GENOMIC AND TRANSCRIPTOMIC INVESTIGATION OF ENDEMIC BURKITT LYMPHOMA AND EPSTEIN BARR VIRUS

A Dissertation Presented by

YASIN KAYMAZ

Submitted to the Faculty of the University Of Massachusetts Graduate School Of Biomedical Sciences, Worcester in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

July 31st, 2017

GENOMIC AND TRANSCRIPTOMIC INVESTIGATION OF ENDEMIC BURKITT LYMPHOMA AND EPSTEIN BARR VIRUS

A Dissertation Presented by
YASIN KAYMAZ

The signatures of the Dissertation Defense Committee signify completion and approval as to style and content of the Dissertation

---

Jeffrey A. Bailey, MD/Ph.D., Thesis Advisor

Ann M. Moormann, Ph.D./MPH, Thesis Co-Advisor

Katherine Luzuriaga, MD, Member of Committee

Lucio Castilla, Ph.D., Member of Committee

Elinor Karlsson, Ph.D., Member of Committee

Andrew M. Evens, MD, External Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

---

Manuel Garber, Ph.D., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical

DEDICATION

I also dedicate this thesis to my family and my lovely fiancée, who have supported me tremendously throughout all these years.

ACKNOWLEDGEMENTS

This thesis would not be possible without the guidance of my thesis advisor, Jeffrey A. Bailey, and co-advisor, Ann M. Moormann, and support from my family, fiancée, friends, colleagues, and collaborators. I am indebted to Zhiping Weng for helping me with the admission process to the program and Nurdan Kasikara Pazarlioglu who is my mentor in Turkey. Throughout my time in the Bailey lab, I have had a great privilege to work with many extremely talented scientists. I cannot express my gratitude enough for the assistance from Ozkan Aydemir, Nicholas Hathaway, and Derrick DeConti in the Bailey lab. I take this opportunity to thank to members of Bailey and Moorman labs, with no special order, Alice Tran, Catherine Forconi, Cliff Oduor, David Mulama, Guang Zhou, Jennifer Moon, Joslyn Foley, Mercedeh J. Movassagh, Patrick Marsh, Priya Saikumar Lakshmi, Srinivas Chandrasekaran, Yves Falanga, and our Kenyan collaborators John M. Ong'echa, Kiprotich Chelimo, and Juliana Otieno. This dissertation is also the result of collaborations with Melissa J. Moore, Leslie J. Berg, Rachel Gerstein, David N. Wald, Ami Ashar-Patel, and Lingtao Peng. I also would like to thank Alper Kucukural, Hakan Ozadam, Alan Derr, Deniz Ozata, Eric Weiss, Yasin Bakis, and Michela Frascoli for useful discussions. I also have to thank my committee members Katherine Luzuriaga, Lucio Castilla, Andrew M. Evens, Elinor Karlsson, and especially the committee chair Manuel Garber for guiding me during my research and helping with the challenges. I was fortunate to know our department director Konstantin Zeldovich, specialists and assistants Heidi Beberman, Christine Tonevski, and Barbara Bucciaglia. Most importantly, I thank the Kenyan children and their families for their participation in the study and the staff at Jaramogi Oginga Odinga Teaching and Referral Hospital and Kenya Medical Research Institute for their assistance with this research. Finally, I would like to state my appreciation to the Turkish Ministry of National Education Graduate Study Abroad Program and various other funding sources from the Government of United States.

ABSTRACT

Endemic Burkitt lymphoma (eBL) is the most common pediatric cancer in malaria-endemic equatorial Africa and nearly always contains Epstein-Barr virus (EBV), unlike sporadic Burkitt Lymphoma (sBL) that occurs with a lower incidence in developed countries. Despite this increased burden the study of eBL has lagged. Additionally, while EBV was isolated from an African Burkitt lymphoma tumor 50 years ago, however, the impact of viral variation in oncogenesis is just beginning to be fully explored. In my thesis research, I focused on investigating molecular genetics of the endemic form of this lymphoma with a particular emphasis on the role of the virus and its variation in pathogenesis using novel sequencing and bioinformatic strategies.

First, we sought to understand pathogenesis by investigating transcriptomes using RNA sequencing (RNAseq) from 30 primary eBL tumors and compared to sBL tumors. BL tumor samples were prospectively obtained from 2009 until 2012 in Kenya. Within eBL tumors, minimal expression differences were found based on anatomical presentation site, in-hospital survival rates, and EBV genome type; suggesting that eBL tumors are homogeneous without marked subtypes. The outstanding difference detected using surrogate variable analysis was the significantly decreased expression of key genes in the immunoproteasome complex in eBL tumors carrying type 2 EBV compared to type 1 EBV. Secondly, in comparison to previously published pediatric sBL specimens, the majority of the

expression and pathway differences was related to the PTEN/PI3K/mTOR signaling pathway and was correlated most strongly with EBV status rather than the geographic designation. Moreover, the common mutations were observed significantly less frequently in eBL tumors harboring EBV type 1, with mutation frequencies similar between tumors with EBV type 2 and without EBV. In addition to the previously reported genes, we identified a set of new genes mutated in BL. Overall, these suggested that EBV, particularly EBV type 1, supports BL oncogenesis alleviating the need for certain driver mutations in the human genome.

Second, we sought to comprehensively define sequence variations of EBV across the viral genome in eBL tumor cells and normal infections, and correlate variations with clinical phenotypes and disease risk. We investigated the whole genome sequence of EBV from primary tumors (N=41) and plasma from eBL patients (N=21) as well as EBV in the blood of healthy children (N=29) within the same malaria endemic region. We conducted a genome wide association analysis study with viral genomes of healthy kids and BL kids. Furthermore, we found that the frequencies of EBV types among healthy kids were at equal levels while they were skewed in favor of type 1 (70%) among eBL kids. To pinpoint the fundamental divergence between viral genome subtypes, type 1 and type 2, we constructed phylogenetic trees comparing to all public EBV genomes. The pattern of variation defined the substructures correlated with the subtypes. This investigation not only deciphers the puzzling pathogenic differences between subtypes but also helps to understand how these two EBV types persist in the population at the same time.

Overall, this research provides insight into the molecular underpinning of eBL and the role of EBV. It further provides the groundwork and means to unravel the complexity of EBV population structure and provide insight into the viral variation that may influence oncogenesis and outcomes in eBL and other EBV-associated diseases. In addition, genomic and mutational analyses of Burkitt lymphoma tumors identify key differences based on viral content and clinical outcomes suggesting new avenues for the development of prognostic molecular biomarkers and therapeutic interventions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

COPYRIGHT INFORMATION

The chapters of this dissertation have appeared in whole or part in publications below:

Kaymaz Y, Odour CI, Yu H, Otieno JA, Ong'echa JM, Moormann AM, Bailey JA, "Comprehensive Transcriptome and Mutational Profiling of Endemic Burkitt Lymphoma Reveals EBV Type-specific Differences", Molecular Cancer Research, January, [doi]

Ashar A, Kaymaz Y, Rajakumar A, Bailey JA, Karumanchi SA, Moore MJ, "FLT1 and transcriptome-wide polyadenylation site (PAS) analysis in preeclampsia", (Submitted to Scientific Reports, in-review).

Odour CI, Movassagh M, Kaymaz Y, Chelimo K, Otieno J, Ong'echa JM, Moormann AM and Bailey JA, "Human and Epstein-Barr virus miRNA profiling as predictive biomarkers for Endemic Burkitt lymphoma", Frontiers Microbiology, March Vol 8, [doi]

Kaymaz Y, Odour CI, Aydemir O, Moormann AM, Bailey JA, "EBV genome sequence variants and associations with endemic Burkitt lymphoma", (Drafted).

Disclaim

Fine needle aspirates (FNA) were prospectively obtained between 2009 and 2012 at the time of diagnosis and before commencing chemotherapy at Jaramogi Oginga Odinga Teaching and Referral Hospital (JOOTRH), a regional referral hospital for pediatric cancer in western Kenya. Written informed consent was obtained from a parent or legal guardian of the child before enrollment. Ethical approval was obtained from the Institutional Review Board at the University of Massachusetts Medical School and the Scientific and Ethics Review Unit at the Kenya Medical Research Institute.

Chapter I. Introduction

1.1 Burkitt Lymphoma

Burkitt lymphoma (BL) is a monoclonal B-cell non-Hodgkin's lymphoma (Burkitt and Denis 1961). It is composed of monomorphic, medium sized cells with basophilic cytoplasm and one of the highest proliferation rates known for human tumors (Armitage and Weisenburger 1998). Its histological appearance is “sky” like with a background of homogeneous tumor cells punctuated by “stars” consisting of macrophages at apoptotic foci (see **Figure 1.1**). BL is characterized by overexpression of the MYC gene, in the vast majority of cases, due to a chromosomal translocation (Hecht and Aster 2000; I. T. Magrath 1991). BL tumor cells usually express GC centroblast markers such as CD10, CD77, and BCL6. Besides, BL tumors are positive for B cell surface markers CD19, CD20 and CD22 and negative for BCL2 (Ferry 2006a).

The World Health Organization recognizes three clinical subtypes of BL: endemic BL (eBL), sporadic BL (sBL), and immunodeficiency-related BL (idBL). eBL has an annual incidence of 5-15 cases in 100,000 children in areas experiencing perennial *Plasmodium falciparum* transmission. Both EBV infection and holoendemic *P. falciparum* are thought to be etiologically linked to the development of this B cell cancer (reviewed in Moormann and Bailey 2016). In contrast to eBL, pediatric sBL is found at a 10-fold lower incidence in developed countries where malaria is not endemic and only associated with EBV in around 10% to 20% of cases. Pediatric sBL tends to afflict a higher proportion of males and adolescents and present in the abdomen often with disseminated disease (Satou et al. 2015). sBL incidence has a bimodal age distribution with peaks in children and older adults suggesting different etiologies. Adults BL tends to have higher rates of EBV positivity, nodal presentation, along with poorer outcome, and often more variable pathologic features leading to designations of plasmacytoid or atypical BL (Ferry 2006a). These differences within sBL have raised the suggestion that adult sBL should be considered a separate entity (Boerma et al. 2004a) as well as EBV-positive and EBV-negative tumors (B.-J. Chen et al. 2016a).

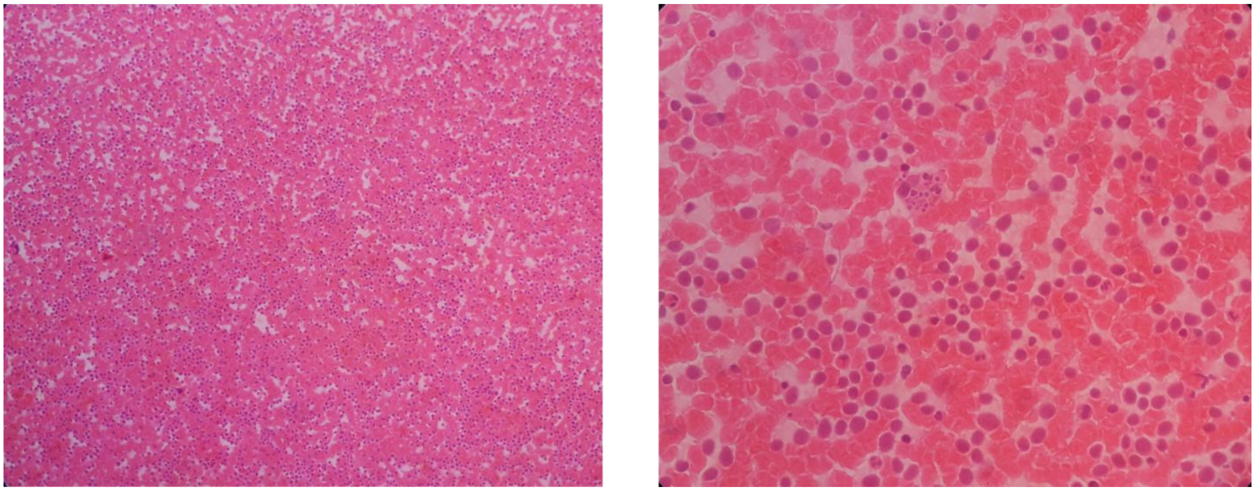


Figure 1.1. Histological appearance of Burkitt Lymphoma under microscope (provided by Dr. Sava Solomon Syanda and Dr. Juliana Otieno, JOORTH, Kenya)

eBL commonly presents in the jaw or facial bones as well as other extranodal sites such as the GI tract, kidneys, and breasts. However, jaw and abdominal tumors are the most common anatomical sites of presentation (50-80% of cases) in pediatric eBL (I. T. Magrath 1991; Buckle et al. 2016c). It has been suggested that there are different epidemiological patterns associated with the clinical presentation of BL. Endemic BL shows distinctive presentation in either the jaw or the abdomen (Mwanda 2004); among Kenyan children within our larger BL cohort, the tumor presentation sites are 43% jaw and 50% abdomen (see **Figure 1.2**) (Buckle et al. 2016a). Children with jaw tumors tend to be younger and mostly males, while abdominal BL tumors present more common among older children and are equally distributed between males and females (Ogwang et al. 2008; Asito et al. 2010a). Differences are also seen in childhood with sporadic BL (that is only associated with EBV in 30-40% of cases) where jaw involvement is rare in favor of abdominal and nodal masses (Mbulaiteye et al. 2009). Despite the observed clinical and pathologic differences, most studies over the years view eBL as a single clinical entity and attribute survival differences to delayed presentation and variability in treatments (Buckle et al. 2016c). Given the epidemiological differences associated with the site of tumor presentation that is incorporated into staging disease, there may be molecular differences underlying eBL tumor tropism that have not been fully elucidated. Clinical features of eBL and response to conventional chemotherapy have not been examined with regards to expression or mutational profile of the tumor. Also, during the study period between 2003 and 2011, 22% of the admitted patients died in-hospital, and 78% completed the course of chemotherapy treatment (Buckle et al. 2016a). In this respect, there was a dramatic difference between the survival rates with 63% of patients with jaw tumors surviving compared to 33% for abdominal tumors. Attempts to associate antibody titers with tumor presentation site and prognosis has shown that anti-Zta IgG levels were elevated in eBL patients with abdominal tumors compared to patients with jaw tumors (Asito et al. 2010b). However, high throughput expression profiling and comparative assays applied here better address the question of distinct molecular features unique to tumor localization or survival outcome.

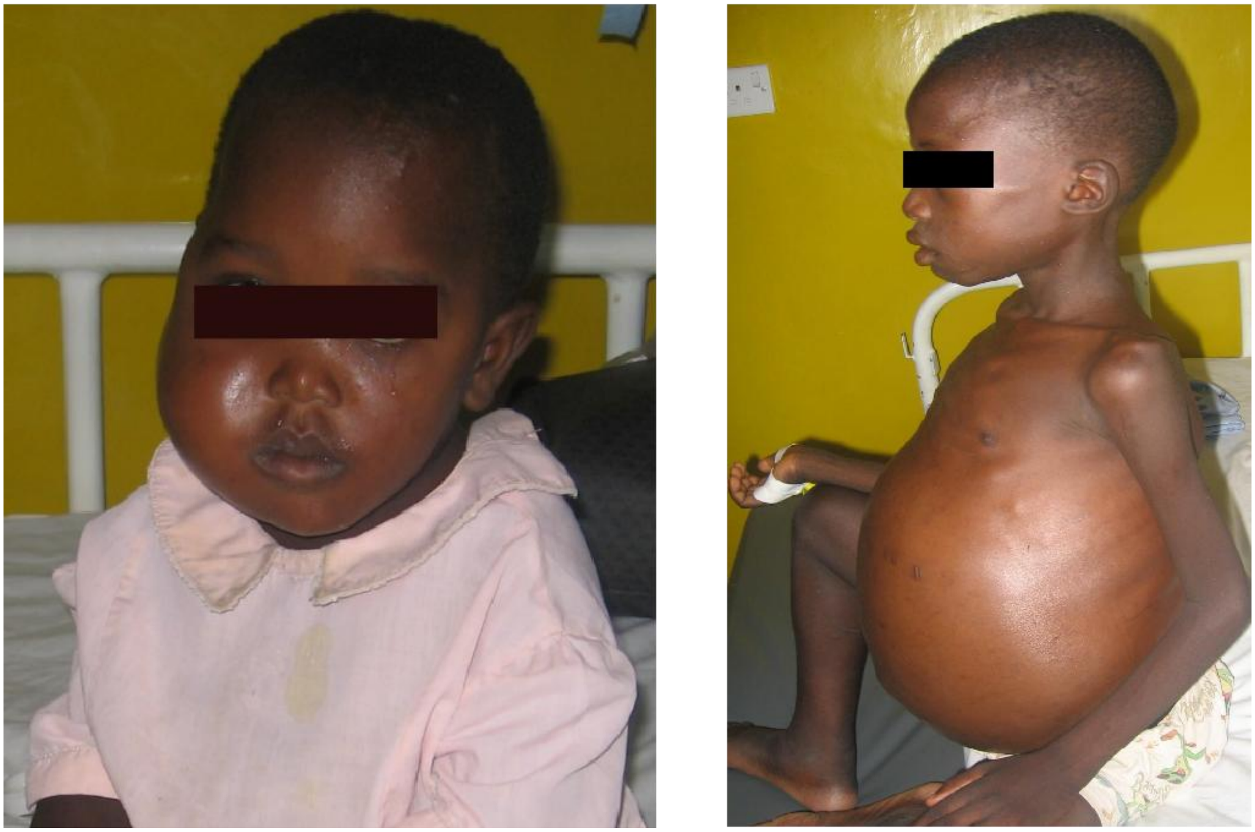


Figure 1.2. eBL patient children with a jaw presentation (on the left) and an abdominal presentation (on the right). Photos have been approved by written consent from the parents to be used for research and educational presentation (JOOTRH, Kisumu, Kenya).

Burkitt lymphoma is the first human cancer for which a translocation associated with an oncogene was identified. Virtually all BL cases share the presence of one of the three reciprocal translocations involving the *MYC* locus on chromosome 8 and either immunoglobulin heavy (IgH) chain locus (chromosome 14) or one of the light chain (kappa, lambda) loci on chromosome 2 or chromosome 22 (I. Magrath 1990). The most common one is the translocation between chromosome 8 and chromosome 14, known as t (8;14). Determining the chromosomal breakpoint locations is important because it may assist to identify in which stage of the B cell differentiation the translocation has occurred. It has been found that the distribution of breakpoint locations and the type of structural alterations differ eBL and sBL tumors (Pelicci et al. 1986). The breakpoints on chromosome 8 localized in the far upstream region of the *MYC* gene in eBLs, and in contrast, within the first exon or first intron of the *MYC* locus in sBLs. The breakpoints on chromosome 14 accumulated differently in eBLs as they mostly appeared the Ig joining region as opposed to sBLs which carried breakpoints in the switch regions. In neither case, the coding sequence is disrupted. However, the translocation of MYC locus causes deregulated expression due to being in close proximity of active Ig promoters. Thus, deregulated expression results in accelerated cell proliferation (Boxer and Dang, 2001).

Studies investigating the potential of using gene expression profiling for accurate diagnosis of Burkitt Lymphoma resulted in molecularly defining the BL (mBL, molecular BL, as named in the studies) (Hummel et al. 2006; Dave et al.

2006). By utilizing the gene expression profiles of validated BL tumors as a classifier, they identified and correctly diagnosed new tumors. This helped to distinguish BL from other aggressive B-cell lymphomas (non-mBL) such as diffuse large B-cell lymphoma. The most important feature of the classifier has reported as the level of MYC gene expression, which is the result of the translocation, to separate BL from other lymphomas. Following gene expression studies have also argued that the three subtypes of BL had distinct pathogenic mechanisms and demonstrated that eBL and idBL had similar gene expression profiles, whereas sBL was relatively more distinct (Piccaluga et al. 2011a).

