Potential Role of CTCF in Differential Papillomavirus Gene Expression

A Major Qualifying Project:

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

By

Alyson Talbot

Date: April 30, 2009

Approved

Professor Destin W. Heilman, Advisor

Abstract

Upon cellular differentiation the late genes of the Human Papillomavirus (HPV) are expressed. Recently, the CTCF DNA-binding protein has been found to be transcription factor and chromatin insulator. We examined if CTCF plays a role in viral transcription and late gene expression. Through ChIP analysis it was found that CTCF binds to two distinct regions of the Bovine Papillomavirus-1 genome. It was observed that the E2 protein and CTCF protein interact by co-immunoprecipitation. Using RNA interference, the effect of CTCF on viral transcription was examined. It was shown that E2 and L1 RNA levels are knocked out with lower levels of CTCF. Further studies should investigate if the type of interactions between E2 and CTCF, and the mechanism in which CTCF controls viral transcription.

Acknowledgements

I would like to thank Elliot Androphy, Carol Jolly, Suzanne Melanson, and all the members of the Androphy Lab (Department of Medicine at UMass Medical) for their non-stop guidance. Also, I owe a thank you to Professor Heilman who helped me with the entire MQP process.

Table of Contents

Tables

Figures

1.0 Background

1.1- Human Papillomavirus Overview

HPV is the most frequently spread sexually transmitted disease, although most forms are not dangerous there are some high risk forms that can develop into cervical cancer, other epithelial cancers, or genital warts (18). Also, high risk HPVs can cause head and neck cancer although it is not as common as cervical cancer (16). The transformation of HPV-infected cells to cancer is relatively rare, but as of 2006 cervical cancer is the fifth deadliest cancer in women worldwide (17). It has also been found that in 99.7% of cervical cancer patients there is a presence of HPV DNA, which means that HPV infection is necessary for the development of cervical cancer (18). On average, genital HPV infections last 12-18 months prior to being cleared by the immune system, but the small number of women whose immune systems do not clear the infection are at risk for cervical cancer (17). This suggests that HPV is able to overcome the immune system's response. One proposed method of how HPV can surpass the body's immune system is that it modifies the body's cell-mediated immune responses and it can also alter the innate immune system (7).

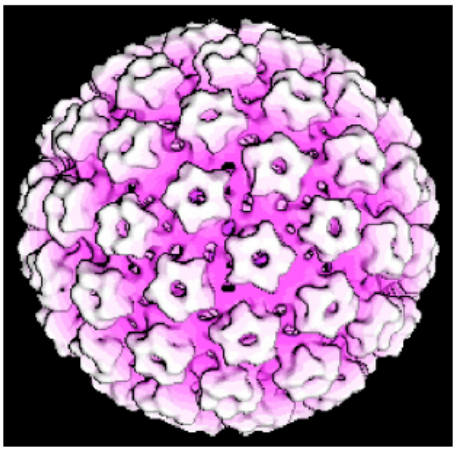


Figure 1: Three-dimensional surface display of the HPV virion (6).

Human Papillomaviruses are small, non-enveloped, icosahedral DNA viruses that induce proliferative lesions, and replicate in the squamous epithelial cells (6,10). The HPV virion particle (Figure 1) is 52 to 55 nm in diameter. The particle

consists of a single molecule of double stranded circular DNA about 8,000 base pairs in size, and is contained within a spherical protein coat known as the capsid (10).

There have been over 100 different types of HPV identified. Different types of HPV are classified as either low-risk or high-risk depending on their potential for malignant transformation. The low-risk types such as HPV-6 and -11 usually cause benign warts and are rarely associated with the formation of cancer. The high-risk HPV such as HPV-16, -18, -31, and -45 have been linked to the formation of cervical cancer, head and neck cancer, and other genital cancers. The difference in the phenotype of the high risk and low risk HPV occur for many different reasons. In high risk HPV there are two major promoters, but in low risk there is only one. In the high risk HPV the early promoter causes E2-mediated transcription to be repressed causing high levels of the oncoproteins E6 and E7. Also, the different forms of HPV have different genomes, where genomes vary in size, composite, and gene organization. Lastly, for the high risk HPV a single promoter is responsible for the expression of E6 and E7 whereas for the low risk HPV the E6 and E7 proteins each have an independent promoter (6).

The genome of HPV has roughly eight open-reading frames (ORFs) (Figure 2) (6). The viral genes can be divided into two categories, early and late, depending on their location within the genome, and their order

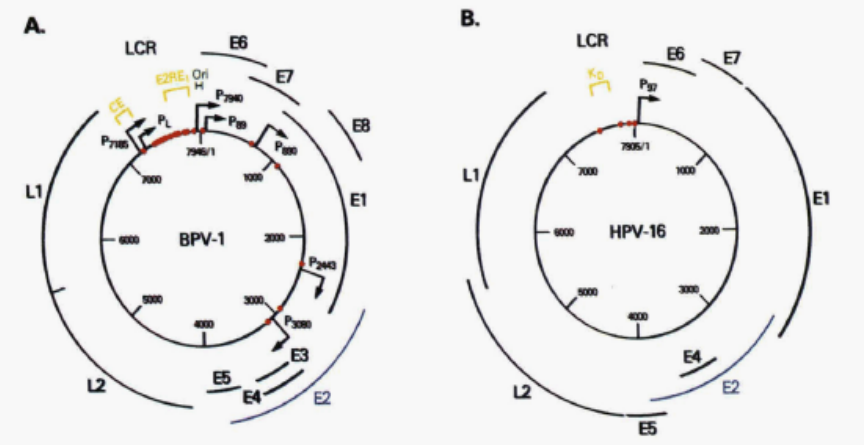


Figure 2: The genomes for BPV-1 and the high risk HPV-16 (8).

of expression (6,10). In the early stages of infection the E6, E7, E1, and E2 proteins are transcribed from the early promoter. During the late stage of the virus high-level transcription of E1, E2, E4, E5, and capsid proteins L1 and L2 occurs (6). The upstream regulatory region (URR) regulates early viral transcription. The URR is approximately 1kb in size and contains binding sites for transcription factors that regulate, induce, or repress HPV transcription (6,10).

The life cycle of HPV is closely associated to epithelial differentiation. HPV enters the cells through an unknown receptor and then migrates to the nucleus of the cells. It has been shown that HPV can bind and infect a wide variety of cell types (10). During the infection of HPV the viral proteins override the cells normal cell cycle that takes place in differentiating cells to allow viral progeny to form. Due to the fact that basal cells are the only cells that can divide in the squamous epithelium the virus must infect the basal cells in a way that allows the virus to thrive (6).

Certain stages of the HPV life cycle only occur in differentiating squamous epithalamium cell such as late gene expression, synthesis of capsid proteins, vegetative viral DNA synthesis, and the assembly of virions. This is due to a specific promoter which becomes active exclusively in differentiated keratinocytes. The vegetative replication of HPV DNA allows the genomes to be generated and packed into virions. The mechanism of the vegetative viral DNA replication is still unknown (10).

The transcription of HPV is complex due to the many promoters located throughout the genome (not shown in Figure 2), which cause alternate and multiple splicing patterns and differential production of mRNA species in different cells. There have been seven different transcriptional promoters identified for bovine papillomavirus

(BPV-1), and their locations are shown on the genome (Figure 2). More than twenty mRNA species and their putative gene products have been identified (6,10).

HPV has three modes of viral DNA replication. The first takes place during the initial infection of the basal keratinocytes by the virus, when there is an amplification of the viral genome to approximately 50 to 100 copies. The next phase involves genome maintenance, which occurs in the dividing basal cells in the lower portion of the epidermis. The DNA is then maintained as a stable multicopy plasmid. During this phase the viral genome replicates on average once per cell cycle during the S-phase in synchrony with the host cell genome, and are then separated evenly into the daughter cells. The last type of DNA replication is vegetative DNA replication, which occurs in the more differentiated epithelial cells. These cells no longer undergo cellular DNA synthesis. The termination of cellular DNA synthesis induces a burst of viral DNA synthesis, generating the genomes to be packaged into progeny virions (10).

