellforschung und Mikroskopische Anatomie*, vol. 2, no. 1, pp. 69–86, 1925.

[224] S. Cajal, "Un sencillo metodo de coloracion seletiva del reticulo protoplasmatico y sus efectos en los diversos organos nerviosos de vertebrados e invertebrados.," *Trab Lab Invest Biol*, vol. 2, pp. 129–221, 1903.

[225] Z. Nizami, S. Deryusheva, and J. G. Gall, "The Cajal body and histone locus body.," *Cold Spring Habour Perspectives in Biology*, vol. 2, pp. a000653–a000653, jul 2010.

[226] V. Lallemand-Breitenbach and H. de Thé, "PML nuclear bodies.," *Cold Spring Harbor Perspectives in Biology*, vol. 2, p. a000661, may 2010.

[227] T. Pederson, "The nucleolus," *Cold Spring Harbor Perspectives in Biology*, vol. 3, pp. 1–15, mar 2011.

[228] D. L. Spector and A. I. Lamond, "Nuclear speckles," *Cold Spring Harbor Perspectives in Biology*, vol. 3, pp. 1–12, feb 2011.

[229] T. Cremer, C. Cremer, T. Schneider, H. Baumann, L. Hens, and M. Kirsch-Volders, "Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments," *Human Genetics*, vol. 62, no. 3, pp. 201–209, 1982.

[230] T. Haaf and M. Schmid, "Chromosome topology in mammalian interphase nuclei," *Experimental Cell Research*, vol. 192, pp. 325–332, feb 1991.

[231] M. Cremer, J. V. Hase, T. Volm, A. Brero, G. Kreth, J. Walter, C. Fischer, I. Solovei, C. Cremer, and T. Cremer, "Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells," *Chromosome Research*, vol. 9, no. 7, pp. 541–567, 2001.

[232] J. A. Croft, J. M. Bridger, S. Boyle, P. Perry, P. Teague, and W. A. Bickmore, "Differences in the localization and morphology of chromosomes in the human nucleus," *Journal of Cell Biology*, vol. 145, no. 6, pp. 1119–1131, 1999.

[233] S. Boyle, "The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells," *Human Molecular Genetics*, vol. 10, pp. 211–219, feb 2001.

[234] D. Zink, M. D. Amaral, A. Englmann, S. Lang, L. A. Clarke, C. Rudolph, F. Alt, K. Luther, C. Braz, N. Sadoni, J. Rosenecker, and D. Schindelhauer,

"Transcription-dependent spatial arrangements of CFTR and adjacent genes in human cell nuclei," *Journal of Cell Biology*, vol. 166, no. 6, pp. 815–825, 2004.

[235] J. O. J. Davies, A. M. Oudelaar, D. R. Higgs, and J. R. Hughes, "How best to identify chromosomal interactions: a comparison of approaches," *Nature Methods*, vol. 14, no. 2, pp. 125–134, 2017.

[236] J. Dekker, "Capturing Chromosome Conformation," *Science*, vol. 295, no. 5558, pp. 1306–1311, 2002.

[237] B. Tolhuis, R. J. Palstra, E. Splinter, F. Grosveld, and W. De Laat, "Looping and interaction between hypersensitive sites in the active B-globin locus," *Molecular Cell*, vol. 10, no. 6, pp. 1453–1465, 2002.

[238] R.-J. Palstra, B. Tolhuis, E. Splinter, R. Nijmeijer, F. Grosveld, and W. de Laat, "The β-globin nuclear compartment in development and erythroid differentiation," *Nature Genetics*, vol. 35, pp. 190–194, oct 2003.

[239] C. G. Spilianakis and R. A. Flavell, "Long-range intrachromosomal interactions in the T helper type 2 cytokine locus," *Nature Immunology*, vol. 5, pp. 1017–1027, oct 2004.

[240] Z. Liu and W. T. Garrard, "Long-Range Interactions between Three Transcriptional Enhancers, Active V Gene Promoters, and a 3 Boundary Sequence Spanning 46 Kilobases," *Molecular and Cellular Biology*, vol. 25, pp. 3220–3231, apr 2005.