MYC oncogene deregulation and ectopic expression by chromosomal translocations is the key molecular driver and hallmark of BL. Even though deregulated expression and subsequent mutations of *MYC* gene severely alter the DNA binding efficiency of this transcription factor, these do not appear to be sufficient for tumorigenesis (Janz, Potter, and Rabkin 2003). The search for additional driver mutations in sBL has yielded several candidate tumor suppressors and oncogenes (Schmitz et al. 2012a; Richter et al. 2012; C. Love et al. 2012). However, eBL primary tumor biopsies have not been studied at a genome-wide level until recently with limited numbers of cases (Abate et al. 2015a). The most common driver mutations in coding regions appear to occur in the transcription factor *TCF3* (E2A) and its inhibitor *ID3*. Cell cycle regulator gene *CCND3* which encodes for cyclin D3 is another gene frequently mutated especially in sBL cases.

Aside from geography differences in incidence, eBL has a polymicrobial etiology involving two ubiquitous pathogens, Epstein Barr Virus (EBV) and *Plasmodium falciparum* (Morrow 1985; Rochford, Cannon, and Moormann 2005). EBV asymptomatically infects more than 90% of adult population world-wide which leads to lifelong persistence (Thorley-Lawson and Gross, 2004). The greatest difference in EBV prevalence in BL tumor classifications is seen between endemic (95%) and sporadic pediatric BL (10%) tumors. EBV positivity is intermediate in id-BL (Ferry 2006b; Satou et al. 2015) and increases with age in sporadic adult cases (30-50%) (Satou et al. 2015), which is expected given that increased age correlates with higher chance of EBV exposure. Unlike other lymphomas such as Hodgkin's (Jarrett et al. 2005) and DLBCL (Park et al. 2007), EBV has not been associated with outcome (Satou et al. 2015). It does appear that EBV-positive tumors may share a similar B cell origin compared to EBV-negative tumors regardless of geographic origin (Navari et al. 2015). Adult sBL tends to have higher rates of EBV positivity, nodal presentation, along with poorer outcome, and often more variable pathologic features leading to designations of plasmacytoid or atypical-BL (Ferry 2006b). These differences within sBL have raised the suggestion that adult sBL should be considered as a separate entity (Boerma et al. 2004b) as EBV-positive and EBV-negative tumors are defined separately (B.-J. Chen et al. 2016b). Supporting this, a causal relationship between EBV positivity and age onset has been reported in a study conducted in southeastern Brazil (Hassan et al. 2008).

P. falciparum is a protozoan parasite causing malaria in >200 million people and resulting in >400,000 deaths annually (WHO 2016). Both early onset of primary EBV infection and high malaria transmission play a significant role in increased risk of BL in Africa. In locations where malaria transmission is stable and intense (i.e. holoendemic malaria), the eBL incidence is high, in contrast to a significantly lower eBL incidence in hypoendemic areas where malaria transmission is unstable and sporadic (Rainey 2005). The precise mechanisms malaria pathogen which may play in the development of this malignancy remains to be elucidated. Two possible mechanisms have been proposed; the first is by suppression of T cell immunity or by activation and expansion of B cells. It has been reported that children diagnosed with eBL were defective for EBNA1 specific IFN-gamma T cell responses (Moormann et al. 2009a). Besides, malaria parasites are powerful polyclonal stimulators of B cell system (Donati et al. 2004). Thus, exposure to a large number of several *Plasmodium falciparum* antigens during multiple infections can cause EBV to reactivate from memory B cells (Chêne et al. 2007). This leads to a higher viral load and increased EBV-infected B cells (Donati et al. 2006). It has been hypothesized that chronic and persistent malaria infection exploits immune regulatory mechanisms that influence EBV control, the high EBV viremia would, therefore, increase the likelihood of B cell transformation by latent EBV, initiating eBL carcinogenesis (Torgbor et al. 2014a). Since high EBV viremia is a characteristic in malaria holoendemic areas and thought to be crucial in the development of eBL, ineffective immune surveillance for the virus or nascent tumor by innate immune responses could increase a child's risk of developing eBL in a malaria holoendemic environment (Wilmore et al. 2015). Activation-induced cytidine deaminase (AID) is responsible for immunoglobulin hypermutation and class-switch recombination during B cell activation (Krieg 2000; Peng 2005; Ruprecht and Lanzavecchia 2006). Inducing AID activation is thought to increase the likelihood of *MYC* translocation, which is the hallmark genetic aberration of eBL tumors (Weiner 2009; Robbiani et al. 2015).

1.2 Epstein Barr Virus

Virus infections are responsible for ~15% of human cancer deaths (McLaughlin-Drubin and Münger 2009). EBV is one of these and known to be responsible for driving the proliferation and survival of infected B cells by expressing multiple viral oncogenes (Young and Murray 2003). However, EBV resides in resting memory B cell compartment of healthy individuals, and they typically carry 1-50 EBV-positive cells out of 1,000,000 B lymphocytes (Khan et al. 1996). Accumulated evidence shows that the virus does not require cell transformation or tumorigenesis for its replication and can persist without them. These are most likely consequences of complex molecular virus-host interactions. On the contrary, some of the viral genes, rather oncogenes, are required to be ubiquitously expressed for tumor cell survival. Even though EBV's role is not entirely understood, it is possible that EBV might contribute to the pathogenesis of BL by a “Hit and Run” mechanism. Supporting this, it has been demonstrated that EBV may not be associated with relapses following treatment (Xue et al. 2002).

Following the entrance to the cell, the viral genome is delivered to nucleus through an unknown mechanism. After delivery, linear genomes become circularized to protect itself from nucleases and is maintained in an episomal state, although there are rare reports of viral genomes integrated into the host genome. EBV does not have an RNA polymerase encoded by itself; thus, it uses cellular transcriptional machinery, RNA pol II. Early viral transcriptional activity is chaotic and not regulated well until the latency is established. Once the latency state is reached, virus controls most of the activity and only allows a limited number of transcriptions. In BL tumor cells, latency I is the predominant state which involves the expression of EBNA1 as the sole protein coding genes in addition to non-coding EBERs, BART region transcripts, and microRNAs (Rowe et al. 1987; Kelly, Bell, and Rickinson 2002a). On the other hand, virus shows almost no expression activity (except EBERs) in PBMCs' of healthy carriers, also known as latency 0 state. In latency II state, viral genes LMP genes are also expressed in addition to latency I genes (Price and Luftig, 2015). Expression profile of virus in cultured BL cells or LCLs in latency III have slight differences as opposed to ones in primary BL cells. This difference mainly originates from the different usage of promoter site Cp (Wp), mostly in LCLs, and Qp in BLs.

The infecting EBV genome in eBL patients can be either of two divergent strains, type 1 or type 2, and comparative genomic studies have demonstrated type-specific divergence (Cohen et al. 1989a; Rowe et al. 1989; Dambaugh et al. 1984a). While type 1 EBV is found globally, type 2 is more commonly found in Africa than other parts of the world (Zimber et al. 1986). Although it has been reported that the transformation efficiency of EBV type 1 is higher compared to type 2 in lymphoblastoid cell line establishments (Rickinson, Young, and Rowe 1987a), both strains are frequently found in African eBL cases and are prevalent within healthy populations in sub-Saharan Africa (Rowe et al. 1989; L. S. Young et al. 1987b). However, the expression and mutational profiles of EBV type 1 and type 2 within primary eBL tumors have not been compared and contrasted to determine if viral variation influences tumorigenesis.

Molecularly, type 1 and type 2 are represented by major latent genes, the coding sequences for EBNA2 and EBNA3-A/B/C genes (Dambaugh et al. 1984b). In the efficiency difference between the two subtypes in making lymphoblastoid cell lines, EBNA2 has been found to be the key determinant (Cohen et al. 1989b). Recently supporting the role of EBNA2, it is reported to be the primary factor differentiating the transformation efficiencies of two types (Lucchesi et al. 2008). Both types also differ in several molecular properties such as their entrance to lytic cycle (Buck et al. 1999) and ability to infect T cells (Coleman et al. 2015). Following their discovery about the transformation efficiency differences between subtypes, Alan Rickinson and his colleagues measured the population frequency of type 2 infections as ~23% among normal individuals in Kenya and New Guinea by generating spontaneous LCLs (L. S. Young et al. 1987a). Originating from these, it is widely known that type 1 EBV is the dominant type all around the world, type 2 is more commonly found in Africa. In contrary, a group of healthy adults in the USA was screened for EBV types using PCR and the study found almost equal levels of type 1 (41%, N=14) and type 2 (50%, N=17) in addition to individuals carrying both (9%, N=3) (Sixbey et al. 1989).

Starting from the 1990s, it has often been hypothesized that EBV genomic variations may contribute to pathogenesis. These studies were initially oriented around subtype-specific genomic regions and pathologic differences between diseases (Tzellos and Farrell, 2012). One of the focuses especially was on immunocompromised patients since the early observations showed increased type 2 infection rates among such individuals. A study conducted in Australia reported that 26 LCLs from HIV infected subjects carried 69% type 1, 19% type 2, and 12% both types (Sculley et al. 1990).

A central challenge is that that establishing LCLs for investigating viral type frequencies is inherently biased given type 1 EBV's better transformation properties. Alternatively, PCR-based methods generated relatively better and reliable results. Another follow-up study with PCR instead of LCL generation showed that 24% (N=15) of the HIV-positive patients were infected with type 1 while 27% (N=17) were with type 2 and 39% (N=24) were with both types. In addition, 39% of the cardiac transplant patients were infected with type 1 while 33% were with type 2 and 28% were with both types (Kyaw et al. 1992). The skew in the typing towards type 1 EBV when LCL method is used as opposed to PCR has been demonstrated. Boyle et al. screened 30 Hodgkin's Disease patients for EBV and found that 7 of these had type 1 and 2 of them had type 2. Interestingly, two patients with type 2 EBV immunocompromised as they were infected with HIV suggesting that type 2 EBV is important pathogen for immunocompromised individuals (Boyle et al. 1993). Supporting this, immunocompromised HIV-positive homosexuals had slightly higher type 2 EBV infection rates compared to other healthy Caucasian individuals in another work (Yao et al. 1998a). On the other hand, among the Taiwanese NPC (Nasopharyngeal Carcinoma), head and neck carcinoma, and saliva from healthy individuals, type 1 was the predominant type (Shu et al. 1992). Another study conducted with Brazilian BL cases found the majority of the tumors as type 1 EBV-carrying (93%, N=13) and 80% of the virus had a 30 bp deletion in their LMP-1 gene (W. G. Chen et al. 1996). A study conducted among healthy individuals from Japan using saliva and throat washings showed that the majority of carriers were infected with type 1 (90%) (Ikuta et al. 2000). Multiple cancer types predominantly carry only type 1 (Peh, Kim, and Poppema 2002). No associations have been found between NPC cases and various genotypes of the virus, including subtypes type 1 and type 2, relative to healthy population (Cui et al. 2011).

Overall these early attempts to measure subtype frequencies were sporadic and involved small study sizes. In addition to concerns regarding the limited samples, researchers have concluded that the generation of LCLs creates a bottleneck for types and skews the results. Alternative methods such as PCR-based assays provided relatively better estimations, however; patient/donor selection criteria relying on viral load levels of individuals also created unbalanced/non-random sampling especially towards immunocompromised people. In other words, picking only patients with high viral loads is a non-randomized sampling of populations.

The desire to determine type frequencies globally lead scientists to conduct many studies around the world. In one of these attempts from Argentina, type 1 was found in 76% of healthy carriers while type 2 was in 15% and 7% of individuals were co-infected with both types (Correa et al. 2004). In Mexico, 33% carried type 1, 57% type 2, and 10% was a mixed infection (Palma et al. 2013). The predominant type was found to be type 1 with 98% frequency in Australia (Lay et al. 2012). These various studies to understand the viral subtype prevalence and attempts to associate with diseases have yet to form a consensus profile as reviewed in Neves et al. (Neves et al. 2017). However, the prevalence of type 2 infections might still be associated with disturbance of immune system (AIDS patients) or chronic immune activation as we observe individuals in malaria-endemic regions. As a summary, the claim that the most of the populations carry type 1 more frequently than type 2 is a flawed statement because such generalization to population levels relies on studies with mostly disease associated cases not all types of individuals including healthy people.

Regarding the mixed type infections, a study conducted by Barlee et al. found no association between mixed (superinfection) or type 2 EBV infection and acquired immunodeficiency syndrome (AIDS)-related non-Hodgkin lymphoma in a study using type specific nested PCR on PBMCs of patients (van Baarle et al. 1999). The study concludes that detecting virus type directly from PBMCs is more sensitive than a cultured virus grown with sLCLs from same PBMCs. Secondly, HIV-infected individuals have high type 2 EBV infection prevalence, however; this does not increase the risk for developing AIDS-related NHL. Also, contrary to previous findings, they found no correlation between type 2 infection and immune system failure. In conclusion, these suggested that type 2 infection was only correlated with

HIV infection but not with immunodeficiency in agreement with an earlier work using PCR assays reporting that the 50% of the EBV-positive HIV-associated non-Hodgkin lymphoma patients carried type 2 (Boyle et al. 1991). Similarly, another study found that HIV-positive patients carry multiple strains both type 1 and type 2 (Yao, Tierney, Croom-Carter, Dukers, et al. 1996). Interestingly, following results have been reported with slightly higher mixed infection rates by Sculley et al as 35%, 27%, and 21%, type 1, type 2, and both, respectively (Apolloni and Sculley 1994) and as 69%, 19%, and 12%, type 1, type 2, and both, respectively (Sculley et al. 1990).

Such first generation PCR-based frequency measurements are also not reliable regarding detecting mixed infection since they can simply return false positive results because of their lack of quantitative properties. The mixed type cases were prone to false results and probably overestimated when overly sensitive PCR was used. Thus, these early studies trying to estimate mixed infection rates should be evaluated cautiously. To determine superinfection cases or accurately assess the level of mixed infections, quantitative PCR (qPCR) assays with multiplexed reactions are required. Specifically for EBV subtypes, Gatto et al. developed such test which can be utilized for better estimation of mixed infections (Gatto et al. 2011). Alternatively, a recently developed method using quantitative sequencing called molecular inversion probes (MIPs) can be utilized with unique molecular identifiers. This targeted capture PCR allows correcting for sequencing errors and properly quantitating initial levels of DNA mixtures in clinical samples.

The viral genome variant association studies involving NPC outnumbers any other diseases. Early attempts investigating sequence variations emerged with restriction site polymorphisms (aka RFLP) and majorly focused on BamHI and XhoI variants. RFLP studies showed that variants of BamHI region are more frequent in Asian strain than Europe, N America, and Africa (Khanim et al. 1996; Cho and Lee 2000). This was the first associated with NPC, but it was in fact just a geographic variant. Similarly, major LMP1 variants were named after countries that they were first found in (Chinese, Alaskan, North Carolina, etc.). LMP1 gene codon deletion has been associated with NPC (Cheung et al. 1998) in addition to the different subgroups of LMP1 sequences (Tiwawech et al. 2008). These studies involved amplicon sequencing with a low satisfactory amplification rate. Loss of XhoI site and 30bp deletion at the C-terminal have been associated with Nasopharyngeal carcinoma or increased tumorigenicity (Hu et al. 1993, 1991; Jeng et al. 1994). An SNV in the RPMS1 coding region has been found to be strongly associated with risk of NPC (Feng et al. 2015). Sequence analysis of immediate early gene BRLF1, Rta, demonstrated significantly different sequence subgroups among various patient and control groups (Jia et al. 2010). Sequence variations in EBNA1 are important for recognition of infected cells by CD8+ T-cells because it is the single protein coding gene in latency state in all EBV-associated malignancies. Thus, it is also targeted by researchers to associate its variations with MHC class I types (Bell et al. 2008). Viral subtype and RFLP polymorphism association study on EBVaGC (Corvalan et al. 2006). EBNA1 and LMP1 variant association case-control study using amplicon sequencing found no association between the variants and multiple sclerosis (Simon et al. 2011). Sandvej et al. and several others found these variants, however, in viral genomes of healthy carriers and patients with non-EBV associated diseases (Sandvej et al. 1997). Despite these various attempts to determine virulent strains and associate these with certain disease types, there is not a clear consensus on results most likely due to limited access to clinical specimens and technical challenges. Especially, such association studies regarding BL are greatly lacking.

1.3 Motivation and Research Goals

Further investigations into the mechanisms EBV-host interactions is warranted to increase understanding of EBV infection and body defense mechanisms that may facilitate the development of novel strategies for controlling EBV infection and reducing eBL carcinogenesis. Thus, a comprehensive study design that will compare gene expression patterns and genomic variations of biopsy samples from eBL and sBL using next-generation sequencing can provide a high-resolution understanding of this divergence in the context of EBV infection.

In this research, my overall goal was to investigate genomic and transcriptomic alterations as well as the known cofactors of tumor initiation and progression. I utilized the next generation sequencing in conjunction with computational techniques to address two specific aims. I proposed: **(1) to determine the genomic and transcriptomic differences between endemic Burkitt lymphoma and sporadic Burkitt lymphoma and correlate transcriptomes of endemic Burkitt lymphoma with clinical/molecular phenotypes; (2) to determine if the EBV genomic variations correlate with geography (confounded by malaria endemicity) and/or eBL diagnosis given the same malaria exposure using a case-control study design.** Results from this research provided new insight into genetic alterations that contribute to eBL etiology and helped to understand roles of pathogens in the disease development.

The hypothesis was that eBL differs from sBL regarding gene expression profiles and mutated gene distributions. Regarding the genomic and transcriptomic differences between endemic BL and sporadic BL, I used high-throughput sequencing to spontaneously measure the expression levels of multiple samples and determine whether there were genes differentially expressed and whether their expression pattern correlated with BL subtypes. I determined the mutational profile and mutated gene rate differences between the BL subtypes. I also investigated the effect of EBV's presence and type on these mutational profile differences by conducting a comparative analysis. In the second part of my research, I worked on the topic of viral genomes in tumors because I wanted to find out possible roles of sequence variations in pathogenesis through molecular interactions so that it can help to develop better therapeutic vaccines against viral infections.

Chapter II. Endemic Burkitt Lymphoma Expression and Mutations

2.1 Summary

Endemic Burkitt lymphoma (eBL) is the most common pediatric cancer in malaria-endemic equatorial Africa and nearly always contains Epstein-Barr virus (EBV), unlike sporadic Burkitt Lymphoma (sBL) that occurs with a lower incidence in developed countries. Given these differences and the variable clinical presentation and outcomes, we sought to further understand pathogenesis by investigating transcriptomes using RNA sequencing (RNAseq) from multiple primary eBL tumors compared to sBL tumors. Here we investigate the transcriptome and mutational profiles of 28 eBL and two sBL primary tumors by deep sequencing and unlike previous studies; we correlated our findings with clinical outcomes. We also explored the viral gene expression activity in EBV positive BL tumors comparing and contrasting type 1 and type 2 virus.

2.2 Methods

To better compare BL subtypes, we also analyzed published RNAseq dataset of sBLs and cell lines (Schmitz et al. 2012b). The sequences in the fastq format were downloaded through the NCBI (SRP009316) for 28 sBL primary tumors and 13 long term BL cultures derived from sporadic and endemic cases. In addition, we also analyzed 89 mRNA sequencing from lymphoblastoid cell lines (LCLs) from healthy individuals involved in the 1000 genome project (Yoruba, YRI)

(ERP001942), which we used to eliminate variant calls likely due to transcript assembly, mapping artifacts or RNA editing.

2.2.1 Sequencing Library Preparation

Briefly, starting with 1-5ug total RNA, we prepared strand-specific RNAseq libraries following the protocol from Zhang et al. (Zhang et al. 2012) combined with mRNA enrichment with oligo-dT using Dynabeads mRNA purification kit (Life Technologies) (see Chapter III for details). Final library qualities were confirmed with Bioanalyzer High Sensitivity DNA kit (Agilent) and sequenced with paired end read (2x100bp) using multiple lanes of Illumina HiSeq 2000 (Illumina Inc., San Diego, CA, USA). Data can be accessed at dbGAP with accession number (phs001282.v1.p1).

