TRANSCRIPTOME ANALYSIS OF LYMPHOMA ASSOCIATED VIRUSES AND ANALYSIS OF VIRAL NONCODING RNAS

AN ABSTRACT

SUBMITTED ON THE 30th DAY OF JULY 2014

TO THE GRADUATE PROGRAM IN BIOMEDICAL SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF MEDICINE

OF TULANE UNIVERSITY

FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

SUBING CAO

APPROVED:

ERIK K. FLEMINGTON, Ph.D.
DIRECTOR

VICTORIA P. BELANCIO, Ph.D.

PRESIDENT

YAN DONG, Ph.D.

TONG WU, Ph.D.
ABSTRACT

During the past fifty years, it has become clear that several viruses are etiologically associated with lymphomas. In this research, we performed a global virome survey in 50 lymphoma cell line models using high-throughput RNA sequencing (RNA-seq) data from the Cancer Cell Line Encyclopedia project. This investigation identified EBV, KSHV and HTLV-1 in cell lines already known to harbor these viruses and it identified active infection with the murine retrovirus, MuLV, in two of these cell lines. In depth EBV and KSHV transcriptome analysis led to the discovery of unexpected transcription in several regions of these viral genomes. Through a de novo assembly approach, we solved the genome sequences of the new HTLV-1 and MuLV strains found in the corresponding cell lines. In addition, we developed a computational method to systematically identify and examine all transcribed viral integration sites in virus infected samples. This analysis identified integration sites for EBV, HTLV-1 and MuLV and provided new insights into integration mediated gene disruption and/or activation mechanisms.

We also carried out an investigation into EBV lytic transcriptome and identified novel bidirectional transcripts derived from the EBV latent origin of replication, oriP. Both the sense and antisense oriP transcripts (oriPtRs and oriPtLs) were expressed with the kinetics of late viral genes and predominantly localized in the nucleus. Structure modeling showed that both oriPtRs and oriPtLs can form stable hairpin structures that were found to be hyper-edited by the adenosine deaminase, ADAR1. These transcripts were associated with the
nuclear paraspeckle component, NonO. oriPtL overexpression reduced the mRNA levels of several heat shock proteins and globally enhanced EBV gene expression. oriPtL knockdown inhibited EBV production and global EBV gene expression and influenced the expression of several immune regulatory genes. These results indicate that oriPtL may have multiple functions in EBV self-regulation and in modulating cellular environment.

In sum, this research provides an atlas of the virome and virus transcriptome in 50 lymphoma cell line models and it sheds light into the functions of novel non-coding RNAs transcribed from the oriP region during EBV reactivation.
TRANSCRIPTOME ANALYSIS OF LYMPHOMA ASSOCIATED VIRUSES AND
ANALYSIS OF VIRAL NONCODING RNAs

A DISSERTATION

SUBMITTED ON THE 30th DAY OF JULY 2014

TO THE GRADUATE PROGRAM IN BIOMEDICAL SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF MEDICINE

OF TULANE UNIVERSITY

FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

SUBING CAO

APPROVED:

ERIK K. FLEMINGTON, Ph.D.
DIRECTOR

VICTORIA P. BELANCIO, Ph.D.

PRESCOTT L. DEININGER, Ph.D.

YAN DONG, Ph.D.

TONG WU, Ph.D.
ACKNOWLEDGMENTS

First and foremost, I want to give my thanks and praises to God, who created this amazing world and lets the curious hearts be satisfied with the beauty and wisdom in His creation. I thank God for that He has been loving me, leading me and providing me strength and courage to face every challenge. I thank God for using this 5-year graduate study to shape my characters and make me more mature in emotion and spirit.

I would like to express my deep and sincere gratitude to my advisor Dr. Erik K. Flemington for giving me the opportunity to do research under his guidance. I really appreciate that he always encourages me and would like to guide me through every step of my research with great patience. His enthusiasm and vision towards research inspired me. I learned from him not only how to do research, but also the characters which a good researcher should possess. All of these will greatly benefit my future career.

I would like to thank my dissertation committee members, Dr. Victoria P. Belancio, Dr. Prescott L. Deininger, Dr. Yan Dong and Dr. Tong Wu, for their valuable input and guidance throughout my graduate study.

I want to say thanks to my lab colleagues, Melody C. Baddoo, Monica Concha, Claire Fewell, Zhen Lin, Hani N. Nakhoul, Christina M. O’Grady, Michael J. Strong, Xia Wang and our lab alumni, Qinyan Yin, Jennifer Cameron and
Adriane Puetter for their generous help and friendship. I also want to thank our collaborators Dr. Walter N. Moss in Dr. Joan A. Steitz lab and Dr. Rolf Renne in University of Florida for their help on my project.

I would like to thank all my friends and my dear brothers and sisters in New Orleans Chinese Baptist Church. They are like my family and make me feel at home.

I would like to thank my fiancé, Diwei Shi, who shares happiness with me and also walks with me through difficulties. I am blessed with our relationship.

Finally, I want to thank my parents for their unconditional love and support. And I am very happy to dedicate this thesis to my best dad, best mom, best brother and loveliest dog in the world. I love them!
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................... ii

LIST OF FIGURES .................................................................................................................. vi

INTRODUCTION ...................................................................................................................... 1

1. Lymphoma epidemiology ......................................................................................... 1

2. Viruses and lymphoma pathogenesis ................................................................. 2

3. Virus-encoded non-coding RNAs (ncRNAs) and tumorigenesis ....................... 14

4. EBV latent and lytic transcriptome analysis ...................................................... 22

CHAPTER 1. High-Throughput RNA sequencing based virome analysis of 50 lymphoma cell lines from the Cancer Cell Line Encyclopedia project ........................................... 25

1.1 Abstract .................................................................................................................. 25

1.2 Introduction .......................................................................................................... 26

1.3 Materials and Methods ...................................................................................... 27

1.4 Results .................................................................................................................. 34

1.5 Discussion ............................................................................................................ 47

1.6 Acknowledgments ............................................................................................... 52

CHAPTER 2. Highly edited novel non-coding RNAs transcribed from the Epstein-Barr virus latent origin of replication (oriP) during reactivation ....................................... 77

2.1 Abstract ................................................................................................................. 77

2.2 Introduction ........................................................................................................... 78
2.3 Materials and Methods .................................................................................. 79
2.4 Results ........................................................................................................... 88
2.5 Discussion ...................................................................................................... 97
2.6 Acknowledgments ......................................................................................... 102
CHAPTER 3. Summary and discussion ............................................................... 114
LIST OF REFERENCES .................................................................................... 116
LIST OF FIGURES

Figure 1.1  HTLV-1, EBV, KSHV and murine type C retrovirus (Mu-C-retro) detection in 50 lymphoma cell lines. ......................................................53

Figure 1.2  EBV read coverage in EBV-positive lymphoma cell lines. ............54

Figure 1.3  Analysis of transcription in the EBV BamHI W intronic and oriLyt regions.................................................................56

Figure 1.4  Analysis of transcription in the EBV oriP-BCRF1/vIL10 gene region. ..................................................................................58

Figure 1.5  Evidence of new splicing of BARTs in the BamHI A region of the EBV genome. .................................................................59

Figure 1.6  Disruption of BACH2 expression by EBV integration in Raji cells occurs through a splice donor sequestration mechanism. ...........61

Figure 1.7  KSHV transcriptome analysis in BCP-1 cells.................................63

Figure 1.8  Evidence of sample cross-contamination. ..................................65

Figure 1.9  Resolution of the murine retrovirus genome in DEL cells. ..........67

Figure 1.10 MuLV integration analysis in DEL cells........................................69

Figure 1.11 Resolution of HTLV-1 genomes in MJ and HuT 102 cells. ..........70

Figure 1.12 HTLV-1 integration analysis in HuT 102 and MJ cells. ...............72

Figure 1.13 The FPKM ratio of EBV lytic genes in Hs 611.T to other analyzed EBV-positive cell lines shows a higher expression level of the BCRF1/vIL10 gene compared to the overall lytic expression level in
the Hs 611.T cell line. ................................................................. 74

Figure 1.14  Raji integration sites. ................................................................. 76

Figure 2.1  oriP transcripts............................................................................ 103

Figure 2.2  OriPt secondary structures............................................................... 105

Figure 2.3  RNA editing analysis. ................................................................. 107

Figure 2.4  oriPt associated proteins. ............................................................. 108

Figure 2.5  Effect of oriPtL overexpression on cellular gene expression. ........ 109

Figure 2.6  Coding potential prediction of oriP transcripts. ............................ 110

Figure 2.7  Effect of oriPtL knockdown on virus production............................. 111

Figure 2.8  Effect of oriPtL knockdown on EBV gene expression..................... 112

Figure 2.9  Effect of oriPtL knockdown on cellular gene expression............... 113
INTRODUCTION

1. Lymphoma epidemiology

Lymphoma is a kind of malignancy originating from lymphocytes of the immune system. It can be classified into two main groups: Hodgkin’s lymphoma (HL), characterized by the presence of giant, malignant Reed-Sternberg cells, and non-Hodgkin’s lymphoma (NHL). NHL can be further divided into B-cell neoplasms and T-cell/natural killer (NK) cell neoplasms with a variety of subtypes according to the World Health Organization lymphoma classifications.

Lymphoma is estimated to account for 3-4% of cancers worldwide (1). In the United States, the annual incidence of NHL and HL was 19.7 and 2.8 per 100,000 people, respectively, in 2013. Both the HL and NHL incidence is higher in males than in females. The age-adjusted incidence of NHL increased by 89.5% from 1975 to 2010, with an average annual increase of 2.6%, while the incidence of HL is consistently lower than NHL and has decreased since the 1970s. NHL incidence increases with age while HL incidence has two peaks: from age 20-24 and from age 80-84. It is estimated that 731,277 people in the US have been living with lymphoma or are in remission. The mortality rate of NHL is seventh among all cancer-related deaths in the US. (2-4).

Chromosome translocations are frequently observed in lymphomas. B-cell lymphomas are usually characterized by chromosome translocation involving the
immunoglobulin loci and proto-oncogenes (5-7), which results in abnormal gene expression and dysregulation of signaling pathways that govern the cell cycle, proliferation and apoptosis. Gene mutations or other genomic alterations also contribute to lymphomagenesis. Constitutively active B-cell receptors as a consequence of chronic antigen stimulus may lead to uncontrolled proliferation. The biggest risk factor for lymphoma development is immune deficiency or suppression. In addition, oncogenic viruses are found to be involved in the pathogenesis of several kinds of lymphomas and their involvement is precipitated by immune suppression.

Traditional treatments include chemotherapy, radiation therapy, stem cell transplant and surgery. Immune therapy is used for some specific subtypes and seems to have great potential. Nevertheless, a better understanding of the etiology of lymphoma subtypes and new therapeutic targets are needed to fight lymphomas.

2. Viruses and lymphoma pathogenesis

It is estimated that viral infections are responsible for about 15-20% of the total incidence of cancer worldwide (8). Lymphoma is one of best studied tumor types that is closely related to several well-characterized viruses. Since the discovery of Epstein-Barr virus (EBV) in Burkitt’s lymphomas, accumulating evidence showed the etiological roles of viruses in lymphoma initiation and progression. These lymphoma-associated viruses can be classified as DNA viruses or RNA viruses based on their genomic materials.
Tumor viruses develop strategies to promote cell survival and proliferation and to evade immune system surveillance in order to establish persistent infection in the host. Due to the small size of a virus genome and the limited number of viral genes, a virus must exploit ways to target the critical and relatively low redundancy cellular pathways to be able to survive in the complex cellular networks. Therefore, the interaction between oncogenic viruses and cellular pathways serves as a good model to study the cellular networks involved in lymphoma pathogenesis. A better understanding of the role of oncogenic viruses may provide insight into the development of new prevention and treatment methods for virus-associated lymphomas.

2.1 Lymphoma-associated DNA viruses

2.1.1 Epstein–Barr virus (EBV)

The first known lymphoma-associated virus was Epstein–Barr virus (EBV), which was discovered in Burkitt's lymphoma (BL) tumors in 1964 (9) and was also the first identified human tumor virus. EBV is a 172 kb, double-stranded DNA virus belonging to the herpesviridae family. It’s the causative agent for a variety of diseases including lymphomas, carcinomas, autoimmune diseases and lymphoproliferative disorders (9-16). The main target of EBV is B-cells, but it also occasionally infects epithelial cells and T-cells. Infection is mediated through the binding of the major viral envelope glycoprotein gp350 to the CD21 receptor, and through gp42 to HLA class II molecule (17). After the entry into cells, the terminal repeats on both ends of the EBV genome join together to form a circular genome,
which is tethered to cellular chromosomes and replicates simultaneously with host DNA (18). The EBV genome can also integrate into the host chromosome but this is not common.

EBV infects 95% of the world’s population but it doesn’t cause any problems to most of the carriers. In fact, EBV infection during childhood is usually asymptomatic while infection during adolescence and adulthood can cause a self-limiting disease called infectious mononucleosis. By taking advantage of normal B-cell differentiation processes, EBV establishes a lifelong latent infection in host cells with intermittent lytic reactivation which is closely monitored by the immune system (19). In some cases, especially in immunocompromised or immunosuppressed people, it can lead to malignant transformations. EBV-associated lymphomas include Burkitt’s lymphoma, Hodgkin’s lymphoma, post-transplant lymphoma and AIDS-associated lymphomas (20). Almost all cases of endemic Burkitt’s lymphoma are EBV-positive. It is associated with 20%-30% of sporadic BL and varying rates with other lymphomas.

Like all herpesviruses, EBV has two major components to its infection cascade, a lytic and a latent component. During the lytic cycle, nearly all viral genes are expressed. Viral DNA replication ensues and virus particles are assembled and secreted for new infection. On the other hand, more restricted viral gene expression occurs during latency, a program that evolved to facilitate persistence of the virus in the host, which is primarily associated with malignant transformation. EBV can exhibit type I, type II or type III latency programs in different lymphomas with distinct sets of latent genes being expressed in the
different latency subtypes. The full repertoire of protein-coding viral latent genes consists of six EBV nuclear antigens (EBNA 1, 2, 3A, 3B, 3C, and LP) and three latent membrane proteins (LMP1, 2A and 2B). In addition, EBV also expresses small non-coding RNAs, EBER1 and 2, and highly alternatively spliced BamHI A rightward transcripts (BARTs) during latency. Lastly, more than 40 miRNAs are also expressed during latent infections.

EBNA1 is expressed in all EBV-infected cells (all three latency subtypes). It is essential for virus episome replication, maintenance and mitotic segregation during cell division, which occurs through binding to the latent replication origin, oriP. EBNA1 also acts to enhance transcription of viral latent genes. Although the transforming activity of EBNA1 is controversial (21,22), at a minimum, it indirectly contributes to cell immortalization by maintaining the EBV genome in the cells. EBNA2 is an essential oncogenic protein in cell transformation process (17). It can transform B-cells in vitro and its deletion leads to the loss of EBV-transforming activity (23,24). EBNA2 is a transactivator protein which can bind to and activate promoters of both viral and cellular genes. The best studied downstream effector of EBNA2 is RBP-Jk (25,26). EBNA2 can bind to this DNA binding protein and activate the Notch signaling pathway (27). It can upregulate c-myc, CD21 and CD23 transcription to promote cell proliferation (28,29). It can also induce expression of the EBV transforming latency genes LMP1 and LMP2. EBNA-LP can cooperate with EBNA2 to relocate transcription repressors from the nucleus to the cytoplasm. EBNA3A, 3B and 3C are also associated with
RBP-Jk and modulate the activity of EBNA2 (30,31). EBNA3A and EBNA3C can transform B-lymphocytes in vitro but EBNA3B is dispensable (32-34).

The transforming capability of LMP1 is implicated in immortalization of rodent fibroblasts and primary B-lymphocytes and plays a tumorigenic role in transgenic mice (23,35,36). LMP1 is a multi-transmembrane protein and can mimic a constitutively active CD40 receptor, which is a marker of B-cell activation (37). It can activate the NF-κB pathway to promoter cell survival and proliferation (37). It can also activate the MAPK and PI3K pathways to induce the production of IL10 (38). LMP2 mimics the active B-cell receptor. It can inhibit the BCR signaling pathways and virus reactivation by blocking the interaction between BCR and LYN, thereby playing a role in EBV latency maintenance (39-42). In cells with deficient BCRs or cells that lack BCR signaling, LMP2 can function as B-cell receptors to transmit survival signals to cells and drive cell proliferation (43). LMP2 can also inhibit apoptosis through interaction with the PI3K/AKT pathway (44).

Collectively, EBV latent genes play important roles in cell transformation by constitutively activating cellular pathways that are normally activated in B-cells by antigens or T-cells.