[241] D. Vernimmen, F. Marques-Kranc, J. A. Sharpe, J. A. Sloane-Stanley, W. G. Wood, H. A. Wallace, A. J. Smith, and D. R. Higgs, "Chromosome looping at the human α-globin locus is mediated via the major upstream regulatory element (HS -40)," *Blood*, vol. 114, pp. 4253–4260, nov 2009.

[242] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat, "Nuclear organization of active and inactive

chromatin domains uncovered by chromosome conformation captureon-chip (4C)," *Nature Genetics*, vol. 38, no. 11, pp. 1348–1354, 2006.

[243] Z. Zhao, G. Tavoosidana, M. Sjölinder, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. Singh Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson, "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions," *Nature Genetics*, vol. 38, no. 11, pp. 1341–1347, 2006.

[244] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker, "Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements," *Genome Research*, vol. 16, no. 10, pp. 1299–1309, 2006.

[245] J. M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker, "Hi-C: A comprehensive technique to capture the conformation of genomes," *Methods*, vol. 58, no. 3, pp. 268–276, 2012.

[246] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science*, vol. 326, no. 5950, pp. 289–293, 2009.

[247] G. Andrey, T. Montavon, B. Mascrez, F. Gonzalez, D. Noordermeer, M. Leleu, D. Trono, F. Spitz, and D. Duboule, "A Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs," *Science*, vol. 340, pp. 1234167–1234167, jun 2013.

[248] E. de Wit, B. A. M. Bouwman, Y. Zhu, P. Klous, E. Splinter, M. J. A. M. Verstegen, P. H. L. Krijger, N. Festuccia, E. P. Nora, M. Welling, E. Heard,

N. Geijsen, R. A. Poot, I. Chambers, and W. de Laat, "The pluripotent genome in three dimensions is shaped around pluripotency factors," *Nature*, vol. 501, pp. 227–231, sep 2013.

[249] L. Pasquali, K. J. Gaulton, S. A. Rodríguez-Seguí, L. Mularoni, I. Miguel-Escalada, . Akerman, J. J. Tena, I. Morán, C. Gómez-Marín, M. van de Bunt, J. Ponsa-Cobas, N. Castro, T. Nammo, I. Cebola, J. García-Hurtado, M. A. Maestro, F. Pattou, L. Piemonti, T. Berney, A. L. Gloyn, P. Ravassard, J. L. G. Skarmeta, F. Müller, M. I. McCarthy, and J. Ferrer, "Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants," *Nature Genetics*, vol. 46, pp. 136–143, feb 2014.

[250] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, and E. Heard, "Spatial partitioning of the regulatory landscape of the X-inactivation centre," *Nature*, vol. 485, pp. 381–385, apr 2012.

[251] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, pp. 1665–1680, dec 2014.

[252] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, J. S. Liu, and B. Ren, "NIH Public Access," *Nature*, vol. 485, no. 7398, pp. 376–380, 2012.

[253] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren, "A map of the cis-regulatory sequences in the mouse genome," *Nature*, vol. 488, pp. 116–120, aug 2012.

[254] B. Mifsud, F. Tavares-Cadete, A. N. Young, R. Sugar, S. Schoenfelder, L. Ferreira, S. W. Wingett, S. Andrews, W. Grey, P. A. Ewels, B. Herman, S. Happe, A. Higgs, E. LeProust, G. A. Follows, P. Fraser, N. M. Luscombe, and C. S. Osborne, "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C," *Nature Genetics*, vol. 47, pp. 598–606, may 2015.

[255] S. Schoenfelder, R. Sugar, A. Dimond, B.-M. Javierre, H. Armstrong, B. Mifsud, E. Dimitrova, L. Matheson, F. Tavares-Cadete, M. Furlan-Magaril, A. Segonds-Pichon, W. Jurkowski, S. W. Wingett, K. Tabbada, S. Andrews, B. Herman, E. LeProust, C. S. Osborne, H. Koseki, P. Fraser, N. M. Luscombe, and S. Elderkin, "Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome," *Nature Genetics*, vol. 47, no. 10, pp. 1179–1186, 2015.

[256] R. Jäger, G. Migliorini, M. Henrion, R. Kandaswamy, H. E. Speedy, A. Heindl, N. Whiffin, M. J. Carnicer, L. Broome, N. Dryden, T. Nagano, S. Schoenfelder, M. Enge, Y. Yuan, J. Taipale, P. Fraser, O. Fletcher, and R. S. Houlston, "Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci," *Nature Communications*, vol. 6, p. 6178, 2015.