2.2.2 Differential Gene Expression Analysis

After quality assessment and preprocessing the raw sequencing reads, we aligned read pairs to a transcriptome index built by RSEM (Li, Bo, and Dewey 2011a) using Gencode v19 protein coding transcript annotations and hg19 genomic sequence. For EBV genes, we used GenBank gene annotations from both the type 1 and type 2 reference genomes (NC_007605 and NC_009334, respectively). To perform differential gene expression test, we used DESeq2 (M. I. Love, Huber, and Anders 2014a) in R computing environment. In order to be able to account for the batch variables and unknown factors while testing for the differential expression, we estimated the number of latent factors for every comparison separately using svaseq (Leek 2014a) while preserving the variation of interest. We then incorporated these surrogate variables into the testing model for DESeq2.

2.2.3 Gene Set Enrichment Analysis

We performed a standard gene set enrichment analysis (GSEA) using the GSEA module implemented by Broad Institute, Cambridge, MA (Subramanian et al. 2005). GSEA was performed on normalized expression data and data after surrogate variable analysis. For a ranking metric, we used the signal to noise value of each gene and performed a permutation test for FDR by permuting sample phenotypes. The analysis included standard gene sets of hallmark and oncogenic signatures as well as the curated C2 gene sets from the Molecular Signatures Database (v5.0 MSigDB) (Liberzon et al. 2011).

2.2.4 Single Nucleotide Variation Detection

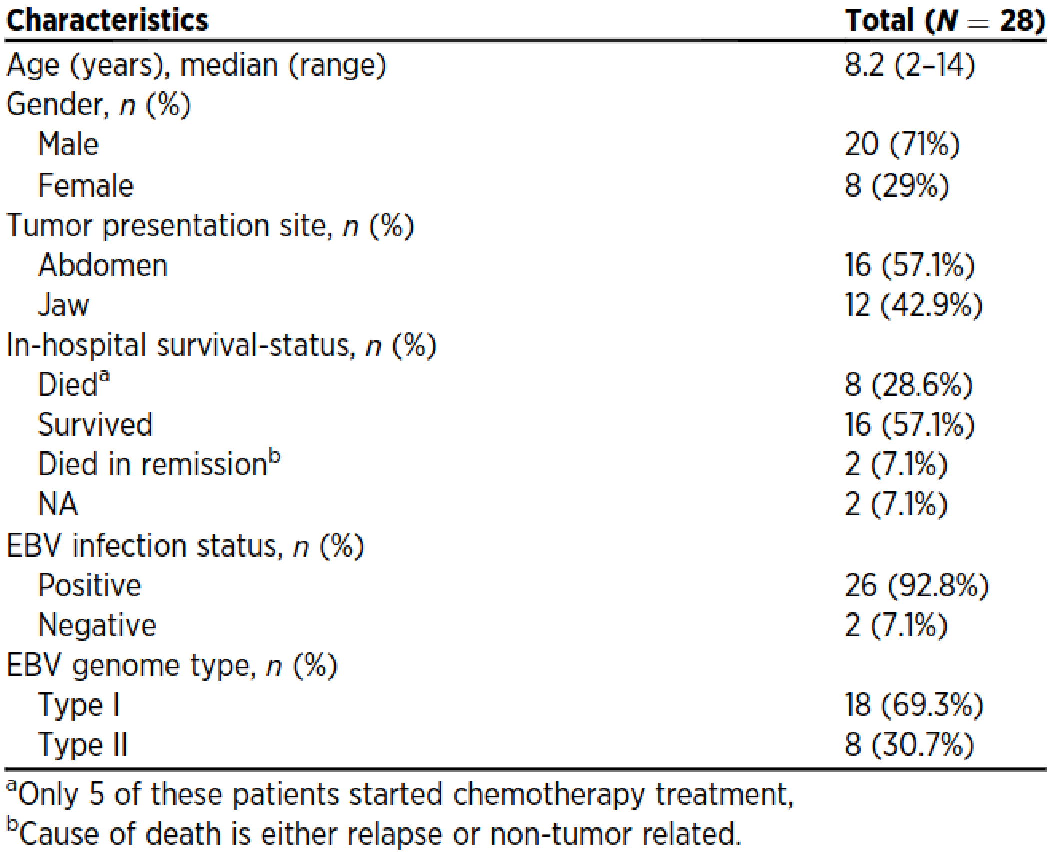
We mapped sequencing reads to human reference genome hg19 using the spliced aligner STAR (Dobin et al. 2013a) after quality trimming and removing the PCR duplicate reads. We followed the standard work flow by GATK (McKenna et al. 2010a) for calling variation within RNAseq data using the HaplotypeCaller module with additional stringency requirements (see Chapter III for details). Variants observed in dbSNP v146 and low-quality calls were excluded. We limited variant calling to translated sequences of protein coding genes in GenCode annotation v19.

2.3 Results

2.3.1 Case Information and Sequencing Summary

To survey the transcriptome of eBL, we sequenced 28 primary histologically-confirmed tumor FNA biopsies collected from Kenyan children with median age 8.2 years old (Table 2.1). We also sequenced two fresh frozen sBL tumors from diagnostic biopsies at the University of Massachusetts Medical School (UMMS). For the eBL patients, the tumor presenting site was 43% (12/28) jaw tumors and 57% (16/28) abdominal tumors. Regarding survival, the eBL samples included three patients who died before receiving any treatment, five patients who died during treatment, and 16 patients who were able to complete the recommended chemotherapy treatment with resolution of their tumor and discharged from hospital (Buckle et al. 2016b). In-hospital survival for the children included in this study was 64% (18/28). For each of the samples, we performed strand-specific RNA sequencing generating on average 14M paired reads per library (range 8.9-53.7M reads). All 30 samples in the sequencing set showed high expression of associated BL markers, including traditional cell surface markers *CD19*, *CD20*, *CD10* and *CD79A/B*, and intracellular markers of *MYC* and *BCL6*, consistent with the molecular phenotype of BL (Ferry 2006b). All samples, including re-analyzed sBLs, showed high proportions of B cell specific expression (Abbas et al. 2005) consistent with adequate aspirates of the tumor cells.

Table 2.1 Summarized clinical information for sequenced endemic BL tumors.



Characteristics	**Total ($N = 28$)**
Age (years), median (range)	8.2 (2–14)
Gender, *n* (%)	
Male	20 (71%)
Female	8 (29%)
Tumor presentation site, *n* (%)	
Abdomen	16 (57.1%)
Jaw	12 (42.9%)
In-hospital survival-status, *n* (%)	
Died[a]	8 (28.6%)
Survived	16 (57.1%)
Died in remission[b]	2 (7.1%)
NA	2 (7.1%)
EBV infection status, *n* (%)	
Positive	26 (92.8%)
Negative	2 (7.1%)
EBV genome type, *n* (%)	
Type I	18 (69.3%)
Type II	8 (30.7%)

[a] Only 5 of these patients started chemotherapy treatment,

[b] Cause of death is either relapse or non-tumor related.

2.3.2 EBV Positive eBL Tumors are Predominantly Canonical Latency I Expression Program

In concert with human transcriptome analysis, we first checked if EBV DNA was present in the tumor isolates using quantitative PCR. As expected, the vast majority of eBLs was positive (93%, 26/28) while only two eBLs and the two sBLs were negative (Lazzi et al. 1998). For EBV positive tumors, viral load assays indicated that tumor cells contained multiple copies of EBV DNA (mean 4,475 copies/ng tumor DNA; ~30 EBV/cell, median 1,542 copies/ng tumor DNA; ~10 EBV/cell). We also determined the virus type using distinguishing primers against viral gene EBNA3C. We found that 31% (N=8) of the EBV positive tumors were infected with type 2 EBV genomes whereas 69% (N=18) of them carried EBV type 1 genomes. We observed no mixed infection of both types of eBL tumors.

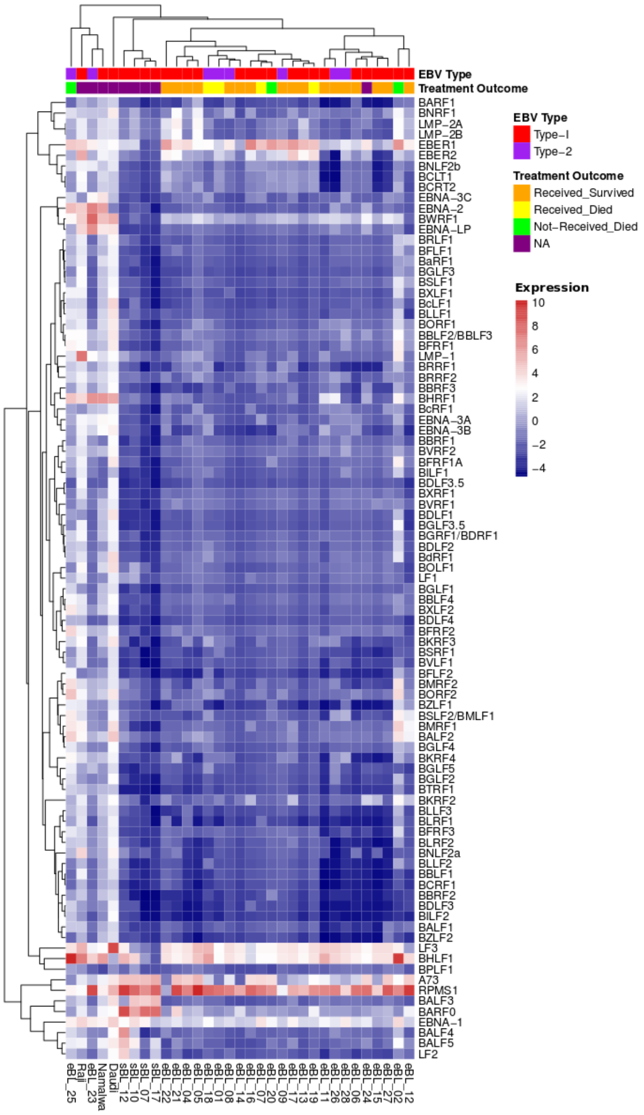


Figure 2.1 Expression heatmap for all known EBV genes for 26 eBL, four sBL and three long-term BL cultures (Daudi, Raji, Namalwa) that were found to be EBV positive. This correlation based clustering heatmap using log2 transformed FPKM values demonstrates a predominant expression pattern resembling latency I for most of the BLs while two eBLs (eBL_23 and eBL_25) and cell lines have elevated expression in other genes. eBL_02 and eBL_20 show intermediate levels lytic genes such as BMRF1, BALF2, and BSLF2/BMLF1 in addition to the two eBLs that cluster with cell lines.

Given that these two types have divergent genomic sequences for several genes, we mapped the RNAseq reads to the appropriate viral transcriptome sequences. EBV positive tumors demonstrated significant viral gene expression, regardless of viral genome type. This expression from the EBV genome ranged across a continuum with the average around 200 RPM (reads per million) (ranging from 10 to 400 RPM, cumulative viral reads per library). The two EBV negative eBL tumors and two sBL tumors did not demonstrate any EBV specific reads supporting the absence of the virus based on qPCR. Interestingly, viral DNA copy numbers did not correlate significantly with overall viral transcriptome activity levels. Along with BHRF1, BHLF1 and several EBV latent genes (EBERs, EBNA-1 and LMP2A/B) are weakly positively correlated with viral DNA levels in individual eBL tumors. On the other hand, BART transcripts, which are the most abundant transcripts including RPMS1, A73, and LF3, demonstrated no correlation with viral DNA levels. This suggests that the observed viral expression levels within tumor cells are for the most part independent of viral load or lytic replication rates and may be dependent on other factors. In addition to the eBL tumors, four of the sBL primary tumors that were re-analyzed were also EBV positive (14%) and carried type 1 EBV genomes. Overall, viral genes in eBL tumors demonstrated a predominant expression pattern consistent with the latency I. However, hierarchical clustering of viral genes revealed several potential subgroups (Figure 2.1). All sBLs clustered separately and showed latency I pattern but increased levels of BALF3 and BARF0, unlike eBLs. The cell lines and three eBL samples showed higher levels of most genes suggestive of increased viral replication and lytic activity. Among these three, two (eBL_02 and eBL_25) with elevated BHLF1 and lytic gene expression were patients who died in-hospital before receiving any treatment.

2.3.3 In-depth Assessment of Correlated Variation with Clinical Features and Viral Type

Our initial question related to the tumor transcriptome was if there were any major expression differences and if any major differences correlated with the features of anatomical tumor presentation site, in-hospital survival or EBV type. After normalization of expression, we performed unsupervised hierarchical clustering based on Pearson correlations on expressed genes with the greatest variation (Figure 2.2). The overall correlations among eBLs were extremely high ($r>0.96$, average). The sporadic tumors which differed in the biopsy collection procedure (surgical biopsy or fine-needle aspirate) and preservation methods (fresh frozen or RNAlater) still showed a high-degree of correlation ($r>0.90$) with eBLs although they distinctly clustered away from eBLs. Similarly, the major principal components showed no discernible separation based on tumor presentation site, treatment outcome, and viral genome type. Overall, this suggests that eBL tumors are a relatively homogeneous group without overt subtypes based on tumor presentation site, survival or EBV type.

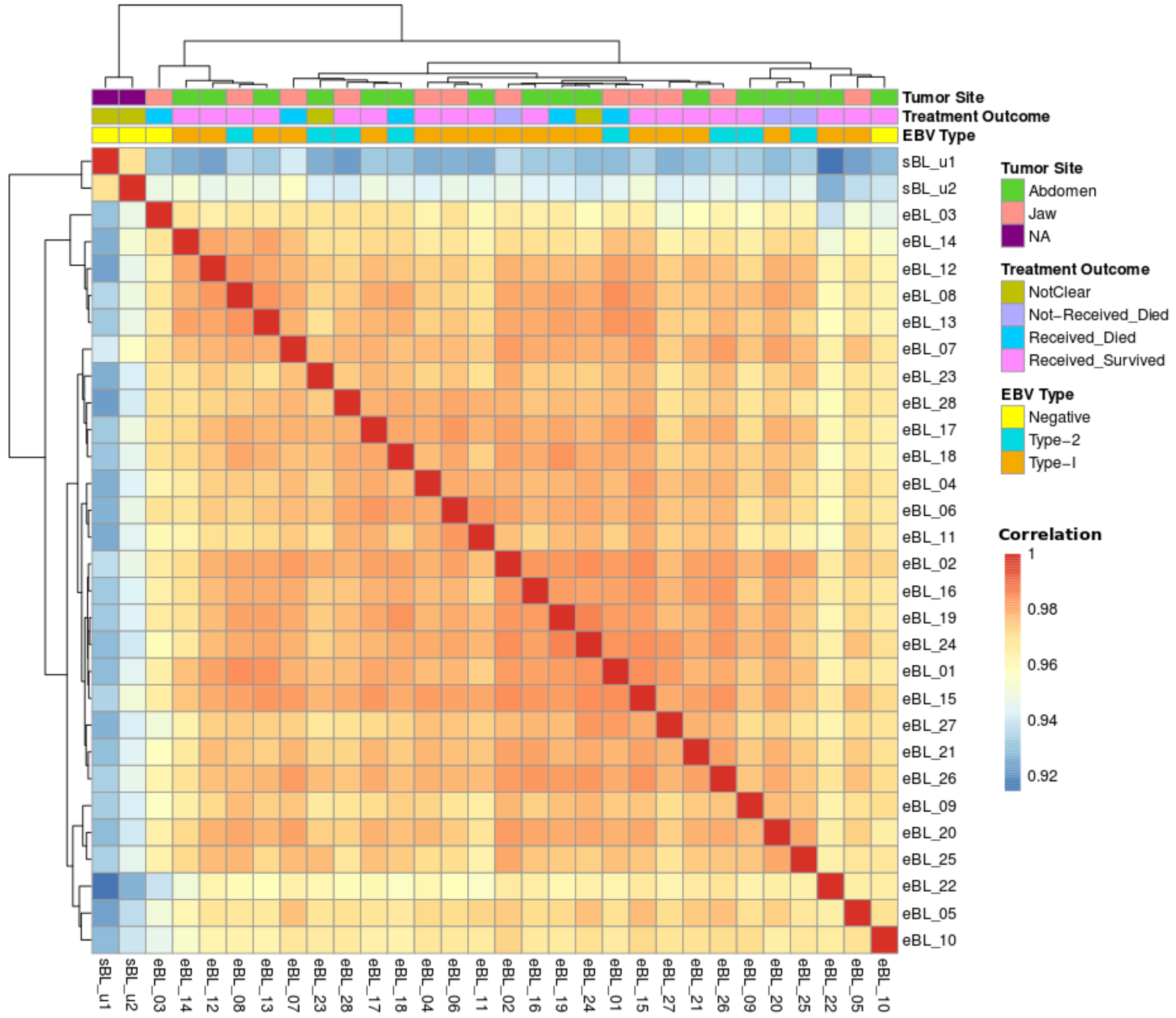


Figure 2.2 Sample to sample clustering of BL tumors based on expression profiles of top genes with the highest correlation of variation (CV) values (calculated using regularized log transformed expression data). While two sBLs (sBL_u1 and sBL_u2) separate out from 28 eBLs, eBL tumor expression profiles demonstrate greater correlation within eBLs (r > 0.95, Pearson correlation; Dark red is 1.0) compared to sBLs, which might be due to differences in biopsy and preservation methods or biology. Overall gene expression correlations between eBLs do not reveal significant clustering consistent with no major underlying molecular subtypes nor clustering correlating with tumor presentation site, treatment outcome, or EBV type.

We then checked individual genes for differential expression between eBL tumors with different clinical features. For tumor site, only *NOS3* showed significantly (~2 fold) higher expression in abdominal tumors. This gene encodes for nitric oxide synthase 3 (aka eNOS) and is known to be more highly expressed in abdominal endothelial (Teng et al. 1998). Given the molecular phenotype of eBL tumors appears relatively homogeneous, it may be that unaccounted variation, biases or stochastic noise may be obscuring the detection of true expression differences. Thus, we used surrogate variable analysis (SVA) to isolate and remove unaccounted variation while preserving the variation associated with the feature of interest (Leek 2014a). As a result, we still failed to determine any significantly differentially expressed genes or pathways between biopsies from two different clinical tumor presentation sites, jaw and abdomen. However, for the in-hospital survival of those who commenced chemotherapy, we detected ten significantly differentially expressed genes between tumors of patients who survived and those who died (Figure 2.3A). *AGPAT3*, *CTSL*, *ISCU*, *CTSD*, and *APOE* showed greater relative expression in tumors of patients who survived while *TUBB6*, *SLC25A24*, *FAM127A*, *HOMER1*, and *SLC12A3* demonstrated greater expression in tumors of patients who died. Gene set enrichment, and pathway analysis between expression profiles of these patient groups suggested several hallmark pathways including hypoxia, *IL2/STAT5* signaling, *MYC* targets, and *TNFα* signaling via *NFκβ*. The leading edge genes (the core of the enrichment signal) mutually shared by these hallmark gene sets are *SERPINE1*, *CD44*, *ENO2*, *PLAUR*, *RHOB*, and *TNFAIP3*.

These genes represent potential prognostic biomarkers requiring further investigation.