In addition to E1, E2, E6, and E7 proteins, the virus is dependent on many host cell factors for viral replication including: DNA polymerase α, thymidine kinase, PCNA,

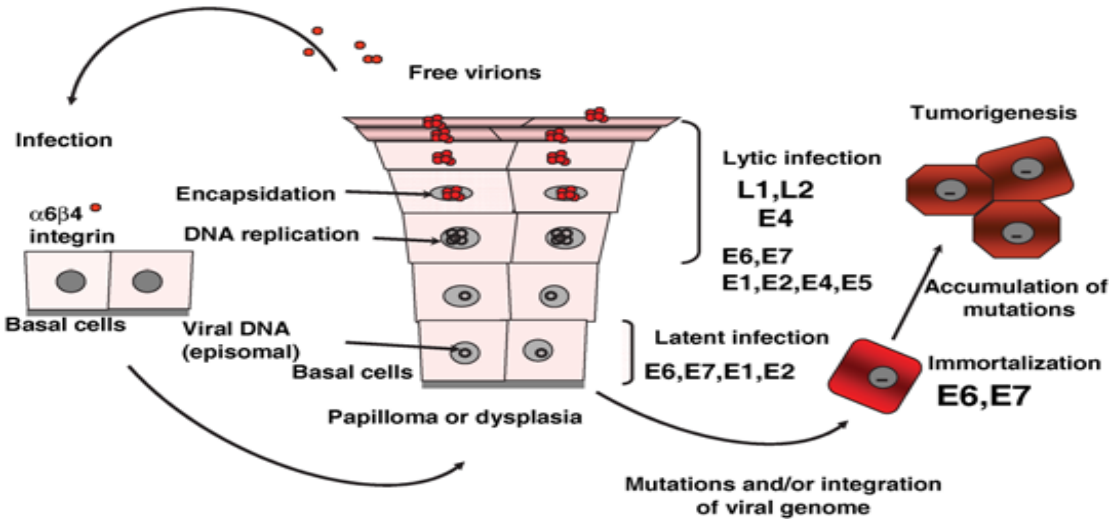


Figure 3: The life cycle of the Human Papillomavirus (12).

and many others. HPV also has evolved a mechanism to activate cellular genes necessary for replication in the vegetative DNA replication. This is important because most of the host cell factors are present only during DNA replication not when the cells are highly differentiated (10).

1.2- HPV Viral Proteins

1.2.1- E6 and E7 Oncoproteins

The E6 protein of the high risk forms of HPV contains approximately 150 amino acids including four Cys-X-X-Cys motifs, which are responsible for binding zinc (6,10,12). The E6 protein is localized to the nucleus and cytoplasm of the infected cells. The main function of the E6 protein is interfering with p53 mediated cell cycle regulation (12). p53 is a tumor suppressor that plays an essential role in the cells response to DNA damage and other cellular stress by activating several regulators of the apoptotic and senescence pathways (5). The E6 protein binds to p53 reducing the steady-state levels of p53. The reduction of steady-state levels of p53 allows viral replication and ends the transcriptional transactivation of p53 (6,10,12). This means that E6 has an anti-apoptotic function. The E6 protein is able to degrade the p53 function by inducing the ubiquitin dependent proteolysis of p53 by forming a complex with E6AP, an ubiquitin-protein ligase, which then can bind to p53. E6 has also been shown to interact with proteins that have PDZ domains, which are involved in cell signaling and cell-cell adhesion. This interaction plays a major role in the HPV life cycle because a mutation to the E6-PDZ binding domain has shown a reduction in growth, episomal maintenance, and early transcription. E6 can also extend the life span of keratinocytes and lead to outgrowth of immortalized clones, which are resistant to terminal differentiation. The E6 protein is

also responsible for the activation of telomerase in infected cells, which causes an increase of length of the telomeres (30). In normal cells telomeres gradual decrease in size that leads to chromosomal instability and cellular senescence and apoptosis (5).

The E7 oncoprotein has approximately 98 amino acids and three conserved regions, CR1, CR2, and CR3. The CR1 domain is part of the amino terminus, CR2, has an LXCXE motif that controls binding of E7 to the retinoblastoma (Rb) tumor suppressor protein family, and CR3 consists of two zinc finger motifs (6). A main role of the E7 protein is the binding and degradation of the Rb proteins (12). The Rb family of proteins are major regulators of the cell cycle. When Rb is hypophoshorylated it controls the transition of the cell cycle at G_1/S phase by binding the E2F family of transcription factors. The E2F transcription factors activate the transcription of many cellular components during S-phase replication. The binding of Rb is regulated by two separate regions of the protein, the LXCXE motif in CR2 is necessary for the binding of the Rb proteins, and the N-terminus has the important residues for the degradation of the Rb proteins. The binding and degradation of the Rb proteins by E7 causes the release of the E2F complex (6). E7 also effects the E2F stimulated transcription by interacting with the class I histone deacetylases (HDACs). HDACs act as transcriptional corepressors, a protein that works with transcription factors to decrease the rate of gene transcription, by inducing chromatin remodeling by the deacetylation of histones. E7 binds the HDACs indirectly by an interaction that is mediated through sequences in the zinc-finger region. The binding of E7 to HDACs has been shown to increase levels of the E2F mediated transcription in differentiated cells that causes them to proceed into S-phase. The expression of E7 also causes genomic instability by causing the infected cells to exhibit

centrosomal irregularities, such as abnormal number of centromeres. The genomic instability is common in many malignancies (6). In the low risk forms of HPV, the E7 protein binds pRb with 10-fold lower efficiency than in the high risk types. Also, low risk forms of HPV E7 are functionally inefficient in cellular transformation (10).

1.2.2-E1 and E2 Protein

E1 is the central and essential replication factor that serves as the origin recognition protein initiating viral DNA replication (25). The E1 protein has a molecular weight of about 70kDa (6). The secondary structures of E1 for both BPV and HPV-11 are made up of 36% α-helices, 24% β-sheets, 31% random coil, and 9% turns. It has been shown that both E1 for HPV and BPV lacking the N-terminal region still support viral DNA replication indicating the necessary functions for viral replication are located on the C-terminus (25). The E1 ORF is the largest ORF in the HPV genome and is well conserved among all types of HPV, which shows the importance of the function of the E1 protein (6, 16). The E1 protein is expressed throughout the HPV life cycle at very low levels, but when the late promoter is active the expression of E1 is increased (6,24). The E1 protein has DNA dependent ATPase, helicase, and nucleotide-binding activities. The ATP and helicase activity plays a role in the origin unwinding fork progression (27). The ATPase and helicase activity of E1 is localized to the C-terminus. The ATPase and helicase activity of E1 makes it the only papillomavirus protein with enzymatic activity (29,32).

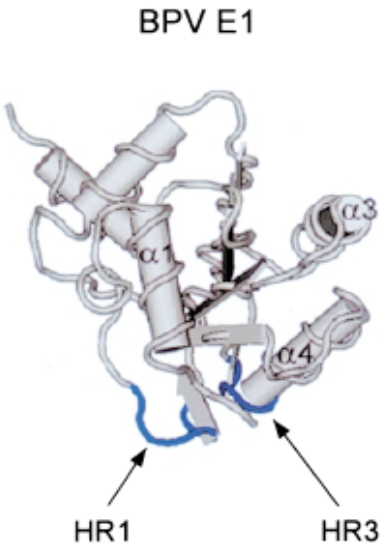


Figure 4: The secondary structure of BPV-1 highlighting the HR1 and HR3 domains(24).

The E2 protein functions both in viral replication and regulation of HPV transcription. The E2 protein has a molecular weight of approximately 50 kDa. E2 consists of a β-barrel that forms dimeric complexes and binds to DNA (6). The N-terminus of E2 has a transactivation domain and the C-terminus is involved in the interaction with the oncoprotein E1 and DNA binding (11). The URRs (upstream regulatory region) of high risk HPV contain four E2 binding sites that have the palindromic sequence $ACCN_6GGT$ (20). Three of those sequences are the E1 recognition sequences at the viral origin, and the fourth site is located near the center of the L1 ORF (6). At low levels E2 activates viral transcription, but at high levels it represses viral transcription. This suggests that E2 serves as a regulator of the E6 and E7 proteins mediated cell cycle activities. Therefore, the loss of E2 expression is associated with increased cellular proliferation induced by E6 and E7 leading to cervical cancer (11). The E2 protein also plays a role in the early promoter expression by altering chromatin remodeling through the recruitment of histone acetyltransferases (HATs). Once differentiation takes places transcription of the viral genes switches from the E2-regulated early promoter to the E2 independent late promoter, this results in high levels of E1 and E2 transcription leading to viral amplification (6). Also, E2 does not play a direct role in the viral DNA replication; it only plays a role in the pre-initiation complex (10). Lastly, E2 has been shown to play a role in episomal maintenance and possibly tether viral genomes to mitotic chromosomes during cell division (6).

1.2.3-L1 and L2 Capsid Proteins

L1 is about 55kDa in size, and L2 has a molecule weight of 70kDa (10). The L1 and L2 capsid proteins are expressed late in the viral life cycle in highly differentiated

cells. After amplification HPV genomes are packaged as chromatin into viral capsids composed of L1 and L2. HPV virions structure is icosahedral, and entails 360 L1 monomers assembled into 72 pentameric structures termed capsomeres. Approximately, 12 copies of the L2 associate with the capsomeres by a domain near the C-terminus of L2. The L1 associates with the capsomeres formed in the cytoplasm and then translocate through the nuclear pore complex into the nucleus. L2 is translocated separately into the nucleus by two nuclear localization signal sequences. The L2 capsid protein may play a role in recruiting viral genomes and initiating encapsidation. L2 can bind DNA and localize to the ND10 domains, which are nuclear bodies and the major site of DNA replication. Therefore, L2 may bind newly replicated viral DNA and recruit L1 to create new virions (6).