2.1.2 Kaposi’s sarcoma-associated herpesvirus (KSHV)

Kaposi’s sarcoma-associated herpesvirus (KSHV) is a double-stranded DNA virus of the gamma herpesvirus subfamily. It was first discovered as the causal agent in Kaposi’s sarcoma (KS) and later was shown to be involved in
primary effusion lymphoma (PEL), multicentric Castleman's disease (MCD) and germinotropic lymphoproliferative disorder (45-49). Like EBV, it can infect B-cells and maintain a lifelong latency infection. PEL is a kind of body cavity, large-cell lymphoma. KSHV-associated PEL largely occurs in HIV-positive men with low CD4 cell counts (50) and often presents with EBV co-infection, suggesting that lymphomagenesis may be due to a combined effect of several virus infections (51). The current model of KSHV in lymphoma initiation and progression involves the oncogenic ability of KSHV latent genes and the effect of lytic genes through paracrine mechanisms. A striking characteristic of KSHV is that it encodes a large number of human orthologs, indicating its close interaction with cellular regulatory networks.

KSHV latent genes expressed in PEL include LANA, vCyclin, vFLIP, vIRF-3 and kaposins which are clustered together in the KSHV genome. LANA is expressed in all KSHV-infected cells and is critical for KSHV genome maintenance, being the counterpart of EBNA1 in EBV. Transgenic mice expressing LANA can develop splenic follicular hyperplasia and lymphomas, indicating the neoplastic potential of LANA (52). It has been shown to inhibit p53-mediated apoptosis (53) and the retinoblastoma protein-E2F transcriptional regulatory pathway (54). v-Cyclin can induce genome instability and subsequent expansion of tumorigenic clones in a mouse model (55). v-Cyclin was also shown to stimulate cell cycle progression (56). vFLIP-expressing transgenic mice have an increased incidence of lymphoma (57). vFLIP can activate the NF-κB pathway, leading to the induction of cytokines and apoptosis inhibitors such as BCL-2.
expression (58-60). vIRF-3 is a B-cell specific latent gene which can inhibit p53 and is required for proliferation and survival of cultured PEL cells (61,62). Kaposins (A, B and C) are produced from alternatively spliced ORF K12. Kaposin A can transform Rat-3 cells and the transformed cells can generate tumors in athymic nude mice (63). Kaposin B activates the p38/MK2 pathway and stabilizes cytokine mRNAs, resulting in elevated levels of pro-inflammatory cytokines (64). Although not all latent genes were shown to have the ability to transform cells in vitro, they can all interact with cellular pathways to promote cell proliferation and survival, thereby indirectly contributing to cell transformation.

Recently, studies have highlighted the role of lytic genes in KSHV oncogenesis. The lytic gene vIL-6 is expressed in most PEL biopsies. vIL-6 is highly expressed in PEL and MCD but not in KS, suggesting it may play a unique role in lymphoproliferative diseases (42). vIL-6 protects the cell from the antiviral effects of IFN-α and also induces cell proliferation in an autocrine manner (65). It can induce cellular IL-6 expression to promote tumor progression (66,67). The lytic genes K1 and vGPCR were also found to be expressed in a few PEL biopsies. vGPCR is a homologue of the human interleukin-8 (IL-8) receptor and it is in a constitutively active state. It has been shown that vGPCR can promote angiogenic lesions in transgenic mice and target a broad range of signaling pathways involved in transcription regulation in PEL (68,69). K1 is a transmembrane protein and functions similarly to the EBV LMP2 gene. It can mimic activated B-cell receptors. It can activate the PI3K/AKT pathway to inhibit FKHR-mediated apoptosis and block Fas signaling by directly binding to Fas
(70,71). In sum, both KSHV latent genes and lytic genes can rewire the cellular signaling pathway and lead to neoplasm transformation.

Although DNA tumor viruses such as EBV and KSHV encode many oncogenes, virus-associated malignant transformation is mostly observed in immunocompromised people or individuals under immune suppression conditions. This indicates that the normal immune system can successfully eliminate or silence the virus-transforming activities. In this way, the tumor viruses can be seen as “opportunists”.

EBV and KSHV take advantage of fragile points of cellular machinery, hijacking cellular pathways for their own benefit. Under appropriate conditions (cytogenic background, cellular microenvironment and host immune status), the interactions with these signaling pathways lead to cell transformation and oncogenesis. The investigation of virus-host interaction can enable us to identify ways to prevent virus infection and educate us on how to interfere with established virus-host interactions to eradicate these viruses and/or their associated diseases.

### 2.2 Lymphoma-associated RNA viruses

RNA viruses have relatively simple genome structures compared to DNA viruses, but they still encode oncogenes that can transform cells. In addition, some viruses have DNA intermediates and can integrate into the host genome, which can be mutagenic and contribute to oncogenesis.
2.2.1 Human T-cell lymphotropic virus type 1 (HTLV-1)

Human T-cell lymphotropic virus type 1 (HTLV-1) belongs to the deltaretrovirus genus of the retroviridae family. It has been implicated in adult T-cell lymphoma/leukemia (ATL), cutaneous T-cell lymphoma (CTCL) and some inflammatory diseases such as myelopathy/tropical spastic paraparesis (HAM/TSP), which develop in only a small fraction of infected individuals (72-74).

The primary target for HTLV-1 is CD4+ T-cells. HTLV-1 has a 9kb, single-stranded RNA genome. After infection, its genome is reverse transcribed into DNA and integrated into the human genome as a provirus. It then uses the cellular transcription and translation machinery to produce viral proteins and assemble new virus. Besides the typical retrovirus gene products gag, pol and env for virus assembly and reverse transcription, HTLV-1 also encodes the tax and rex genes which have been shown to contribute to the transformation process (75-77). Tax can not only enhance virus gene expression, but can also interact with multiple cellular pathways controlling DNA replication, gene transcription and the cell cycle (50,78). It can activate the NF-κB pathway to increase the expression of IL-2 and the IL-2 receptor, which promotes the expression of the anti-apoptotic genes Bcl-xL and survivin to prevent cell death (79,80). Tax also contributes to genomic instability. HTLV-1 rex seems dispensable for cellular immortalization in vitro, but it is required for viral spread and persistence in vivo (77). In addition to viral oncogene-mediated transformation activity, HTLV-1 integration may also play a role in cell transformation. Previous studies showed the HTLV-1 integration sites are not
random but have preference for gene regions or transcriptionally active regions (81-84). However, the mechanism of how integration disrupted and altered the function of these genes has scarcely been addressed. Although the rigorous secondary infectious spread of HTLV-1 in the genome results in dozens of integration sites which may blur the role of integration in oncogenesis, systematic investigation of this disruption mechanism may give us some insight into the interaction networks between HTLV-1 and its host cells.

2.2.2 Human immunodeficiency virus (HIV)

Human immunodeficiency virus (HIV) is well known as the causal agent for acquired immunodeficiency syndrome (AIDS). It is a lentivirus of the retroviridae family. A significantly elevated incidence of B-cell lymphomas is observed in AIDS patients, indicating the context of HIV infection may contribute to the development of lymphomas. Studies have shown that HIV can induce malignant transformation of human B-cells (85) and that the transformation process may be carried out by the HIV tat via interaction with pRb2/p130 (86). Expression of tat in lymphoid tissues of transgenic mice can lead to the development of B-cell lymphomas (85). Although this evidence supports the hypothesis that HIV can contribute to lymphomagenesis, in most patients the frequent co-infection of HIV with other viruses, such as EBV and KSHV, makes it difficult to define the exact role of HIV in the process of transformation. HIV is more likely to be only one of the many contributing factors to lymphomagenesis in AIDS patients.
2.2.3 Hepatitis C virus (HCV)

Hepatitis C virus (HCV) is a sense, single-stranded RNA virus of the family *Flaviviridae*. It doesn’t have a DNA intermediate and, therefore, cannot integrate into the human genome as retroviruses do. HCV is known to be the etiological agent of liver diseases such as hepatitis and hepatocellular carcinoma. Large-scale, case-control studies showed a higher HCV infection ratio in NHL (non-Hodgkin’s lymphoma) compared with controls (87,88). Some studies also showed a higher incidence of lymphoma in HCV-infected patients. However, the association between HCV infection and lymphoma is still not solid because of the regional variations involved (51). The role of HCV in lymphomagenesis is unknown, although there is some evidence that the binding of HCV E2 protein to CD81 may activate B-cells and stimulate their proliferation (89). Some studies suggested that the role of HCV in lymphomagenesis is indirect (51). The current evidence cannot sufficiently support the etiological role of this virus. To establish the causal relationship, more studies need to be carried out to investigate the prevalence of lymphoma by following up those chronic HCV-infected patients.

2.2.4 Summary of human tumor viruses and lymphomas

There are several pathways, such as the NF-κB and p53 pathways, which are frequently targeted by different viruses, indicating these pathways play critical roles in lymphomagenesis. The fact that most virus-infected individuals don’t develop tumors and that for many cancer patients the lapse between primary virus infection and onset of malignancy could be several decades, suggests virus
infection is not the only factor that is responsible for tumorigenesis. Other factors, such as chronic inflammation and acquisition of gene mutations, may also contribute to the cell transformation (90). Tumor development is a long-time process during which dynamic virus-host interactions and selection pressures from the outer environment of infected cells work together to preserve a mutually beneficial relationship between the virus and host cells. The virus may play a necessary but insufficient role in the initiation of tumorigenesis. At a certain stage of tumor development, the virus-encoded molecules are incorporated into cellular signaling pathways and function as critical signaling molecules for lymphoma cell proliferation. The tumor phenotype is highly dependent on virus gene expression, which is supported by the evidence that the loss of virus from lymphoma cells will lead to the loss of tumorigenic potential (91). However, during proliferation cells may acquire some new mutations that can take over the role of the virus oncogenes. This can explain why we observe deletions of critical oncogenes in some viruses, such as in the case of EBNA2 deletion in HL. Based on this knowledge, targeting the critical host-virus interaction at the time when the host and virus is dependent on each other would be a promising strategy to cure the disease.

2.3 Role of next generation sequencing in virome and new virus detection studies in lymphomas

The discovery of tumor-related viruses is often guided by observation of virus prevalence in endemic cancers using some traditional methods, such as
immunohistochemistry, in situ hybridization and PCR. However, these methods suffer from a lack of sensitivity in detecting low abundance viruses and are difficult to be applied to global virome and gene expression analysis. The application of next-generation sequencing (NGS) to virome studies greatly facilitates the identification of known or previously unknown tumor-related viruses and their relative abundance in human tissue and cell samples. De novo assembly of sequencing reads can help construct the genome of new virus species that are not in our current databases. All of these tools will aid in the discovery and establishment of new associations between viruses and lymphomas.

The co-infection of several viruses in single lymphoma cases may indicate the synergetic effects of these viruses in promoting neoplastic process. Therefore, it is important to evaluate how virus communities as a whole interact with host machinery and lead to cancers. NGS can be used for unbiased assessment of all organisms in a single assay and reveal the unique virome in a sample or a specific type of lymphoma. Unbiased global analysis of the virus gene expression and virus-host interactomes can help us to identify new cellular pathways usurped by viruses to facilitate their life cycle and provide insights into the role of viruses in oncogenesis.

3. Virus-encoded non-coding RNAs (ncRNAs) and tumorigenesis

Viruses interact with host signaling pathways not only via oncoproteins, but they also develop strategies to use RNA as regulatory molecules to usurp RNA-
mediated pathways in the host. RNAs are less immunogenic than proteins, which is beneficial to the virus to evade host immunosurveillance. ncRNAs can be generated from intergenic or intronic regions which is an efficient use of viral genomic space. Viral ncRNAs add a new layer to virus-host interactions involved in lymphomagenesis.

3.1 Virus-encoded microRNAs (miRNAs)

After the discovery of EBV miRNAs in EBV-infected human B-cells in 2004 (92), hundreds of virus-encoded miRNAs were identified which can regulate both viral and human gene expression. Many DNA viruses, especially herpesviruses, express a significant number of miRNAs, while RNA viruses generally don’t encode miRNAs (93). In the EBV genome there are two miRNA clusters: one cluster of 25 miRNAs is generated in the intronic region of the BamHI A rightward transcripts (BARTs) and the other cluster of 3 miRNAs is from the BHRF1 transcript. KSHV encodes 12 microRNAs clustered together in the latency locus. Understanding the functions of these miRNAs requires the identification of their mRNA targets. By combining bioinformatic prediction based on miRNA-mRNA base-pairing, luciferase reporter assay and high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation, a number of viral miRNA targets were identified and validated. It was found that viral miRNAs interact with multiple cellular pathways and perform a variety of functions such as immune evasion, apoptosis regulation, lytic reactivation, angiogenesis and cellular transportation (94). Interestingly, EBV and KSHV can target a set of similar genes through
different target regions, indicating these viral miRNAs may have similar functions although they don’t share the same seed sequences (94). There has been some evidence of the oncogenic role of viral miRNAs. It was reported that EBV BART miRNAs can sustain BL cells in the absence of other viral oncogenes by inhibiting apoptosis and that they can promote the transformation of primary B-lymphocytes (95). Previous studies showed that KSHV miRNAs can target IkBα and the NF-κB pathway to rescue cell cycle progression and inhibit apoptosis. A KSHV mutant with deletions of these miRNAs failed to transform cells and caused cell cycle arrest and apoptosis (96). Viruses can also encode human miRNA orthologs to mimic the function of these miRNAs to regulate host gene expression. For example, KSHV encodes miR-K12-11, an ortholog of human miR-155, which can induce splenic B-cell expansion and potentially lead to KSHV-associated lymphomagenesis by targeting C/EBPβ (97). Similarly, EBV encodes miR-BART5 which shares a similar seed sequence with human oncomiR miR-18, indicating its role in oncogenesis (98).

3.2 Virus-encoded small non-coding RNAs

The VA (viral-associated) RNAs are a class of highly abundant small ncRNAs identified in adenovirus (99). They are approximately 160 nt long with stable stem-loops and functional central regions. They are transcribed by RNA polymerase III (Pol III) and are non-polyadenylated with predominant cytoplasmic localization. Adenoviruses can either encode one or two VA RNAs, VAI and VAII, which are considerably different in sequence and expression levels. They are
transcribed during the early phases of infection, accumulating to very high levels by the latent phase (100). VAI can also regulate the stability of ribosome-bound RNA and is required for virus growth and efficient viral protein synthesis (101). VAI RNA was shown to counteract host antiviral response by preventing the activation of dsRNA-activated protein kinase PKR. PKR can be induced by interferon and activated by binding of dsRNAs, which can subsequently lead to mRNA degradation via RNase L and translation inhibition via eIF2. VA RNAs were found to act as decoys to saturate the binding sites of PKR to prevent its activation by dsRNAs generated during virus infection. VA RNAs are also involved in another host defense pathway: the RNAi pathway. Previous studies showed that a terminal stem region of VAI RNA can be processed into 22nt-long small RNAs which may function as siRNA or miRNA. Given the high abundance of VA RNAs, they are competitive substrates to bind to Dicer and quench its activity (102,103). Taken together, VA RNAs play important roles in protecting the virus from immunosurveillance by suppressing the interferon/PKR and RNAi pathways. Although there is no evidence linking this RNA to oncogenesis, it serves as a good model for virus ncRNA study.

EBV-encoded small RNAs (EBERs) are the most abundant EBV transcripts in EBV-infected cells (104) and are used as diagnostic markers for EBV infection. Like VA RNA, they are non-polyadenylated Pol III transcripts. They are about 170 nt in size and are predominantly localized in the nucleus (105). EBERs are highly conserved across different EBV strains (106) with stable stem-loops which serve as good protein binding sites. EBERs can form RNP with several cellular
proteins such as the La protein, ribosomal protein L22 and AUF1/hnRNPs\(^{(104,107,108)}\), but the biological significance is unclear. Previous studies have shown that EBERs can also bind to PKR and function like VA RNAs to inhibit PKR activation to rescue cells from translation inhibition mediated by the PKR pathway \(^{(109)}\), which counteracts the host antivirus response. They can protect BL cells from interferon-\(\alpha\) induced apoptosis \(^{(110)}\). Stable expression of the EBERs in EBV-negative Akata BL cells enhanced the tumorigenic potential of EBV-negative BL cells in SCID mice \(^{(111,112)}\). EBERs significantly enhance the growth potential of transformed B-cells \(^{(113)}\). Collectively, EBERs may contribute to tumorigenesis by interacting with cellular pathways in control of host defense, apoptosis and cell growth. In addition to EBERs, a small nucleolar RNA (snoRNA) was reported in the nucleolus of EBV-infected B-cells \(^{(114)}\), which interacts with three canonical snoRNA proteins. It is very abundant and can be processed into 24nt RNA species \(v\)-snoRNA\(^{124pp}\). The viral DNA polymerase BALF5 is a potential target for \(v\)-snoRNA\(^{124pp}\), indicating this RNA may play a role in EBV self-regulation. Recently, a study showed the presence of an EBV-encoded stable intronic sequence RNA (sisRNA) in the intronic region of \(W\) repeats \(^{(115)}\). Its expression level is comparable to EBERs. It is 81nt-long with stable secondary structure, and is predominantly localized in the nucleus. However, the function of this sisRNA is still unknown. All of these EBV-encoded RNA species are abundantly expressed and well-conserved, suggesting they may play important roles in the EBV life cycle and lymphomagenesis.
3.3 Virus-encoded long non-coding RNAs (IncRNAs)

In the past few years, deep sequencing of the human transcriptome has revealed a new kind of gene expression and epigenetic regulator, long non-coding RNAs, which constitutes a large portion of the human transcriptome. There is now accumulating evidence showing that IncRNAs have functions in almost all stages of gene expression. It has been observed that cancer patients have abnormal IncRNA expression patterns and some studies have shown evidence of specific IncRNAs that are associated with cancer development. While more and more evidence has accumulated for the important function of human long non-coding RNAs, virus-encoded IncRNAs have scarcely been addressed. Given the fact that viruses can encode small non-coding RNAs and utilize cellular pathways to facilitate their life cycle, it is possible that viruses may also exploit the strategy of using long non-coding RNAs to regulate gene expression and contribute to oncogenesis.