[257] T. Nagano, C. Várnai, S. Schoenfelder, B.-M. Javierre, S. W. Wingett, and P. Fraser, "Comparison of Hi-C results using in-solution versus in-nucleus ligation," *Genome Biology*, vol. 16, no. 1, p. 175, 2015.

[258] S. Wingett, P. Ewels, M. Furlan-Magaril, T. Nagano, S. Schoenfelder, P. Fraser, and S. Andrews, "HiCUP: pipeline for mapping and processing Hi-C data," *F1000Research*, vol. 1310, pp. 1–12, 2015.

[259] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2.," *Nature methods*, vol. 9, pp. 357–9, mar 2012.

[260] I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, and B. R. Lajoie, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, sep 2012.

[261] B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. Fernandez Banet, K. Billis, C. García Girón, T. Hourlier, K. Howe, A. Kähäri, F. Kokocinski, F. J. Martin, D. N. Murphy, R. Nag, M. Ruffier, M. Schuster, Y. A. Tang, J.-H. Vogel, S. White, A. Zadissa, P. Flicek, and S. M. J. Searle, "The Ensembl gene annotation system," *Database*, jun 2016.

[262] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, oct 1990.

[263] 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 Research*, vol. 19, pp. 1639–1645, sep 2009.

[264] B. Mifsud, I. Martincorena, E. Darbo, R. Sugar, S. Schoenfelder, P. Fraser, and N. M. Luscombe, "GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data," *PLoS ONE*, vol. 12, no. 4, 2017.

[265] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, pp. 2460–2461, oct 2010.

[266] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Molecular Systems Biology*, vol. 7, pp. 539–539, apr 2014.

[267] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences*, vol. 74, pp. 5463–5467, dec 1977.

[268] N. C. Durand, J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, and E. L. Aiden, "Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom," *Cell Systems*, vol. 3, pp. 99–101, jul 2016.

[269] N. C. Durand, M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, and E. L. Aiden, "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments," *Cell Systems*, vol. 3, pp. 95–98, jul 2016.

[270] F. A. Klein, T. Pakozdi, S. Anders, Y. Ghavi-Helm, E. E. Furlong, and W. Hu-

ber, "FourCSeq: Analysis of 4C sequencing data," *Bioinformatics*, vol. 31, pp. 3085–3091, oct 2015.

[271] J. R. Hughes, N. Roberts, S. McGowan, D. Hay, E. Giannoulatou, M. Lynch, M. De Gobbi, S. Taylor, R. Gibbons, and D. R. Higgs, "Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment," *Nature Genetics*, vol. 46, no. 2, pp. 205–212, 2014.

[272] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, "Formation of Chromosomal Domains by Loop Extrusion," *Cell Reports*, vol. 15, no. 9, pp. 2038–2049, 2016.

[273] N. H. Dryden, L. R. Broome, F. Dudbridge, N. Johnson, N. Orr, S. Schoenfelder, T. Nagano, S. Andrews, S. Wingett, I. Kozarewa, I. Assiotis, K. Fenwick, S. L. Maguire, J. Campbell, R. Natrajan, M. Lambros, E. Perrakis, A. Ashworth, P. Fraser, and O. Fletcher, "Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C," *Genome Research*, vol. 24, no. 11, pp. 1854–1868, 2014.

[274] S. Schoenfelder, M. Furlan-Magaril, B. Mifsud, F. Tavares-Cadete, R. Sugar, B. M. Javierre, T. Nagano, Y. Katsman, M. Sakthidevi, S. W. Wingett, E. Dimitrova, A. Dimond, L. B. Edelman, S. Elderkin, K. Tabbada, E. Darbo, S. Andrews, B. Herman, A. Higgs, E. LeProust, C. S. Osborne, J. A. Mitchell, N. M. Luscombe, and P. Fraser, "The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements," *Genome Research*, vol. 25, no. 4, pp. 582–597, 2015.

[275] S. M. Tan-Wong, J. D. French, N. J. Proudfoot, and M. a. Brown, "Dynamic interactions between the promoter and terminator regions of the mammalian BRCA1 gene," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 5160–5165, apr 2008.