1.3- HPV Late Gene Activation

The HPV late genes L1 and L2 have been shown to activate only during high levels of cellular differentiation, and linked to cellular differentiation for late gene expression is the use of a transcription factor to trigger the expression of L1 and L2. The cohesin complex has recently been shown to interact with the E2 viral protein. The cohesin complex also associates with the known transcription factor CTCF. For this reason it was proposed that the CTCF protein would be a good candidate as the transcription factor controlling late gene expression.

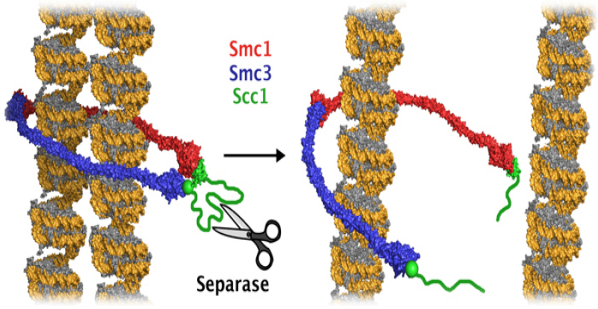
1.3.1- Cohesin Complex

One of the most important factors for a virus to survive is its ability to replicate its viral genome and distribute it evenly into the host's daughter cells. The cohesin complex is essential for this due to its role in chromosome segregation. Cohesin is a protein

complex that forms a ring around the sister chromatids to hold them together during DNA replication at telephase(14, 23). The cohesin protein complex is composed of four subunits Smc-1, Smc-3, Scc-1 (Rad-21), and Scc-3. The Smc proteins are heterodimers ATPases that are responsible for structural maintenance of the complex (2,28). The formation of the ring has the Smc-1 and Smc-3 proteins linked at one end by a hinge interaction and by the Scc-1 protein at the other end. The Scc-1 protein holds the cohesin complex together (28).

The cohesin complex forces the sister chromatids to be handled as a pair then segregated away from each other. After the chromosomes are properly aligned in the mitotic spindle, cohesin is cleaved to allow separation of the sister chromatids into the daughter cells. The cleavage of cohesin occurs by the cysteine protease, separase, at the Scc-1 protein during anaphase. Cohesin has also been shown to play a role in gene regulation, DNA repair, and chromatin structure and development (28).

Figure 5: Cleavage of cohesin by seperase (21).



Cohesin contributes to gene regulation by influencing enhancers, silencers, and insulators. For example, the cleavage of cohesin is required for establishing silencers and it influences the ability of insulators to separate the regions of active and inactive chromatin (28).

1.3.2- Cohesin and CTCF Association

The first connection that was found between cohesin and CTCF was that many of the cohesin binding sites that were mapped in the human genome overlap the binding

sites of the CTCF protein (8,23). Two of these regions are the *H19* imprinting control region and the LCR region. Also, it was found that the CTCF and Scc-1 sequences are very similar to one another (8,19).

CTCF is required for the localization of cohesin at its binding sites. This was shown by the effect of removing CTCF and the result was disruption of the positioning of the Scc-1 and Smc-3 proteins of the cohesin complex (23). It was revealed that CTCF is not required for loading of cohesin onto DNA, showing that both cohesin and CTCF can associate with DNA independently. Although CTCF is required for cohesin localization, cohesin is not required for CTCF localization (3,14,23).

Even though cohesin recognizes CTCF for positioning through the genome, it has been shown that cohesin can interact with chromatin fibers independently of CTCF. The significance of this is that CTCF does not play a role in sister chromatids cohesion, but that cohesin could determine the function of CTCF at specific sites (4,18,23). Specifically, cohesin has been shown to be required for CTCF's insulator function (23). A mechanism was proposed that cohesin stabilizes CTCF chromosome loops during chromatin insulation. This was shown by the fact that genes within 25 kb of cohesin sites had a higher tendency to be upregulated. This mechanism has the CTCF bound to the DNA in a loop, and then the cohesin

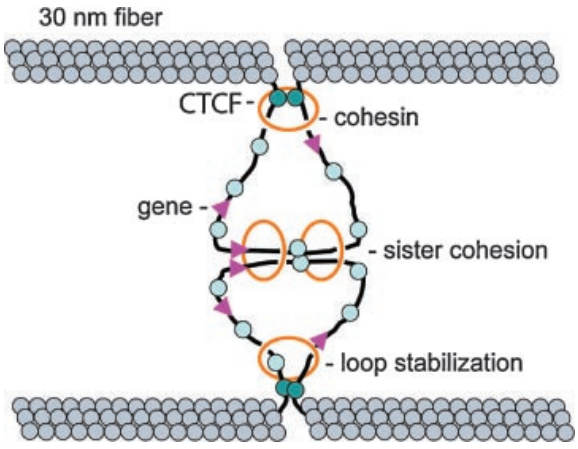


Figure 6: Loop mechanism for CTCF and cohesin (19).

encircles the two CTCF loops to hold the CTCF to the DNA, similar to how cohesin holds sister chromatids together (19).

1.3.3-E2 Interactions with Cohesin

In mitosis the separation of the papillomavirus genes, is mediated by the C-terminal DNA binding domain of the E2 protein. The E2 protein binds and tethers the viral genome to the host cell chromosomes by protein-protein interactions. Those protein-protein interactions require N-terminal transcactivation domain of E2, which E2 binds with the Scc-1 protein of the cohesin complex. The binding appears to be independent of any other cellular factors (21).

1.3.4-CTCF Protein

CTCF is a ubiquitously expressed nuclear protein with an 11 zinc-finger DNA-binding domain. It was first identified by its ability to bind a number of dissimilar regulatory sequences in the regions close to the promoters of the chicken, mouse, and human MYC oncogenes (1,8). CTCF is a highly conserved protein, between avian and mammalian 93% of the amino acids of CTCF are identical, but 100% of the amino acids are identical in the 11 zinc-finger region. Also, the ORFs of CTCF have not changed in over 300 million years, and are conserved through a variety of species. This indicates that the CTCF protein is adapted and has had no environmental need to mutate and change it sequence (1,13).

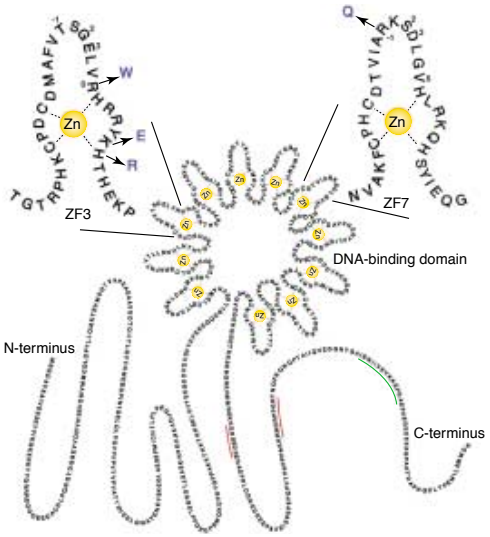
CTCF is present as a nuclear extract at 130kDa on SDS-PAGE, but the theoretical molecular weight is 82kDa. This is because CTCF is encoded in a 4.1kb mRNA, with the largest ORF predicting a 728 amino acid protein. Possibilities for this contradiction are: a missing exon in the cDNA due to alternative splicing, posttranslational processing, and

certain amino acid compositions could lead to CTCF abnormal electrophoretic migration. Also, there is a truncated version of CTCF at 70kDa, which represents only the N-terminal domain of the protein due to premature termination of translation (9).

1.3.5-CTCF Structure

The CTCF protein is organized into three domains: the N-terminal domain, zinc-finger domain, and the C-terminal domain. The N- and C-termini account for two-thirds or the protein, but the zinc-finger domain is where most of the functionality of CTCF occurs. The zinc fingers of CTCF are capable of binding either DNA or proteins (13,19). The CTCF zinc fingers are unique in that they can bind different DNA sequences by each zinc finger group or one zinc finger motif can bind DNA independently while another zinc finger can bind a protein independently.

Figure 7: Structure of CTCF zinc-fingers (13).



The first ten zinc-fingers are approximately 30 residues long containing a pair of cysteine residues, which are separated by 12 amino acids from a pair of histidine residues. These amino acids are coordinated with zinc to form an α-helix structure that recognizes DNA. The α-helix structure allows the zinc-fingers to insert into the major groove of the DNA. The recognition of the CTCF protein to different DNA sequences is mediated by varying zinc-fingers; therefore, certain sets of zinc-fingers are necessary for binding to certain sequences of DNA, but are dispensable for binding to a different sequence. Also, not all zinc-fingers that bind to DNA behave in a similar manner. For example some zinc-fingers

will bind to the major groove to contact the base pairs, while others will migrate to the minor groove making few or no contacts with the DNA base pairs. It has also been shown that the same zinc finger can behave differently when CTCF is bound to certain target sequences. When CTCF is bound to DNA both their polypeptides allosterically customize their conformation to engage different zinc-fingers, either for making base contacts or to make a target-specific surface that determines interactions with other nuclear proteins (13).