KSHV encodes a polyadenylated nuclear RNA called PAN RNA. It is a highly abundant, 1.2kb long non-coding RNA (116). It was first found in a human body cavity lymphoma cell line 48 hours after butyrate induction of KSHV lytic replication (116). It accumulates in the nucleus and accounts for 80% of the total polyadenylated mRNA in KSHV-infected cells during reactivation (116). PAN may also be expressed during latency based on a recent study (117) which indicated that PAN may play a role in both latent and lytic phases. PAN has been shown to be necessary for the production of late lytic viral proteins from polyadenylated mRNAs (118). A previous study also indicated that PAN RNA may play a role in
virus replication. A recombinant BACmid with a deleted PAN RNA locus failed to produce supernatant virus and showed a general decrease in the accumulation of viral lytic mRNAs (119). It may also contribute to the latent to lytic phase transition by activating the KSHV immediate early gene Rta (ORF50), through interacting with demethylases JMJD3 and UTX. These demethylases can decrease the repressive H3K27me3 mark at the ORF50 promoter that controls Rta expression (119). PAN RNA may also function as an immune modulator and decrease the expression of gamma interferon, interleukin-18, alpha interferon 16, and RNase L (120). All these discoveries suggest that PAN RNA interacts with both viral and cellular proteins and plays multiple roles in the KSHV life cycle. Although there is no evidence showing a direct role of PAN RNA in oncogenesis, it is possible that it plays an indirect role in tumor development by regulating lytic reactivation.

HIV is reported to encode an antisense long non-coding RNA (121). This transcript spans the majority of the viral genome and overlaps the 5’ LTR. It lacks a polyA tail and is localized in the nucleus. It is involved in epigenetic regulation of HIV gene expression. By recruiting the chromatin epigenetic modification proteins DNMT3a, HDAC-1 and EZH2, it leads to the formation of heterochromatin around HIV gene promoters, resulting in transcriptional silencing of these genes. Inhibition of this antisense IncRNA by small RNAs prevents the formation of heterochromatin and results in elevated viral gene expression and replication. This evidence suggests a function for this HIV antisense IncRNA in
HIV latency maintenance. Further investigation is required to discern the role of this transcript in HIV-associated malignancies.

During latency, a group of alternatively-spliced rightward transcripts from the EBV BamHI A region called BARTs are expressed. Although these transcripts contain several putative ORFs, their protein products have not yet been identified. They are expressed at a considerable level in EBV-associated malignancies. It is possible that they may function as regulatory RNAs. A genome-wide study of the EBV genome identified regions that may produce novel RNAs and RNA structures (115). These structures are conserved across different EBV strains, indicating they are probably functional. During the EBV lytic stage, the non-coding RNA BHLF1 is expressed at a high level. Previous investigations showed it is required to form an RNA-DNA triplex in the oriLyt region to initiate lytic replication (122). In addition, an EBV lytic transcriptome study from our lab revealed global bidirectional transcription of the EBV genome during lytic reactivation (123). Most of the newly identified transcripts show little coding potential and are localized in the nucleus and we have suggested that many of these transcripts may facilitate virus production process as regulatory RNAs (123). Since latent cell populations always contain a variable subpopulation of cells that undergo lytic reactivation, it is possible that these lytic transcripts may also contribute to cell transformation and tumor phenotype maintenance. The exact functions of these non-coding transcripts await further investigation.
4. EBV latent and lytic transcriptome analysis

EBV associates with a relatively broad spectrum of lymphoma subtypes compared with other lymphomas-associated viruses, suggesting its importance in lymphoma development. In different lymphoma subtypes EBV can exhibit one of three latency programs: latency I, latency II, or latency III, each expressing a distinct set of viral gene products. The distinct EBV gene expression patterns may result from fine-tuning by the differentiation stage and the genetic background of the infected cells as well as the microenvironment outside of the cells. Transcriptome analysis of different EBV-associated lymphomas may reveal the general rules of interaction between EBV and the host and provide a basis for customized treatment of different lymphoma subtypes.

In BL cells, EBV exhibits a very restricted type I latency: besides EBERs and BARTs miRNAs, only EBNA1 is expressed, which implies relatively harsh immunosurveillance. A hallmark of BL cells is the translocation of the proto-oncogene Myc into one of the immunoglobulin loci, leading to high expression of Myc (5). It has been shown that EBERs can counteract the pro-apoptotic effect of Myc (124), therefore playing a role in cell transformation. Serial passage of BL cell lines in vitro leads to the expression of other EBNAs and LMPs. The phenotypes of these cells were shifted towards EBV in-vitro transformed lymphoblastoid cells, suggesting the microenvironment is an important cofactor in shaping EBV gene expression patterns. In NPC and HL, EBV exhibits type II latency during which it expresses LMP1 and LMP2 as well as all the type I latency genes. Since HRS cells in HL often bear a deficiency in B-cell receptor
expression, which is critical for B-cell survival, the resultant expression pattern of EBV in these diseases suggests a role in rescuing the cells from apoptosis. This task is fulfilled by LMP1 and LMP2 which can mimic CD40 and BCR, respectively, to transmit survival signals to the cells. EBV expresses a full spectrum of latent genes (type III latency) in post-transplant lymphoma. These latent genes can be expressed and contribute to cell proliferation without eliciting immune responses, likely due to immune suppression. All the above expression patterns reflect the dynamic interactions among EBV, host cells and the microenvironment, which ultimately leads to lymphoma development. In addition to these typical EBV expression patterns, some unique EBV expression patterns were also observed which cannot be represented by any of the three latency types, indicating that the EBV transcriptome should be analyzed in a cell-specific context.

In the past 50 years of EBV research, extensive studies have been carried out on EBV oncogenes and have greatly enriched our knowledge of this DNA virus. It is becoming clear that the reciprocal relationship between EBV and its host is established through intensive virus-host-microenvironment interactions. This process is critical for lymphomagenesis and requires comprehensive system-level interpretation. High-throughput sequencing greatly facilitates the global investigation of the dynamic interaction networks. Both the cellular and viral transcriptome, and possibly interactome, can be quantified and analyzed in a single RNA-seq assay. It can also be applied to the EBV secretome analysis, which may provide insights into the communication between EBV and the microenvironment. Transcriptome, interactome and secretome analysis may
provide the basis for anti-viral therapies by targeting key signaling molecules to prevent EBV latency establishment and develop new prevention methods.

It is clear that EBV latency gene expression is essential for cell immortalization. However, the evidence from KSHV studies that lytic genes may contribute to tumorigenesis through a paracrine mechanism (42) leads us to think about similar and/or other scenarios involving the lytic cascade in EBV-associated malignancies. This hypothesis is supported by the observation that there is always a fraction of cells undergoing lytic reactivation in vitro and in vivo and that some latent genes can be induced during lytic reactivation. It is possible that the lytic gene products are not solely for new virus production, but are also involved in tumor etiology. Lytic reactivation is a highly organized cascade, during which hundreds of EBV genes are expressed. In addition to those known lytic genes, EBV lytic transcriptome analysis also revealed many novel transcripts across the EBV genome without coding potential (123). Understanding the EBV lytic transcriptome and its interplay with the latent transcriptome may open new perspectives into the etiological role of EBV in lymphoma development and progression.
CHAPTER 1. High-Throughput RNA sequencing based virome analysis of 50 lymphoma cell lines from the Cancer Cell Line Encyclopedia project

1.1 Abstract

Using high-throughput RNA sequencing (RNA-seq) data from 50 common lymphoma cell culture models from the Cancer Cell Line Encyclopedia project, we performed an unbiased global interrogation for the presence of a panel of 740 viruses and strains known to infect human and other mammalian cells. This led to the findings of previously identified infections by Epstein-Barr virus (EBV), Kaposi’s sarcoma herpesvirus (KSHV), and human T-lymphotropic virus 1 (HTLV-1). In addition, we also found a previously unreported infection of one cell line (DEL) with a murine leukemia virus. High expression of MuLV transcripts was observed in DEL cells and we identified 4 transcriptionally active integration sites, one being in the TNFRSF6B gene. We also found low levels of MuLV reads in a number of other cell lines and provided evidence for cross-contamination during sequencing. Analysis of HTLV-1 integrations in two cell lines, HuT 102 and MJ, identified 14 and 66 transcriptionally active integration sites with potentially activating integrations in immune regulatory genes including IL15, IL6ST, STAT5B, HIVEP1, and IL9R. While KSHV and EBV do not typically integrate into the genome, we investigated a previously identified integration of EBV into the BACH2 locus in Raji cells. This analysis identified a BACH2 disruption mechanism involving splice donor sequestration. Through viral gene
expression analysis, we detected expression of stable intronic RNAs from the EBV BamHI W repeats that may be part of long transcripts spanning the repeat region. We also observed transcripts at the EBV vIL10 locus exclusively in the Hodgkin’s lymphoma cell line, Hs 611.T, the expression of which were uncoupled from other lytic genes. Assessment of the KSHV viral transcriptome in BCP-1 cells showed expression of the viral immune regulators, K2/vIL6, K4/vIL8-like vCCL1, and K5/E2-ubiquitin ligase 1 that was significantly higher than expression of the latency-associated nuclear antigen, LANA. Together, this investigation sheds light into the virus composition across these lymphoma model systems and provides insights into common viral mechanistic principles.

1.2 Introduction

Over the past fifty years, it has become well established that viruses are a significant cause of a variety of human malignancies (90). Throughout this time, a large number of highly varied experimental methods ranging from electron microscopy to polymerase chain reaction (PCR) have been important for the study of virus-tumor associations and the underlying mechanisms. From this work, we have gained a great appreciation for many of the virus-cancer associations as well as many of the mechanisms driving the virus infection cycle and virus mediated oncogenesis. Despite the substantial advances using these methods, next generation sequencing (NGS) has the potential to further our understanding of viral oncogenesis in new ways. First, NGS can be used to investigate infectious agents without the aid of prior knowledge of the infectious
agents. At the same time, there are diverse kinds of information that can be derived from NGS studies - ranging from global transcriptome data, chromatin association and configuration data, to viral integration information - that expand beyond the simple virus-tumor associations to teach us new aspects of viral infection and oncogenic mechanisms.

Human viruses such as the Epstein-Barr virus (EBV), Kaposi’s Sarcoma Herpesviruses (KSHV) and human T-lymphotropic virus 1 (HTLV-1) are important contributors to B-cell and T-cell lymphomas. Despite some common themes, there is great diversity in the ways that these viruses interact with the host to achieve productive infections and in some cases, oncogenesis. Here, we utilized lymphoma RNA-seq data sets to perform a global assessment of viral involvement in a panel of 50 routinely used lymphoma cell line models. We also took advantage of the richness of RNA-seq data to inform us about the viral transcriptomes and mechanisms of virus-host interactions in these model systems.

1.3 Materials and Methods

**RNA-seq data acquisition.** RNA-seq data (in BAM format) from 50 lymphoma cell lines was obtained from the Cancer Genomics Hub (CGHub) (https://cghub.ucsc.edu/). This data was generated by the Broad Institute for The Cancer Cell Line Encyclopedia (CCLE) project (125) and was deposited under “Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC)”. PolyA-selected cDNA libraries were prepared using total RNA from each cell line. The libraries
were subjected to paired-end 101 base sequencing reactions using the Illumina platform. The downloaded BAM files were converted to paired FASTQ files using the Picard samtofastq script (default parameters; http://picard.sourceforge.net). Under the default condition, Picard samtofastq outputs only the primary alignments. This is important because the BAM files contain secondary alignments for repeat regions which results in multiple copies of repeat aligned reads. The inclusion of these copies during conversion to FASTQ format would cause gross over-representation of read coverage at repeat regions.

**Virome analysis.** The paired-end FASTQ files for each of the 50 cell lines were aligned to a custom reference genome using the sequencing aligner STAR (Spliced Transcripts Alignment to a Reference) version 2.3.0 (command options: clip5pNbases 6, outFilterMultimapNmax 1000) (126). The custom reference genome was built to include the human reference genome, hg19 (Genome Reference Consortium GRCH37), human ribosomal DNA sequences, 740 mammalian virus genomes from NCBI and bacterial phage genomes from the DNA database of Japan (DDBJ). The numbers of reads that mapped to human hg19, ribosomal DNA, viruses and phages were quantified using in-house scripts. Viral read numbers were normalized by calculating viral reads per million unique mapped reads (VPMM).

**Transcriptome and splicing analysis.** The paired-end FASTQ files were aligned to the human hg19 genome and the genomes of the identified viruses in our virome analysis using the junction aligner, TopHat version 2.0.9 (command options: -p 10, --solexa-quals, -g 1, --no-coverage-search, --fusion-search)
Aligned viral reads were extracted from the subsequent TopHat output. Splice junction reads with counts lower than 5 were discarded. Viral reads and annotated junctions were loaded onto the Integrative Genomics Viewer (IGV) (128,129) for visualization and analysis.

Hierarchical clustering analysis of EBV gene expression. The expression of EBV genes was quantified using SAMMate (130). Reads aligned to each EBV gene were counted and FPKMs were calculated by normalizing the read numbers to the kilobases of total exon length and millions of total mapped EBV reads. The oriLyt unique regions (right and left), W repeat intronic regions and W repeat exonic regions were included as additional annotated gene features in our annotation file. Lytic genes that overlap with latent genes were excluded from the clustering analysis to avoid ambiguous read allocation. In addition, quantification data for the Raji cell line was excluded from the clustering analysis because the deletion of a number of lytic genes may cause biases in determining the basal lytic transcription level for this cell line. Hierarchical clustering analysis was performed using MeV (MultiExperiment Viewer) based on EBV read counts and FPKM (Fragments per kilobase per million mapped reads) values (Metric: Pearson Correlation, Average linkage clustering) (131).

Quantification of KSHV gene expression. The paired FASTQ files from the KSHV-positive BCP-1 cell line were run through RSEM-1.2.12 (command options: rsem-calculate-expression -p 24 --paired-end --no-bam-output) (132) to quantify gene expression. The RSEM reference file was built with genome sequences and annotation for hg19, EBV, KSHV, MuLV and HTLV-1.
Cell culture. Hs 611.T (ATCC® CRL7373™) cells were cultured in DMEM/High Glucose medium (Thermo Scientific, Cat. #SH30243) supplemented with 10% fetal bovine serum (FBS; Invitrogen-Gibco, Cat. #16000-069) and 0.5% Penicillin-Streptomycin (pen/strep; Invitrogen-Gibco, Cat. #15070-063). EBV positive Akata, Raji and JY cells were cultured in RPMI 1640 medium (Thermo Scientific, Cat. #SH30027) supplemented with 10% FBS and 0.5% pen/strep. All cell lines were cultured at 37°C in a humidified, 5% CO₂ incubator.

vIL-10 transcription analysis. Total RNA was isolated from Hs 611.T and Akata cells using TRlzol® reagent (Life technologies, Cat. #15596-018) following the vendor’s protocol. Nuclear and cytoplasmic RNA was isolated using a cytoplasmic and nuclear RNA purification kit (Norgen Biotek Corp., Cat. #21000) following the vendor’s protocol. All RNA preparations were subjected to DNase treatment using the DNA-free™ kit (Life technologies, Cat. #AM1906). Standard reverse transcription was performed using an iScript™ cDNA synthesis kit (Bio-Rad, Cat. #170-8891) according to the vendor’s protocol. Strand-specific reverse transcription was carried out with modified gene specific primers using the method as described by Feng et al. (133). cDNA was synthesized at 65°C for 50 min using ThermoScript reverse transcriptase (Life Technologies, Cat. #12236-022) and treated with exonuclease I to remove the excess gene specific primers. qPCR analysis was performed using iQ SYBR green supermix (Bio-Rad, Cat. #170-8882) on a Bio-Rad CFX96 instrument as follows: 2 µl cDNA (5ng/µl) and 1 µl 10µM primers were mixed with 10 µl of SYBR green Supermix and 7 µl nuclease free H₂O was added to a final reaction volume of 20 µl. Polymerase
was activated and cDNA was denatured at 95°C for 5 min. cDNA was then amplified for 40 cycles with 15 s denaturation at 95°C, 60 s annealing/extension and plate reading at 60°C. Melting curve analysis was performed at temperatures from 60°C to 90°C with 0.5°C increment per 5 s. Expression fold changes were calculated using the comparative CT method ($2^{-\Delta\Delta CT}$).