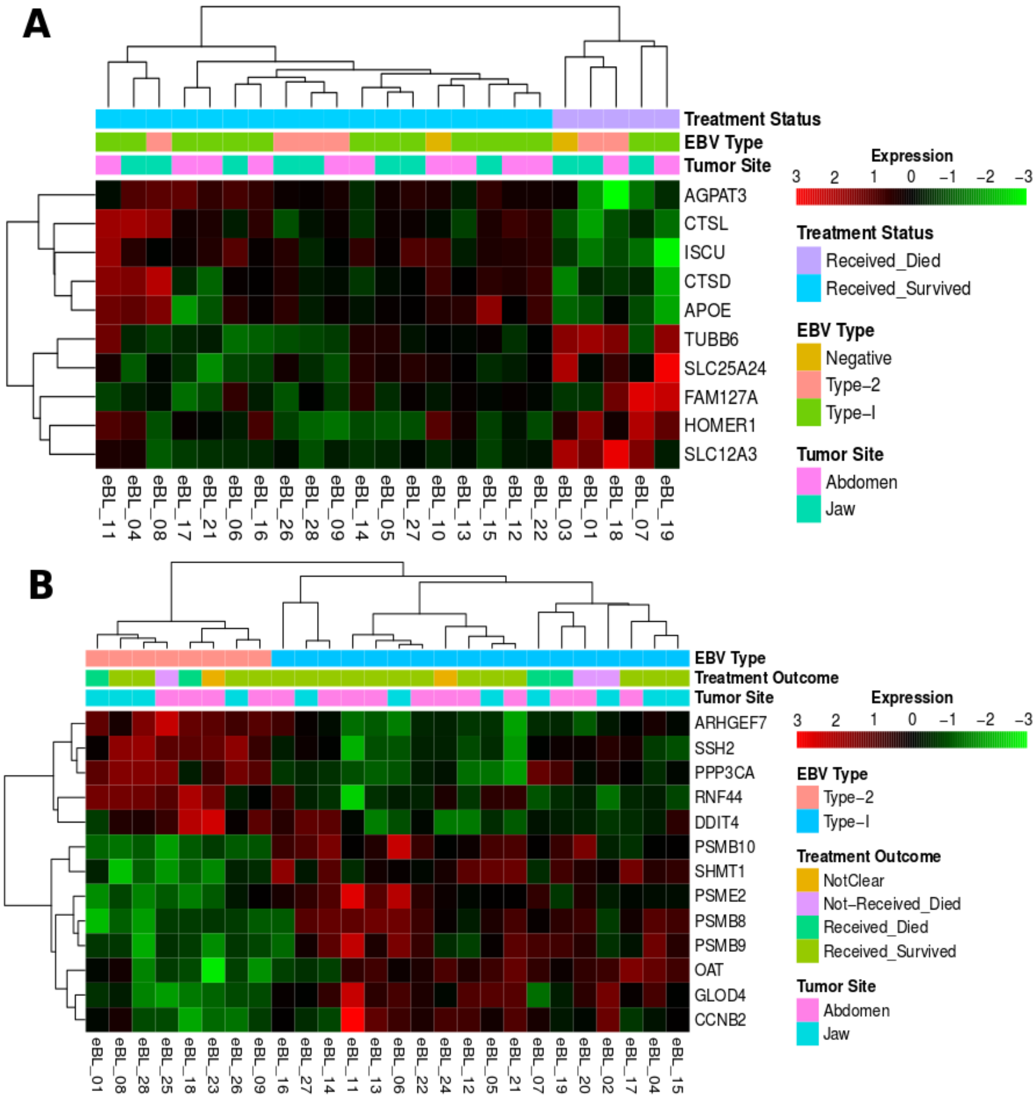


Figure 2.3 Clustering heatmap of significantly differentially expressed human genes between factors of phenotypes. Sample wise scaled log2 expression values range between lowest as light green and highest as dark red. Clustering dendrogram based on Pearson correlation demonstrates tumor grouping proper to the phenotype of interest. **A)** Significantly differentially expressed genes between Survivors and Nonsurvivors (BH $P_{\text{adj}} < 0.1$). **B)** Significantly differentially expressed human genes between eBL tumors carrying EBV Type 1 and eBL tumor with EBV Type 2 (BH $P_{\text{adj}} < 0.1$).

Comparison between EBV type 1 and type 2 viral-containing eBL tumors revealed 13 significant, differentially expressed genes (Figure 2.3B). Four out of 8 genes that have significantly higher expression in eBL tumors with type 1 EBV are coding for the required components of immunoproteasome complex formation; *PSMB9* (*β1i*), *PSMB10* (*β2i*), *PSMB8* (*β5i*), and *PSME2* (*PA28β*). In addition, all of the other proteasome gene transcripts showed increased expression on average in eBLs with type 1 EBV. Consistent with this, our gene set enrichment and pathway analysis revealed several significant differential gene sets involving MHC class I antigen presenting cascades, ubiquitination, and proteasome degradation, and antigen cross presentation altered between type 1 and 2. This difference in expression of IFN-gamma inducible immunoproteasome complex genes and enriched pathways suggest that type 1 and type 2 genomes of EBV might differ in the pathogenesis of infection as well as in their roles promoting oncogenesis.

2.3.4 Human Gene Expression Appears to be more Differentiated Based on EBV Status rather than eBL and sBL Geographic Designation

We next investigated whether sBLs differ from eBLs regarding human gene expression profiles. While it is inherently challenging to compare samples that have been collected and experimentally processed with non-identical methods, we attempted to control for collection and processing differences by accounting for them in the comparison sets using SVA. We included 7 BL cell cultures as well as two sBL primary biopsies from our sequencing set and observed proper clustering according to their eBL and sBL designations. Differential gene expression analysis comparing the only eBL and sBL primary biopsy samples resulted in 504 genes with significantly different expression profiles based on geographic BL subtype classification. Leading edge analysis following the GSEA of differentially expressed genes reoccurring in multiple Gene Ontology (GO) gene sets demonstrated that the genes involving biological processes such as vasculature or blood vessel development (BH $P_{\mathrm{adj}} = 6.0 \times 10^{-24}$) and angiogenesis (BH $P_{\mathrm{adj}} = 4.0 \times 10^{-23}$) are the major variation source between our eBL biopsy collections with FNA and sBLs with FFB. These dominant enrichment sets are likely associated with the different biopsy collection techniques rather than pathological distinctions. On the other hand, the differentially expressed genes between eBL tumors and sBL tumors also resulted in significant enrichments in Hallmark gene sets including apoptosis, *IL2/STAT5* signaling, Notch signaling, *KRAS* signaling, and *TNFα* signaling via *NFκβ*. Leading edge genes in these hallmark sets point to strong differential expression of the *PI3K-Akt* signaling pathway (BH $P_{\mathrm{adj}} = 3.0 \times 10^{-23}$) which plays a central role in BL pathogenesis/oncogenesis (Kawauchi et al. 2009; Rickert 2013).

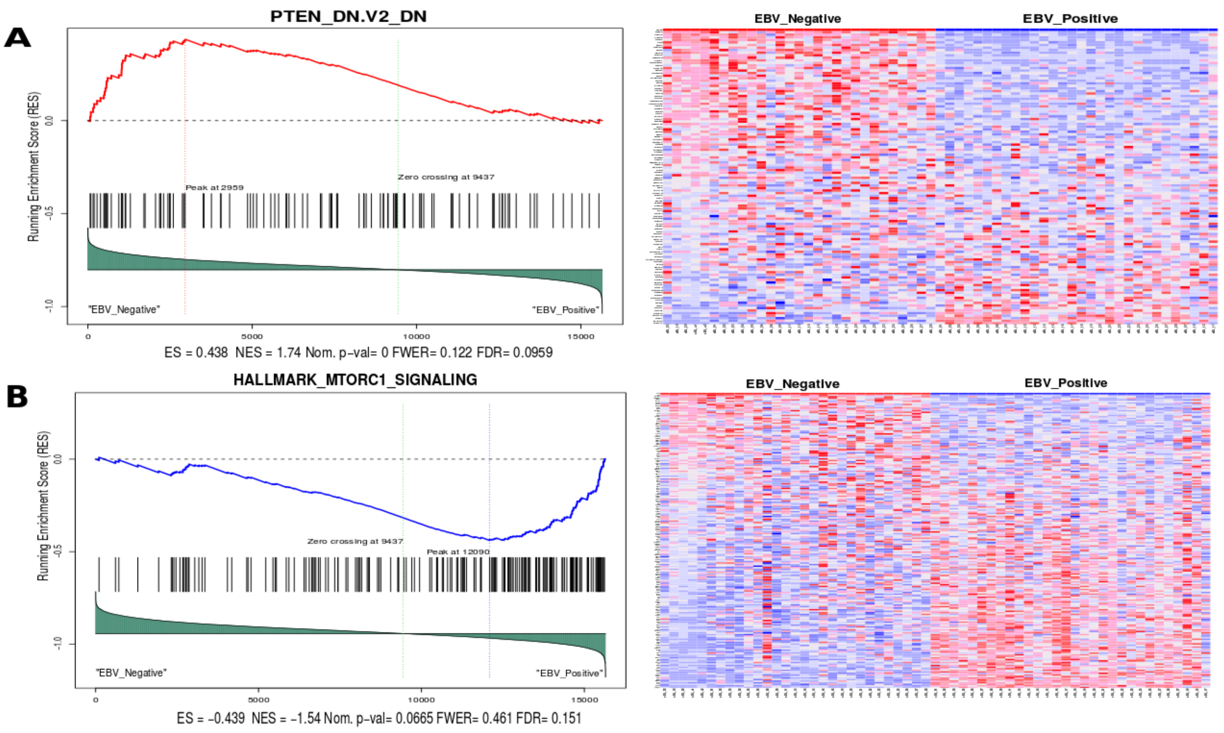


Figure 2.4 Gene set enrichment plot and expression heatmap of corresponding genes in the enriched gene set. Left panels include the running enrichment score throughout the gene set, and projection of genes in the gene set to the complete list of genes rank ordered based on the signal to noise ratio. Leading edge genes that build up the enrichment score of the geneset (RES at the peak) are the most important genes for these tumor sample comparison. On the expression heatmap (columns are tumors, rows are genes in the gene set), dark red represent higher expression while dark blue lower expression. **A)** Genes in this enrichment have been shown to be downregulated upon *PTEN* knockdown, and are observed to be downregulated in EBV positive BLs relative to EBV negative BLs (ES = 0.438, Nominal P = 0.00, FDR q = 0.0959) and **B)** Hallmark gene set enrichment showing mTOR complex 1 signaling genes to be relatively more activated in EBV positive BLs compared to negative BLs (ES = -0.439, Nominal P = 0.0665, FDR q = 0.151). Enrichment of genes associated with mTOR activation supports the enrichment of genes linked to *PTEN* inhibition.

Given that EBV presence is highly correlated with eBL tumors, we hypothesized that EBV might be a major determinant affecting differential expression between BL tumor subtypes. Therefore, we stratified our samples sets by their EBV content. Hierarchical clustering of the sample correlations demonstrates that we successfully preserved the variation associated with only BL tumors' EBV status and removed other unwanted covariates. As a sign of this, three EBV positive BL cell cultures as well as 4 EBV positive sBL tumors clustered with the rest of the EBV positive eBL tumors. Confirming this stratification, two EBV negative eBLs and two sBLs from our sequencing set properly clustered with the rest of the negative sBLs. We then performed differential gene expression and pathway enrichment analysis between the primary BL tumors, excluding the cell lines. This resulted in 1658 significantly differentially expressed genes between EBV positive and negative BL tumors. Increased number of significantly differentially expressed genes suggests that EBV presence in BL tumor affects host expression profile more dramatically than subtype designations based on geography. These differentially expressed genes highlighted functions in biological processes involving DNA replication, mismatch repair as well as cell cycle regulation pathways. Interestingly, gene set enrichment of the differentially expressed genes between EBV positive and negative BL tumors resulted in a significant enrichment in one of the oncogenic signature gene sets which consists of genes down-regulated when *PTEN* was experimentally knock-down (FDR q=0.096). Figure 2.4A shows the genes that have higher expression in EBV negative BLs compared to EBV positive BLs. This suggests that EBV positive BLs, regardless of their geographic origin, share a common mechanism in which *PTEN* is suppressed. Supporting this, enrichment of another gene set in which genes are upregulated through activation of mTORC1 (mTOR complex 1) suggests the loss of the regulatory role of *PTEN* in this signaling pathway in EBV positive BL tumors (Figure 2.4B). Combined these suggest increased activity of PI3K and subsequently AKT/mTOR pathway driving cell cycle and proliferation.

2.3.5 Examination of Transcript Mutations

We next explored the transcriptome for somatic mutations in eBL and compared it to previously sequenced sBL in order to investigate whether gene mutation frequencies diverge as well as gene expression profiles. After excluding the known genomic variants (SNPs), a total of 2728 putative somatic mutations were determined across the 56 tumor samples. Our carefully controlled variant detection allowed us to compare the gene mutation frequencies between eBL and sBL and clinical correlates (see Chapter III for details). This resulted in a total of 21 genes mutated in 4 or more (>7%) of the sporadic and endemic tumors. Interestingly, the number of mutations did not differ significantly between sBL and eBL with an average of per tumor 3.6 vs. 4.1 genes mutated in eBL and sBL respectively (P = 0.24, t-test, 2-tailed). However, for the top ten most commonly mutated genes the difference was significant with 2.5 and 3.5 genes mutated per tumor in eBL and sBL, respectively (P = 0.017, t-test, 2-tailed). Two of the top 3 genes were equally mutated genes including *MYC* and *DDX3X* (Figure 2.5A, pink-cyan bars). The mutation rates of the genes *ID3*, *CCND3*, *TCF3*, and *SMARCA4* were notably less frequently mutated in eBL tumors accounting for the difference between eBL and sBL. *ID3* was significantly different, mutated in 32% of the eBL compared to 67% of the sBL (P = 0.007, Fisher's Exact). Also, 36% of the sBLs carried mutations in *CCND3* compared to 14% of eBLs (P = 0.061, Fisher's Exact).

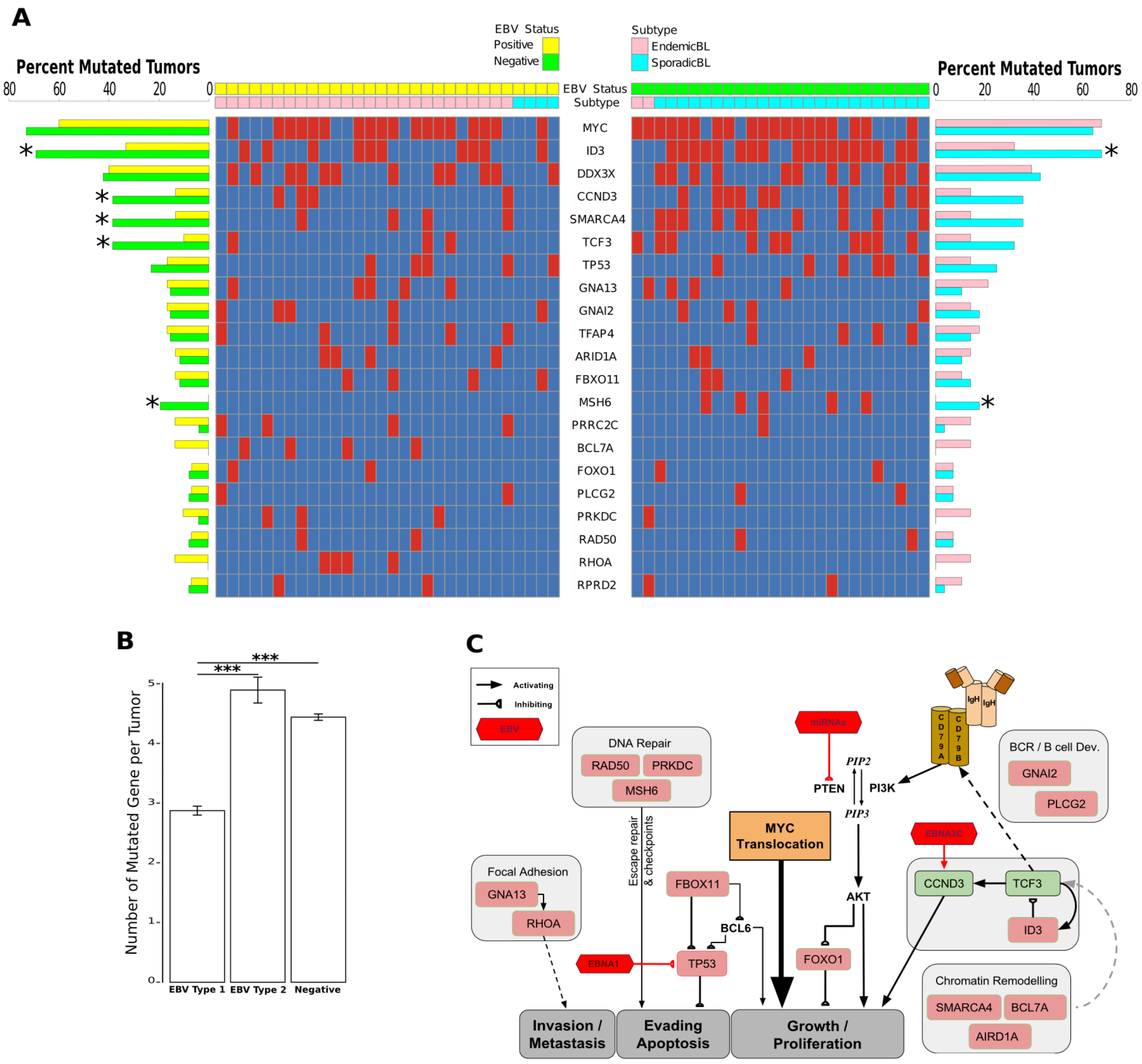


Figure 2.5 Mutational landscape of BL tumors. **A)** This panel demonstrates mutated gene distribution in each tumor sample (columns) are tumors, and rows are top frequently mutated genes (>10% of samples mutated at least once). Tumor samples were grouped based on their EBV content and second color bar shows the subtype of the tumor. Red squares represent mutated gene while blue is for no mutation detected. Barplot on the right measures the frequency of mutated tumor samples and compares regarding the subtype of BLs (Percent frequency). Similarly, bar plot on the left compares the mutated tumor frequencies for each gene stratified by EBV status (* $P < 0.05$, Fisher's Exact). **B)** An average number of mutated genes

per BL tumor by EBV type. Error bars represent standard error (*** $P < 0.01$, t-test). **C)** Schematic overview of the proposed key pathways and frequently and mutated genes in eBL pathogenesis. Genes in the boxes are found to be frequently mutated (gain of function and likely loss of function, represented by green and light-red boxes, respectively). Likely key interactions with EBV components are shown in red connections. Possible interactions are shown in gray.

Given that *ID3*, *TCF3* and *CCND3* are thought to be key drivers for sBL oncogenesis, and eBLs do not carry mutations as frequent as sBLs, we hypothesized again that EBV might be already affecting these pathways abrogating the need for additional mutation. While EBV status is strongly correlated with endemic versus sporadic status, we re-analyzed the tumors based on the viral content to see the effect on the mutational spectrum. Interestingly, the difference in the frequency of the ten most common genes differentiated further to 2.4 to 3.7 for EBV positive and negative tumors, respectively. Only 196 out of 10,000 permutations, in which two eBLs and four sBLs were randomly assigned as virus negative and positive, respectively at each iteration, equaled or exceed the difference observed between EBV positive and negative tumors (P = 0.0198, t-test, 2-tailed). While individual genes in the simulation did not reach significance, *CCND3*, *SMARCA4*, and *TCF3* gene mutation frequencies showed further differentiation and reached significance by Fisher's exact comparing EBV positive and negative tumors ($P < 0.05$ for each) (Figure 2.5A, yellow-green bars). Overall, these findings suggest that the molecular feature of gene mutations correlates better with the presence or absence of EBV within the tumor. Further supporting this, EBV negative BL tumors carried significantly more *TCF3* or *ID3* or *CCND3* mutations (P = 0.0021), and there was only a single occurrence of more than one of these genes being mutated in EBV positive tumor compared with 11 EBV negative tumors.