The regions flanking the zinc-finger domain are the N- and C-terminus. The C-terminus has a role in the trans-repressing function of CTCF, whereas the role of the N-terminus domain is not as well classified. These domains show no significant similarities to any previously described protein modules, except for three short motifs that are all located in the C-terminus. The first motif expressed is KRRGRP-type AT-hook that possibly has a role in DNA binding and protein-protein interactions in chromatin. Another motif that is strictly conserved is the SKKEDSSDSE motif in the trans-repressor region. In this region the protein is phosphorylated on four serines by casein kinase II. A third motif is located between the AT-hook and phosphorylation sites. This motif contains two repeats of the PXXP-signature characteristic of the SH3-domain binding proteins (13).

1.3.6-CTCF Functions

CTCF is a versatile protein that has been shown to have many different cellular functions. The best described and studied functions of CTCF are its ability to be a transcription factor, chromatin insulator protein, and its role in epigenetics and cancer. Other areas where CTCF could possibly play a role in are gene activation and as a tumor

suppressor (8,13).

CTCF was first characterized as a transcription factor by its ability to bind to the metazoan silencing elements. For example, binding of CTCF to the promoters and upstream silencer elements of the chicken lysozyme gene and the chicken and human MYC gene results in transcriptional repression of the lysozyme and MYC gene. Further studies of CTCF found that other CTCF target sites identified transcriptional response elements that function in gene repression and activation. Also, CTCF shares many common traits with enhancers: it independently functions of position and orientation, it compromises functional modules that synergize in transcriptional control, and it acts directly on a promoter. An example is that CTCF can act as transcription activator or repressor on the lysozome gene. CTCF's role depends on the presence of the thyroid hormone receptor, if this hormone is bound along with CTCF it leads to synergistic activation of the lysozome gene, but if it is not bound CTCF causes synergistic repression of the gene. CTCF also contains an independent silencing domain that mediates transcriptional repression. The discovery of this domain led to the identification of CTCF-interacting co-repressors that recruit histone deacetylase (HDAC) activity. Silencing by CTCF and its co-repressors could act directly on the transcriptional start site. The ability of CTCF to recruit HDACs leads to CTCF inhibiting transcription by, interfering with the transcription initiation complex and/or modifying the promoter nucleosomes (13).

CTCF has a pivotal role in many chromatin insulators, and is the only major protein implicated in establishment of insulators in vertebrates. Protein insulators operate by blocking the communication between pivotal *cis* regulatory elements, gene promoters,

and enhancers or silencers. *Cis*-regulatory DNA elements are a region of DNA that controls the expression of genes on the same strand. These elements mediate the level of gene expression by recruiting *trans*-acting factors that influence transcription. The regulatory elements used for the insulation are often distant from one another on the linear genome. This makes the insulators position-dependent, where the insulator must be positioned between the enhancer and its target promoter (8). This suggests that the insulator prevents propagation of signals along the chromatin fiber without continuously engaging enhancer or silencing factors. CTCF may mediate the chromatin interaction without the need of co-factors due to the finding that CTCF can dimerize when bound to DNA and connect two separate DNA molecules. An example of CTCF as a chromatin is its interaction with the core insulator sequences in vertebrate insulators, β-globin FII insulator, *Xenopus* repeat organizer elements, and the BEAD-A insulator. CTCF flanks the entire region of the β -globin gene cluster, and it presumably protects this domain from effects of the adjacent regulatory elements (13).

One of the most characterized interactions of CTCF describes its role in both chromatin insulation and epigenetics. CTCF plays a central part in the H19 imprinting control region (ICR) where CTCF affects the downstream Igf2 gene. The region is maternally unmethylated and paternally methylated. It regulates the expression of the maternal allele of the Igf2 gene by blocking the Igf2 promoters and enhancer from communicating with each other. CTCF associates only with the maternal unmethylated allele of H19 ICR. The CTCF protein prevents the activation of the Igf2 gene in this circumstance. When this region is methylated CTCF cannot bind though, allowing the Igf2 gene to be expressed. Also, the loss of imprinting on the H19 and Igf2 locus have

been found to cause Beckwith-Wiedemann Syndrome with increased tumor formation (1,8,13).

Due to CTCF's ability to read epigenetic marks and the common occurrence of epigenetic disturbances in cancer, the function of CTCF in cancer has also been studied. Gene mutations to CTCF were found in patients with breast, prostates, and Wilm's tumor's. These mutations were located in the zinc-finger finger region of the protein, which affects the binding and interaction of DNA with other genes. Also, in the PXXP motif region a common tumor suppressor was found, the MYC-binding protein BIN1. It was shown that if mutations are made to the prolines in this region, it causes the elimination of binding to BIN1, and affects the trans-repressing activity of the C-terminus region (1,13).

CTCF is a unique protein with roles in transcription, insulation, tumor suppression, and gene activation. All of these roles could be essential in the HPV genome. Also, CTCF and cohesin have been shown to interact with one another and be necessary for one another to function, and E2 and cohesin have been shown to interact within HPV. Therefore, it was examined if CTCF is not only found within the HPV genome, but also if it plays a role in late gene expression and viral transcription.

2.0 Methods

Tissue Culture: ID13 and A3 BPV transformed cells were grown at 37° C in DMEM media with 10% Fetal Bovine Serum and 1% pen-strep. Cells were grown up to $1.5x10^6$ cells/plate before harvesting. After 24 hours cells were harvested using Phosphate Buffered Saline (PBS).

Prediction Program: The prediction program "CTCFBSDB: a CTCF binding site database for characterization of vertebrate genomic insulators" (22) was used for the prediction of CTCF sites in the BPV-1 genome.

Chromatin Immunoprecipitation: Formaldehyde (37%) was added to ID13 plated cells in media and were then incubated at 37°C for 10 minutes. Cells were then harvested with PBS and protease inhibitors. Cell pellets were lysed (SDS Lysis Buffer- 5% Tris 1M pH=8.1, 2% EDTA, 10% SDS, and protease inhibitors) in 550µL and incubated on ice for 10 minutes. Samples were sonicated at 30% in 5 second intervals for one minute. Samples were then diluted 10 fold (SDS dilution buffer- 1% of 10% SDS, 11% of 10% Triton, 0.04% EDTA, 1.67% of 1M Tris pH=8.1, 3.34% of 5M NaCl and protease inhibitors) except for the input which included 100µL of sample and was placed directly into -20°C. A bead slurry of 1:2 Protein A to Agarose was made and pre-cleared with the cell lysates for half an hour. The sample was then added to a new set of beads along with a respective anti-body, CTCF (rabbit polyclonal Millipore), Rad-21 (rabbit polyclonal Abcam), II-1 (rabbit polyclonal Androphy Lab), and II-1 Pre-Bleed. These were immunoprecipitated overnight at 4°C. Samples were then washed with Low Salt (1% of

10% SDS, 10% of Triton, .4% of 0.5M EDTA, 2% IM Tris pH=8.1, 3% of 5M NaCl), High Salt (low salt buffer with 10% 5M NaCl), LiCl, and 2x TE buffers respectively. Samples and input were eluted from beads with buffer (10% SDS and 0.084g of Sodium Bicarbonate) and NaCL. Lysates were transferred off the beads. Samples were heated at 65°C for four hours. Protein digest was preformed for the samples (EDTA, Tris-HCl pH=6.5, and Proteinase K) and incubated for an hour at 45°C. Samples were then prepared for PCR using Qiagen PCR clean-up kits. Samples were prepared for amplification with *Taq* polymerase and buffer, 10mM dNTPs, water, and primers that flank CTCF prediction sites at 6682 and 3721 (Table 1).

Table 1: Primers used for ChIP Analysis

Primer Name	Forward Sequence (5'→3')	Reverse Sequence (5'→3')
Flanking 6531	TTCAAGCACAGAGGGCATAAGTC	CAAAATGGCTGAGGACGCTG
Flanking 3682	TCAATGTTTTCTCCTGTATCC	TTACTATTCTCGCTTTGGTGACG

Co-Immunoprecipitation: IP Buffer (20 mM Tris pH=8.1, 125mM KCl, 0.5% Triton, 20% Glycerol, 10mM NaF, 2mM Na_3VO_4, 5mM EDTA, 100mM $MgCl_2$, 1x protease inhibitors) was added to the plated A3 cells and left at -80°C overnight. Cells were then harvested. Lysates were sonicated at 30% for 5 seconds. A 1:1 solution of sample and IP buffer was made with a 10% input. Antibodies were added to sample, CTCF, Rad-21, II-1, II-1 PB, and Igg (rabbit Sigma). Protein A beads were added and the samples were immunoprecipitated overnight. Samples were washed with IP buffer. 2x Laemmli SDS Loading Buffer was added to samples and boiled for 10 minutes. Samples were then run on a 10% SDS-PAGE gel.