Primers used for qPCR:

- **P1_F**: ATAGCACAATGCCACCACACTGAACC
- **P1_R**: GCCTGCTTTCATTCCTCTCGT
- **P2_F**: GTATGGAGCAGAAGGTTAGTG
- **P2_R**: GTCACATTGGTCTGTACCTC
- **P3_F**: GAAGGTTAGTTGGTCACTCTG
- **P3_R**: GACTGAAGGCTCTCTCTTAGG
- **ACTB_F**: CACTCTCAGGCTCCTTCTC
- **ACTB_R**: GTACAGGTCTTTGCGGATGT

Modified gene specific primers used for strand-specific RT (lower-case refers to mismatches):

- **P1_mGSP_L**:
  ATAGCACAATGCCACCACACTGAACCaattaTTCAAATTTTA
- **P1_mGSP_R**:
  GCCTGCTTTCATTCCTCTCGTggcGCgctAGAATAACTG
- **P3_mGSP_L**:
  GAAGGTTAGTTGGTCACTCTGtAtTaTaTTGTTGCTGCTTTTA
- **P3_mGSP_R**:
De novo assembly of virus genomes. For the virus genome assembly in the DEL cell line, all reads that mapped to viruses were pulled out from the SAM alignment file and de novo assembled using the assembler, Trinity (134). Three long contigs of about 6,800 bp (representing alternatively spliced transcripts) and a single overlapping short contig were used to assemble a nearly full length sequence that spanned most of the genome except for a few hundred LTR sequences at each end. The missing flanking LTR sequences were substituted with the corresponding sequences from the close relatives, EKVX and DG75 to generate a chimeric genome. Alignment of reads from DEL cells against this reconstructed genome yielded 3 mismatches at the 5’ inserted DG75 derived sequence. This sequence was corrected in our genome after which alignment of reads from the DEL cell line showed no mismatches. The HTLV-1 genomes in the HuT 102 and MJ cell lines were assembled through a similar process.

Virus integration analysis. Raji integration sites were first detected by analyzing the chimeric reads that span both the EBV genome and the human genome as described in the following paragraph. The result of this analysis was consistent with the finding in a previous report by Takakuwa et al. (135). To study the aberrant transcription and splicing caused by EBV integration, a chimeric EBV-chr6 chromosome was generated representing the integration and Raji FASTQ files were aligned to this chimeric chromosome along with all the other
human chromosomes using TopHat. Splice junctions with read counts lower than 5 were discarded. The read coverage and filtered junctions were loaded onto IGV to visualize the transcription and assess the consequences of integration on BACH2 and EBV gene splicing.

Viral integration analysis was carried out using the following in-house protocol and scripts. To increase the sensitivity for detection of virus integrations and the junction spanning potential of each read pair, truncated read pairs contained the first 27 bases from each of the original paired-end reads were generated and the read partners were aligned separately to the human hg19 genome plus the appropriate virus genome using Novoalign (Novocraft) (http://www.novocraft.com). Read pairs in which one of the pair aligned to the EBV genome and the other aligned to the human genome were pulled out. To dismiss ligation chimeras, clusters of reads were collated and counted based on the number of consecutive reads that mapped to within 300 bases of each other in the human genome. Read clusters with greater than 10 unique (non-PCR duplicate) reads were visually inspected on a genome browser to assess uniform read orientations (for ligation chimeras, read orientation is random whereas for true integrations, reads are uniformly aligned toward the integration site).

**EBV integration and splicing validation in Raji cells.** Total Raji RNA was isolated using TRIzol® reagent. cDNA was made using the iScript™ cDNA synthesis kit. To validate the integration sites and splicing junctions, PCR was performed using a Taq PCR kit (NEB, Cat. # E5000S) following the vendor’s protocol. Initial denaturation of cDNA templates was carried out at 95°C for 30 s.
The templates were then amplified for 30 cycles: denaturation at 95°C for 30 s, annealing at 55-60°C for 30 s and elongation at 68°C for 1 min. Final extension was at 68°C for 5 min. Annealing temperatures were optimized for each primer pair. The PCR products were run on a 1% agarose gel.

Primers used for PCR:

BACH2 exon1-exon2_F: CTTGGTCCCAAATGATGTCT
BACH2 exon1-exon2_R: CGAGCTGCCATGTGATG
BGRF1-BACH2 exon1_F: CTTCACTTAGGTGGCGAAATG
BGRF1-BACH2 exon1_R: TTGTGGACGAGGCTAACT
Exon1-EBNAs_F: CTTCACTTCGCTCTCCTCA
Exon1-EBNAs_R: GAACGAGCTGCCATGTG

RNA sequencing. Total RNA was extracted from JY cells using TRizol® reagent according to the vendor's instructions. PolyA-selected or ribodepleted (Ribo-Zero; Epicentre, Cat. # MRZH11124) cDNA libraries were prepared using the TruSeq stranded protocol (Illumina, Cat. # RS-930-2001). PolyA-selected samples underwent 101 base paired-end sequencing using an Illumina HiSeq 2000 instrument. Ribodepleted samples underwent 101 base single-end strand-specific sequencing using an Illumina HiSeq 2000 instrument.

1.4 Results

Virome analysis of 50 lymphoma cell lines. To assess the viromes of lymphoma cell lines sequenced through the Cancer Cell Line Encyclopedia (CCLE) project, we downloaded a 50 lymphoma RNA-seq dataset from the
Cancer Genomics Hub (CGHub; https://cghub.ucsc.edu/). This panel of cell lines represented a range of B- and T-cell lymphoma subtypes. The average number of reads per cell line was approximately 164 million, ranging from 74.6 million to 209.8 million. Based on our previous investigations of microbial agents in biological specimens and cell lines (136-138) these read depths were deemed to be well above the minimum levels required to detect meaningful pathological infections.

The virome for each cell line was assessed by aligning each RNA-seq dataset against a reference genome containing the human hg19 genome plus the genomes of 740 virus strains that are known to infect human and other mammalian cells. The reads aligned to each virus from each cell line were counted and the data for all detected viruses was collated into a single excel spread sheet. Although low numbers of reads aligned to hepatitis C in most samples, the corresponding reads were found to be composed of homopolymer sequences (polyT tracts) that mapped to a polyT tract in the hepatitis C genome. With a lack of any other read evidence for hepatitis C, these polyT-containing sequences most likely represent polyA tracts from human transcripts. Low numbers of reads were also attributed to hepatitis B, human herpesvirus 7, Merkel cell polyomavirus, SV40, human adenovirus, and parainfluenza virus 5 in several samples. Manual BLAST analysis of these reads showed true matches to the corresponding viruses. Nevertheless, the low read numbers in these cases suggest that they are most likely due to cross contamination during sequencing rather than transcription from endogenous viruses (see Strong et al., in press and
below). In contrast, more than 10,000 reads were found for EBV, KSHV, HTLV-1 and murine retroviruses in some samples and these viruses were investigated in more detail.

**Detection of EBV.** For each of the seven cell lines known to harbor EBV, more than 10,000 reads were detected, corresponding to more than 400 reads per million unique mapped reads (Fig. 1.1). An additional cell line, DOHH-2, showed a relatively low 838 EBV reads. The DOHH-2 cell line was originally categorized as EBV negative (139) and a cell line vendor, Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ), classifies it as EBV negative. Nevertheless, a previous study showed the presence of EBV in DOHH-2 cells by PCR analysis with EBV episomes detected in only a small fraction of cells by FISH (140). Since the DOHH-2 cell line was derived by the clonal expansion of a mixed EBV negative and EBV positive culture as described by DSMZ, it is possible that there was carryover of a small percentage of EBV positive cells during the selection of EBV negative cells. Lastly, very low numbers of EBV reads were detected in another 29 samples. Manual BLAST analysis of the corresponding EBV reads excluded the possibility of misalignment of human reads to the EBV genome. We therefore suspect that similar to our contention regarding low read numbers detected for other viruses, these low numbers of EBV reads most likely reflect cross contamination during sample processing and/or sequencing (Strong et al., in press).

**EBV transcriptome analysis.** To comprehensively explore the EBV transcriptome in EBV positive cell lines, we aligned the corresponding RNA-seq
reads to the Akata EBV genome (141). To allow for the detection of LMP2 splicing across the terminal repeat region, we split the genome between the lytic genes, BBRF2/3 and BGLF3.5, rather than at the terminal repeats (Fig. 1.2). As additional comparisons, we included the RNA-seq data generated in our lab for the type I latency Akata cell line (123) and the type III latency JY cell line (see materials and methods). For the most part, gene expression reflected type III latency-like transcription profiles and was in line with previously reported gene expression profiles and genomic deletion data for these cell lines (Fig. 1.2) (142-144). Nevertheless, unique expression patterns were detected in the EBNA BamHI W repeat intronic regions, the oriLyt loci and the oriP-BCRF1/viL10 region.

**EBV BamHI W repeat intronic transcripts.** A recent study identified a short (81nt) stable intronic sequence (sis)RNA between the W1 and W2 exon of the EBV EBNA BamHI W repeat region and a large RNA hairpin structure upstream from the W1 exon (Fig. 1.3) (115). The large hairpin is part of an extensively-structured RNA predicted throughout the long W repeat intron which was similarly suggested to be an independent sisRNA (145). In all type III or type III-like latency cell lines, we noted marked coverage across most of the intronic regions of the EBNA BamHI W repeat sequences, although the levels in the Namalwa cell line were found to be low (Fig. 1.2 and 1.3B). In contrast, no expression was observed in the type I latency cell line Akata. This is in keeping with the hypothesis that these transcripts are derived from the Cp and/or Wp promoters that drive EBNA latency gene expression in latency type III cells (as opposed to type I latency cell lines where EBNA1 expression is derived from the
downstream Qp promoter). Strand-specific RNA sequencing of the JY cell line indicated that transcription of the intronic regions is in the rightward direction (Fig. 1.3B), an observation that is likewise consistent with the sisRNAs being generated from the Cp or Wp promoters.

sisRNAs typically arise from excised intronic sequences and are therefore not polyadenylated. Expecting that excised intronic BamHI W RNAs were not polyadenylated, we were surprised to find abundant coverage in the data from the CCLE cell lines and the non-strand-specific JY cell line, all of which was derived from polyA-selected RNA. Furthermore, the level of BamHI W intronic coverage was similar in data from ribodepleted (strand-specific experiment) and polyA (non-strand-specific data) selected JY RNA (Fig. 1.2 and 1.3B). In contrast, 20,717 and 0 reads mapping to the non-polyadenylated EBV EBER1 gene were found in ribodepleted and polyA-selected JY RNA, respectively. This suggests that the BamHI W intronic RNAs are linked to polyA tracts or tails.

Considering the possibility that cryptic polyA signals might exist in the BamHI W region, we searched EBV reads for evidence of polyA tails using the bioinformatic approach that we had used previously (123). While polyA-containing reads were well correlated with other known polyA signals in the EBV genome, we saw no clear evidence of polyA reads mapping to the BamHI W region. This raises the possibility that much of the coverage that we observe in the BamHI W intron regions results from long, relatively unspliced, EBNA transcripts containing their normal 3’ terminal polyA sequences. Clearly, some of these transcripts were normally spliced because we detected junction reads
across this region (data not shown). Nevertheless, the high levels of intronic coverage suggest a complex mixture of spliced, partially spliced, sisRNA, and possibly unspliced transcripts that perhaps extend across the entire BamHI W repeat region.

**EBV oriLyt transcripts.** We have previously observed bidirectional transcription of the oriLyt region in reactivated Akata cells ([123] and SC, TO’G, and EKF, unpublished). Unexpectedly, however, we observed read coverage at the lytic origins of replication in the type III latency cell lines that was not observed in uninduced Akata cells (Fig. 1.3C). At least in JY cells, these transcripts are oriented in the rightward direction as determined by strand-specific sequencing (Fig. 1.3C). Due to the homology between the left and the right oriLytS, however, we were not able to determine whether the transcripts originated from one or the other or both regions of the genome. There was no noticeable correlation between oriLyt coverage and lytic gene expression (Fig. 1.2) raising the possibility that these reads were derived from latency transcripts. Further, hierarchical clustering showed that these transcripts cluster more closely to latency genes with the closest relationship being to the intronic W repeat transcripts and BHRF1 transcripts (Fig. 1.3A). These oriLyt transcripts likely represent latency RNAs that are derived from novel promoters and/or are isoforms of BHRF1 transcripts (146).

**Transcription of the EBV vIL10 locus in a Hodgkin’s lymphoma cell line.** While the coverage at most of the lytic genes roughly correlated with each other across the different cell lines, we noted unusually high coverage at the EBV
encoded vIL10 (BCRF1) locus in the Hodgkin’s cell line, Hs 611.T, that didn’t correlate with the coverage levels of other lytic genes (Fig. 1.4A). In line with the disconnection between vIL10 coverage and other lytic genes, vIL10 clustered separately from other lytic genes (Fig. 1.3A) and was expressed at higher relative levels in Hs 611.T cells (Fig. 1.13). qRT-PCR analysis showed higher transcription levels in Hs 611.T than Akata cells (Fig. 1.4B) and strand-specific qRT-PCR showed that these are transcribed in the rightward direction (Fig. 1.4C). Together, these data suggests that expression of vIL10 transcripts in Hs 611.T cells is disconnected from the expression of other lytic genes, perhaps driven instead by tissue specific cellular factors in this Hodgkin’s lymphoma milieu.

**New splicing evidence in the BamHI A region of the EBV genome.**

Alternative splicing of a group of rightward non-coding latency transcripts from the BamHI A region (BamHI A rightward transcripts (BARTs)) has been extensively studied (147-149). In addition to giving rise to apparently stable viral IncRNAs, introns from some of these transcripts are further processed to generate the bulk of the EBV encoded microRNAs. We analyzed splicing across the EBV genome using TopHat and detected all of the previously identified canonical splicing events for these EBV IncRNAs (Fig. 1.5A). A genome deletion of part of this region in JY cells (corresponding to the deletion found in the B95-8 strain of virus) gives rise to the splicing of upstream exons to the most proximal splice acceptors of the last few exons (Fig. 1.5A).

In addition to previously annotated splice junctions, we also found evidence of novel splicing. Evidence for a new junction originating from the intron
downstream from exon 4 and spliced to the exon 5 splice acceptor was found in seven of the cell lines. And exon 6 was spliced to a new acceptor in its 3’ intron (Fig. 1.5B). Additional new splice junctions were found within exons 5 and 7, some of which were observed in multiple cell lines (Fig. 1.5). This data adds to the already complex splicing known for this region and implies the presence of extensive fine-tuning mechanisms that generate a complex set of isoforms with refined functions in regulating EBV and/or cellular gene expression and signaling.

In Raji cells, EBV integration disrupts BACH2 expression through a splice donor sequestration mechanism. Although rare, EBV has occasionally been found to integrate into the host cell genome (150-156). To assess integration events in all EBV infected cell lines, we developed a pipeline to identify chimeric transcripts and their associated genome locations from NGS data. This method uses truncated read ends (27 bases each) from each paired-end read set to align separately to the human and EBV genomes. Chimeric read pairs are then analyzed for high-density clustering on the human genome (see materials and methods for details). Using this approach, we saw no evidence of integration in the Granta-519, Daudi, CI-1, Namalwa, Hs 611.T or P3HR-1 cell lines (DOHH-2 data was not analyzed due to the low read numbers). Consistent with previous studies, however, we found two clusters of 209 and 235 chimeric read pairs mapping to the first intron of the BACH2 tumor suppressor gene in Raji cells (135,150). Alignment using full length Raji reads showed robust coverage of BACH2 exon 1 but minimal coverage at all downstream exons (Fig. 1.6) whereas Namalwa cells, which don’t have a BACH2 integration, showed
high coverage across all exons. This suggests that the integration in intron 1 is disruptive to BACH2 expression.

To address the mechanism of BACH2 disruption, we generated a chimeric chromosome 6 containing the EBV genome in silico. We then aligned full-length Raji reads to this chimeric genome using the junction mapper, TopHat. Whereas no BACH2 exon 1 to 2 splicing was observed in the alignment to the wild type genome, 135 reads mapped to a chimeric junction corresponding to splicing from the BACH2 exon 1 to the BamHI W exon, W1 (Fig. 1.6). BACH2 exon 1 to W1 splicing but not the BACH2 exon 1 to exon 2 splicing was validated in Raji cells by RT-PCR (Fig. 1.6, bottom panel). This data showed that disruption of BACH2 expression occurs through sequestration of the exon 1 splice donor. Conversely, these findings also show that expression of EBV latency genes in Raji cells is directed in part by the BACH2 promoter.

**KSHV transcriptome analysis.** Like EBV, the KSHV genome principally resides in the host cell as an episome rather than through integration into the host genome. In line with this, we saw no evidence of transcriptionally active integrations for KSHV indicating that this is likely the case in BCP-1 cells.