[276] B. R. Lajoie, J. Dekker, and N. Kaplan, "The Hitchhiker's guide to Hi-C analysis: Practical guidelines," *Methods*, vol. 72, no. C, pp. 65–75, 2015.

[277] J. Dekker, "The three 'C' s of chromosome conformation capture: controls, controls, controls," *Nature Methods*, vol. 3, no. 1, pp. 17–21, 2006.

[278] B. M. Javierre, S. Sewitz, J. Cairns, S. W. Wingett, C. Várnai, M. J. Thiecke, P. Freire-Pritchett, M. Spivakov, P. Fraser, O. S. Burren, A. J. Cutler, and J. A. Todd, "Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters," *Cell*, vol. 167, no. 5, pp. 1369–1384.e19, 2016.

[279] J. Schiemer, "Understanding Illumina TruSeq Adapters," *Tufts University Core Facility*, pp. 1–5, 2011.

[280] Illumina Inc., "HiSeq ® 2500 Sequencing System," no. Figure 2, pp. 8–9, 2013.

[281] L. Wang, S.-Z. Dai, H.-J. Chu, H.-F. Cui, and X.-Y. Xu, "Integration sites and genotype distributions of human papillomavirus in cervical intraepithelial neoplasia.," *Asian Pacific journal of cancer prevention : APJCP*, vol. 14, no. 6, pp. 3837–41, 2013.

[282] Y. Liu, Z. Lu, R. Xu, and Y. Ke, "Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology," *Oncotarget*, vol. 7, no. 5, pp. 5852–64, 2015.

[283] Y. Liu, Z. Lu, R. Xu, Y. Ke, Y. Liu, Z. Lu, R. Xu, Y. Ke, Y. Liu, Z. Lu, R. Xu, and Y. Ke, "Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology," *Oncotarget*, vol. 7, pp. 5852–5864, feb 2016.

[284] C. Bodelon, M. E. Untereiner, M. J. Machiela, S. Vinokurova, and N. Wentzensen, "Genomic characterization of viral integration sites in HPV-related cancers," *International Journal of Cancer*, vol. 139, no. 9, pp. 2001–2011, 2016.

[285] A. Adey, J. N. Burton, J. O. Kitzman, J. B. Hiatt, A. P. Lewis, B. K. Martin, R. Qiu, C. Lee, and J. Shendure, "The haplotype-resolved genome and

epigenome of the aneuploid HeLa cancer cell line," *Nature*, vol. 500, no. 7461, pp. 207–211, 2013.

[286] Y. Liu, C. Zhang, W. Gao, L. Wang, Y. Pan, Y. Gao, Z. Lu, and Y. Ke, "Genome-wide profiling of the human papillomavirus DNA integration in cervical intraepithelial neoplasia and normal cervical epithelium by HPV capture technology," *Scientific Reports*, vol. 6, no. 1, p. 35427, 2016.

[287] B. Xu, S. Chotewutmontri, S. Wolf, U. Klos, M. Schmitz, M. Dürst, and E. Schwarz, "Multiplex Identification of Human Papillomavirus 16 DNA Integration Sites in Cervical Carcinomas," *PLoS ONE*, vol. 8, no. 6, p. e66693, 2013.

[288] I. K. Christiansen, G. K. Sandve, M. Schmitz, M. Durst, and E. Hovig, "Transcriptionally active regions are the preferred targets for chromosomal HPV integration in cervical carcinogenesis," *PLoS ONE*, vol. 10, no. 3, p. e0119566, 2015.

[289] J. M. Doolittle-Hall, D. L. Cunningham Glasspoole, W. T. Seaman, and J. Webster-Cyriaque, "Meta-analysis of DNA tumor-viral integration site selection indicates a role for repeats, gene expression and epigenetics," *Cancers*, vol. 7, no. 4, pp. 2217–2235, 2015.

[290] A. A. McBride and A. Warburton, "The role of integration in oncogenic progression of HPV-associated cancers," *PLoS Pathogens*, vol. 13, no. 4, 2017.

[291] H. X. Jin, Z. Q. Liu, Z. C. Hu, and Y. G. Zheng, "Biosynthesis of (R)-epichlorohydrin at high substrate concentration by kinetic resolution of racemic epichlorohydrin with a recombinant epoxide hydrolase," *Engineering in Life Sciences*, vol. 13, no. 4, pp. 385–392, 2013.