RNA interference: CTCF mouse shRNAs were packaged in a retrovius, pSM2c vector, and transformed into *E. Coli* (UMass Medical Core Facilities). The bacterial DNA was then isolated by Maxi-Prep (Qiagen Kit). A restriction digest was performed for the samples using the sites BamHI and EcoR1 enzymes (Promega) to ensure the vector was isolated in the bacterial pellets. The shRNAs were then transfected with lipofectamine into ID13 cells in varying concentrations of 0-3µg (Invitrogen). Cells were harvested, and a BCA assay was then conducted on the lysates. 20µg were run on an 10% SDS –PAGE gel with actin used as the loading control.

Reverse Transcription: ID13 cells were transfected as previously described with 0µg of shRNA, 1µg of shRNA, 1µg of non-silencing shRNA. Cells were then treated for RNA isolation using RNAeasy kit (Qiagen). The isolated RNA was then DNA digested (Promega 10x buffer, Promega DNase, Water) and incubated at 37°C for half an hour. DNase stop solution (Promega) was added and incubated at 65°C for 10 minutes. The samples then underwent reverse transcription using Promega ImpropII enzyme and buffer, RnasIN, and Random primer. Samples were then amplified using the previously describe PCR master mix with primers designed for E6, E2, and L1.

Table 2: Primers used for analysis of CTCF knockdown on viral transcription

Primer Name	Forward Sequence (5'→3')	Reverser Sequence (5'→3')
E2	TGCAGTTGTCTTTGCAGGAG	AGCACCGTTTAGGTTCTGACAT
E6	ATGGACCTGAAACCTTTTGC	CAGCCTTCCCGAATTACAAC
L1	GCCTGTTTGTTTCCTGTCATCTG	ATCTCCCTCCAACCCCTGTAAG

3.0 Results

3.1- Location of CTCF binding sites within the BPV Genome

To determine if CTCF is involved in BPV it was first examined *in silco* if there were any CTCF binding sites on the BPV-1 genome. To accomplish this an online prediction program for CTCF binding sites was used (22). The prediction site is a statistical analysis of the core motifs of CTCF that are represented by position weight matrices (PWM). The program focuses on the four PWM that provide the strongest range of conserved motifs.

Two predicted sites were found in the BPV-1 genome, one in the Pre-E5 region and the other in the L1 region (Figure 8C). The Pre-E5 region corresponds to the LM7 PMW and the L1 region is the LM2 PMW. These sequences received high scores of 10.2766 and for LM7 5.86352 for LM2, where a 3.0 is a suggestive match (Figure 8A). The two PMW for BPV-1 were then studied for homology with the actual BPV-1 sequence. For LM2 4 of 19 predicted nucleotides were direct matches to the BPV-1 sequence. The LM7 site had 3 of 20 nucleotides that directly matched the respective BPV-1 sequence. Xie et al. examined the properties of the top ten discovered motifs of CTCF, which included the LM2 and LM7 motifs. The conserved nucleotides within these PMW were mapped, allowing for the creation of a motif profile for LM2 and LM7. LM2 has 16 conserved nucleotides of the 19 nucleotide motif, and LM7 has 17 conserved nucleotides within its 20 nucleotide motif. These conserved nucleotides were then compared to the predicted PMW. The predicted LM2 motif had 16 conserved nucleotides of which three were also found within the BPV-1 sequence. The LM7 motif had 13 conserved nucleotides, but none were matches to the BPV-1 sequence. (Figure 8B)

3.2-CTCF Binding to the BPV-1 DNA

As previously shown CTCF was found to bind the BPV genome *in silco*. The CTCF prediction sites were shown to occur in two specific regions, L1 and Pre-E5. With this evidence it was further examined if CTCF binds to the BPV genome within the specified regions *in vivo.* Also, due to the known interaction of CTCF with cohesin and E2 with cohesin, it was investigated if E2 and/or Rad-21, a subunit of cohesin, were bound to the same regions of the BPV-1 genome. The BPV transformed cells were cross-

linked with 37% formaldehyde. First, Rad-21 was immunoprecipitated with its specific antibody. The DNA was then amplified for two specific genomic locations, L1 and Pre-E5. For the L1 region Rad-21 was shown to bind to genomic sequence, but it did not bind to the Pre-E5 region. Next, CTCF was immunoprecipitated and its DNA amplified for the two specific genomic areas. CTCF was found to bind to the L1 and Pre-E5 regions. This shows CTCF does bind to genome at both predicted locations. Also, it shows that CTCF can bind to the genome independently of the cohesin complex since Rad-21 only bound in the L1 region of the virus. Lastly, the E2 protein was immunoprecipitated and its DNA amplified in the two separate regions of the genome, L1 and Pre-E5. The E2 protein bound to the genome in the L1 region, but not in the Pre-E5 region (Figure 9).

3.3- Interactions of CTCF and E2

Having shown that E2 and CTCF all bind to the L1 region of the BPV-1 genome and that CTCF and E2 both interact with the cohesin complex it was examined if the CTCF and E2 protein associate with one another. The interactions of E2 and CTCF were investigated through co-immunoprecipitation. Two different circumstances were studied: one involved the immunoprecipitation of the CTCF protein with its respective antibody and the other was involved the immunoprecipitated of the E2 protein. The immunoprecipitation are specific for that protein due to the use of its antibody. This ensures that only that protein and proteins that associated with it are pulled down into the beads. First, the E2 protein was immunoprecipitated and both E2 and CTCF were blotted with their respective antibodies. Unfortunately the heavy and light chain backgrounds disrupted the bands for the proteins and it could not be clearly determined if both E2 and

CTCF was present. Next, CTCF was immunoprecipitated and again both E2 and CTCF were blotted with their respective antibodies. This showed that both E2 and CTCF interact with one another within the BPV-1 genome even though they are not found to bind all the same regions of the genome (Figure 10).

3.4- CTCF Effect on Viral Transcription

As previously shown, CTCF is located on the BPV-1 genome in two distinct locations. One of the locations, Pre-E5, is a location that could be responsible for regulation of late gene expression. As CTCF has been shown to be a transcription factor it was examined if CTCF regulates late gene expression in the virus. Also, the CTCF protein was shown to interact with the transcriptional regulator protein E2. This interaction led to possibility that CTCF plays a role in viral transcription. Therefore it was investigated the affect of CTCF protein knockdowns on viral transcription and late gene expression.

The first step in determining the role of CTCF in viral transcription and late gene expression is ensuring that shRNAs could effectively knockdown CTCF levels *in vivo.* The shRNAs work by cutting out the mRNA responsible for the expression of CTCF from the cells (Figure 11). In order to package the shRNAs into the cells it was genetically engineered into the pSM2c retrovirus vector. The vector was then transformed into *E. Coli*. The DNA was transfected into BPV-1 transformed cells in ranging concentrations of 1-3µg. The lysates were then analyzed for effective knockdown of CTCF by western blots with actin as a loading control. The actin levels were not even; therefore, the bands were normalized by integrated density on the ImageJ program. As

shown clearly in the bar graphs there was still a significant knockdown observed in the cells ranging to near a 50% knockdown of CTCF in the cells (Figure 12).

Since the CTCF protein was affected by shRNAs, CTCF's role in viral transcription and gene expression could be studied. This was investigated by using shRNAs to knockdown CTCF. The BPV transformed cells were transfected with 0µg, 1µg of shRNAs and 1µg of non-silencing shRNAs respectively. The RNA was then isolated from the cells. The isolated RNA then underwent a reverse transcription. First, the affect of low levels of CTCF on the E6 oncoprotein was examined. It was shown that there was little to no change in the RNA levels of E6. If any change occurred with the CTCF knockdown it was actually a rise in the E6 levels. Next, influences of the CTCF knockdown on E2 RNA levels was examined, and as Figure 13 shows with low levels of CTCF the E2 levels were knocked out. Lastly, it was investigated how CTCF knockdown would influence L1. As Figure 13 shows the L1 levels were also completely knocked out. This result is contrary to the idea that CTCF is regulating late gene expression, but L1 was also degraded by the non-silencing shRNAs. This could be due to the fact that L1 is knocked out due to the small hairpin shape and not the sequence it encodes (Figure 13).

Motif PWM	Input Sequence	Motif Sequence	Motif Start Location	Motif Length	Score
MIT_LM2	BPV-1 Complete Genome	TTAACAGTGGGGGACAATA	6682	19	5.86352
MIT_LM7	BPV-1 Complete Genome	GAACCAGGTGGTGGTGCAGT	3721	20	10.2766

A.