The classic KSHV latency locus is located towards the right side of the linear genome and encodes genes such as Kaposin A, vFLIP, vCyclin, LANA as well as the viral microRNAs (Fig. 1.7). Nevertheless, it is becoming clear that other genes, such as the non-coding PAN and the vIL6 genes located in the left cluster of the genome (Fig. 1.7) are expressed in some tissues in the absence of reactivation (117,157-161). Notably, in BCP-1 cells, many genes in the left
cluster of the genome including vIL6 and PAN (Fig. 1.7) are expressed at levels that are considerably higher than genes in the classic rightward latency cluster. For example, the most highly expressed gene in the rightward cluster was found to be LANA which showed an expression value of 39 FPKM (Fragments per kilobase per million mapped reads). In contrast, expression of K2/vIL6 was 840 FPKMs, K4/vIL8/vCCL3 was 145 FPKMs, K5/E3-ubiquitin ligase was 351 FPKMs, and K6/vIL8/vCCL-1 was 54 FPKMs. Expression of some of these genes, such as K4/vIL8/vCCL3 were shown previously to be partly dependent on Rta expression (161). Nevertheless, even if expression of some of these genes occurs in a small percentage of reactivating cells, the high expression levels of secreted cytokines, for example, might be expected to not only have an intrinsic impact on lytic cells but may also have an extrinsic impact on proximal latently infected cells.

In addition to this cluster, we also noted the expression of the IRF cluster containing viral IRF1, 2, 3, and 4, at levels similar to those of the LANA containing rightward cluster (Fig. 1.7). Therefore, in addition to the expression of cytokines, the virus is programmed to modulate innate immunity in mostly latently infected cell populations.

**A murine type C-like retrovirus in the DEL cell line.** Over the past several years, we and others have unexpectedly discovered the presence of murine leukemia viruses or related viruses in a number of routinely used human cell lines (138,162-168). In many cases, this has been thought to occur during propagation of human cell lines in mouse xenografts (167). As seen in Fig. 1.1B
and 1.8A, two anaplastic large cell lymphoma cell lines, DEL and Ki-JK, were found to have more than 2,000 reads that aligned to the murine type C retrovirus, with the DEL cell line having nearly 600,000 virus matching reads. To our knowledge, the presence of a murine retrovirus has not previously been reported for these cell lines and at least for DEL cells, this needs to be considered for future experiments utilizing this cell line (see discussion).

When the murine type C retroviral reads were displayed in the order of file names, we found that the DEL and the Ki-JK cell lines are positioned in two separate sample clusters with cell lines that all contain lower but detectable numbers of reads matching the murine type C retrovirus (Fig. 1.8A). We also noted that the alignment profile of all murine type C retrovirus reads within the DEL cluster are similar and that single nucleotide differences with the aligned genome are nearly identical across all samples. We have previously addressed the impact of sample cross-contamination during the sequencing process (Strong et al., in press). Because the file names are sequential, we suggest that they were likely processed in a batch and that samples may have been contaminated with DEL sequences at some point in the sequencing pipeline. The samples in the cluster at the bottom of Fig. 1.8A have homology to murine type C retrovirus but have a better match to the Moloney murine leukemia virus (Fig. 1.8B). Here it can be seen that there are similar mismatch profiles with the Moloney murine leukemia virus genome, suggesting a possible common contaminating source. In this case, however, it seems likely that the contamination occurred from a
separate cell line that is not in our study set because all read numbers are relatively low.

**Resolution of the murine retrovirus genome in DEL cells.** To resolve the genomic structure of the murine retrovirus in DEL cells, we first performed *de novo* assembly of all DEL reads that matched the viruses in the viral genome index. *De novo* assembly was performed using the Trinity assembler and resulted in a long genomic sequence. BLASTing this assembled sequence against the NCBI NT database showed close homology to the EKVX genome (NCBI accession number JF908817) and the DG75 virus genome (NCBI accession number AF221065). Approximately 1510 bases of flanking LTR sequences missing from our *de novo* assembled contig were taken from the EKVX and DG75 virus genomes and appended to the ends of the DEL genome. DEL reads were then aligned to this chimeric genome and three mismatches were corrected to give the final DEL murine retrovirus genome (Fig. 1.9). This analysis showed that the DEL retrovirus has similarity to the murine retrovirus originally identified in human prostate samples, XMRV (xenotropic murine retrovirus) (167), but has closest homology to the murine retroviruses found in EKVX and DG75 cells (Fig. 1.8).

**Integration of the DEL murine retrovirus into the human TNFRSF6B gene.** Using our integration analysis pipeline, we identified 4 candidate sites of integration for the DEL retrovirus, three of which were within the introns of human coding genes (TNFRSF6B, OPA1, and ZNF767). The integration at TNFRSF6B, which encodes a factor that suppresses FasL and LIGHT mediated cell death,
showed the greatest chimeric read depth with 1607 junction spanning reads (Fig. 1.10). Notably, greater TNFRSF6B expression was observed at the first two exons than the third exon which is downstream from the integration site (Fig. 1.10) raising the possibility that the integration disrupts TNFRSF6B expression.

**Resolution of HTLV-1 genomes in MJ and HuT 102 cells.** Alignment of reads from MJ and HuT 102 cells to the HTLV-1 (NCBI accession number NC 001436.1) genome showed close homology but with numerous mismatches (Fig. 1.11). We resolved the viral genomes harbored in these cells using the *de novo* assembly approach outlined above for the DEL MuLV genome. Aligning MJ and HuT 102 reads to the resolved genomes showed near exact matching although a small number of partially penetrant differences were observed. The presence of these partially penetrant differences suggests genome heterogeneity due to mutations or modifications in these cell lines. It may also represent the nucleotide variation caused by RNA editing. Nevertheless, the consensus genomes show close relationships with a number of members of the HTLV-1 family (Fig. 1.11, bottom panel).

**Multiple HTLV-1 integrations in HuT 102 and MJ cells.** Previous studies have shown that primary HTLV-1 induced T-cell tumors generally have a single integrated provirus but that during propagation in tissue culture, the number of integrations can increase significantly (169,170). Whether these extra integrations influence cell line evolution and/or cell growth in culture is unclear.

We globally assessed HTLV-1 integrations in HuT 102 and MJ cells through the identification of chimeric transcripts. This analysis revealed 14
transcriptionally active integrations in HuT 102 cells and 66 transcriptionally active integrations in MJ cells. Most of these were found in known gene loci but a few were localized to unannotated regions of the genome. Visualization of all integrations shows an expected strong bias for introns over exons; likely due to the general substantially greater length of introns which makes them much larger targets.

Although integration into coding exons clearly predicts functional disruption, intronic insertions are also likely in most cases to have an impact on the respective gene function/expression. Some of these intron insertions are likely to alter the corresponding gene expression through disruption whereas others are likely to enhance expression of the cognate gene. Interestingly, in each cell line, there are a number of insertions that occur within introns but which are nonetheless upstream from the exons containing the translation initiation codon. In these cases, we observed little coverage of the upstream non-coding exon(s) but detected expression of exons downstream from the integration site with evidence of splicing from the HTLV-1 genome to the expressed downstream exons (for examples, see Fig. 1.12). These examples are suggestive of a mechanism for integration mediated expression of otherwise silent cellular genes.

1.5 Discussion

**Virus findings in lymphoma cell lines.** For nearly all of the cell lines tested here, our virome survey revealed no previously unknown infections. It is comforting to know that there are likely no other viral infectious agents in these
“workhorse” cell lines that may potentially confound data interpretation. In contrast, the previously unknown finding of MuLV in DEL cells may potentially cause unforeseen alterations in cell signaling pathways that can cause data misinterpretation. For example, investigations into pathways related to TNFRSF6B might be affected by the integration of MuLV into this locus. More generally, the robust expression of retroviral transcripts in these cells may alter a number of signaling pathways and possibly influence overall metabolism. Outside of these concerns, the apparent ease with which murine leukemia viruses can infect human immune cells raises a potential hazard to unknowing investigators working with these cell lines.

**Low-level virus read findings.** Theoretically, the presence of any viral reads in a sequencing data set should indicate the presence of the respective virus in the biological specimen. In practice, however, the high sensitivity of next generation sequencing in which hundreds of millions of reads can be obtained from a single run sets the stage for the detection of low level contaminants. We have recently reported on the pervasiveness of bacterial read contamination in RNA-seq data that likely arises from the sequencing pipeline itself rather than from the source material (Strong et al., in press). Some contamination likely occurs through the presence of bacterial DNA in water and library preparation reagents, such as enzymes, nucleotides and other reagents that are derived from bacterial sources. Nevertheless, we also provided evidence for contamination across samples (Strong et al., in press). Though it is unlikely for eukaryotic virus RNA or DNA to be present in library preparation reagents, it is certainly possible
for virus cross-contamination in samples that are processed in parallel and/or in succession. Though not proof, the finding of clusters of samples with the same genome subtype intimates this possible source of viral reads in some of these cases. Care must therefore be taken to avoid reading too much into the finding of low levels of viral reads in biological samples using next generation sequencing.

**Nonconformists in gammaherpesvirus latent/lytic gene categorization.**

In the EBV and KSHV fields, evidence has been accumulating over the past several years that blurs the lines between latent and lytic gene classification. In 2006, Keiff’s lab (171) and Lu et al. (172) reported that EBV latent genes are induced during EBV reactivation. Conversely, a number of groups have reported that a number of KSHV genes that were considered to be lytic genes are, in fact, also regulated/expressed independently of reactivation. This indicates a role for cellular tissue specific transcription factors in controlling the expression of these genes in some milieu (117,157-161). Notably, many of these KSHV genes encode immune regulatory factors such as interleukins (e.g. vIL6 and vIL8) and interferon response factors (e.g. vIRF3). As shown in Fig. 1.3, we observed expression of transcripts at the EBV vIL10 region that is similarly disassociated from other lytic genes, being unique to the Hodgkin’s lymphoma cell line. This is in line with another study in which IL10 expression was noted in Hodgkin’s lymphoma biopsies (173). Like the observed tissue specific expression of viral immune regulatory genes in KSHV, EBV similarly encodes at least one lytic immune modulatory factor that is also regulated in a tissue specific manner.
EBV BamHI W intronic RNAs. Although read coverage at intronic regions is commonly found in ribo-depletion-based RNA-seq experiments, the level is typically low in polyA-selected data (data not shown). Because transcripts mapping to the BamHI W intronic sequences are apparently not filtered out by polyA selection, we have proposed that they may be associated with a 3’ polyA sequence. The lack of polyA-containing reads mapping to within the BamHI W repeats raises the possibility that these intronic sequences are associated with a polyA signal that is downstream from the repeats. Notably, the high level of intronic coverage at the BamHI W repeats is discordant with the low coverage at the longer downstream introns that are spliced out in EBNA1 and EBNA3A, B, and C transcripts, suggesting that the splicing of these downstream introns is much more efficient. Possibly, transcripts containing the BamHI W introns utilize the EBNA2 or BHRF1 polyA signals or some other novel polyA signals. Nonetheless, our findings raise the possibility that there are large, possibly up to 90kb unspliced intron-containing transcripts that serve certain functions during latency such as regulating chromatin structure and/or controlling promoter activity and usage.

Alteration of gene function by virus integration. Although exonic viral integrations typically have strong ablative impacts on gene function, the integration into intron regions is much more prevalent due to the substantially larger sizes of introns relative to exons. Nevertheless, there are many ways that intronic integration can have an impact on gene function. As nicely illustrated for EBV in Raji cells, integration can cause disruption by sequestering the splice
donor from the upstream cellular gene exon through viral splice acceptors. In this case, the tumor suppressor gene, BACH2, is disrupted, potentially promoting cell growth and/or tumorigenesis. Conversely, this exact splicing configuration may be a way to support constitutive expression of the EBV latency genes, which would similarly be predicted to promote cell growth and survival through the function of viral latency genes. This integration may therefore serve a dual purpose.

A second interesting mechanism for altering gene function through intron integration was observed for a number of cases of HTLV-1 integration. Specifically, integration was observed in genes in which the translation initiation codon is located in downstream exons with the virus having integrated into an intron that is upstream from these AUG containing exons. This configuration allows highly-expressed viral transcripts to splice to cellular gene splice acceptors in a manner that retains the full coding capacity of that particular gene. In these cases we typically noted little to no expression of exons upstream from the integration site and high expression of downstream exons. The high transcriptional activity of the virus therefore likely causes expression of genes that are normally silent. As mentioned above, several of these scenarios for HTLV-1 integrations occurred in cytokines or cytokine regulatory factors that are predicted to promote cell proliferation or survival. For example, IL15 (likely induced by integration in HuT 102 cells, Fig. 1.12) is a pro-inflammatory cytokine that promotes T-cell growth and survival in vivo (174). IL6ST, also likely induced by HTLV-1 integration in HuT 102 cells, promotes cytokine signaling that may
provide a survival advantage to these cells in culture. In MJ cells, integrations such as those in the cellular genes, STAT5B, HIVEP1 and IL9R, may similarly play a role in promoting cell growth. While most of these integrations likely occurred during culture and were not germane to the genesis of the original tumor, they may have had an evolutionary benefit in culture.

It is perhaps not a coincidence that in the gammaherpesviruses, EBV and KSHV, which don’t typically integrate into the genome, we observed high expression of virus encoded immune modulatory factors in lymphomas. With all of these findings being observed in lymphomas, they may speak to the importance of hijacking immune regulatory pathways in supporting lymphoid cell growth and survival.

1.6 Acknowledgments

This study was supported by National Institutes of Health grants R01CA138268, R01AI101046 and R01AI106676 to E.K.F., F30CA177267 to M.J.S., F31CA180449 to T.O., R01CA119917 to R.R. and P20GM103518 to Prescott Deininger. Data analysis was carried out in the Tulane Cancer Center Next Generation Sequence Analysis Core using core computational resources and in the University of Florida High Performance Computing Center.
Figure 1.1  HTLV-1, EBV, KSHV and murine type C retrovirus (Mu-C-retro) detection in 50 lymphoma cell lines.

(A) Heat map shows the number of detected viral reads per million unique mapped reads (VPMM) in the 50 cell lines. Color intensity represents relative VPMM across all cell lines. (B) Histogram of VPMM for each virus in the respective virus-positive cell lines.
Figure 1.2  EBV read coverage in EBV-positive lymphoma cell lines.

The vertical axis represents the number of reads aligning to each nucleotide position. The linear EBV annotation was split between BBLF2/3 and the BGLF3.5 lytic genes instead of the terminal repeats to facilitate the analysis of coverage and splicing for the LMP2 gene. Blue bars represent lytic genes; red bars represent latent genes; green bars represent non-coding genes; aquamarine bars represent microRNAs; black bars represent non-gene features.
Figure 1.3  Analysis of transcription in the EBV BamHI W intronic and oriLyt regions.

(A) Hierarchical clustering analysis of EBV gene expression shows expression in the BamHI W intronic region and oriLyt regions more closely resembles latency gene expression than lytic gene expression. Overlapping genes were excluded from analysis due to uncertainty of read mapping. The Raji cell line was excluded from the analysis due to deletion of a number of lytic genes. The top horizontal axis shows distance between cell lines based on EBV transcriptome patterns. (B) Read coverage in the BamHI W repeat region. The vertical axis represents the number of reads aligned to each nucleotide position. Only the beginning of the repeat region is shown to better illustrate coverage. Ribo-depleted strand-specific JY RNA-seq data (JY-se for transcription in sense direction and JY-as for antisense direction) suggests the BamHI W intronic region is transcribed from the sense direction. (C) Read coverage in the right oriLyt (oriLyt-Rt) and left oriLyt (oriLyt-Lt) regions. Transcription of oriLyt regions is in the sense direction in JY cells.
Figure 1.4  Analysis of transcription in the EBV oriP-BCRF1/vIL10 gene region.

(A) Transcription level at the oriP-BCRF1 region is higher in Hs 611.T than other EBV-positive cell lines. The vertical axis represents the number of reads aligned to each nucleotide position. BBRF3 is used as a reference gene to illustrate the discordance of oriP-BCRF1 expression with other lytic gene expression. (B) Transcription in oriP-BCRF1 region in Hs 611.T cells is validated by qRT-PCR analysis. (C) Strand-specific qRT-PCR analysis shows that the oriP-BCRF1 region is predominantly transcribed in the sense direction. The primers used for analysis are indicated in (A).
Figure 1.5  Evidence of new splicing of BARTs in the BamHI A region of the EBV genome.

(A) Canonical splicing of BARTs is shown for each cell line. (B) Novel splicing events identified in each cell line. Each bar represents a spliced-out intron with color intensity (black to red) reflecting read abundance for each splicing event (total read number was shown below each bar). Only introns with more than 5 reads are shown.
Figure 1.6 Disruption of BACH2 expression by EBV integration in Raji cells occurs through a splice donor sequestration mechanism. The vertical axis represents the number of reads aligned to each nucleotide position. The top panel shows read coverage and splicing data for the wild type BACH2 gene in Namalwa and Raji cells (only canonical splicing is shown). The bottom panel shows the alignment of RNA-seq data to EBV-chr6 chimeric genome and the splicing events across the EBV-chr6 junction sites. Gel pictures show the presence of chimeric transcripts in Raji cells but not in the negative control Akata cells by RT-PCR analysis.
Figure 1.7  KSHV transcriptome analysis in BCP-1 cells.

The vertical axis represents the number of reads aligned to each nucleotide position. Coverage across entire genome is represented in the bottom panel and expanded coverage views for the left, middle, and right expressed gene clusters are shown in upper panels.
Figure 1.8 Evidence of sample cross-contamination.