[292] W. Deng, J. W. Rupon, I. Krivega, L. Breda, I. Motta, K. S. Jahn, A. Reik, P. D. Gregory, S. Rivella, A. Dean, and G. A. Blobel, "Reactivation of developmentally silenced globin genes by forced chromatin looping," *Cell*, vol. 158, pp. 849–860, aug 2014.

[293] L. J. Conway, L. Riley, L. Saiman, B. Cohen, P. Alper, and E. L. Larson, "Implementation and impact of an automated group monitoring and feedback system to promote hand hygiene among health care personnel," *Joint Commission Journal on Quality and Patient Safety*, vol. 40, no. 9, pp. 408–417, 2014.

[294] C. Paris, I. Pentland, I. Groves, D. C. Roberts, S. J. Powis, N. Coleman, S. Roberts, and J. L. Parish, "CCCTC-Binding Factor Recruitment to the Early Region of the Human Papillomavirus 18 Genome Regulates Viral Oncogene Expression," *Journal of Virology*, vol. 89, pp. 4770–4785, may 2015.

[295] Y. Satou, P. Miyazato, K. Ishihara, H. Yaguchi, A. Melamed, M. Miura, A. Fukuda, K. Nosaka, T. Watanabe, A. G. Rowan, M. Nakao, and C. R. M. Bangham, "The retrovirus HTLV-1 inserts an ectopic CTCF-binding site into the human genome," *Proceedings of the National Academy of Sciences*, vol. 113, no. 11, pp. 3054–3059, 2016.

[296] A. E. Zacapala-Gómez, O. Del Moral-Hernández, N. Villegas-Sepúlveda, A. Hidalgo-Miranda, S. L. Romero-Córdoba, F. O. Beltrán-Anaya, M. A. Leyva-Vázquez, L. d. C. Alarcón-Romero, and B. Illades-Aguiar, "Changes in global gene expression profiles induced by HPV 16 E6 oncoprotein variants in cervical carcinoma C33-A cells," *Virology*, vol. 488, pp. 187–195, 2016.

[297] D. Hnisz, A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, and R. A. Young, "Activation of proto-oncogenes by disruption of chromosome neighborhoods," *Science*, vol. 351, no. 6280, pp. 1454–1458, 2016.

[298] P. H. L. Krijger and W. de Laat, "Can We Just Say: Transcription Second?," *Cell*, vol. 169, pp. 184–185, apr 2017.

[299] I. Kraus, C. Driesch, S. Vinokurova, E. Hovig, A. Schneider, M. V. K. Doeberitz, and M. Dürst, "The majority of viral-cellular fusion transcripts in cer-

vical carcinomas cotranscribe cellular sequences of known or predicted genes," *Cancer Research*, vol. 68, no. 7, pp. 2514–2522, 2008.

[300] S.-M. Huang and D. McCance, "Down Regulation of the Interleukin-8 Promoter by Human Papillomavirus Type 16 E6 and E7 through Effects on CREB Binding Protein/p300 and P/CAF," *Journal of Virology*, vol. 76, no. 17, pp. 8710–8721, 2002.

[301] T. J. Stasevich, Y. Hayashi-Takanaka, Y. Sato, K. Maehara, Y. Ohkawa, K. Sakata-Sogawa, M. Tokunaga, T. Nagase, N. Nozaki, J. G. McNally, and H. Kimura, "Regulation of RNA polymerase II activation by histone acetylation in single living cells," *Nature*, vol. 516, no. 7530, pp. 272–275, 2014.

[302] N. Ahuja, A. R. Sharma, and S. B. Baylin, "Epigenetic Therapeutics: A New Weapon in the War Against Cancer," *Annual Review of Medicine*, vol. 67, no. 1, pp. 73–89, 2016.

[303] K. Struhl and E. Segal, "Determinants of nucleosome positioning," *Nature Structural & Molecular Biology*, vol. 20, pp. 267–273, mar 2013.

[304] H. X. Jin, Z. Q. Liu, Z. C. Hu, and Y. G. Zheng, "Biosynthesis of (R)-epichlorohydrin at high substrate concentration by kinetic resolution of racemic epichlorohydrin with a recombinant epoxide hydrolase," *Engineering in Life Sciences*, vol. 13, no. 4, pp. 385–392, 2013.