LM7							
Motif PMW	GAA	***CCA***	***G***GT	***GGT***	***GG***T	***GC***A	***GT***
BPV-1 seq.	GAC	AAG	CAC	AAA	TAC	TGA	TC

B.

LM2							
Motif PMW	TTA	A***CA***	***GT***G	***GGG***	***G***A***C***	***A***A***T***	***A***
BPV-1 seq.	AAT	GTA	TAC	CAT	AGA	CAT	A

Key:
Blue= Sequence Match
Bold+Italics= Conserved Match

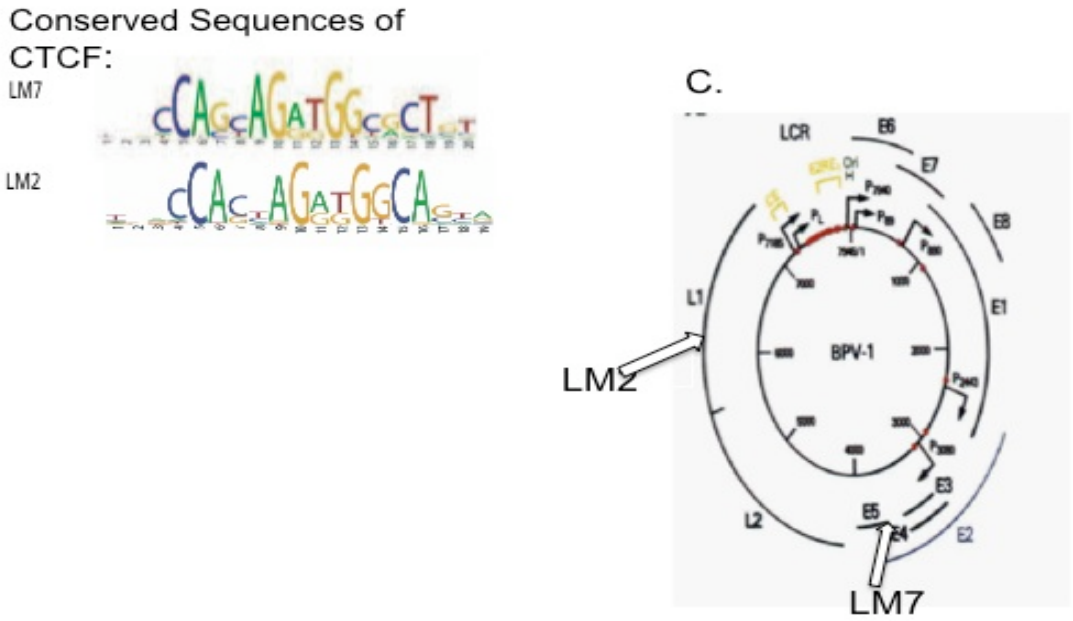


Figure 8: The prediction site provided two possible CTCF binding sites LM2 and LM7. These sites both received high scores because anything above a 3.0 is a suggestive match (A). These sites are located in the L1 and Pre-E5 regions of the BPV-1 genome respectively (C). The predicted sequences were then compared to the equivalent BPV-1 sequence and to the common conserved nucleotides where blue is a match to the BPV-1 sequence and bold and italics is a match to the conserved nucleotides (B).

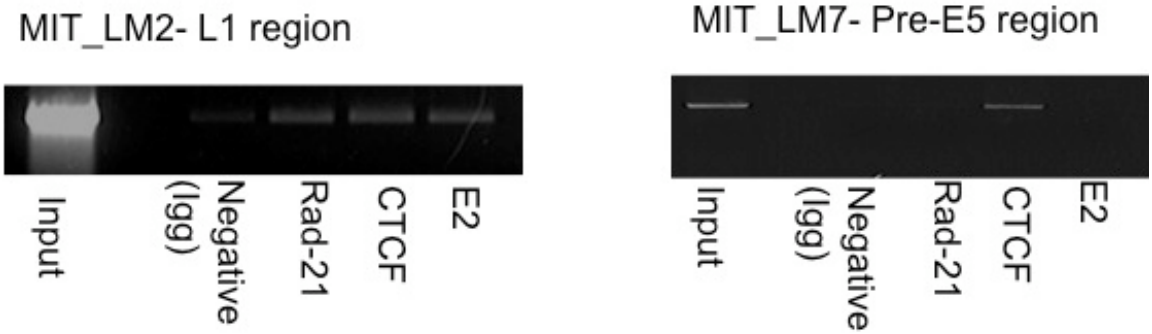


Figure 9: ChIP analysis was used to determine if CTCF, E2, and cohesin bind to the LM2 and LM7 regions of the BPV-1 genome. Each ChIP had a 10% input of the entire cross-linked cellular lysates, and a negative control of Rabbit Igg was used. For the LM2 site E2, CTCF, and Rad-21 all were found to bind to the genome. For the LM7 predicted site only CTCF was found on the genome.

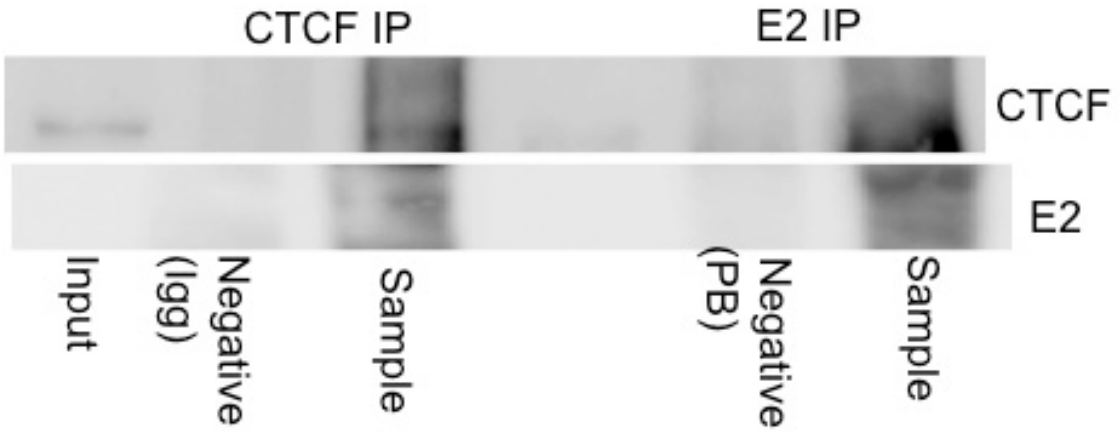


Figure 10: Co-immunoprecipitation was completed to determine if CTCF and E2 interact with one another within the BPV. Both CTCF and E2 were immunoprecipitated with a 10% input of total cellular lysate and their respective negative controls, Rabbit Igg and Pre-bleed. For the CTCF IP both E2 and CTCF were observed on the membrane proving a protein-protein interaction. For the E2 IP the heavy and light chain background bands interfered in the same area where the protein bands should be seen.