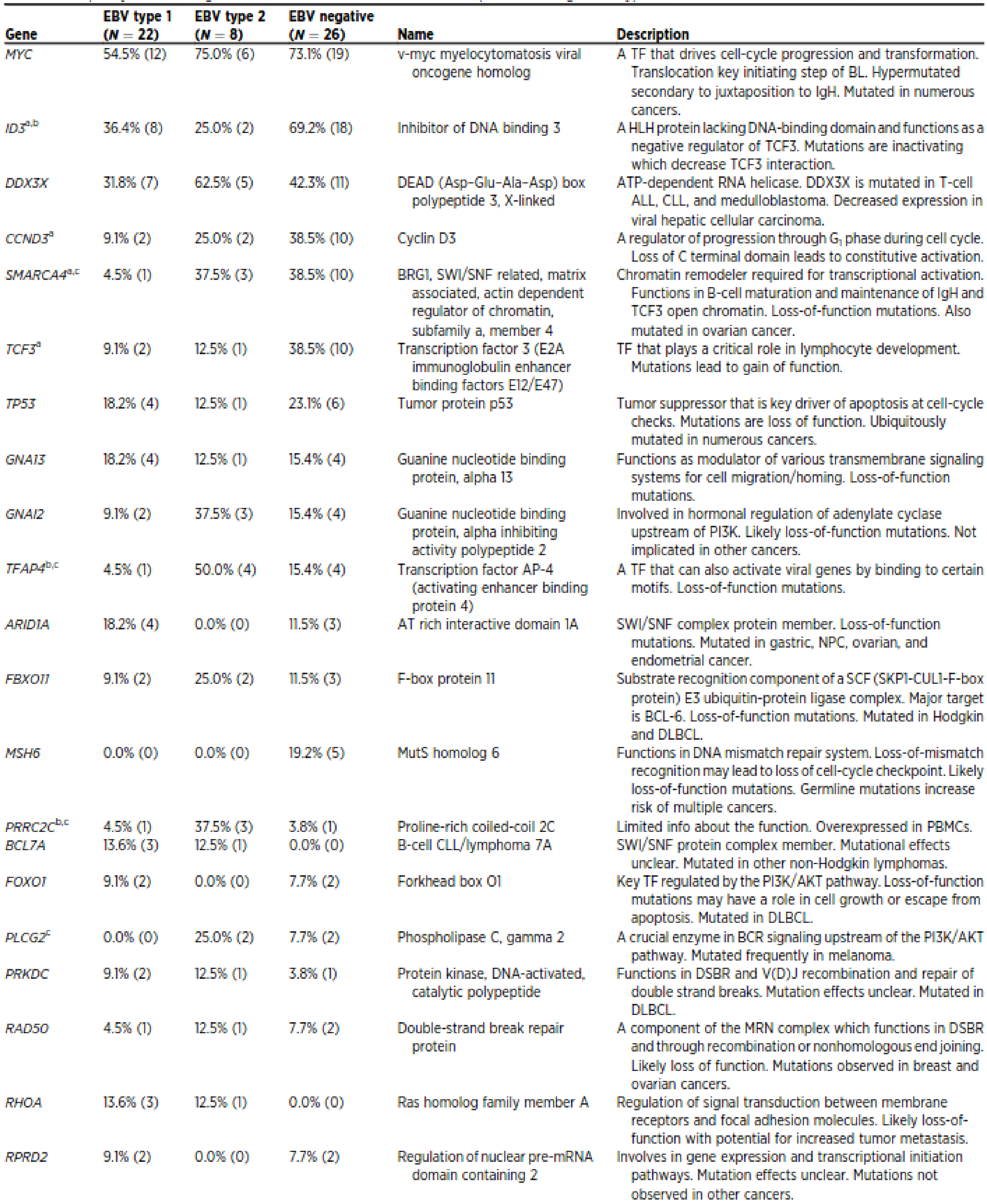
Apart from *ID3*, *TCF3,* and *CCND3*, two key genes *SMARCA4* and *ARID1A* involved in chromatin and nucleosome remodeling as part of the SWI/SNF complex are highly mutated. Interestingly, *ARID1A* was mutated in roughly equivalent levels in either categorization. In contrast, *SMARCA4* demonstrates decreased gene mutation when comparing either eBL and sBL or EBV positive and negative. Furthermore, Ras homolog family member A, *RHOA*, the gene was mutated in 14% of eBL tumors in a mutually exclusive manner with *TCF3*/*ID3* mutations. In addition, when we examined the mutations of two small GTPases, *GNA13* and *GNAI2*, which are also recurrently mutated in DLBCL (diffuse large B-cell lymphoma) (Morin et al. 2013), this mutually exclusive mutation pattern among the BL tumors could be further extended to include mutations in *GNAI2* ($P = 0.005$, Fisher's Exact) but not *GNA13*. While this potentially suggests an alternative path for tumor drives other than deregulated *ID3*/*TCF3*, *CCND3* and *RHOA* were not mutually exclusive potentially suggesting that the effects of *CCND3* may be independent of *TCF3*'s drive.

2.3.6 Novel mutated genes in eBL and clinical correlates with mutational status

We identified several new genes that appear somatically mutated (Table 2.2). *TFAP4*, transcription factor AP-4, whose down regulation can protect against glucocorticoid-induced cell death (Tsujimoto et al. 2005) and appears to be involved in p21-myc regulation of the cell cycle (Jung and Hermeking 2009). *TFAP4* carried five mutations in its HLH (helix-loop-helix) domain of likely deleterious effect, in addition, to stop gain and start loss mutations. A new category of mutations seen in our analysis are genes involved in DNA repair including *RAD50*, *PRKDC* and *MSH6*. While such mutations are common in other cancers such as colorectal cancer, breast and ovarian cancer, and several autoimmune diseases (Mathieu et al. 2015; Okkels et al. 2012; Heikkinen et al. 2003), they have not been reported previously in BL. Another previously undocumented gene showing somatic mutation was *BCL7A*. It has only been previously implicated in lymphoma through the observation of complex rearrangements (Zani et al. 1996) and its overexpression was associated with Germinal Center (GC) phenotype in DLBCL (Blenk et al. 2007). Interestingly, it is also a member of the SWI/SNF complex further highlighting the complex's importance (Kadoch et al. 2013). Mutations in *BCL7A*, *SMARCA4*, and *ARID1A* were not observed to co-occur in either eBL or sBL. In addition, we detected mutations in the *PRRC2C*, *RPRD2*, *FOXO1*, *PLCG2* that are normally overexpressed in peripheral blood mononuclear cells (PBMCs) and lymph nodes.

We also examined the correlation of given mutations to clinical and molecular features. Supporting the lack of expression difference between tumor presentation sites, we observed no suggestive associations with mutations. A single gene *GNAI2* has a potential association with in-hospital survival ($P = 0.021$) where one out of 18 survivors had a mutation in *GNAI2* compared to three of five who died during initial hospitalization.

Table 2.2 Frequency of mutated genes in BL tumors classified based on EBV presence and genome type.



Gene	EBV type 1 (*N* = 22)	EBV type 2 (*N* = 8)	EBV negative (*N* = 26)	Name	Description
MYC	54.5% (12)	75.0% (6)	73.1% (19)	v-myc myelocytomatosis viral oncogene homolog	A TF that drives cell-cycle progression and transformation. Translocation key initiating step of BL. Hypermutated secondary to juxtaposition to IgH. Mutated in numerous cancers.
ID3[a,b]	36.4% (8)	25.0% (2)	69.2% (18)	Inhibitor of DNA binding 3	A HLH protein lacking DNA-binding domain and functions as a negative regulator of TCF3. Mutations are inactivating which decrease TCF3 interaction.
DDX3X	31.8% (7)	62.5% (5)	42.3% (11)	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked	ATP-dependent RNA helicase. DDX3X is mutated in T-cell ALL, CLL, and medulloblastoma. Decreased expression in viral hepatic cellular carcinoma.
CCND3[a]	9.1% (2)	25.0% (2)	38.5% (10)	Cyclin D3	A regulator of progression through G_1 phase during cell cycle. Loss of C terminal domain leads to constitutive activation.
SMARCA4[a,c]	4.5% (1)	37.5% (3)	38.5% (10)	BRG1, SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4	Chromatin remodeler required for transcriptional activation. Functions in B-cell maturation and maintenance of IgH and TCF3 open chromatin. Loss-of-function mutations. Also mutated in ovarian cancer.
TCF3[a]	9.1% (2)	12.5% (1)	38.5% (10)	Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)	TF that plays a critical role in lymphocyte development. Mutations lead to gain of function.
TP53	18.2% (4)	12.5% (1)	23.1% (6)	Tumor protein p53	Tumor suppressor that is key driver of apoptosis at cell-cycle checks. Mutations are loss of function. Ubiquitously mutated in numerous cancers.
GNA13	18.2% (4)	12.5% (1)	15.4% (4)	Guanine nucleotide binding protein, alpha 13	Functions as modulator of various transmembrane signaling systems for cell migration/homing. Loss-of-function mutations.
GNAI2	9.1% (2)	37.5% (3)	15.4% (4)	Guanine nucleotide binding protein, alpha inhibiting activity polypeptide 2	Involved in hormonal regulation of adenylate cyclase upstream of PI3K. Likely loss-of-function mutations. Not implicated in other cancers.
TFAP4[b,c]	4.5% (1)	50.0% (4)	15.4% (4)	Transcription factor AP-4 (activating enhancer binding protein 4)	A TF that can also activate viral genes by binding to certain motifs. Loss-of-function mutations.
ARID1A	18.2% (4)	0.0% (0)	11.5% (3)	AT rich interactive domain 1A	SWI/SNF complex protein member. Loss-of-function mutations. Mutated in gastric, NPC, ovarian, and endometrial cancer.
FBXO11	9.1% (2)	25.0% (2)	11.5% (3)	F-box protein 11	Substrate recognition component of a SCF (SKP1-CUL1-F-box protein) E3 ubiquitin-protein ligase complex. Major target is BCL-6. Loss-of-function mutations. Mutated in Hodgkin and DLBCL.
MSH6	0.0% (0)	0.0% (0)	19.2% (5)	MutS homolog 6	Functions in DNA mismatch repair system. Loss-of-mismatch recognition may lead to loss of cell-cycle checkpoint. Likely loss-of-function mutations. Germline mutations increase risk of multiple cancers.
PRRC2C[b,c]	4.5% (1)	37.5% (3)	3.8% (1)	Proline-rich coiled-coil 2C	Limited info about the function. Overexpressed in PBMCs.
BCL7A	13.6% (3)	12.5% (1)	0.0% (0)	B-cell CLL/lymphoma 7A	SWI/SNF protein complex member. Mutational effects unclear. Mutated in other non-Hodgkin lymphomas.
FOXO1	9.1% (2)	0.0% (0)	7.7% (2)	Forkhead box O1	Key TF regulated by the PI3K/AKT pathway. Loss-of-function mutations may have a role in cell growth or escape from apoptosis. Mutated in DLBCL.
PLCG2[c]	0.0% (0)	25.0% (2)	7.7% (2)	Phospholipase C, gamma 2	A crucial enzyme in BCR signaling upstream of the PI3K/AKT pathway. Mutated frequently in melanoma.
PRKDC	9.1% (2)	12.5% (1)	3.8% (1)	Protein kinase, DNA-activated, catalytic polypeptide	Functions in DSBR and V(D)J recombination and repair of double strand breaks. Mutation effects unclear. Mutated in DLBCL.
RAD50	4.5% (1)	12.5% (1)	7.7% (2)	Double-strand break repair protein	A component of the MRN complex which functions in DSBR and through recombination or nonhomologous end joining. Likely loss of function. Mutations observed in breast and ovarian cancers.
RHOA	13.6% (3)	12.5% (1)	0.0% (0)	Ras homolog family member A	Regulation of signal transduction between membrane receptors and focal adhesion molecules. Likely loss-of-function with potential for increased tumor metastasis.
RPRD2	9.1% (2)	0.0% (0)	7.7% (2)	Regulation of nuclear pre-mRNA domain containing 2	Involves in gene expression and transcriptional initiation pathways. Mutation effects unclear. Mutations not observed in other cancers.

Fisher exact test $P < 0.05$ was denoted by * for type 1 vs type 2; † for type 1 vs EBV negative; ‡ for type 2 vs EBV negative BLs. Mutated BL tumor counts are in parenthesis. Abbreviations TF; Transcription Factor, BCR; B-cell Receptor, PBMCs; Peripheral blood mononuclear cells, DSBR; Double-stranded break repair, ALL; Acute lymphoblastic leukemia, CLL; Chronic lymphocytic leukemia, DLBCL; Diffuse large B cell lymphoma, NPC; Nasopharyngeal carcinoma. Gene description and functions are from NCBI/GenBank database. Mutations in other cancers are from the COSMIC database (http://cancer.sanger.ac.uk/cosmic).

2.3.7 Rates of Gene Mutations vary based on EBV genome type

Regarding EBV type, *PRRC2C*, *SMARCA4*, *PLCG2*, and *TFAP4* differed significantly ($P<0.05$). Interestingly, for these genes, type 2 EBV tumors always had the higher proportion of mutations compared to eBL tumors containing EBV type 1 (Table 2). All of the mutations that *TFAP4* carried were either deleterious (Met1?, Gln15*, Arg127*, Pro185Leu) or accumulated in the DNA-binding domain (Arg50Trp, Arg58Trp, Arg60Cys). These suggest a possible loss of function for the protein AP4 encoded by this gene mutated mostly in type 2 carrying eBLs (50%). *SMARCA4* mutation rates in groups of BLs with type 1, type 2, and negative were roughly 5%, 38%, and 39%, respectively. Mutations in this gene were located in its important domains SNF2, helicase, and HSA domains. In general, the average number of mutated genes per tumor (including all 21 genes) was 2.9 in BL tumors infected with type 1 while 4.9 in BLs with type 2 EBV ($P < 0.01$, t-test, 2-tailed) (Figure 2.5B). This was significant even excluding type 1 tumors that were sporadic and also the overall type 2 mutation rate was on par with that of EBV negative tumors. The only genes that appear to have significantly lower mutation rates in type 2 tumors compared to EBV negative tumors were *ID3* and *TCF3*, which were on par with type 1. Overall, functions of genes with distinct mutation frequencies in these groups, in addition to the significantly different general mutation rates, supports type 1 EBV's reputation regarding better transformation ability compared to type 2, which had almost equivalent levels of mutated genes per tumor as EBV negative BLs (4.9 and 4.4, respectively).

2.4 Discussion

Our major finding is that eBL is a homogenous tumor with highly correlated expression profiles, regardless of tumor location within the body. This means there is no need to create diagnostic subdivisions based on tumor presentation site or refine staging based on gene expression profiles. This contradicts previous expression analyses that suggested greater heterogeneity within eBL compared to sBL or id-BL (Piccaluga et al. 2011b). Our unsupervised hierarchal clustering did not reveal major clades based on the virus, viral type, in-hospital survival or tumor presentation site. Our analysis of tumor presentation site suggested minimal differences that could be explained by associated cellular microenvironment (e.g., differences in endothelium presence, *NOS3*). On the other hand, gene expression comparison of tumors based on survival status of the patients revealed several candidate genes and gene sets providing potential prognostic biomarkers which are currently lacking for eBL. Survival rate difference could be attributed to delayed time to diagnosis of abdominal cases compared to more apparent facial tumors (Kazembe et al. 2003). Further studies are required to validate the clinical utility of such markers.

The analysis presented here of BL patients from Kisumu, Kenya was underway when a similar study was published involving 20 Ugandan patients (Abate et al. 2015b). Therefore, we re-evaluated our analysis to determine if we could validate their findings. We found similar mutation rates in *ID3* and *TCF3* genes and associated these with the lack of EBV positivity. However, we observed less frequent *RHOA* and no *CCNF* mutations in Kenyan eBL tumors. In contrast to the Ugandan study, we also failed to detect any significant trace of other herpes viruses such as KSHV or CMV other than EBV although this may be attributable to our FNAs which decrease the sampling of the connective tissue where these viruses were mainly present. Their study was also limited in its ability to examine EBV types for which we found significant differences in expression within our Kenyan tumors. The mutated genes we observed in BL tumors are also dysregulated or mutated in other cancer types with viral etiologies. *DDX3X* is activated by Hepatitis C Virus (HCV) leading to alteration of host cellular gene expressions (Ariumi et al. 2007). *RHOA* and *CCND3* are dysregulated by Human T lymphotropic virus type I (HTLV-I) (Marriott and Semmes 2005). For the DNA tumor viruses, KSV tumors appear to be driven by viral programming in settings of immunocompromised with only a few described driver mutations including interleukin 1 receptor-associated kinase (*IRAK1*) in primary effusion lymphoma (D. Yang et al. 2014). In HPV-associated squamous cell cervical carcinoma, the most common driver mutations are *PIK3CA*, *EP300*, *TP53*, *FBXW7*, and *MAPK1* *(Ojesina et al. 2014).* PI3K mutations are common in epithelial derived EBV-positive nasopharyngeal carcinoma, with the most commonly somatically altered genes being *TP53*, *CDKN2A/B*, *ARID1A*, *CCND1*, *SYNE1* and *PI3KCA* *(Lo, Chung, and To 2012)* (D.-C. Lin et al. 2014). While these are comparable with BL the targeting of the PI3K pathway, SWI/SNF, and p53, the overall differences suggest that even between EBV malignancies the major factor in determining what genes are the lynch pins between normalcy and malignancy is the cell lineage and state. This concept is supported by the greater mutational commonality with other lymphomas and the fact that both virus positive and negative BL tumors have mutational commonality differing mainly in degree.

Acknowledgements

We thank the children and their families for their participation and the study staff at Jaramogi Oginga Odinga Teaching and Referral Hospital for their assistance with this study. We thank the Director, Kenya Medical Research Institute for approving this manuscript for publication.

Chapter III. Experimental and Computational Framework for Studying Comparative Gene Expression and Mutations using RNA Sequencing Data

3.1 Summary

In this chapter, I cover major experimental and computational techniques unique to RNAseq for comparative gene expression and mutational profiling studies. I briefly summarize important aspects of experimental designs, possible biases, and proposed approaches to correct them. I also explain the experimental and computational methods we used in our projects in detail with their reasonings for why we chose them. In addition to gene expression quantification, I also cover methods to accurately call spontaneous mutation sites in expressed protein-coding genes using RNAseq data. Furthermore, I provide a unique data analysis pipeline for PASseq (Polyadenylation site sequencing) datasets and how we approached this to utilize it for determining the polyA site isoform level expressions. This computational framework also allows users to identify significant alternative polyA site switches between different conditions. Finally, I provide the scripts and tools mentioned in these frameworks via GitHub to users who are willing to apply to their projects.

3.2 Accurate Measurement of Gene Expression

3.2.1 RNA Isolation for Quantifying Expression Level

Transcriptional processes produce a variety of RNA molecules which are classified based on their potential for coding for a protein sequence. Non-coding RNAs (ncRNAs) play distinct functional roles in various ways and mainly consist of transfer RNA (tRNA), ribosomal RNA (rRNA), long intergenic non-coding RNAs (lincRNAs), as well as small size RNAs such as small nucleolar RNAs (snoRNAs), microRNAs (miRNAs), or Piwi-interacting RNAs (piRNAs). Messenger RNAs (mRNAs), on the other hand, carry genetic information for assembling a protein which will function as an enzyme or a building block of a cell. In eukaryotes, precursor mRNA molecules go through post-transcriptional processing by splicing out intronic sequences in the nucleus. During the transcription, 5'-end of the molecules are modified with a methylated guanine nucleotide (5'-capping) to form a stable molecule and protect it from degradation. Another modification which affects almost all mRNAs is a process called polyadenylation which involves the addition of many adenosine nucleotides to the 3'-end of mRNAs. A subset of histone protein coding mRNAs, replication-dependent histones, are the known exceptions of polyA tailing and their 3'-end forms a step-loop instead (Marzluff, Wagner, and Duronio 2008).