[305] G. Gao, S. H. Johnson, G. Vasmatzis, C. E. Pauley, N. M. Tombers, J. L. Kasperbauer, and D. I. Smith, "Common fragile sites (CFS) and extremely large CFS genes are targets for human papillomavirus integrations and chromosome rearrangements in oropharyngeal squamous cell carcinoma," *Genes Chromosomes and Cancer*, vol. 56, pp. 59–74, jan 2017.

[306] M. K. Jang, D. Kwon, and A. a. McBride, "Papillomavirus E2 proteins and the host BRD4 protein associate with transcriptionally active cellular chromatin.," *Journal of Virology*, vol. 83, pp. 2592–2600, mar 2009.

[307] M. K. Jang, K. Shen, and A. A. McBride, "Papillomavirus Genomes Associate with BRD4 to Replicate at Fragile Sites in the Host Genome," *PLoS Pathogens*, vol. 10, no. 5, p. e1004117, 2014.

[308] Y. Zhu, J. Dai, P. G. Fuerst, and D. F. Voytas, "Controlling integration specificity of a yeast retrotransposon.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 5891–5, may 2003.

[309] R. Zhang, C. Shen, L. Zhao, J. Wang, M. McCrae, X. Chen, and F. Lu, "Dysregulation of host cellular genes targeted by human papillomavirus (HPV) integration contributes to HPV-related cervical carcinogenesis," *International Journal of Cancer*, vol. 138, no. 5, pp. 1163–1174, 2016.

[310] L. S. Qi, M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, A. P. Arkin, and W. A. Lim, "Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression," *Cell*, vol. 152, pp. 1173–1183, feb 2013.

[311] I. B. Hilton, A. M. D'Ippolito, C. M. Vockley, P. I. Thakore, G. E. Crawford, T. E. Reddy, and C. A. Gersbach, "Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers," *Nature Biotechnology*, vol. 33, pp. 510–517, apr 2015.

[312] P. I. Thakore, A. M. D'Ippolito, L. Song, A. Safi, N. K. Shivakumar, A. M. Kabadi, T. E. Reddy, G. E. Crawford, and C. A. Gersbach, "Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements," *Nature Methods*, vol. 12, pp. 1143–1149, dec 2015.

[313] J. I. McDonald, H. Celik, L. E. Rois, G. Fishberger, T. Fowler, R. Rees, A. Kramer, A. Martens, J. R. Edwards, and G. A. Challen, "Reprogrammable CRISPR/Cas9-based system for inducing site-specific DNA methylation," *Biology Open*, vol. 5, pp. 866–874, jun 2016.

[314] A. Vojta, P. Dobrinic, V. Tadic, L. Bockor, P. Korac, B. Julg, M. Klasic, and

V. Zoldos, "Repurposing the CRISPR-Cas9 system for targeted DNA methylation," *Nucleic Acids Research*, vol. 44, pp. 5615–5628, jul 2016.

[315] P. H. L. Krijger and W. de Laat, "Regulation of disease-associated gene expression in the 3D genome," *Nature Reviews Molecular Cell Biology*, vol. 17, no. 12, pp. 771–782, 2016.

[316] M. J. Fullwood and Y. Ruan, "ChIP-based methods for the identification of long-range chromatin interactions," *Journal of Cellular Biochemistry*, vol. 107, pp. 30–39, may 2009.

[317] J. O. J. Davies, J. M. Telenius, S. J. McGowan, N. A. Roberts, S. Taylor, D. R. Higgs, and J. R. Hughes, "Multiplexed analysis of chromosome conformation at vastly improved sensitivity," *Nature Methods*, vol. 13, no. 1, 2015.

[318] R. A. Beagrie, A. Scialdone, M. Schueler, D. C. A. Kraemer, M. Chotalia, S. Q. Xie, M. Barbieri, I. de Santiago, L.-M. Lavitas, M. R. Branco, J. Fraser, J. Dostie, L. Game, N. Dillon, P. A. W. Edwards, M. Nicodemi, and A. Pombo, "Complex multi-enhancer contacts captured by genome architecture mapping," *Nature*, vol. 543, pp. 519–524, mar 2017.