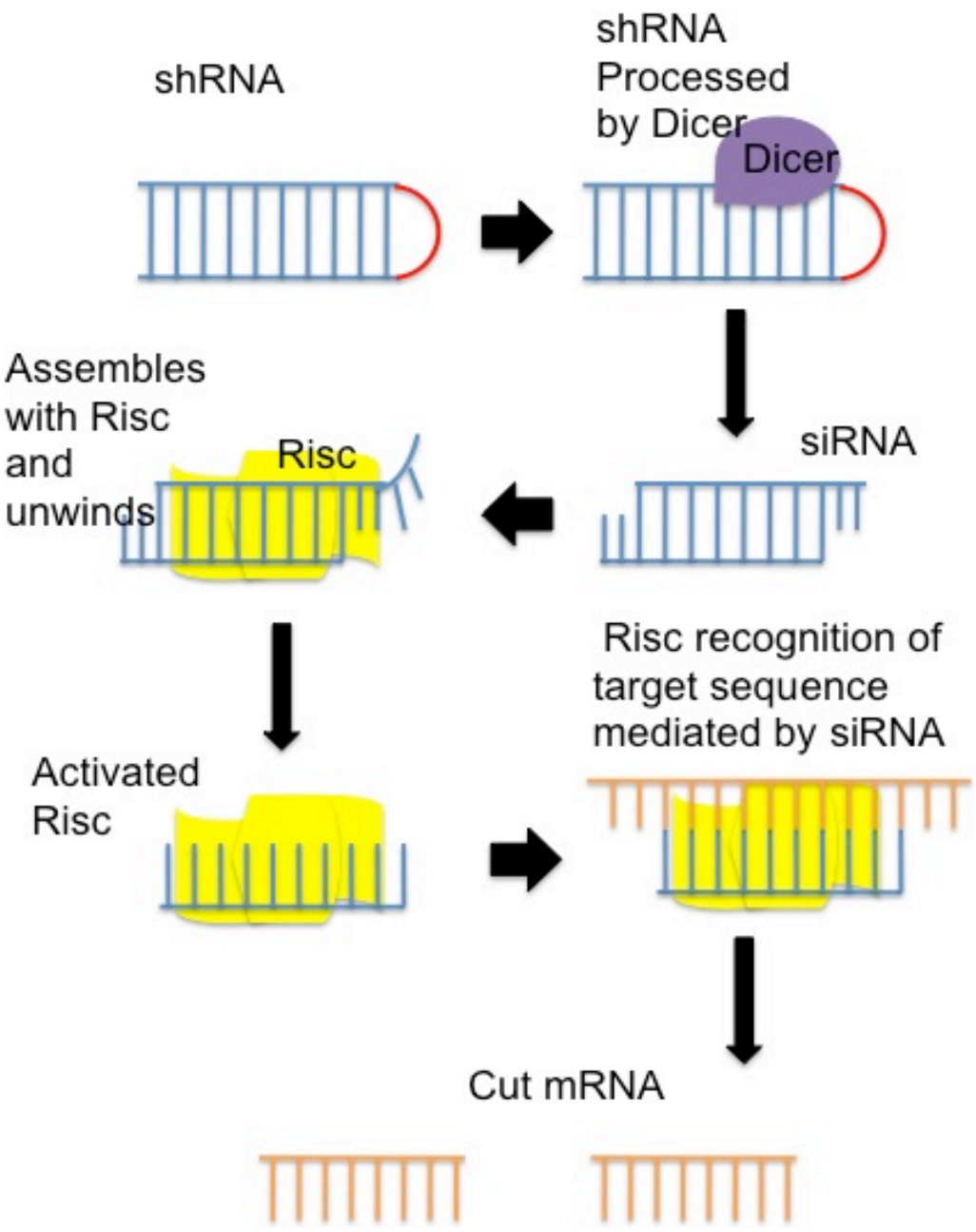


Figure 11: Process of how shRNAs work to knockdown expression of proteins within systems. The shRNA is processed by the DICER complex, which prepares it for the RISC complex. The RISC complex unwinds the RNA to activate it, and it guides it through until it recognizes a target sequence. Upon recognition of a target sequence the mRNA is cut out preventing it from being expressed.

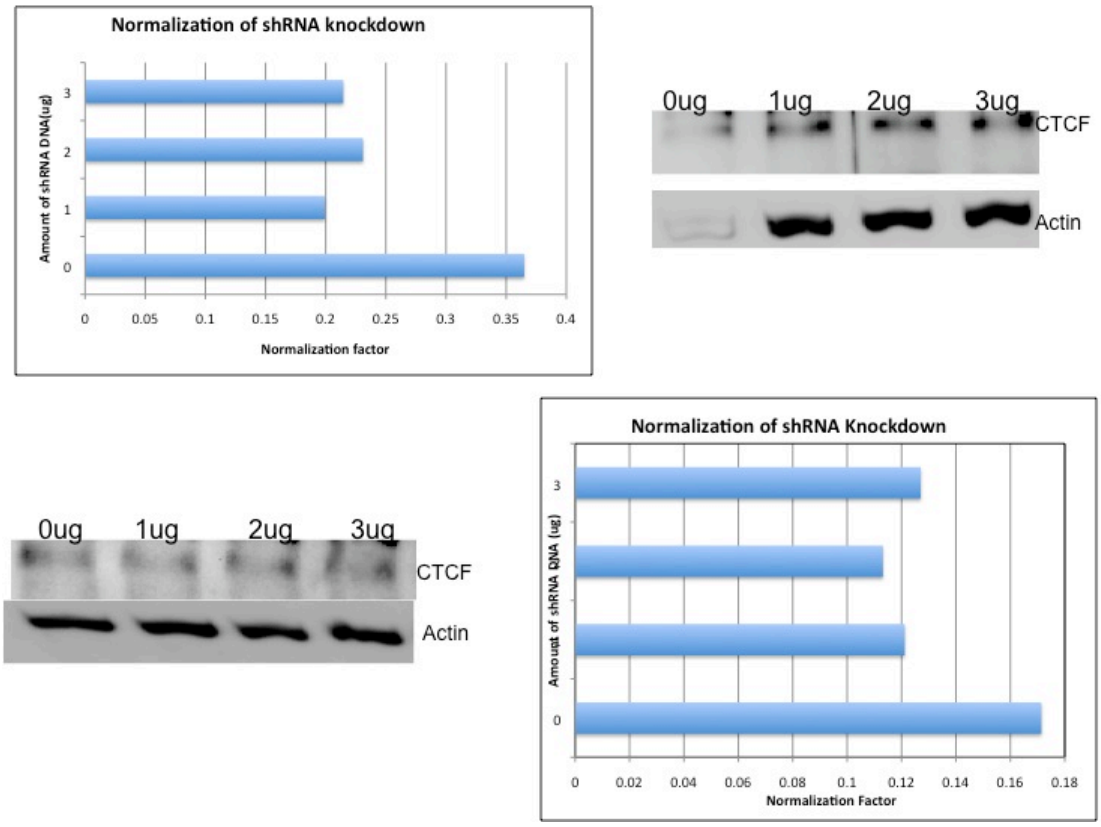


Figure 12: shRNA knockdown of CTCF levels in BPV transformed cells. The blots show the different concentration dosing of the shRNAs, 0-3µg. The actin levels were low for the loading control; therefore, blots were normalized by integrated density. The graphs correspond to the normalization values of the shRNA knockdown, which shows successful knockdown to approximately 50%.

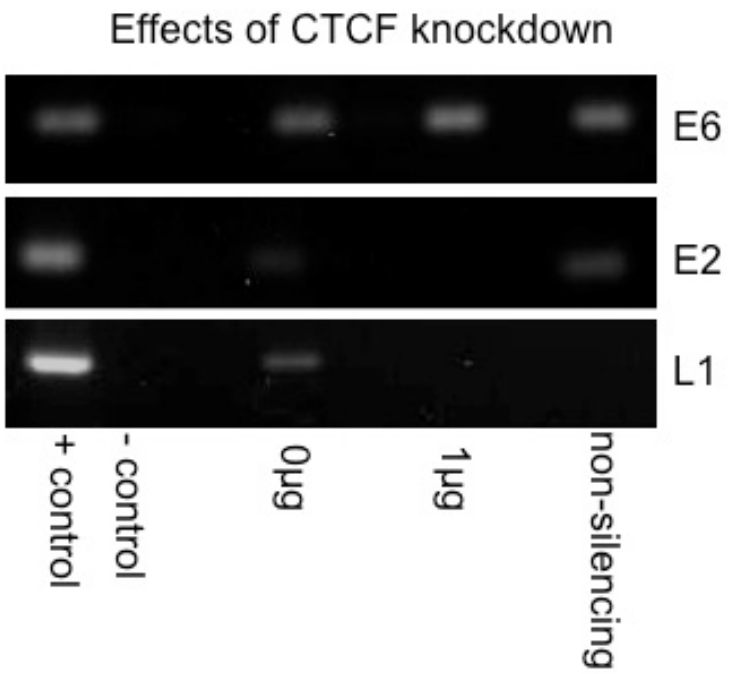


Figure 13: Reverse Transcription of CTCF shRNAs transfected cells were examined for its affect on viral transcription. The BPV genome was used as a PCR positive control and water was used a PCR negative control. The lanes correspond to 0µg, 1µg, of CTCF shRNAs and 1µg of non-silencing shRNAs respectively. With low levels of CTCF both E2 and L1 were knocked out where E6 levels possibly rose. Also, L1 levels were affected by the non-silencing shRNAs leading to a possible influence of the shRNA shape on L1.

4.0 Discussion

CTCF has been shown in many different systems to play a role as a transcription factor and/or a chromatin insulator. CTCF has been shown to associate with the cohesin complex, and the cohesin complex has recently been revealed to interact with the E2 viral protein of HPV. This study was the first to find that the CTCF protein associates with HPV.

The first part of this study showed that CTCF has possible binding sites in the BPV-1 genome *in silico* (Figure 8A-C). The predicted site provided guidance on where to look for possible CTCF sites *in vivo*, but it doesn't mean that those are the only possible CTCF binding sites within the BPV-1 genome. It is possible that it binds to many other locations within the genome or that within that one region it could bind numerous times. Therefore, CTCF binding sites should be mapped throughout the genome this would provide us with a better understanding of where CTCF binds and other possible mechanisms for the protein.