Coverage and single nucleotide variation spectrum exhibits different patterns that cluster according to their file names. (A) Left panel shows the numerical ordering of file names of cell line data with total number of reads mapping to the murine type C retrovirus. Alignments are shown in the right panel with the vertical axis representing the number of reads aligned to each nucleotide position. (B) Left panel shows the numerical ordering of file names of cell line data with total number of reads mapping to the moloney murine leukemia virus. Alignments are shown in the right panel with the vertical axis representing the number of reads aligned to each nucleotide position.
Figure 1.9  Resolution of the murine retrovirus genome in DEL cells.

DEL RNA-seq data was aligned to several closely-related retrovirus genomes and the \textit{de novo} assembled DEL retrovirus genome. The vertical axis represents the number of reads aligned to each nucleotide position. Phylogenic tree (generated by Lasergene 10 MegAlign) shows the distance of sequence divergence between the DEL retrovirus and other analyzed retroviruses.
Figure 1.10  MuLV integration analysis in DEL cells.

Top panel shows coverage for chimeric read pairs aligning to the TNFRSF6B gene (their mates mapped to the MuLV genome). The vertical axis represents the number of reads aligned to each nucleotide position. Mapped reads are shown below coverage frames with pink representing rightward oriented reads and blue representing leftward oriented reads. The MuLV integration site is located in the second intron of the TNFRSF6B gene. Bottom panel shows the coverage of total reads aligned to the TNFRSF6B gene in the DEL cell line.
Figure 1.11 Resolution of HTLV-1 genomes in MJ and HuT 102 cells.

Top panel shows the alignment of MJ RNA-seq reads to the HTLV-1 reference genome and the *de novo* assembled HTLV-1 genome. Middle panel shows the alignment of HuT 102 RNA-seq reads to the HTLV-1 reference genome and the *de novo* assembled HTLV-1 genome. Phylogenetic tree in the bottom shows the distance of sequence divergence between different HTLV-1 strains.
Figure 1.12  HTLV-1 integration analysis in HuT 102 and MJ cells.

In HuT 102 and MJ cells, 14 and 66 integrations were detected respectively, with 3 and 8, respectively, occurring upstream from exons containing AUG initiation codons. Coverage of chimeric reads and total reads are shown on upper and lower tracks for the integration into the IL15 (HuT 102 cells) and the STAT5B (MJ cells) genes. The vertical axis represents the number of reads aligned to each nucleotide position. Mapped reads are shown below coverage frames with pink representing rightward oriented reads and blue representing leftward oriented reads.
Figure 1.13  The FPKM ratio of EBV lytic genes in Hs 611.T to other analyzed EBV-positive cell lines shows a higher expression level of the BCRF1/vIL10 gene compared to the overall lytic expression level in the Hs 611.T cell line.
Figure 1.14  Raji integration sites.

Top panel shows the coverage of chimeric reads aligned to EBV and the integration joints in the BGRF1/BDRF1 gene and W repeats. Bottom panel shows the coverage of chimeric reads aligned to chr6 and the integration site in intron 1 of the BACH2 gene.
CHAPTER 2. Highly edited novel non-coding RNAs transcribed from the Epstein-Barr virus latent origin of replication (oriP) during reactivation.

2.1 Abstract

We have previously shown that EBV likely encodes a significant number of viral non-coding RNAs during reactivation. Here we show that the EBV latency origin of replication (oriP) is transcribed bi-directionally and that both leftward and rightward transcripts (oriPtLs and oriPtRs) are largely localized in the nucleus. While the oriPtLs are most likely non-coding, at least some of the oriPtRs contain the BCRF1/vIL10 open reading frame. Nonetheless, the oriPtRs with long 5'UTRs may partially serve non-coding functions. Both oriPtL and oriPtR transcripts are expressed with late viral genes and their expression is inhibited by phosphonoacetic acid. Structural predictions showed thermodynamically stable large hairpins within the family of repeat (FR) regions of oriPtLs and oriPtRs. The double-stranded RNA binding protein and RNA editing enzyme, ADAR (adenosine deaminase) was found to bind to oriPtL and oriPtR and the FR regions of these transcripts were found to be “hyper-edited”. Further, the paraspeckle protein, NonO, which has a propensity for edited RNAs, was found to bind to oriP transcripts, raising the possibility that these transcripts may play a role in paraspeckles. Functional studies showed that oriPtL is required for efficient EBV production and can influence global EBV gene expression. oriPtL
may also play roles in modulating cellular environment by regulating the expression of heat shock proteins and immune-related genes.

2.2 Introduction

EBV is an oncogenic herpesvirus that infects 95% of the world’s population. It usually maintains a persistent asymptomatic infection in B-lymphocytes of immunocompetent individuals, but is associated with several types of lymphomas, carcinomas and lymphoproliferative disorders in immunocompromised individuals. EBV exhibits a biphasic infection cascade. After initial infection, it shuts down most of its genes and executes strict “latency” viral gene expression programs. Under certain conditions latent virus can reactivate, inducing the expression of hundreds of viral genes that are required for the production of infectious virions which go on to infect new cells or hosts.

In the past few years, deep sequencing of the human transcriptome has revealed a new kind of gene expression and epigenetic regulator, long non-coding RNAs (lncRNAs), which constitutes a large portion of the human transcriptome. Accumulating evidence indicates that lncRNAs are important regulators (175) in almost all the aspects of gene expression (176,177). Altered expression of lncRNAs has been associated with a number of diseases including cancer. For example, the lncRNA HOTAIR has been shown to recruit transcriptional repressors to silence tumor suppressor genes, thereby contributing to cancer metastasis (176,178).
While evidence has accumulated indicating essential roles for human lncRNAs in cell signaling and disease, virus-encoded lncRNAs have scarcely been addressed. Given the fact that EBV can encode small non-coding RNAs and utilize cellular pathways to facilitate their life cycle, it seems reasonable that EBV may also exploit lncRNA-mediated regulatory mechanisms to regulate gene expression which may contribute to viral oncogenesis. Using strand-specific RNA-seq, we have recently reported the extensive bidirectional expression of the EBV genome during reactivation (123) and raised the possibility that EBV encodes a large number of viral non-coding RNAs. Here, we show novel bidirectional transcription from the EBV latency origin of replication (oriP) and provide evidence that these transcripts may function in the nucleus and play a role in regulating gene expression and viral DNA replication.

### 2.3 Materials and Methods

**Cell culture.** EBV-positive Akata and Mutu I cells were grown in RPMI 1640 medium (Thermo Scientific, Cat. #SH30027) supplemented with 10% fetal bovine serum (FBS; Invitrogen-Gibco, Cat. #16000-069) and 0.5% Penicillin-Streptomycin (pen/strep; Invitrogen-Gibco, Cat. #15070-063). 293T cells were grown in DMEM/High Glucose medium (Thermo Scientific, Cat. #SH30243) supplemented with 10% FBS and 0.5% pen/strep. All the cells were cultured at 37°C in a humidified, 5% CO₂ incubator.

**EBV lytic cycle induction.** Akata and Mutu I cells were spun down and resuspended at a concentration of 10⁶ cells/ml in fresh RPMI 1640 medium (10%
FBS, 0.5% pen/strep). Anti-human IgG (Sigma-Aldrich, Cat. #I2136) or anti-human IgM (Sigma-Aldrich, Cat. #I0759) was added to the Akata and Mutu I cell suspensions, respectively, to a final concentration of 10 µg/ml. Treated and untreated cells were harvested 24 hours later for RNA and protein isolation.

**RNA preparation.** Total RNA was isolated using TRizol® Reagent (Life technologies, Cat. #15596-018) or a miRNeasy Mini Kit (Qiagen, Cat. #217004) following the respective vendor’s protocols. Nuclear and cytoplasmic RNA was isolated using a cytoplasmic and nuclear RNA purification kit (Norgen Biotek Corp., Cat. #21000) following the vendor’s protocol. All RNA preparations are subject to DNase treatment using the DNA-free™ Kit (Life technologies, Cat. #AM1906).

**Standard qRT-PCR and strand-specific qRT-PCR.** Standard reverse transcription was performed using an iScript™ cDNA synthesis kit (Bio-Rad, Cat. #170-8891) according to the manufacturer’s protocol. Strand-specific reverse transcription was carried out using modified gene specific primers as per the method described by Feng et al. (133). cDNA was synthesized at 65°C for 50 min using ThermoScript reverse transcriptase (Life Technologies, Cat. #12236-022) and treated with exonuclease I (NEB, Cat. # M0293L) to remove the excess gene specific primers. qPCR analysis is performed using iQ SYBR green supermix (Bio-Rad, Cat. #170-8882) on a Bio-Rad CFX96 instrument as follows: 1 µl cDNA and 1 µl 10 µM primers were mixed with 10 µl of SYBR green supermix and 8 µl nuclease free H₂O to a total of 20 µl reaction volume. Polymerase was activated and cDNA was denatured at 95 °C for 5 min. cDNA
was then amplified for 40 cycles with 15 s denaturation at 95°C, 60 s annealing/extension and plate reading at 60°C. Melting curve analysis was performed at temperatures from 60°C to 90°C with 0.5°C increment per 5 s. Expression fold changes were calculated using the comparative CT method (2^ΔΔCT).

Modified gene specific primers used for strand-specific RT:
oriPtL: ATAGCACAATGCCACCACCCTGaattaTTCAAATTTTA
oriPtR: GCCTGCTTCTTCATTCTCCTTCGTggcGCggcTAGAATAACTG
ACTB: GTACAGGCTTTTGCAGGATGTtAtaTaAACTTCATG

Primers used for qPCR:
OriPt_F: ATAGCACAATGCCACCACCCTGAACC
OriPt_R: GCCTGCTTCTTCATTCTCCTTCGT
ACTB_F: CACTCTCCAGCCTTCTTC
ACTB_R: GTACAGGCTTTTGCAGGATGT
Zta_F: TGGGCACATCTGCTTCAA
Zta_R: AATGCCGGGCAAGTTTAA
Rta_F: CGAGACCATAGTCTGGAACATC
Rta_R: CTGAACAGATGGGAGAAGGAAG
BCRF1_F: GTATGGAGCGAAGGTTAGTG
BCRF1_R: GTACACATTGGTCTGTACCTC
hIL10_F: ACCAAACCACAAGACACAG
hIL10_R: GTGCAGCTTTCTCAGAC
KCNQ1OT1_F: TACCGGATCCAGGTTTTGAGTACA
**oriPt knockdown by GapmeRs.** Customized GapmeRs targeting oriPtL were purchased from Exiqon. oriPtL and negative control GapmeRs were transfected into Akaka cells by Amaya electroporation using Nucleofector® Solution R (Lonza, Cat. #VCA-1001) and program G-16 on a Nucleofector™ II device. Briefly, Akata cells were placed in antibiotic-free RPMI media two days before electroporation. For each transfection, $2 \times 10^6$ cells in 100 µl solution R were electroporated with 3 µl negative control or oriPtL targeting GapmeRs (50 pmole/µl) and transferred to a 6-well plate with 1.5 ml media per well. Transfection was performed in triplicates for the negative control and two oriPtL targeting GapmeRs. 24 hours later, 1.5 ml fresh RPMI medium and anti-human IgG (to a final concentration of 10 µg/ml) was added to each well to induce EBV lytic reactivation. Cells were harvested for RNA isolation 24 hours post-induction. Ribo-depleted cDNA libraries were prepared and subjected to 100 bp single-end RNA-seq using Illumina TrueSeq protocol on an Illumina HiSeq 2000 instrument. Triplicate samples were multiplexed using one sequencing lane.

**RNA immunoprecipitation.** 20 million anti-human IgG treated (24 hours) Akata cells were harvested and washed once with PBS. Cell pellets were resuspended in 20 ml 1% formaldehyde/PBS solution and incubated for 10 min at room temperature on an end-to-end rotator. Crosslinking reactions were quenched with 1/20th volume of 2.5M glycine for 5 min. Cells were collected and
washed once with 20 ml cold PBS. 20 million cells were resuspended in 1 ml cold PBS and transferred to 1.5 ml tubes. Cells were spun down and lysed in 1.2 ml RIPA buffer (Santa Cruz, Cat. #SC-24948) with supplements (12 µl PMSF solution, 12 µl sodium orthovanadate solution, 12 µl PIC and 6 µl SUPERase•In™ RNase Inhibitor (Life technologies, Cat. #AM2696)). Cell suspensions were split equally into two 1.5 ml tubes and sonicated for 3×10 s with 30 s interval at 30% amplitude using a Branson Digital sonifier 250. Sonicated samples were centrifuged at 16,100 RCF for 10 min at 4 °C and pellets were discarded. Each vial of cell lysate was pre-cleared with 50 µl Dynabeads® Protein G (Life technologies, Cat. # 10004D) overnight at 4 °C on a rotator. Pre-cleared lysate pairs for each sample were pooled together and aliquoted into 1.5 ml tubes with 0.6 ml per tube. 50 µl lysate was saved as the INP sample for qPCR analysis. 5 µg of antibody-coated beads were then added to 0.6 ml of cell lysate. After 3 hours’ incubation at 4 °C, the beads were precipitated and washed five times with RIPA buffer supplemented with 0.1 U/ml SUPERase•In. To reverse crosslinking, beads were resuspended with 200 TE buffer (supplemented with 1 µl 5 M NaCl, 2.5 µl 20 mg/ml proteinase K, 1 µl SUPERase•In) and the supplemented TE buffer was also added to the INPUT sample to a volume of 200 µl. All samples were incubated at 65 °C for 2 hours. The samples were then chilled on ice and subjected to RNA isolation using Trizol. All RNA preparations were treated with DNase using the DNA-free™ Kit. qRT-PCR analysis was performed on these RNA samples.
Primary antibodies used for RNA IP: RIPAb+ p54nrb/NonO (Millipore, Cat. #03-113), rabbit anti-ADAR1 antibody (Abcam, Cat. #ab168809), rabbit IgG control antibody (Abcam, Cat. #ab37415) and mouse ANTI-FLAG® M2 antibody (Sigma, Cat. # F1804).

RNA fluorescence in situ hybridization (FISH) and Immunofluorescence (IF). RNA Stellaris FISH probes with CAL Fluor Red 610 fluorophore were designed and purchased from Biosearch Technologies. OriPt FISH was performed according to the Biosearch Technologies online protocol for suspension cells. For IF, mouse anti-p54nrb was added to the fixed and permeabilized cells at the same time as OriPt probes and incubated overnight at 37°C. Cells were washed on the following day and incubated with DAPI and Alexa Fluor® 488 goat anti-mouse IgG (H+L) secondary antibody (Life technologies, Cat. #A31619) for 1 hour in the dark. Images were taken with a 100/1.35 oil objective lens on a Leica DMRXA2 deconvolution upright microscope using Slidebook software.

RNA editing analysis. Ribob-depleted induced Akata total RNA was sequenced using Illumina TrueSeq protocol on an Illumina HiSeq 2000 instrument (123). Strand-specific RNA sequencing data was aligned to human reference genome hg19 (Genome Reference Consortium GRCH37) and Akata EBV genome (GenBank: KC207813.1) using Novoalign (Novocraft) (http://www.novocraft.com). Aligned EBV reads were extracted from the output files and separated into sense and antisense read files. Pile-ups for each file were generated using SAMtools (command options: mpileup –f) (179) and run
through VarScan v2.3.5 (180) to call the variants at each nucleotide position and to calculate the frequency of variants. Variants with frequency above 0.01 were reported. Adenosine-to-Guanine (A-to-G (I)) editing data was extracted from the output files for downstream RNA editing analysis. A-to-I variations within the repeat regions, IR1, IR2 and IR4 were omitted due to a comparatively high general variation rate (relative to A-to-I) likely due to genomic variants from one repeat to another.

**OriPt secondary structure analysis.** The secondary structure of the oriP FR region was initially analyzed using RNAalifold (181) to compute the minimum energy structure formed by a set of aligned sequences from different EBV strains. The consensus structure was then manually refined to maximize conservation and identify compensating mutations using a covariance model built by INFERNAL (182). Z-scores were calculated by comparing the normalized folding free energy of native OriPts with sets of dinucleotides randomized in silico mutants using RNAz program (https://www.tbi.univie.ac.at/~wash/RNAz/).

**Plasmid constructs and transient transfection.** oriPtL sequences spanning nucleotides 7148 to 9452 of the Akata genome (141) were amplified by PCR and inserted into the pMSCV-puro expression vector. ADAR1 (Cat. #SC119438) and HDLBP (Cat. #SC116829) cDNA clones were purchased from OriGene and their ORFs were inserted in-frame into a 3xFLAG containing pMSCV-puro expression vector. Transfection of 293T cells was performed using calcium phosphate method. Briefly, 293T cells were seeded on 10 cm plates in 10 ml DMEM/High Glucose medium (supplemented with 10% FBS and 0.5%
pen/strep) one day before transfection. On the day of transfection, the 10 ml medium was replaced with 8 ml fresh supplemented DMEM medium. For each transfection, 500 µl HBS (pH 7.2) was mixed with 10 µg constructed plasmid and 20 µg pUHD10 plasmid in one FACS tube. 30 µl 2.5 M CaCl₂ per transfection was added to each tube and then immediately mixed by vortex on a low setting. DNA/HBS/CaCl₂ mixtures were incubated for 20 min at RT. DNA/HBS/CaCl₂ complexes were then added to 293T cells dropwise and the plates were gently rocked back and forth and put in a 37°C incubator. The next day, the medium were aspirated and replaced with 10 ml fresh DMEM media to stop the transfection. Cells were harvested 48 hours later.