Although mRNAs and proteins are completely different molecules, the functional consequence of the proteins can be extrapolated by determining the level of mRNAs in a cell. Both of these have a highly dynamic activity and are controlled via strict regulatory mechanisms. In addition to variation from gene to gene, cell type specific regulations and environmental factors play a major role in fluctuations of expression levels. Thus, gene expression measurement is not a trivial task. An expression level of a gene, X_{ijt}, should be defined with its constraints as the number of mature mRNA molecules produced from a gene i, in cell j, at time point t. Proper measurement of this will allow us to predict functional consequences of proteins and to infer the faith of the system or phenotypic outcome. Therefore, methods for estimating expression levels have to be precise and produce results closer to reality. In order to determine mRNA levels, very first step is to reach molecules by disrupting the cellular structures and isolating RNA from the lysate. This delicate process is essential to maintain the integrity of RNA composition and characteristics. The quality of isolated RNA is usually assessed by checking the abundance ratio of 28S to 18S rRNA, which should roughly be 2:1, respectively. Departure from this proportion might indicate poor quality and degradation in isolated RNA. Most of the current RNA isolation techniques require enormous amounts of starting material from cells and tissues. As a result, isolated mRNA levels no longer pertain to individual cells composing tissue bulk rather represent an approximate average of them. Therefore, the expression level of a gene should be re-defined as $X_{i\mu t}$, where μ represents the average of all cells used in that isolation.

One of the major issues faced with during the detection of expression levels of target RNA populations, generally mRNAs, is sampling inequality because of naturally high abundant rRNA transcripts. rRNAs function as the key component of ribosome complex and comprise of approximately 80% of total RNA strains depending on cell type since they are produced by multiple repetitive and highly conserved loci of the genome. Abundant rRNA strains reduce sensitivity in expression measurements because of their dominance, and they are less informative unless specifically targeted for microbiome studies. Therefore, rRNA transcripts should be eliminated. There are two widely used techniques; the first one is selecting only RNA strains desired for expression measurement. Reverse transcriptase quantitative PCR (RT-qPCR) conducted using specific primers targeting unique gene sequences can be considered in this class. Microarray chips consist of many, as opposed to RT-qPCR, gene-specific hybridization probe sequences attached to a solid surface is another technique which has been used for expression measurement. Both of these are highly successful for eliminating rRNA and reaching the desired signal. An alternative method is a negative enrichment by removing with magnetic beads carrying hybridization probes on the surface designed specifically as complementary to rRNAs by utilizing conserved nature. However, this increases the cost of assay undesirably.

For Burkitt lymphoma expression quantification study, we extracted genomic DNA and total RNA from FNA eBL biopsies stored in RNAlater and from sBL frozen biopsies stored in liquid nitrogen using Qiagen Allprep DNA/RNA/Protein mini kit (QIAGEN Sciences, Germantown, MD) according to manufacturer's instructions. Following the extractions, concentrations of nucleic acids were measured with picoGreen (Thermo Fisher Scientific Inc.), and quality checked using NanoDrop. To further assess input total RNA quality and integrity, we ran 1ul extract on Agilent Bioanalyzer RNA Nano chip (Agilent Technologies, Waldbronn, Germany) and only included samples with RNA Integrity Number (RIN) higher than 7.0.

3.2.2 Sequencing as a Method for Expression Measurement

The emergence of sequencing technologies and reduced costs opened areas for new applications. Sequencing RNA transcripts (RNAseq) and determining the abundance levels is one of most common practices. Although there are multiple alternatives, Illumina short reads sequencing is widely preferred for this purpose. Sequencing DNA or cDNA using Illumina instrument requires a preparation of a library composed of fragmented pieces of sequences. Two primary methods are used to remove rRNA, as mentioned above, one of them is rRNA depletion technique, and the other one is polyA selection which enriches almost all mature polyadenylated mRNA transcripts using oligo-dT attached magnetic beads. Library preparation protocols also vary depending on the purpose of sequencing. However, the difference becomes apparent when it comes to preserving the transcription strand of RNA. Early standard Illumina protocols widely used produced non-strand specific libraries while improvements in protocols allowed maintaining original strand information (Parkhomchuk et al. 2009). The most common strand specific RNAseq protocol is based on utilizing dUTP in the second strand synthesis instead of dTTP; therefore it is called dUTP-method (details of this will be given later in this chapter) (Levin et al. 2010). One of the advantages of using a strand-specific RNAseq protocol is better expression quantification of, especially overlapping genes. Depending on the technique used during the library preparation, various biases can occur. Coverage biases towards the ends of transcripts emerge by using cDNA in fragmentation rather than RNA molecule (Z. Wang, Gerstein, and Snyder 2009). 3'-end bias also occur especially when polyA tail selected low quality or degraded RNA isolates are used in preparation. Both of these result in non-uniform distribution of read coverage of sequenced transcripts. Besides, extreme GC content of transcripts also affect their sequencing coverage and result in under-estimated expression levels compared to other moderate GC transcripts because of multiple PCR steps during the library preparation.

If the desired experimental measurement is to determine gene expression levels, short-read Illumina sequencing, 36-100 nt, is commonly used. Although it has been found that there was little improvement in quantification when reads used longer than 50 nt, sequencing libraries with paired end significantly increased the isoform quantification quality because of the decrease in mapping ambiguity (Chhangawala et al. 2015). Target sequencing read depth per sample depends on experimental questions. For human gene expression quantification, 30 million reads/sample is enough, while for discovering novel transcripts and isoforms, sequencing depth should be at least more than 50 million per sample (Sims et al. 2014).

Starting with 1-5 ug total RNA, we isolated mRNA with oligo-dT using Dynabeads mRNA purification kit (Life Technologies). mRNA was fragmented with metal ions, and after the first strand synthesis with Superscript III, we marked the second strand by incorporating dUTP instead of dTTP. Then end-repair and A-tailing steps were followed by Y-shaped Illumina adapter ligation step which maintains strand directionality. We size-selected libraries using XP Ampure magnetic beads (Beckman Coulter Inc.) selecting for a 330 bp insert size. Before low cycle PCR with High Fidelity Phusion polymerase (NEB) for amplification and adapter sequence completion, the second strand marked with dUTP was degraded. During the sequencing library preparation, we checked DNA quality and fragment size distribution using Bioanalyzer Agilent 1000 kit. Final library qualities were confirmed with Bioanalyzer Agilent High Sensitivity DNA kit and sequenced with paired-end read (2x100bp) using an Illumina HiSeq 2000 (Illumina Inc., San Diego, CA, USA).

3.2.3 RNAseq Data Processing and Expression Quantification

Qualities of raw sequencing reads were assessed by using FastQC (Andrews 2010). Reads were sorted by sample based on unique sample barcodes identified by Novobarcode (Novocraft Technologies, Malaysia). Residual Illumina adaptor sequences on the 3' end of reads were trimmed using cutadapt (M. Martin and Marcel 2011). Reads that mapped to ribosomal RNA using bowtie2 were removed (Langmead, Ben, and Salzberg 2012).

There are various approaches for specific questions to answer with RNAseq, and each has its own challenges (Garber et al. 2011). The first step towards determining the expression level of each gene is normalization by total sequencing depth which is the total amount of reads generated by the instrument. RNAseq method naturally produces more sequencing reads for longer transcripts. This property causes underestimation to occur towards shorter transcripts. Thus, read counts of each gene need to be normalized to the gene exonic length before comparing genes to each other. This method of normalization is abbreviated as FPKM (Fragments per Kilobase exon per Million reads) or RPKM (Reads instead of Fragments in the case of single-end). If we define number of read counts that gene i gets as Xi, and total number of reads that instrument generated for the library is N, then;

$CPM_i = \frac{Xi}{N/10^\wedge 6}$, gives the library size normalized read count for gene i (Counts per Million), and

$FPKM_i = \frac{Xi}{(Li/10^\wedge 3)/(N/10^\wedge 6)}$ where Li represents the effective length of the gene i in Kb. On the other hand, there is an alternative normalization method proposed by Bo Li and Colin Dewey (Li, Bo, and Dewey 2011b) called TPM (Transcripts per Million) which can be derived as

$$TPM_i = \frac{FPKMi}{\sum_j \quad FPKMj} 10^6$$

Although these correction methods take care of library sequencing depth and transcript length bias of RNAseq, they do not take GC content of transcript sequences into consideration. GC content becomes a factor needs to be considered especially when gene to gene expression comparison is required. Cqn R package provides functions that can be incorporated during FPKM calculations (Kasper Daniel Hansen and Wu 2012). This way, generated gene expression matrix can be much more reliable especially when it is used for gene set enrichment analysis in which gene ranking plays an important role.

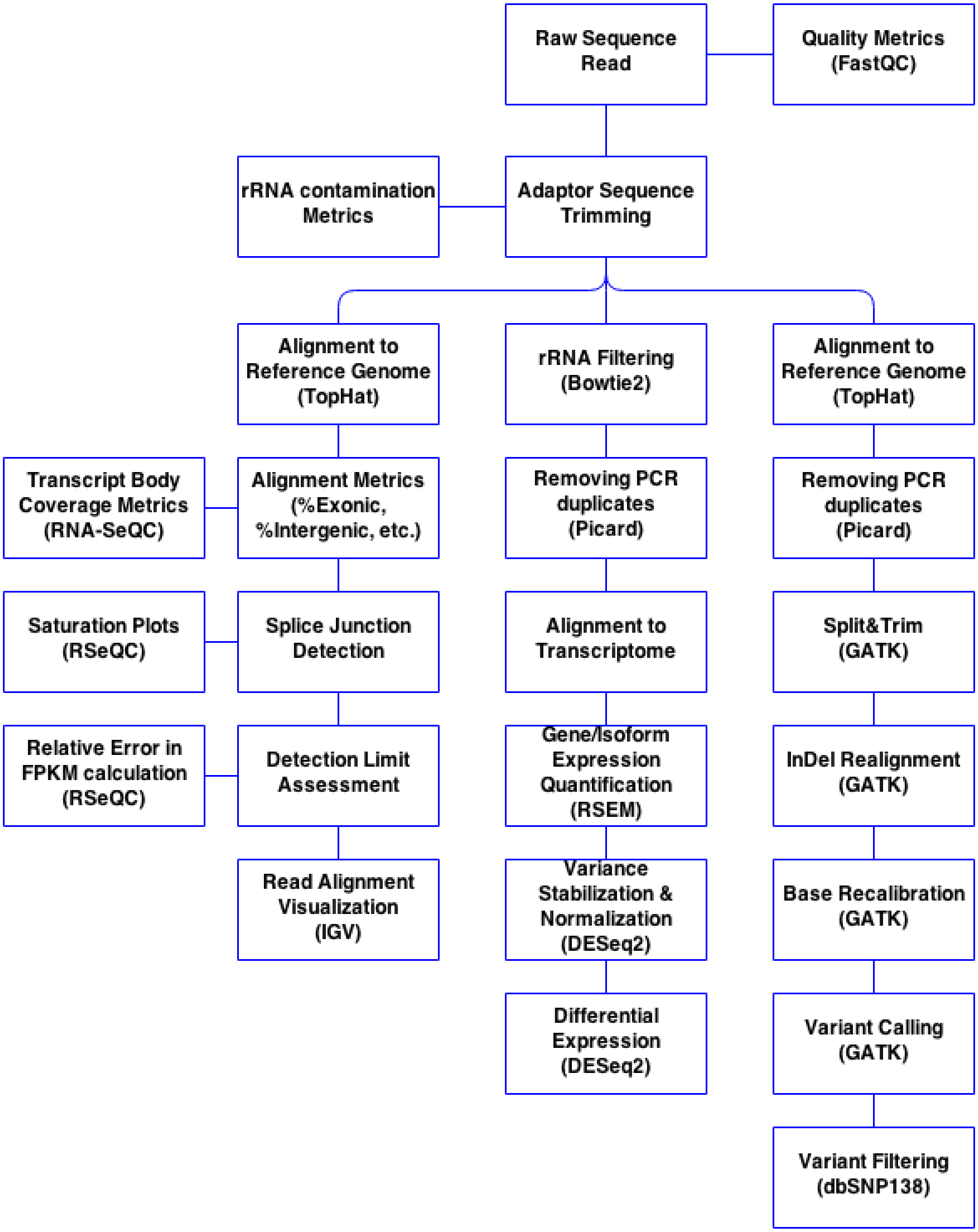

Raw Sequence Read
Quality Metrics (FastQC)
rRNA contamination Metrics
Adaptor Sequence Trimming
Alignment to Reference Genome (TopHat)
rRNA Filtering (Bowtie2)
Alignment to Reference Genome (TopHat)
Transcript Body Coverage Metrics (RNA-SeQC)
Alignment Metrics (%Exonic, %Intergenic, etc.)
Removing PCR duplicates (Picard)
Removing PCR duplicates (Picard)
Saturation Plots (RSeQC)
Splice Junction Detection
Alignment to Transcriptome
Split&Trim (GATK)
Relative Error in FPKM calculation (RSeQC)
Detection Limit Assessment
Gene/Isoform Expression Quantification (RSEM)
InDel Realignment (GATK)
Read Alignment Visualization (IGV)
Variance Stabilization & Normalization (DESeq2)
Base Recalibration (GATK)
Differential Expression (DESeq2)
Variant Calling (GATK)
Variant Filtering (dbSNP138)

Figure 3.1 Overview of the RNAseq data analysis workflow.

3.2.4 Differential Gene Expression and Dealing with Batch Effects

For differential gene expression analysis between two groups of samples or conditions, length or GC bias corrections are not required because, in such analysis, expression levels of different samples for each gene are compared rather than a comparison of genes to each other. Thus, most of the widely used testing approaches model gene expression by accepting raw read counts. After library size normalization, they fit read count distribution of each sample to Negative Binomial and perform a statistical test to determine significant expression changes.

After preprocessing, we aligned read pairs to a transcriptome index built by RSEM (Li, Bo, and Dewey 2011b) using Gencode version 19 protein-coding transcript annotations and hg19 genomic sequence. We calculated the expected read counts for each gene with strand-specific settings of rsem-calculate-expression. We used a union gene model which considers all possible isoforms and unionize to an inclusive gene model, which has the longest transcript structure. Fragments of transcripts that have varying GC contents tend to have different amplification efficiencies during the library enrichment. Therefore, after removing genes with all zero counts, genes were subjected to GC and transcript size normalization using functions in R package cqn (conditional quantile normalization) (Kasper D. Hansen, Irizarry, and Wu 2012). These corrections were applied to FPKM calculations to generate normalized gene expression values. Genes that do not reach to median 1.0 FPKM value for both of the conditions were excluded from that particular comparison. We then assessed the libraries separate batches and sample collection with principal component analysis and further corrected gene FPKM values for known covariates such as different batches using R package ComBat (H. S. Parker et al. 2014). To perform differential gene expression analysis, we used DESeq2 (M. I. Love, Huber, and Anders 2014b) which fits the read counts to a negative binomial distribution and tests for significance using Wald test. Genes that are differentially expressed between conditions with were considered as significant with 10% FDR cutoff.

To be able to account for the batch variables and unknown factors while testing for the differential expression, we estimated the number of underlying factors for every comparison separately using svaseq (Leek 2014b) while preserving the variation of interest. This process involves identifying the surrogate variables using log-transformed expression data. **Figure 3.2** shows the sample to sample clustering of primary BL tumor and cell line transcriptome profiles after removing the batch effect.

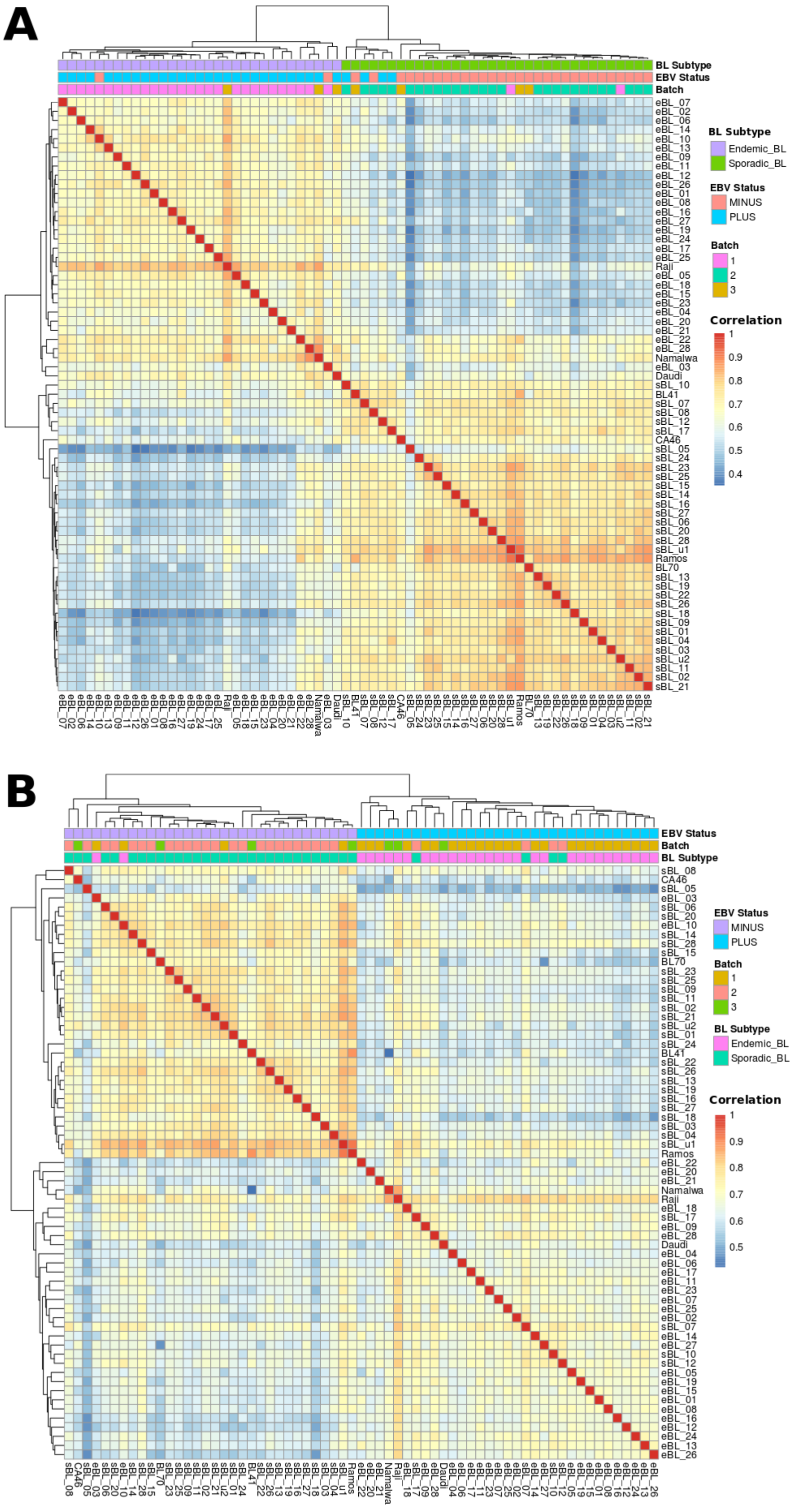

A
BL Subtype
EBV Status
Batch
eBL_07
eBL_02
eBL_06
eBL_14
eBL_10
eBL_13
eBL_09
eBL_11
eBL_12
eBL_26
eBL_01
eBL_08
eBL_16
eBL_27
eBL_19
eBL_24
eBL_17
eBL_25
Raji
eBL_05
eBL_18
eBL_15
eBL_23
eBL_04
eBL_20
eBL_21
eBL_22
eBL_28
Namalwa
eBL_03
Daudi
sBL_10
BL41
sBL_07
sBL_08
sBL_12
sBL_17
CA46
sBL_05
sBL_24
sBL_23
sBL_25
sBL_15
sBL_14
sBL_16
sBL_27
sBL_06
sBL_20
sBL_28
sBL_u1
Ramos
BL70
sBL_13
sBL_19
sBL_22
sBL_26
sBL_18
sBL_09
sBL_01
sBL_04
sBL_03
sBL_u2
sBL_11
sBL_02
sBL_21
BL Subtype
Endemic_BL
Sporadic_BL
EBV Status
MINUS
PLUS
Batch
1
2
3
Correlation
1
0.9
0.8
0.7
0.6
0.5
0.4
B
EBV Status
Batch
BL Subtype
sBL_08
CA46
sBL_05
eBL_03
sBL_06
sBL_20
eBL_10
sBL_14
sBL_28
sBL_15
BL70
sBL_23
sBL_25
sBL_09
sBL_11
sBL_02
sBL_21
sBL_u2
sBL_01
sBL_24
BL41
sBL_22
sBL_26
sBL_13
sBL_19
sBL_16
sBL_27
sBL_18
sBL_03
sBL_04
sBL_u1
Ramos
eBL_22
eBL_20
eBL_21
Namalwa
Raji
eBL_18
sBL_17
eBL_09
eBL_28
Daudi
eBL_04
eBL_06
eBL_17
eBL_11
eBL_23
eBL_07
eBL_25
eBL_02
sBL_07
eBL_14
eBL_27
sBL_10
sBL_12
eBL_05
eBL_19
eBL_15
eBL_01
eBL_08
eBL_16
eBL_12
eBL_24
eBL_13
eBL_26
EBV Status
MINUS
PLUS
Batch
1
2
3
BL Subtype
Endemic_BL
Sporadic_BL
Correlation
1
0.9
0.8
0.7
0.6
0.5