The predicted sites in the L1 and Pre-E5 regions were then examined *in vivo.* These sites were examined for the binding of the Rad-21 protein, subunit of cohesin, and the E2 viral protein. It was shown that for L1 CTCF, E2, and Rad-21 all bind but for Pre-E5 only CTCF bound (Figure 9). This is interesting because it provides the suggestion that CTCF has two independent functions or mechanisms within the genome. The Rad-21 protein has been shown to be necessary for CTCF's insulator function. The E2 protein has been shown to associate with the cohesin complex throughout the cell cycle, providing evidence E2 is essential for cohesin role in gene regulation. Also, E2 is responsible for transcriptional repression of many viral proteins, which could aid in the

insulation process for CTCF. Therefore, it is has been presumed that E2 and cohesin are important for the insulation function of CTCF. This provides significant evidence of CTCF acting as an insulator in the L1 region and not in the Pre-E5 region. The role of CTCF acting independently in the Pre-E5 region would need to be further examined due to the fact CTCF is a versatile protein with a range of functions.

The CTCF protein was also found to associate with the viral protein E2 (Figure 10). This is interesting because as previously stated E2 did not bind to the Pre-E5 region but CTCF did. Therefore this means that although the proteins do associate it is not always the case. This provides further evidence that although the proteins do interact it could be just for certain viral functions, such as the previously described insulation. Also, due to CTCF known roles in transcription and E2 acting as the viral protein responsible for transcription it is probable that CTCF plays a role in the HPV transcription. This could mean that only when CTCF is affecting viral transcription, such as chromatin insulation, do the CTCF and E2 protein interact with each other. The reason this is plausible that the interactions could occur only at specific locations is due to the fact the CTCF's zinc fingers are do unique. The zinc fingers are able to allosterically change for specific sequences. This could mean that when CTCF is bound to certain genomic regions its binding region for E2 is experiencing a conformational changed and the binding can't occur, or it could work the other way so that when bound to E2 the zinc fingers are in a conformation that the CTCF protein can only bind to distinct genomic regions. Although, if E2 directly binds to CTCF or the association is by an indirect mechanism stills needs to investigated.

With the ability to lower levels of CTCF through shRNA transfection the role of CTCF in late gene expression was studied (Figure 12-13). With the knockdown of CTCF levels, L1 levels were knocked out. This is contradictory to the idea of CTCF as an insulator for late gene expression, which would have shown an increase in L1 with CTCF knockdown. This is further contradicted by the presumption that for CTCF to be able to function as an insulator both Rad-21 and E2 must be present, but in the Pre-E5 area that precedes the late genes, only CTCF was bound to the genome. Another issue is that L1 was also affected by the non-silencing shRNAs. This could mean that the shape of the shRNA degrades L1 independently of the sequence of the shRNA. Therefore, a further study needs to pursue if the small hairpin shape of the shRNAs affect L1 levels independently of their sequence.

CTCF ‘s role in viral transcription was also examined through shRNA transfection. When CTCF levels were lowered E2 was knocked out where E6 levels rose. This is because E2 controls the levels of E6, so with high levels of E2 the E6 protein is regulated but with low levels of E2 the E6 is expressed in high levels. A possible reason for this is CTCF is necessary for E2 transcription and possibly an enhancer for E2. The reason CTCF can be an enhancer for E2 is because enhancers can act independently of their location. Therefore, in the Pre-E5 region where only CTCF is bound it could be acting with the E2 promoter for transcription. This it still a preliminary mechanism that needs to be further studied for how CTCF and E2 associate during viral transcription.

A mechanism has been proposed for CTCF’s function in the L1 region, and it entails CTCF insulating anti-sense transcription of the virus. This mechanism has significant evidence starting with the binding of both E2 and Rad-21 in the L1 region,

which as previously discussed could be vital for CTCF chromatin insulation. Therefore it is possible that when the CTCF levels are low the virus undergoes anti-sense transcription, which leads to the pKR cellular death pathway. This could also explain the knockout of L1 when there are low levels of CTCF because the anti-sense transcription would lead to rapid degradation of L1. A cellular death assay should be completed to see the affect of shRNA CTCF knockdown on the cells. This mechanism provides an explanation for the role of CTCF in the L1 region taking into account all of the findings of this study, but it still needs to be further examined.

References

1. Bartkuhn, M., Renkawitz, R. (2008). Long range chromatin interactions involved in gene regulation. *Biochimica et Biophysica Acta*, 1783, 2161-2166.
2. Gause, M., Schaaf, C., Dorsett, D. (2008). Cohesin and CTCF: Cooperating to control chromosome conformation?. *Bioessays*, 30 (8), 715-718.
3. Gondor, A., Ohlsson, R. (2008). Chromatin insulators and cohesins. *EMBO reports , 9* (4), 327-329.
4. Hagstrom, K. A., Meyer, J. B. (2003). Condensin and Cohesin: More Than Chromosome Compactor and Glue. *Nature Reviews , 4*, 520-534.
5. Hanahan, D., Weinberg, R. (2000). Hallmarks of cancer. *Cell, 100*, 57-70.
6. Hebner, C., Laimins, L. (2006). Human papillomaviruses: Basic mechanisms of pathogenesis and oncogenicity. *Reviews in Medical Virology, 16*, 83-97.
7. Kanodia, S., Fahey, L., Kast, M. (2007). Mechanisms used by human papillomaviruses to escape the host immune response. *Current Cancer Drug Targets, 7*(1), 79-89.
8. Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., et al. (2007). Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell , 128*, 1231-1245.
9. Klenova, E., Nicolas, R., U, Sally., Carne, A. (1997). Molecular weight and abnormalities of the CTCF transcription factor: CTCF migrates aberrantly in SDS-PAGE and the size of the expressed protein is affected by the UTRs and the sequences within the coding region of the CTCF gene. *Nucleic Acid Research*, 25 (3), 466-473.
10. Knipe, D., Howley, P. (Eds.). (2001). *Fields Virology* (Fourth ed.). Philadelphia, PA: Lippincott Williams and Wilkins.
11. McBride, A., Romanczuk, I. (1991). The Papillomavirus E2 Regulatory Proteins. *Journal of Biological Chemistry, 266*(28), 18411- 18414.
12. Narisawa-Soto, M., Kiyono, T. (2007). Basic mechanisms of high-risk human papillomavirus-induced carcinogenesis: Roles of E6 and E7 proteins. *Cancer Science, 98*(10), 1505-1511.
13. Ohlsson, R., Renkawitz, R., Lobanenkov, V. (2001). CTCF is a uniquely varsatile transcription regulator linked to epigenetic and disease. *Trends in Genetics , 17* (9), 520-527.
14. Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., et al. (2008). Cohesin Functionally Associates with CTCF on Mammalian Chromosome Arms. *Cell , 132*, 422-433.
15. Parish, J., Melanson, S., Bean, A., Androphy, E., (Submitted for Publication). Association of cohesin with papillomavirus E2 protein and episomal viral genomes.
16. Psyrri, A., DiMaio, D. (2008). Human papillomavirus in cervical and head-and-neck cancer. *Nature Clinical Practice: Oncology, 5*(1), 24-31.
17. Schiffman, M., Castle, P., Jeronimo, J., Rodriguez, A., Wacholder, S. (2007). Human papillomavirus and cervical cancer. *Lancet, 370*, 890-907.
18. Steben, M., Duarte-Franco, E. (2007). Human papillomavirus infection: Epidemiology and pathophysiology. *Gynecologic Oncology, 107*, S2-S5.
19. Stedman, W., Kand, H., Lin, S., Kissi, J. L., Bartolomei, M. S., Lieberman, P. M. (2008). Cohesin localizes with CTCF at the KSHV latency control region and at cellular c-myc and H-19/Igf2 insulators. *The EMBO Journal*, 1-13.
20. Sun, Y., Lu, J., McCance, D. (1995). Mapping of HPV-11 Binding Site and Determination of Other Important *cis* Elements for Replication of the Origin. *Virology, 216,* 219-222.
21. Uhlmann, F. (2008). Cohesin Branches Out. *Nature , 451*, 777-778.
22. University of Tennessee Health Science Center (April 2009). "CTCFBSDB: a CTCF binding site database for characterization of vertebrate genomic insulators." Retrieved on 2009-04-09.
23. Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., et al. (2008). Cohesin Mediates Transcriptional Insulation by CCCTC-binding factor. *Nature , 451*, 796-803.
24. Wilson, V., West, M. (2002). Papillomavirus E1 proteins: Form, function, and features. *Virus Genes, 24*(3), 275-290.
25. World Health Organization (February 2006). "Fact sheet No. 297: Cancer". Retrieved on 2008-09-12.)

26. Xie, Xiaohui., Mikkelsen, Tarjei., Gnirke, Andreas., Lindblad-Toh, Kerstin., et al. (2007). Systematic Discovery of regulatory motifs in conserved region of the human genome, including thousands of CTCF insulator sites. *PNAS*, *104* (17), 7145-7150.
27. Yang, L., Mohn, I., Fouts, E., Lim, D., Nohalie, M., & Botchan, M. (1993). The E1 protein of bovine papilloma virus 1 is an ATP-dependent DNA helicase. *PNAS, 90,* 5086-5090.