**oriPtL overexpression.** Stable Mutu I oriPtL-overexpressing cell lines were generated using retroviral infection followed by puromycin selection. Briefly, 293T cells were seeded on 10 cm plates in 10 ml DMEM/High Glucose medium (supplemented with 10% FBS and 0.5% pen/strep) one day before transfection. 10ug of each packaging plasmid (pVPACK-Gp-dl and pCl-VSV-G) as well as the oriPtL or control plasmid (for a total of 30ug DNA) were mixed and precipitated in NaOAc and ethanol at -20°C overnight. On the day of transfection, the 10 ml medium was replaced with 8 ml fresh supplemented DMEM medium. Plasmids were spun down and washed with 70% ethanol and resuspended with 30 µl nuclease- free H₂O. Plasmids were added to 500 µl HBS (pH 7.2) in one FACS tube per transfection. 30 µl 2.5 M CaCl₂ was added to each tube and mixed by vortex on a low setting.
DNA/HBS/CaCl$_2$ mixtures were incubated for 20 min at RT. DNA/HBS/CaCl$_2$ complexes were then added to 293T cells dropwise and the plates were gently rocked back and forth and put in a 37°C incubator. The next day, medium was aspirated and replaced with 10 ml fresh medium to stop the transfection. Virus-containing supernatant was collected 3 days post-transfection and filtered through a 0.45 µm SFCA filter to eliminate any remaining cells. Mutu I cells were spun down and resuspended in RPMI (with 10% FBS and 0.5% pen/strep) with 24 µg/ml polybrene at 2-4 million cells/ml. The resuspended Mutu I cells were then seeded on a 6-well plate with 1 ml per well. 1 ml virus particles, either control or oriPtL, were added to each well. Cells were spun down and incubated in a 37°C incubator for 4 hours. Cell suspensions were pipetted up and down and transferred to 15 ml tubes. Cells were spun down and resuspended in 4 ml fresh RPMI (with 10% FBS and 0.5% pen/strep) and transferred to a T25 flask. Two days later, 4 ml fresh RPMI (with 10% FBS and 0.5% pen/strep) was added and puromycin was added to a final concentration of 1 µg/ml to start selection. Cells were maintained in RPMI (with 10% FBS, 0.5% pen/strep and puromycin) until selection was completed. After stable oriPtL-overexpressing and control cell lines were generated, the cells were treated with 10 µg/ml anti-IgM for 24 hours to induce EBV reactivation. RNA was isolated using Trizol. Ribo-depleted cDNA libraries were made and subjected to 100 bp single-end RNA-seq using Illumina TrueSeq protocol on an Illumina HiSeq 2000 instrument.

**RNA-seq data analysis.** The FASTQ files were aligned to human reference genome hg19 and Akata EBV genome (141) using sequence and
junction aligner, TopHat (127). Aligned EBV reads were pulled out from the Tophat output. EBV reads and junctions were loaded on Integrative Genomics Viewer (IGV) (128,129) for visualization. Gene expression was quantified by SAMMate (130).

2.4 Results

Bidirectional transcription of oriP during reactivation. We recently showed extensive bidirectional transcription of the EBV genome during reactivation of the EBV-positive Akata cell line following B-cell receptor (BCR) crosslinking (123). Detailed analysis of our strand-specific RNA-seq data from induced Akata cells (24 hours) revealed extensive sense and antisense coverage across the EBV latency origin of replication (oriP) which was not observed in uninduced cells (Fig. 2.1A). To begin characterization of these transcripts, we compared the oriP read coverage in RNA-seq data derived from a polyA-selected RNA library relative to a ribo-depleted RNA library to assess the extent of polyadenylation of these transcripts (Fig. 2.1A). Most of the oriP sense transcription was slightly enriched in the polyA-selected sample versus the ribo-depleted sample, indicating that most of the rightward transcripts (oriPtRs) are likely polyadenylated (possibly using the BCRF1 polyA site) (Fig. 2.1A). Antisense transcription was clearly detected across the oriP and EBER regions in the polyA-selected sample, indicating the presence of polyadenylated leftward transcripts (oriPtLs) (Fig. 2.1A). Nonetheless, there was also a modest
enrichment in coverage in the ribo-depleted data suggesting the presence of non-polyadenylated oriPtLs.

**Potential non-coding functions of oriP transcripts.** To assess the coding potential of these transcripts, we used an online coding potential calculator (183). As expected, the BCRF1/vIL10 open reading frame was detected for oriPtRs with 35 homologues identified in UniProtKB protein database (Fig. 2.6). As such, oriPtRs were classified as protein coding transcripts. Previous studies reported several BCRF1 isoforms derived from multiple promoters during reactivation (184-187). While the classic BCRF1 transcript is around 0.8 kb and is initiated from the most proximal promoter, three other isoforms containing very long 5' UTRs were also reported (1.6 kb, 3.5 kb and 4.5 kb). Although these long transcripts contain BCRF1 ORF, it is unclear whether they can efficiently produce BCRF1 protein or perform non-coding functions but contain the BCRF1 ORF by virtue of using the BCRF1 polyA site. For oriPtLs, no significantly long reading frames were detected and no homologous protein sequences were identified for the short reading frames (Fig. 2.6). These transcripts were, therefore, classified as non-coding.

**oriP transcripts are “late”.** To investigate the stage of the viral replication cascade where the oriP transcripts may function, we assessed their expression kinetics. Strand-specific qRT-PCR was used to analyze the expression of oriPtLs and oriPtRs at 0 min, 5 min, 30 min, 1 h, 2 h, 4 h, 8 h, 24 h, 48 h and 72 h after BCR crosslinking in Akata cells. Neither oriPtLs nor oriPtRs were detected at 0 min, but notable increases were detected at 8 h, and peaks were observed at 24
h (Fig. 2.1B). This kinetics pattern closely mimics that of the late genes VCA and gp350 (Fig. 2.1B). In contrast, the immediate early lytic gene, Zta peaked at 4 h and the early lytic genes, Rta and BMLF1 peaked near or at 8 h (Fig. 2.1B). This analysis indicated that the oriPts are likely “late” transcripts.

Since the expression of most late lytic genes is dependent on viral DNA replication, we tested whether the expression of the oriPts was inhibited by the viral polymerase inhibitor phosphonoacetic acid (PAA). Akata cells were treated with anti-IgG to activate BCR signaling in the absence or presence of PAA and oriPt expression was analyzed using strand-specific qRT-PCR. Compared to the immediate early lytic gene Zta, oriPtLs and oriPtRs were found to be more sensitive to PAA treatment, indicating that they are likely true late transcripts (Fig. 2.1C).

**Nuclear localization of oriP transcripts.** To determine the cellular compartments where the oriP transcripts may function, we investigated their spatial distribution. First, nuclear and cytoplasmic RNA fractions were isolated from induced Akata cells and the expression levels of oriPts were analyzed by strand-specific qRT-PCR analysis. oriPtLs showed a much higher nuclear-to-cytoplasmic ratio than the cytoplasmic ACTB transcript (Fig. 2.1D). Further, this ratio is comparable to the nuclear-to-cytoplasmic ratios observed with the nuclear-localized non-coding RNA KCNQ1OT1 (188) and ANRIL (189). Interestingly, using primers for oriPtRs which were located in the oriP region similarly showed a relatively significant nuclear-to-cytoplasmic ratio (Fig. 2.1D) while using primers designed to assess rightward transcription in the BCRF1
region showed lower nuclear-to-cytoplasmic ratios (although these ratios were still higher than that of ACTB) (Fig. 2.1D). This suggests that oriPtLs are primarily localized in the nucleus. The intermediate results observed using primers in the BCRF1 region suggest a mix of distribution patterns, possibly with transcripts initiated from the proximal promoter being localized in the cytoplasm but longer transcripts extending to oriP region being localized in the nucleus.

We further investigated the subcellular distribution of oriPts by fluorescence in situ hybridization (FISH). Unfortunately, we were unable to reliably detect oriPtRs using the customized probe set. oriPtL fluorescence signals were detected in only a few cells in uninduced Akata cells but in a significantly higher fraction of anti-IgG treated Akata cells. Consistent with the qRT-PCR based cell fractionation results, oriPtLs were found to be localized exclusively in the nucleus with fluorescence signals often detected in replication compartment-like regions (Fig. 2.1E) (190,191).

Based on the qRT-PCR and FISH results, the oriPtLs are primarily localized in the nucleus and we propose that a least a substantial fraction of oriPtRs are similarly localized in the nucleus. This suggests that these transcripts may play non-coding RNA functions in human and/or viral genome replication and transcription during EBV reactivation.

**Stable hairpin structures in FR region of oriPts.** Unlike coding transcripts where their functions can be investigated from the standpoint of the predicted amino acid sequences of the reading frames, predicting functions from the primary sequences of non-coding transcripts is more challenging.
Nevertheless, non-coding transcripts often have a propensity to fold into thermodynamically stable and conserved secondary or higher-order structures that can provide clues about their functions. Previous studies showed a variety of RNA structures, such as hairpins, stem-loops, pseudoknots and G-quadruplexes, that are involved in almost every stage of gene regulation (192). In a previous study, Moss et al. (115) predicted several regions of EBV genome that are likely to produce functional RNAs by using computational modeling combined with RNA-seq analysis of latent EBV-infected cells. We similarly utilized this pipeline to search for the stable structures in oriP transcripts. Interestingly, predicted giant hairpin structures were mapped to the sense and the antisense sequences in FR region (Fig. 2.2). The homologous sequences were found in several other EBV strains and the hairpin structures are similarly conserved. The hairpin (-255.1 kcal/mol) in the sense direction is predicted to be more thermodynamically stable than the hairpin (-222.3 kcal/mol) in the antisense direction based on folding free energy scores. To control for sequence bias, we compared the normalized folding free energy of native oriP hairpins with sets of randomized dinucleotide in silico mutants by Z-scores calculated by the RNAz program. Both the sense and antisense hairpins were found to be about 4 standard deviations more stable than random sequences, indicating a high probability of forming this hairpin structure in the cells. To investigate hairpin conservation in related viruses, we built a covariance model for the EBV oriP hairpins and searched for this motif in other herpesviruses. One match was found in Macacine herpesvirus 4 (MHV4). The average APSI for alignment is 70.2% for MHV4 compared with other EBV
strains. The conservation of the hairpin structure in this relatively distant virus indicates that oriPts may have important functions in virus infection cycle.

**Extensive RNA editing of oriP transcripts.** To globally assess A-to-I RNA editing along the EBV genome for both sense and antisense transcription, we aligned induced Akata strand-specific RNA-seq data to EBV genome (GenBank accession No. KC207813) and A-to-I (read as G in sequencing data) alterations across the EBV genome were quantified using VarScan (193). We then calculated the editing frequency at each nucleotide position for each strand. A frequency of 1% A-to-G changes was used as a cutoff for reporting. Repeat regions (IR1, IR2, IR3 and IR4) in EBV genome were excluded from the editing analysis because the difficulty to tell true editing events from the mismatches due to higher genomic heterogeneity in repeat regions (probably resulting from replication errors called slipped-strand mispairing).

Using this approach, several editing clusters were observed across the EBV genome (Fig. 2.3A). The most substantial RNA editing was observed at the oriP region. Both oriPtLs and oriPtRs were hyper-edited, with 69.3% and 67.5% of adenosines in the FR region being edited, respectively. The frequency of editing for each adenosine varied greatly, ranging from 1.3% to 76.5% in oriPtRs and 1.1% to 67.8% in oriPtLs.

In addition to the editing of oriP transcripts, two of the editing clusters correspond to the viral microRNA clusters, BamHI A microRNA and the BHRF1 microRNA regions (Fig. 2.3C). Most of the editing events in the microRNA cluster regions were observed on the forward strand in the same orientation as the
microRNAs. Nevertheless, small numbers of editing events were also observed in the antisense orientation in the left-most cluster of the BamHI A microRNA region (Fig. 2.3C, left panel) which occurs in leftward transcripts expressed during reactivation (Strong et al., unpublished).

**oriP transcripts are associated with the RNA editing enzyme, ADAR1.** To further address ADAR-mediated RNA editing of oriPtTs, we assessed whether ADAR1 is bound to these transcripts in the cells. First, we co-transfected a control or a FLAG-tagged ADAR1 expression vector with an oriPtL expression vector into 293T cells. ADAR1 complexes were pulled down using an anti-FLAG antibody and RNA was isolated from transduced cells. qRT-PCR using three different primer pairs across the oriP region showed enrichment of oriPtL in FLAG-ADAR1 transfections but not in control transfections. Enrichment was not observed in transfections with an expression vector for another FLAG-tagged RNA-binding protein, HDLBP, further indicating a specific association between ADAR1 and oriPtL. Notably, this experiment shows the binding of ADAR1 to oriPtL in the absence of oriPtR transcripts, suggesting that the binding of ADAR1 is likely due to intramolecular duplex structures rather than heteroduplexes formed between oriPtR and oriPtL transcripts.

To test whether endogenous ADAR1 can bind to OriPtTs during EBV reactivation, we immune-precipitated ADAR1 using an ADAR1-specific antibody in induced Akata cells. oriPtTs were enriched in ADAR1-associated RNA fractions relative to the control antibody fractions, demonstrating the endogenous association of expressed ADAR1 with oriP transcripts during EBV reactivation.
OriPts are associated with paraspeckle proteins. Previous studies have shown that hyper-edited transcripts can be recognized by NonO (p54NRB), a RNA binding protein with high affinity and specificity to inosine-containing RNAs. NonO is localized in the nucleus and forms a complex with the splicing factor PSF and the inner nuclear matrix structural protein matrin 3. Binding of NonO to hyper-edited RNAs leads to the nuclear retention of these RNAs (194).

Gene expression analysis in Akata cells showed that both the non-coding RNA NEAT1 and protein components NonO and SFPQ are abundantly expressed throughout the time-course induction (not shown), indicating the presence of paraspeckle complexes during EBV reactivation. Since oriPts are hyper-edited RNAs that are predominantly localized in the nucleus, we investigated whether OriPts are associated with paraspeckles. NonO immunoprecipitation was performed in induced Akata cells and RNA fractions were isolated and then tested for the presence of oriPts by qRT-PCR. oriPts were found to be enriched in NonO pull-down relative to control (IgG) pull-down fractions, indicating the association between oriPts and NonO and a possible association of oriPts with paraspeckle compartments in the cells.

Functional analysis of oriPts. To investigate the role of oriPtLs in EBV lytic cascade and host-virus interactions, we stably overexpressed oriPtL in the Burkitt’s lymphoma cell line, Mutu I (EBV latency type I cell line), using a pMSCV-oriPtL retroviral expression vector. Mutu I cell line is another model system for assessing BCR-mediated viral reactivation and was used here because of the relative ease of retroviral transduction in these cells (compared to Akata cells).
To study the functions of oriPtL during EBV reactivation, the stable oriPtL-overexpressing and control cells were either left untreated or treated with anti-IgM to induce BCR signaling. RNA-seq based gene expression analysis showed oriPtL overexpression globally enhances EBV gene expression, indicating oriPtL is involved in EBV self-regulation. In addition, oriPtL overexpression decreases the mRNA levels of heat shock protein 70 (Hsp70) family, such as HSPA8, HSPA1A and HSPA1B, in both uninduced and induced Mutu I cells. The consistent decrease of stress-induced heat shock proteins during oriPtL overexpression suggests a role for oriP in combatting cellular stress response to virus reactivation. This analysis also showed that transcription of the MET oncogene was also decreased, which raised the possibility that oriPtL may be involved in oncogenic pathways.

To test the hypothesis that oriP may regulate virus production, we knocked down oriPtL in reactivated Akata cells using GapmeR antisense oligos. Cells were harvested at either 24 hours or 96 hours post-induction and EBV genomic DNA amount were quantified by qPCR (Fig. 2.7B). EBV DNA amount was significantly decreased at 96 hours post-induction in oriPtL knockdown cells, suggesting that knockdown of oriPtL inhibits virus production. Combined with the result that oriPtL overexpression inhibited Hsp70 family expression, it is likely that oriPtL may promote virus production through suppressing ER stress.

We also performed RNA-seq analysis on the control and oriPtL knockdown samples during EBV reactivation. Initial data analysis suggested that oriPtL knockdown has a global inhibition effect on EBV lytic genes (Fig. 2.8), which is
consistent with oriPtL overexpression results. oriPtL knockdown also affected the expression of several immune-related cellular genes. For example, FCRL proteins, which are known as the B-cell immunomodulators and are shown to inhibit activation of B-cells (195), were unregulated in oriPtL knockdown cells during EBV reactivation (Fig. 2.9). The antiviral chemokine CCL3 (196) was also increased after oriPtL knockdown (Fig. 2.9). These results indicate that oriPtL may protect EBV from host immune defense during reactivation.