Figure 3.2 After removing the batch effect, clustering of samples correlate with their phenotypes. **A)** The sample to sample correlation heatmap of eBL, sBL tumors, and long-term BL cultures after removing all unwanted covariates except tumor origins. Gene expression profiles group into two clusters based on endemic and sporadic origins. 2 sBL biopsies that we sequenced (sBL_u1 and sBL_u2) and four originally sBL cultures group with the primary sBL biopsies while our 28 eBL biopsies cluster with Namalwa, Daudi, and Raji which have African origins. This clustering demonstrates that the expression variation preserved in the data set is solely unique to the origin of the BL tumor rather than different batches or EBV presence in the sample. Batch 1; primary biopsy samples that we sequenced, Batch 2; primary sBL biopsies previously sequenced, Batch 3; previously sequenced long-term BL cultures. **B)** The sample to sample correlation heatmap of eBL, sBL tumors, and long-term BL cultures after removing all unwanted covariates except EBV status of the samples. This time gene expression profiles result in two clusters based on EBV's presence or absence. EBV-negative BL cultures Ramos, BL41, BL70, and CA46 group with EBV-negative 26 sBL samples including two of our sBLs (sBL_u1 and sBL_u2). EBV-positive BL culture Raji, Namalwa, and Daudi cluster with EBV-positive eBL tumors as well as 4 EBV-positive sBL tumors. Again, this sample-wise clustering demonstrates that we have successfully removing unwanted covariants while preserving variable of interest, such as EBV status in this case.

We then adjusted the expression data by regressing the surrogate variables via matrix reconstruction. In parallel, to test individual genes for significant differential expression, we constructed the design of DESeq2 by incorporating these surrogate variables into the testing model. This enables to account for unwanted covariates without disturbing the assumptions of gene expression based on read count distributions in the process of testing for significance.

3.3 Determining Mutations in Expressed Genes using RNAseq

Although sequencing technologies improve rapidly, whole-genome DNA sequencing is still a costly method to detect genomic alterations and mutations. Because the majority of disease-associated single nucleotide variations (SNVs) occur in protein-coding regions, it is an appropriate approach to search for mutations in expressed exons using RNA-seq data.

When a comparative analysis is performed regarding the mutation frequency differences using RNAseq, the expression level of the gene needs to be controlled. Since different cell types might have different expression profiles, this might occur as a differential coverage over exonic regions of genes. Thus, this should be controlled to avoid false negative calls from genes that are not naturally expressed under certain conditions or phenotypes. Besides, spliced aligners sometimes fail to align reads which span splice junctions correctly. Such reads create "dangling" read stacks at exon edges which often mismatch with the intronic sequence. Similarly, RNA-editing sites are changes on transcribed mature mRNA molecules and often end up being called as mismatches relative to genomic reference DNA. All of these issues need to be carefully identified and appropriately handled when RNAseq data is used for SNV calling.

In BL mutational profiling study, we started with preprocessing the RNAseq data. Redundant read pairs as a result of PCR amplification of libraries prior to sequencing were detected and removed using Picard tools (Wysoker, Tibbetts, and Fennell 2013). Then, reads were mapped to reference genome hg19 using STAR spliced read aligner (Dobin et al. 2013b) with a 2-pass approach to improve alignment across splice junctions. In the first pass, all reads are aligned to genome just to determine the splice junctions. Then, in the second pass, all of the reads are aligned to the same genome again with the help of detected splice junctions in order to increase read alignment rate, the sensitivity of spliced alignments, and overall coverage. In addition, for improved variation calling, we trimmed trailing bases with Phred quality scores less than 20 starting from 3' end of each read. To control for the differences in the sequencing depth between libraries, we randomly paired one eBL and one sBL tumor, and sub-sampled reads from the higher depth sample of a pair, so that the average depth per base for a pair was equivalent. We then called the single nucleotide variations (SNVs) using GATK. To assess the sub-sampling performance, to ensure comparability, we examined the re-discovery rates of known germline SNPs included in dbSNP v146, which showed equivalent sensitivity for SNP call rates across the paired samples, with the average number being slightly higher in eBLs compared to sBLs (6228 SNPs vs. 5690 SNPs, respectively). The frequency of detected non-synonymous mutations was also comparable.

Following the standard workflow by GATK (McKenna et al. 2010b) for calling variation with RNAseq data, we prepared alignments by processing reads that span intronic regions, realigning reads that carry InDels and recalibrating base qualities. After realignment and calibration of bases, we excluded reads that had low mapping quality (<20). Then, we called variations by utilizing the UnifiedGenotyper module in GATK with options -stand_call_conf 20.0, -stand_emit_conf 20.0. In addition standard to remove SNP clusters of at least 3 variants within 15nt window, we applied filter settings variance quality/confidence (QD < 2.0), mapping quality (MQ < 20.0), read depth (DP < 5.0), Phred-scaled p-value using Fisher's exact test detecting strand bias (FS > 30.0). We also enforced conservative fragment depth counts instead of per read. Variants included in dbSNP version 146 and low-quality calls (mapping or base) were excluded. We limited variant calling to translated sequence of protein coding genes in GenCode annotation version 19. Due to their highly variable structures, we excluded immunoglobulin (IG_V), histocompatibility (HLA) genes, uncharacterized proteins, and known pseudogenes to reduce false calls. Splice region variation calls caused by failed junction spanning read alignments were filtered out in addition to variants overlapping with the repetitive regions (RepeatMasker, UCSC). Variant calls reoccurring at the exact genomic location with high frequencies also in the lymphoblastoid cell line (LCL) RNAseq datasets were considered as possible artifacts due to mismapping, undocumented variation, or RNA editing and were excluded. The effect of the variations was predicted as “low,” “moderate,” and “high” using snpEFF (Cingolani et al. 2012) based on protein sequence changes.

3.4 End Sequencing as an Alternative to Full-length RNAseq

As new sequencing technologies emerged, new techniques for targeting specific molecules and sequencing their content also specialized. For example degradome-seq (aka PARE; parallel analysis of RNA ends), RACE (rapid amplification of cDNA ends), etc. are modified specific sub-versions of full-length RNA sequencing. These methods not only enrich specific molecular strains but also produce higher coverage over the targeted transcript regions. Depending on the experimental questions, these methods provide sequencing read slightly different than conventional RNAseq methods. Thus, the data generated using such methods might require special care or alternative analysis methods. Polyadenylation Site sequencing (PASseq) is also a method for sequencing only the 3’-end of mRNA molecules for expression measurement purposes (Shepard et al. 2011). As it is evident with its name, this unique sequencing library preparation method utilizes polyadenylation of mRNAs and targets the tails with oligo-dT for enrichment. Sequencing often is done with single-end and reads usually span mRNA cleavage site. Thus, poly-A stretches in sequencing reads need to be trimmed before processing further. One of the major issues with this protocol is the enrichment approach. Genomic DNA or transcribed RNA body sequence can also contain poly-A stretches. These cause problems when hybridization oligo-dTs are used to pull down the molecules. These false enrichment products need to be identified in silico after sequencing and processed with cautious. One of the other purposes of choosing this method is to determine precise locations of the cleavage site, aka poly-A site. This allows researchers to study alternative poly-A site usage under certain conditions and determine site switch events which can be associated with any other phenotype. However, determining these locations with high precision requires special data processing and analysis. Here, we developed an analytical framework to investigate alternative polyadenylation site changes. We implemented this analysis pipeline in multiple programming languages, named it as "**PoolPASS,**" and applied it to several projects.

3.4.1 PASSeq Library Preparation and Sequencing

Although there are various versions of PASseq library preparation, one example is as follows which mostly adapted a previously published method for making libraries from RNA fragments (Heyer et al. 2015). Total RNA integrity was first confirmed by agarose electrophoresis, and polyA tailed RNA was enriched by hybridization with oligo-dT. Following the enrichment, RNA samples were then fragmented to 60-80 nt size via chemical hydrolysis and reverse transcribed with oligonucleotides containing forward and reverse Illumina sequencing primer sites separated by a hexa-ethylene glycol spacer (Sp18) linker. At the 5′ end, each oligonucleotide began with 5'p-GG to promote ligation, followed by five random nucleotides (unique molecular index, UMI) to enable PCR duplicate removal. Each primer also harbored a unique five nt Hamming barcode (BC), allowing for sample multiplexing. Following cDNA circularization with CircLigase I, libraries were PCR amplified (12-14 cycles) and subjected to single-end 100 nt sequencing on the Illumina HiSeq platform.

3.4.2 Data Analysis with PoolPASS

The primary goal of the analysis framework is to accurately measure expression levels of each mRNA isoform using PASseq data. This is achieved by three major steps; first is to preprocess sequencing reads and cleaning the background noise. The second step is to determine and remove internal priming locations. Then, the third step is clustering candidate locations in proximity to putative cleavage sites and reporting normalized expression levels of each location. **Figure 3.3** demonstrates an overview of the PoolPASS analysis framework. The novelty of our approach is to process all samples in parallel and then to pool all candidate polyA locations together for further evaluations. This allows capturing all possible sites to be considered and prevents to miss any rare site switches. For example, a polyA site can be used in only one cell type or under certain conditions and not used at similar levels in another. This pooling scheme we implemented allows lining up all putative locations for every sample and accurately comparing their expression levels.

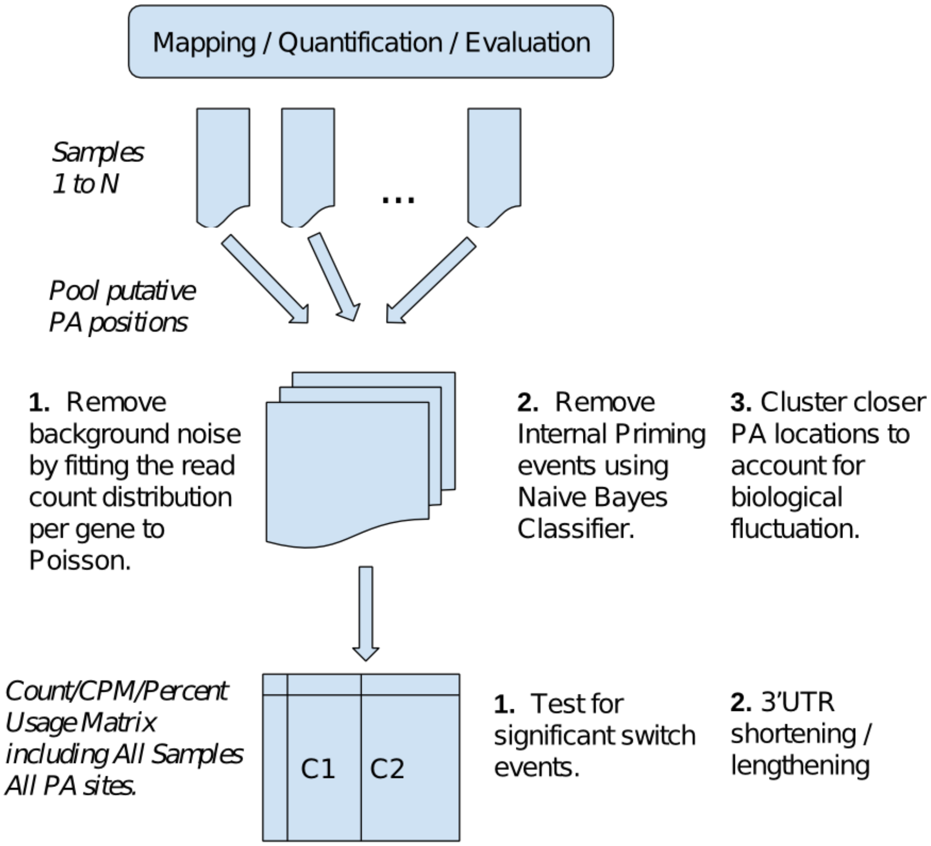

Mapping / Quantification / Evaluation
Samples 1 to N
...
Pool putative PA positions
1. Remove background noise by fitting the read count distribution per gene to Poisson.
2. Remove Internal Priming events using Naive Bayes Classifier.
3. Cluster closer PA locations to account for biological fluctuation.
Count/CPM/Percent Usage Matrix including All Samples All PA sites.
C1
C2
1. Test for significant switch events.
2. 3'UTR shortening / lengthening

Figure 3.3 Overview of PoolPASS analysis framework.

After quality controlling with FASTQC (Andrews 2010), we removed PCR duplicate reads using their unique molecular indexes (UMI). Using cutadapt (M. Martin 2011), we then trimmed residual adaptor/spacer sequences and poly-A stretches from read 3' ends and mapped the reads to the hg19 reference genome using Bowtie2 (with parameters –m –best –p4). Only uniquely mapping reads with high-quality scores (>20) were used for further analysis. We defined the polyadenylation cleavage site as the base closest to the poly-A stretch captured within read sequence, thus, very last base at the 3'-end site after the trimming. Based on this, we calculated read coverage of every genomic location considering only 3'-end of the aligned reads. This forms a sharp peak of read coverage around the cleavage sites with an average width 40nt (ranging between 1nt-120nt). We then clustered sites that are in proximity closer than the 40nt window and summed the counts of clustered sites. In order to determine all possible PAS, which might differ from sample to sample, we pooled all candidate locations of all samples and reiterated the clustering using the same window length. Since genomic DNA containing poly-A stretches can be hybridized and pulled with the oligo-dT primers, aka internal priming, we used a Naïve Bayes classifier (Sheppard, Lawson, and Zhu 2013) based software to determine the likelihood of all sites being false priming sites. In addition to filtering internal priming locations based on this, we also removed background noise caused by widespread base level read alignments by fitting the count distribution of each gene to Poisson distribution. After determining precise PA locations, we then calculated expression/read counts of every PAS for each sample and annotated using Gencode v19. We also marked known polyA sites with PolyA DB (http://exon.umdnj.edu/polya_db/). For differential gene expression tests, we used the sum of all PAS counts of each gene and calculated library sizes based on this. The raw read counts for genes were used as an input for differential gene expression analyses using DESeq2 (M. I. Love, Huber, and Anders 2014c) in R (https://www.R-project.org). The default normalization using 'estimateSizeFactors' function was used. Adjusted p-values were calculated using the Benjamini and Hochberg (BH) method (Benjamini and Hochberg 1995). In order to determine significance levels of alternative polyA site switches, we constructed a two by two contingency table composed of mean normalized read counts of one site and the average of rest of the sites, if there are multiple, in each of the conditions. We tested the significance with Chi-square test for given site and iterated through all sites of the gene. We then corrected p-values for multiple hypotheses testing using BH method.

Code can be reached at https://github.com/yasinkaymaz/PoolPASS.git

Chapter IV. Studying EBV genomes in clinical specimens: from wet lab to in silico

4.1 Summary

This chapter covers commonly used DNA sequencing methods and a variety of approaches developed to solve technical challenges faced with during whole genome assemblies. We established a workflow comprising of the targeted hybrid capture enrichment of viral DNA from the clinical specimens and high-throughput sequencing with next-generation sequencing technology. Analysis of individual EBV genomes consisted of de novo genome assemblies, variation and recombination detection followed by comparative phylogenetic tree constructions.

4.2 Introduction

Applications of sequencing methods in clinical and translational sciences are wide. For example, a patient can be screened through time for diagnostic purposes during or after chemotherapy treatment by sequencing DNA circulating in the bloodstream (Rapisuwon, Vietsch, and Wellstein 2016). A pathogenic outbreak can be resolved by tracking infected patients and finding patient zero with the help of sequencing (Gire et al. 2014). Another example is that sequencing many pathogen genomes assist vaccine development efforts by deciphering divergent or conserved sequences.

Sanger sequencing method has opened many opportunities. The human genome project was completed with shotgun sequencing. Sanger sequencing is a reliable technique with a low error rate. However, sequenceable DNA fragment length, which is a couple kb/run, is limiting. Given average length of a genome or a target genomic region is much bigger, the cost of using this method for one sample is very high. Here, next generation sequencing (NGS) comes into play which solves the problem with massively parallel sequencing. In addition, sequencing cost with NGS rapidly reduces day by day allowing users to generate more data and create new applications.

Targeted sequencing is one of these applications developed for reducing the cost further down and sequencing only the DNA of interest rather than generating redundant data. One advantage of using targeted approach is that it significantly improves the sensitivity of sequencing which allows users to design experiments involving metagenomic screens and to study heterogeneity in clinical samples. Tumor evolution and heterogeneity is one of these areas improving in the light of targeted sequencing approaches and sensitive assays. Researchers are often interested in determining the multiplicity of infection (or complexity of infection) levels when dealing with an infectious disease. A targeted approach with increased sensitivity allows us to detect multiple strains infecting the same individual. We could not even imagine this when the first human genome project was completed at the end of 20th century.