Taken together, oriPtLs may promote virus production and modulate the cellular environment by suppressing ER stress response and immune surveillance during EBV reactivation. In addition, oriPtLs may enhance virus gene expression during reactivation to facilitate EBV infection cycle.

2.5 Discussion

oriPt secondary structure. Previous studies on several IncRNAs (XIST, antisense UCHL1 and TERRA) (197-199) showed IncRNA can function, at least in part through repeat sequences. The functional roles of repeats are likely derived from their secondary structures, which can be recognized by RNA binding proteins. The predicted stable hairpin structures in the oriP transcripts, which are formed in the repeat region (FR), may similarly serve as protein binding platforms. One RNA binding protein that we considered as a candidate is the EBV latency replication protein EBNA1, which binds to the FR and DS elements within oriP. In addition to its ability to bind DNA, EBNA1 has also been shown to bind to a relatively broad spectrum of RNA with a higher affinity for the
GU-rich targets. Because the OriPt hairpin structure is also GU rich, we tested whether EBNA1 could bind to either the oriPtLs or oriPtRs. However, immunoprecipitation of FLAG-tagged EBNA1 from pMSCV-oriPtL transient transduced 293T cells showed no enrichment of oriPts. Nonetheless, it is still possible that EBNA1 and oriPts may interact endogenously and the hypothesis awaits further investigation using efficient available EBNA1 antibodies.

**RNA editing in EBV miRNA and EBERs.** RNA editing is a kind of common post-transcriptional modification in the human transcriptome, which elicits diversification of transcript coding capacity and/or function. A few studies also showed evidence of editing of viral RNAs. For example, Kumar *et al.* (200) showed that extensive A-to-I editing of early-strand mRNA in polyoma virus is induced by late-strand transcripts, causing the nucleus retention of early-strand mRNA. Another study reported A-to-I editing of the KSHV K12/Kaposin RNA eliminates the transforming activity of K12 transcript (201). Editing of the EBV-encoded BART6 miRNA primary transcript suppresses the processing of miR-BART6 and counteracts miR-BART6 function in controlling latency (202).

We found that the most extensive RNA editing in EBV lytic transcriptome is in the FR portion of oriPts. More than 60% adenosines in the FR portion are edited with varying frequency up to more than 70%. Further investigation into the edited sites revealed the preference of ADAR targeting. It is found that uracils are the most frequent 5’ and 3’ neighbors of edited adenosines. This observation is different from that of the EBV miRNA primary transcripts, which indicates that ADAR may use different targeting rules for small non-coding RNAs and long non-
coding RNAs. The most common tri-nucleotide for editing is UAU for oriPtRs and UAG for oriPtLs. Because UAG is a stop codon and A-to-I editing converts it to UGG coding for tryptophan, we wonder if the elimination of these stop codons would change the coding potential of the oriPtLs. We evaluated the coding potential of the edited oriPtLs using a coding potential calculator and the result showed that editing has no effect on the non-coding property of oriP transcripts. However, editing may affect the secondary structure of oriPts by changing the canonical Watson-Crick A-U base pairs to G-U wobbles, which may alter oriPt stability or functions. A relatively low level of editing was found in the DS region, in which 41.4% of the adenosines in oriPtRs and 8.1% in oriPtLs were edited, with the frequency varying from 1.16% to 16.09%. Besides, 5 adenosines in oriPtRs near 3’ end of FR were edited at a very low frequency below 2%. These low efficiently edited sites may indicate the formation of low affinity dsRNA structures for ADAR binding.

We also investigated the edited sites in EBV miRNA primary transcripts and found 32 edited adenosines within mature viral miRNA sequences and their complementary sequences. 20 out of these 32 sites were located in miRNA seed regions (2~8 nt at the beginning of miRNA 5’ end). The frequency of editing ranges from 1.08% to 56.57%. We observed a maximum number of three adenosines can be edited in a miRNA (miR-BART17-5p and miR-BART1-3p), although all the three editing events may not coexist in a single transcript. To investigate whether the editing is random or selective, we compared the 5’ neighbor for each edited adenosine and found that the frequency of 5’ neighbors
from high to low is uracil (19), adenosine (10), cytosine (2) and guanine (1) while the 3’ neighbor is guanine (19), uracil (5), adenosine (5) and cytosine (3). 20 out of 32 edited adenosines have uracils as its 5’ neighbors and guanines as its 3’ neighbors and the tri-nucleotide UAG is the most frequently edited site. This finding is consistent with a previous study by Dennis et al. (203), showing that the tri-nucleotides UAG and AAG were relatively favored for editing in Alu elements in human transcriptome. Editing of miRNA primary transcripts increases the diversity of miRNA species and their targets. Our study indicates that EBV may encode more miRNA species than we have currently known and have a much wider range of targets. Besides, editing may regulate the processing and RISC-loading of EBV miRNAs.

The small noncoding EBER RNAs are known to form stable stem-loop structures (204). Despite their secondary structures, we didn’t detect any editing events in EBER1 or EBER2. Previous study showed RNA-editing enzyme requires dsRNA substrates with at least 15-20 base pairs (205). It is likely that short stem-loop structures in EBERs are poor substrates for ADAR recognition. Surprisingly, editing was found in the opposite direction of EBER2. 15 edited sites were identified in the antisense direction in EBER2 region with frequency ranging from 1.22% to 15.79%. This may indicate that during EBV lytic reactivation, transcripts in the opposite direction of EBER2 may form hairpin structures that can be recognized by ADAR.

*oriP* transcripts are associated with *NonO*. NonO and PSF have been shown to be components of the paraspeckle, a relatively newly identified
subnuclear body localized in the interchromatin nucleoplasmic space (206). Recent studies showed that the non-coding RNA hNEAT1 is required for paraspeckle integrity and can cause nuclear retention of edited mRNAs (207,208).

Previous studies showed that NEAT1 can be induced by virus infection (209) and that NEAT1 expression levels are associated with paraspeckle formation (208). This suggests that paraspeckles may play a role in virus-associated diseases. With this in mind, we tested whether retention of high amounts of oriPts during EBV reactivation might affect the morphology and number of paraspeckles in induced Akata cells compared with uninduced cells. We performed oriPtL FISH with NonO immunofluorescence but no significant changes in paraspeckles were observed between latent and lytic conditions, indicating that oriPt retention has no effect on paraspeckle assembly.

Paraspeckles have been shown to play a role in regulating gene expression by RNA nuclear retention. They can temporally sequester mRNA with hyper-edited 3'UTRs and release the cleaved protein-coding RNA for translation under stress conditions (210). At least some of the oriPtRs contain the BCRF1 ORF and a long hyper-edited 5'UTR. It is possible BCRF1 expression may be also regulated through this retention-release mechanism whereby stress induces cleavage of the long 5' UTR allowing translation of the BCRF1 reading frame. For oriPtLs, probably the most direct effect caused by NonO binding is sequestration in the nucleus.
The function of oriPtLs. Based on gene expression analysis in oriPtL overexpression and knockdown samples, oriPtL is likely to play roles in both EBV self-regulation and host gene regulation. oriPtL knockdown globally inhibits EBV gene expression, which is consistent with the result that oriPtL overexpression globally enhances EBV gene expression. It is also observed that EBV production is decreased after oriPtL knockdown, indicating oriPtL may facilitate EBV production. One hypothesis that can explain these results is that oriPtL may indirectly affect the overall EBV transcription through regulating the EBV genome copy number inside the cells during lytic reactivation. For cellular genes, oriPtL overexpression inhibits stress-induced Hsp70 family. And several immune surveillance related genes are unregulated after oriPtL knockdown. These results indicate that oriPtL can modulate cellular environment during EBV reactivation by dampening ER stress and immune defense, which may indirectly contribute to EBV production. Our study suggests that EBV can not only interact with host immune signaling pathways via EBV protein coding genes, but may also use long-noncoding RNAs to modulate host immune responses during reactivation.

2.6 Acknowledgments

This study was supported by National Institutes of Health grants R01CA138268, R01AI101046 and R01AI106676 to E.K.F., F30CA177267 to M.J.S., F31CA180449 to T.O., and P20GM103518 to Prescott Deininger.
Figure 2.1  oriP transcripts.

(A) Transcription in oriP region. Strand-specific RNA-seq coverage data for libraries generated from polyA-selected RNA and ribo-depleted RNA. Sense transcription is presented in blue and antisense in orange. (B) Time course analysis of selected EBV lytic genes and oriP transcripts (oriPts) during lytic reactivation by qRT-PCR at 5 min, 10 min, 30 min, 1 h, 2 h, 4 h, 8 h, 24 h, 48 h and 72 h. Strand-specific qRT-PCR was performed for oriPtRs and oriPtLs. (C) oriPtRs and oriPtLs are more sensitive to PAA treatment than the immediate early/early transcript, Zta, indicating their transcription depends on EBV replication. (D) The relative expression levels of oriPtRs and oriPtLs in nucleus and cytoplasm by strand-specific qRT-PCR analysis. ACTB is used as a cytoplasmic gene control and KCNQ1OT1 and ANRIL are used as nuclear gene controls. (E) FISH showing the subcellular localization of oriPtLs.
Figure 2.2 oriPt secondary structures.

(A) Covariance hairpin structures were predicted in the FR region of oriP across 5 EBV strains (Akata, EBV1, EBV2, Mutu and GD1) and Macacine herpesvirus 4 (MHV4). The average pairwise sequence identity (APSI) was calculated for the regions where the hairpin structures were identified in the 6 virus strains. The Z-scores were calculated for the hairpin structures in sense (Fwd) and antisense (Rev) transcripts in each virus. (B) The hairpin structure predicted in oriPtR and the folding free energy. The edited adenosines and the editing frequencies in the hairpin are shown. (C) The hairpin structure predicted in oriPtL and the folding free energy. The edited adenosines and the editing frequencies in the hairpin are shown.
## Table A

<table>
<thead>
<tr>
<th>EBV genome</th>
<th>Coordinates</th>
<th>APSI</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Akata</td>
<td>B95-8</td>
</tr>
<tr>
<td>Akata (K207813.1)</td>
<td>7458-7927</td>
<td>1</td>
<td>0.75</td>
</tr>
<tr>
<td>B95-8 (NC_007605.1)</td>
<td>7492-8148</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>AG872 (NC_008333.1)</td>
<td>7455-7915</td>
<td>0.82</td>
<td>0.83</td>
</tr>
<tr>
<td>Mutu (K207814.1)</td>
<td>7545-8014</td>
<td>0.76</td>
<td>0.88</td>
</tr>
<tr>
<td>GD1 (AY961628.3)</td>
<td>7363-8087</td>
<td>0.74</td>
<td>0.88</td>
</tr>
<tr>
<td>MHV4 (NC_008146.1)</td>
<td>7512-7936</td>
<td>0.68</td>
<td>0.73</td>
</tr>
</tbody>
</table>

## Diagram B

- **FWD (OriPφR)**
  - Un-Edited
    - \( \Delta G = -255.1 \text{ kcal/mol} \)
  - Edited

- **REV (OriPφL)**
  - Un-Edited
    - \( \Delta G = -222.3 \text{ kcal/mol} \)
  - Edited
Figure 2.3  RNA editing analysis.

(A) The frequency of RNA editing (A-to-I) for each adenosine across the whole EBV genome in Akata cells. Editing of sense transcripts was presented in green and editing of antisense transcripts was presented in red. (B) The frequency of RNA editing for each adenosine in oriPts. (C) The frequency of RNA editing for each adenosine in BARTs microRNA and BHRF1 microRNA regions.
Figure 2.4  oriPt associated proteins.

(A) Either FLAG-tagged ADAR1 or HDLBP expression vector was co-transfected with pMSCV-puro-oriPtL expression vector into 293T cells. Immunoprecipitation (IP) of FLAG-tagged ADAR1 and HDLBP showed oriPtL was enriched in ADAR1 IPs but not in HDLBP IPs. (B) Immunoprecipitation of endogenous ADAR1 in induced Akata cells. oriPts were enriched in ADAR1 IPs compared with IgG control IPs. (C) Immunoprecipitation of endogenous NonO in induced Akata cells. oriPts were enriched in NonO IPs compared with IgG control IPs.
Figure 2.5  Effect of oriPtL overexpression on cellular gene expression.

The effect of oriPtL overexpression on cellular gene expression by RNA-seq (A) and qRT-PCR analysis (B). oriPtL overexpression decreased the mRNA levels of Hsp70 family as well as MET gene.
Figure 2.6  Coding potential prediction of oriP transcripts.

Red bars with arrows indicated the oriPts with the transcription direction.
Figure 2.7 Effect of oriPtL knockdown on virus production.

oriPtL knockdown has a negative effect on virus production during lytic reactivation. (A) oriPtL was efficiently knocked down by GapmeR antisense oligos during lytic reactivation. (B) The relative EBV copy number in oriPtL knockdown cells after 24 hour or 96 hour post-induction. oriP-G1 and oriP-G4 are two different GapmeRs targeting different regions of oriPtls.
Figure 2.8 Effect of oriPlk knockdown on EBV gene expression.

Heat map showing the global inhibition effect of oriPlk knockdown on EBV gene expression during lytic reactivation. oriP-G1 and oriP-G4 are two different GapmeRs targeting different regions of oriPlks.
Figure 2.9  Effect of oriPtL knockdown on cellular gene expression. oriPtL knockdown upregulated FCRLs and CCL3Ls in reactivated Akata cells. (A) RPKMs of FCRLs in control and knockdown samples during EBV reactivation. (B) RPKMs of CCL3Ls in control and knockdown samples during EBV reactivation. oriP-G1 and oriP-G4 are two different GapmeRs targeting different regions of oriPtLs.
CHAPTER 3. Summary and discussion

It is estimated that viral infections contribute to 15-20% of the total incidence of cancer worldwide (8). Lymphoma is one of the tumor types that are closely associated with virus infections. Extensive studies have been carried out to investigate the role of viruses in lymphomagenesis based on single-gene hypothesis testing. With the aid of next generation sequencing, we are standing on a vantage point that can allow the discovery of new lymphoma-associated viruses and globally assessing virus transcriptomes and virus-host interactomes in an unbiased manner.

In our research, we took advantage of the publicly available lymphoma RNA-seq data from Cancer Cell Line Encyclopedia project. Using some bioinformatic tools, we identified virus infections in these cell lines, de novo assembled the genomes of new virus strains, investigated into virus transcriptomes and examined the viral integration status and the subsequent effects. All of these work generated an atlas of virome and virus-host interactions in these 50 lymphoma cell lines, which will be a good reference for future cell line utilization and experiment designs.

To establish persistent infection in immune cells, tumor viruses must develop strategies to evade immune surveillance. Summarized from previous and our studies, there are several ways that lymphoma-associated viruses
exploit to interact with host immune signaling pathways. The DNA viruses, EBV and KSHV, “stole” immune regulatory genes from host cells and incorporated them into their own genomes, for example, the human orthologs IL10, IL6, IL8/CCL3. Based on our study, these genes are not only expressed during lytic reactivation but also in some special cellular contexts during latency. For certain RNA viruses, they can modulate immune-related gene expression through integration mediated gene activation or disruption, as we saw in HTLV-1 and MuLV infected cell lines. Importantly, our research suggests that the interactions between viruses and hosts are not only through protein-coding genes but also via non-coding RNAs. We showed the evidence that oriPtL may play roles in modulating cellular immune responses during EBV lytic reactivation. These results underscore the importance of immune-based targeted therapy for lymphomas.

For future research, virome analysis is needed to be performed in lymphoma biopsies to determine the critical virus-host interactomes in different lymphoma subtypes for better targeting. With the aid of single-cell sequencing, we may gain some insights into the interplay between latency and lytic phases of viruses in vivo and have a clearer picture of the dynamic virus-host interactions during reactivation cascade. In addition, it is interesting to know the molecular mechanisms of how viral non-coding RNAs interact with cellular machinery and search for possible therapeutic targets for lymphoma treatment.
LIST OF REFERENCES


<table>
<thead>
<tr>
<th>Citation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>109.</td>
<td>Bhat, R.A. and Thimmappaya, B. (1983) Two small RNAs encoded by Epstein-Barr virus can functionally substitute for the virus-associated...</td>
</tr>
</tbody>
</table>


143. Jones, M.D., Foster, L., Sheedy, T. and Griffin, B.E. (1984) The EB virus genome in Daudi Burkitt's lymphoma cells has a deletion similar to that observed in a non-transforming strain (P3HR-1) of the virus. The EMBO journal, 3, 813-821.


of the National Academy of Sciences of the United States of America, 106, 11667-11672.


BIOGRAPHY

Subing Cao was born in Boxing, Shandong Province, People’s Republic of China on March 21\textsuperscript{th}, 1986. She received her Bachelor of Science degree in Biochemistry and Molecular Biology at Ocean University of China, Qingdao, in June 2008. Subing was enrolled in the Biomedical Sciences Ph.D. program in August 2009 and joined the laboratory of Dr. Erik K. Flemington in pursuit of a Ph.D.