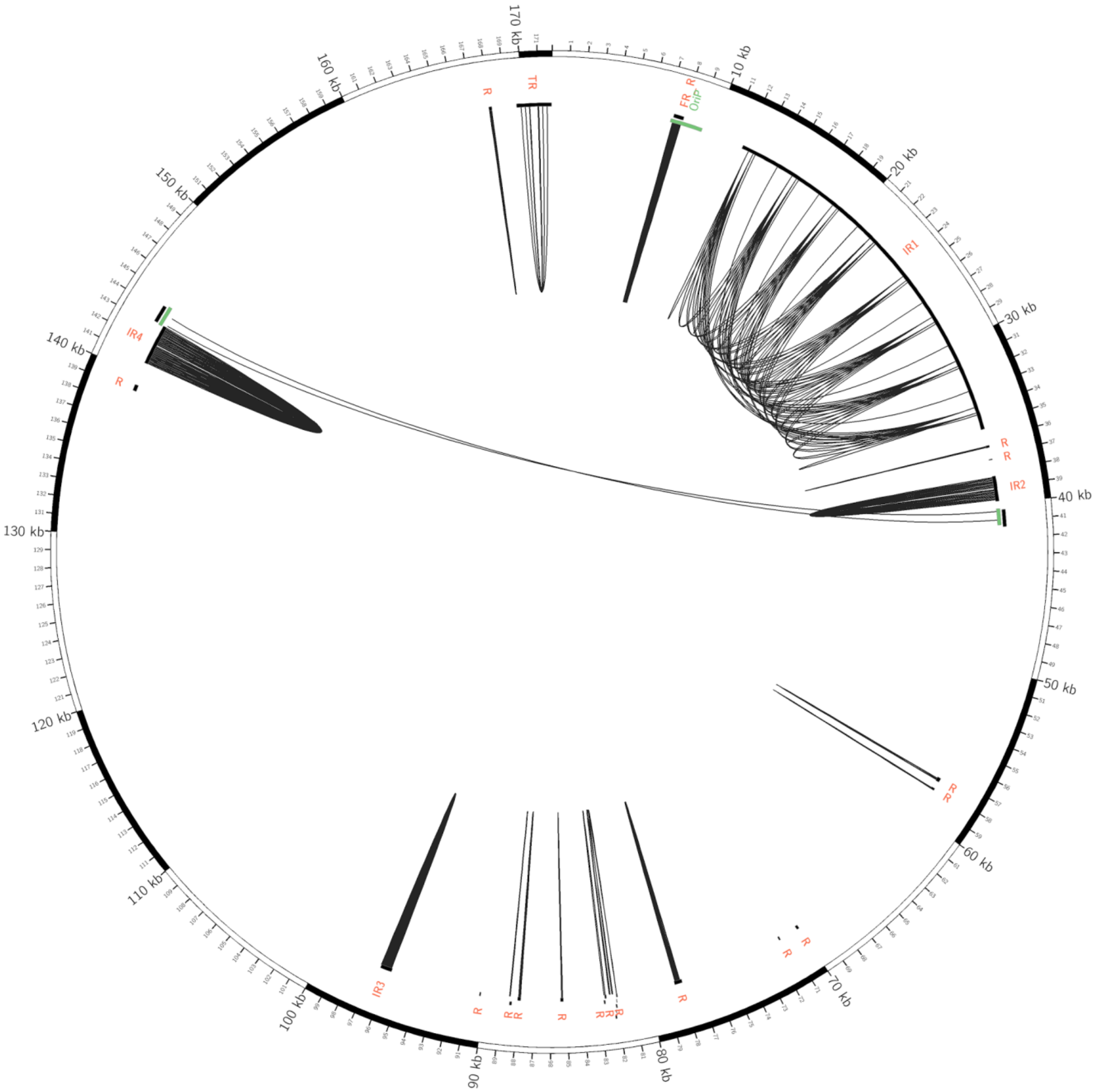


Figure 4.1. Repetitive Structure of EBV genome.

EBV genome is a 172 Kb DNA which is in linear form when encapsidated. Upon infection, the linear genome becomes circularized at the terminal repeats region to protect itself from cellular exonucleases. It has four large repeats regions IR1 to IR4 and short repeat regions (see **Figure 4.1**). Other than being repetitive, these regions reach up to 80% GC content elevating the overall genomic GC percent to 60%. On the other hand, although the multiplicity of infection is commonly observed, the viral genome abundance is still too low compared to the human genome. Unless viral replication is triggered through lytic reactivation, viral DNA is at negligible levels in the cell. In addition, EBV mainly persists in B cells of healthy individuals, and they typically carry 1-50 EBV positive cells out of 1,000,000 B lymphocytes (Khan et al. 1996), which makes abundance even smaller when whole blood is considered. All of these make EBV genome studies challenging and sometimes impossible. Therefore, here in this chapter, I outlined an experimental and computational methodological framework for studying pathogen genome sequences in the context of Epstein Barr virus and provided solutions to overcome several challenges (see **Figure 4.2**).

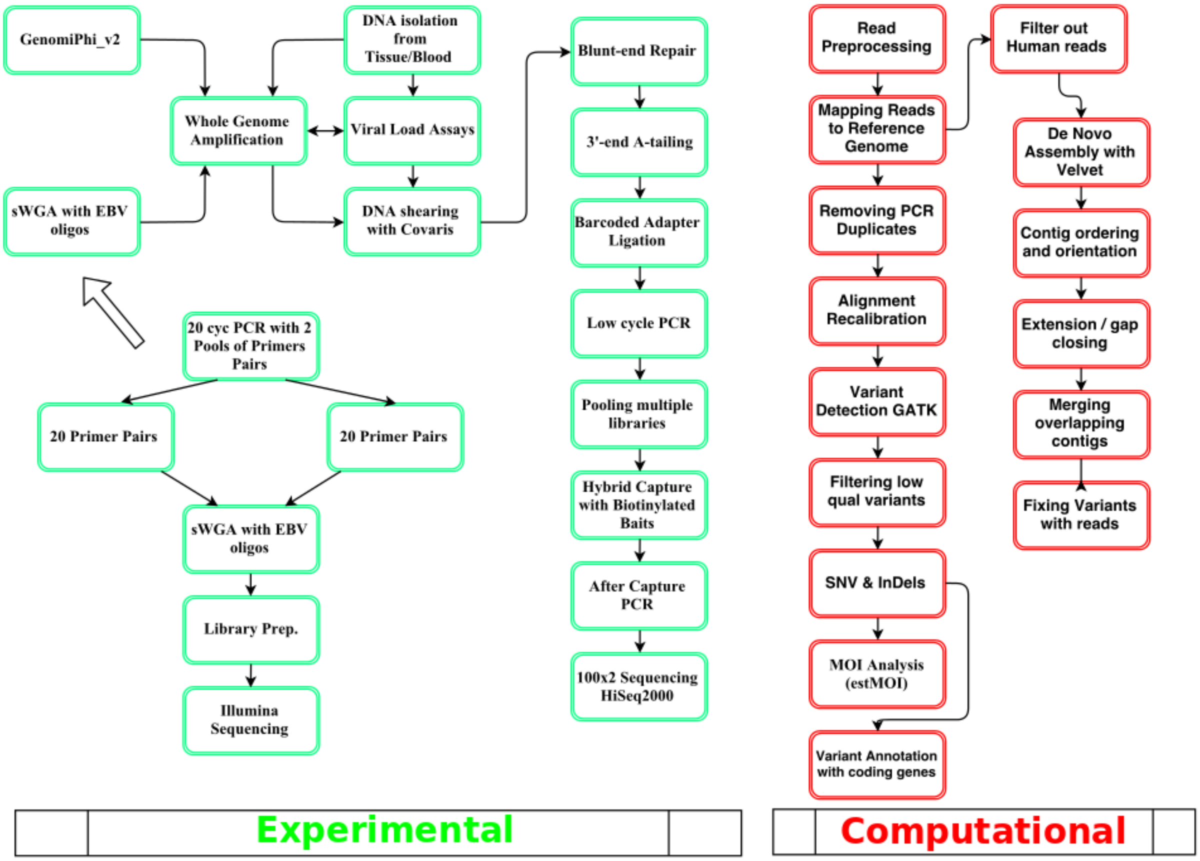

GenomiPhi_v2
DNA isolation from Tissue/Blood
Blunt-end Repair
Whole Genome Amplification
Viral Load Assays
3'-end A-tailing
sWGA with EBV oligos
DNA shearing with Covaris
Barcoded Adapter Ligation
20 cyc PCR with 2 Pools of Primers Pairs
Low cycle PCR
20 Primer Pairs
20 Primer Pairs
Pooling multiple libraries
sWGA with EBV oligos
Hybrid Capture with Biotinylated Baits
Library Prep.
After Capture PCR
Illumina Sequencing
100x2 Sequencing HiSeq2000
Read Preprocessing
Filter out Human reads
Mapping Reads to Reference Genome
De Novo Assembly with Velvet
Removing PCR Duplicates
Contig ordering and orientation
Alignment Recalibration
Extension / gap closing
Variant Detection GATK
Merging overlapping contigs
Filtering low qual variants
Fixing Variants with reads
SNV & InDels
MOI Analysis (estMOI)
Variant Annotation with coding genes
Experimental
Computational

Figure 4.2 EBV sequencing Pipeline overview.

4.3 Targeted Pathogen Genome Sequencing

There are multiple ways to sequence pathogen genomes. In molecular biology, newly developed techniques let us manipulate DNA composition of samples and specifically work with certain DNA types. One of the sequencing approaches is metagenome sequencing which involves sequencing all types of DNA molecules present in the sample. However, depending on their relative abundance, the likelihood of getting a good quality sequence of every DNA strains varies. Thus, sensitivity decreases dramatically to low levels for some pathogen genomes because of highly abundant other, mostly host, genomic DNAs. To overcome this issue, culturing cells and pathogens organisms for producing abundant DNA is one option. This approach relatively increases pathogen DNA abundance, however; it might introduce bias to an experimental design by promoting only replication competent strains thus by losing natural mixture frequencies. One other approach is conventional PCR method to increase target DNA amount before sequencing, aka amplicon sequencing. PCR requires prior information about the target sequence to design primers this it can only be applied to already sequenced genomes. The most significant limitation of conventional PCR, which requires abundant template material, is the size of the target region that can be amplified. In addition, it is difficult to target and amplify highly variable and polymorphic DNA regions. This limits sequencing to certain regions and requires multiple reactions in case whole genomes are desired. Optimization and balancing all PCR products make conventional amplicon sequencing approach very challenging and labor intensive.

Recent developments in molecular techniques improved enrichment methods. These enrichment techniques were initially applied in exome capture sequencing studies in which only known exonic regions of genomes, roughly 1%, are sequenced. Short single stranded RNA molecules carrying covalently attached biotin called RNA baits or probes complementary to exonic regions in a tiling fashion allows hybridizing genomic DNA in a solution. Then, a pulldown with a streptavidin attached magnetic bead is used to select only DNA:RNA hybrids with biotin specifically. This technique dramatically increases sensitivity and read coverage of the targeted regions. Although, it also requires prior sequence and content information about the targeted genomes, the dynamic hybridization of RNA baits to target DNA can tolerate mismatches or polymorphic sites. This way, more variation can be captured in the DNA mixture.

4.4 An Algorithm for Designing Hybrid Capture Bait Sequences

Although hybrid capture probes are more tolerant for mismatches as opposed to conventional PCR primers, they also have a limit for mispairing. Highly divergent sequences can possibly be missed or at least captured with low efficiency. Therefore, the optimization of such hybrid capture probe sequence design is essential. For this purpose, we developed an algorithm and implemented it as a tool for users to design their own probe sequence set for desired genomes with increased sensitivity. We developed this with certain improvements in communication with Agilent Inc. representatives. Here, the purpose is to create custom design RNA bait sequences targeting specific genomes or regions. In addition to the reference sequence, another source of sequences for the same target can be included to compensate for rare sequence variations like mutations, indels, etc. For example, these probes can also be produced using the genomes of cultured lab strains. The main steps implemented in the algorithm are as follows;

1. Input multiple target genome sequence as a fasta file.
2. Check cross-hybridization with Blast and mask regions homologous to human.
3. Create candidate probe sequences with user defined parameters such as tiling (2X, 3X, 4X, etc.), length of the probes (default: 120 nt)
4. Calculate GC content and an annealing temperature of each candidate probe sequence.
5. Map candidate probes back to target genome.
6. Check for mapping and alignment quality (repetitive regions or mismatch rate)
7. Filter candidate probes with mismatch rates above given threshold.
8. Calculate expected coverage of target window.
9. Create new probe sequence for low coverage regions.
10. Update the set of probe sequences
11. Repeat 2.-10. Iterate through all input target genomes.
12. Boost up probe numbers depending on GC content and annealing temperature.
13. Output bait sequence set.

Code is available at

https://github.com/yasinkaymaz/TackleBox/TargetProbeDesign_pipeline.py

This tool can be used for any pathogen genome especially when increased sensitivity for mixed infections is desired. We first tested it with EBV strains. In addition to type I (NC_007605) and type I reference genomes (NC_009334), we included recently sequenced Mutu I, Akata, GD1 and GD2 sequences in in-silico probe design. The overlapping probes were 120 nt in length tiling across the genomic sequences at least four times. As a result, we created 6,564 unique probe sequences with a various number of representations based on GC content of the target regions. A total of 55,000 biotinylated RNA oligos were ordered for capture reactions and produced by MyBait Mycroarray Inc.

4.5 Experimental

4.5.1 Whole Genome Amplification

One of the applications of targeted deep sequencing is the detection of all known drug-resistant pathogen strains in clinical samples as frequently practiced with CMV or influenza. Whole genome sequencing provides various useful information such as antigen epitope sequences, evidence for recombination between sub-strains, or novel drug-resistant variants. Studying viral genomic evolution in patients over time and changes to epitopes can be feasible with whole genome sequencing. Deep sequencing refers to a genomic sequencing with the extremely high depth of coverage. Since the number of reads generated by the instrument is proportional to the content of the DNA in the library, increasing the sequencing depth allows detecting low frequent strains with better sensitivity. In some situations, pathogen strains carrying variants with less than 1% can be quickly dominant under certain selective pressures such as drug treatment. Therefore, detecting all variants ultimately in the mixture for example at early stages of treatment would be vital.

Some species are hard to isolate and culture. Often, culturing is not desired since there is always a chance to bottleneck and select only sub-strains that are replication competent. Therefore, alternative molecular techniques are explored to recover or increase the abundance of desired genomes. Whole genome amplification with multiple strand displacement does not necessarily increase the sensitivity in favor of pathogen genomes since it happens randomly. Besides, genomic regions with high GC content might be underrepresented. Thus, amplification of EBV which has extreme GC content regions (reaching up to 80%) is problematic. Illumina sequencing relies on cluster generation with amplification and extension based sequencing which also affect read representation of such regions. Selective whole genome amplification (sWGA) is a method that utilizes target specific oligos rather than random hexamers. This method has been suggested by Leichty et al. 2014 and takes advantage of more or less specific motif sequences by using them as primers for amplification. The method also uses phi29 polymerase which works at isothermal temperature with high fidelity, 100 times less error rate than regular Taq polymerase, which also makes it appealing for whole genome sequencing. Strand displacement activity of phi29 allows it to be progressive, and it can amplify up to 70kb products. Multiple strand displacement with target specific oligos leads to rapid branching and quick amplification.

4.5.2 Preamplification PCR and Selective WGA

In order to evaluate whole genome amplification techniques and optimize for EBV genome sequencing, we conducted multiple experiments with various reaction conditions. We conducted experiments to improve and optimize reactions. Then, we tested this method on cultured cell line DNA and primary clinical specimens. BL culture cells, Namalwa, Daudi, Raji, and Jijoye were grown in a complete growth medium, RPMI 1640 (Life Technologies), with 2mM L-glutamine adjusted to contain 1.5g/L sodium bicarbonate, 4.5g/L glucose, 10mM HEPES, and 1.0mM sodium pyruvate, 92.5%; fetal bovine serum, 7.5%. Blood samples were collected on the admission of the patients before treatment with chemotherapy, and the samples were stored at -80C prior to DNA extraction. FNA biopsies were transferred into RNAlater immediately after collection and stored at -20C. Genomic DNA was extracted from cultured cells and FNA biopsies stored in RNAlater using Qiagen Allprep DNA/RNA/Protein mini kit (QIAGEN Sciences, Germantown, MD) according to manufacturer's instructions. DNA from plasma samples were extracted using Qiagen Qiaamp DNA Kit. Prior to library prep or amplification, DNA extracts were purified with 2x XP-Ampure magnetic beads to remove residual inhibitory chemicals that may be introduced during the extraction process.

As standard whole genome amplification, we majorly followed the manufacturer’s instructions for whole genome amplification with Genomiphi v2 kit. We used 20 ng template DNA for amplification and incubated at 30C for 16h. Before and after genomic amplification of samples, the ratio of viral genomic copy number to human genomic DNA was checked with bi-plex qPCR using EBV specific primers against BALF5 gene and human beta-actin gene. Overall DNA quality and quantities were assessed with NanoDrop and Picogreen. The challenge was to increase EBV amplification yield by increasing the EBV to human DNA copy ratio after whole genome amplification. In order to overcome this challenge, we replaced the random hexamers required for MDA with EBV genome specific oligos designed specifically for isothermal temperatures. We designed the oligo sequences using a script provided by Leichty et al. using human genome as a background and EBV as the foreground. We ordered the oligos with 3’-end modification to protect from exonuclease activity of Phi29. See appendix for protected oligo sequences.

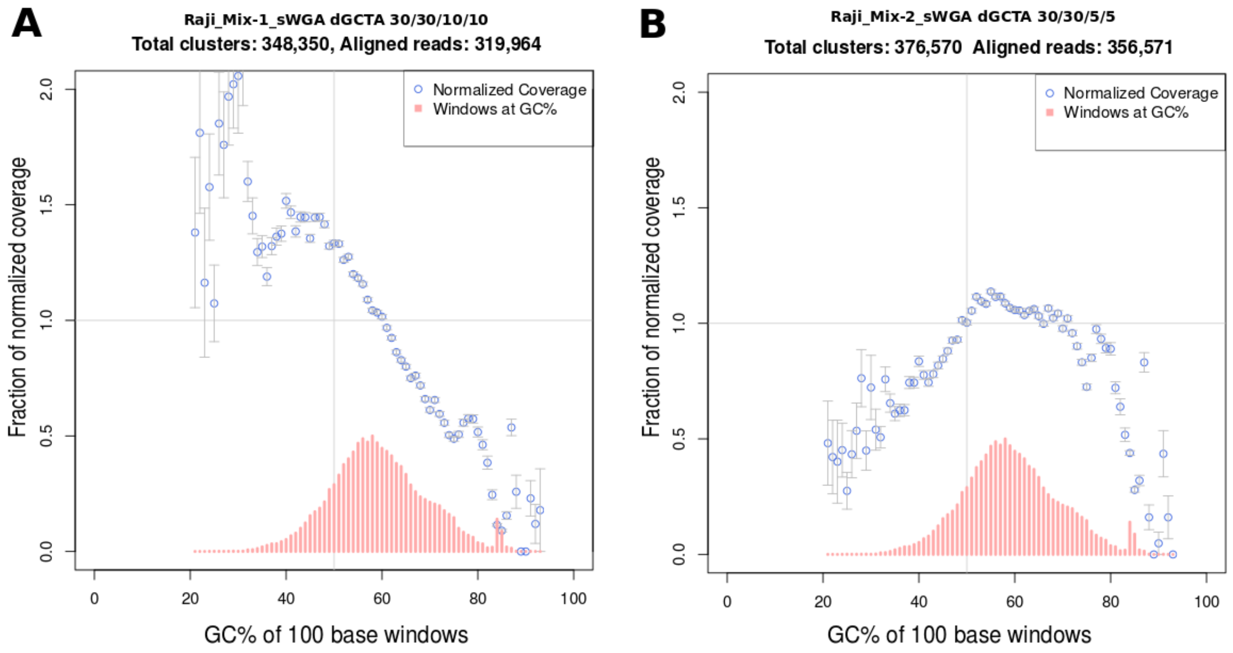


Figure 4.3 The effect of modified dNTP composition on GC high region coverage. Modifying the composition of deoxyribonucleotides in dNTP increases the EBV genomic coverage over GC high regions.

Table 4.1 Optimization of selective EBV whole genome amplification reaction conditions. Various factors such as dNTP composition, the amount of template DNA, denaturation buffer, incubation time and temperature were tested.

Optimization	EBV Copy input per 20ng DNA for sWGA	Denaturation Buffer	Input EBV copy/uL (Average)	Input B-actin/uL (Average)	dNTP composition (G/C/T/A)	Incubation	Output EBV copy/ ng DNA	Output B-actin/ng DNA	(EBVpost / EBVpre) / (DNApost / DNApre)	(Humanpost / Humanpre) / (DNApost / DNApre)
Sensitivity, dNTP Composition	10,000	TE	62,448	3,456	30/30/10/10	16h at 30C	4,573	7	1.06	0.03
	1,000	TE	10,038	13,911	30/30/10/10	16h at 30C	505	59	1.01	0.08
	100	TE	467	10,883	30/30/10/10	16h at 30C	55	70	2.28	0.12
	10	TE	10	11,326	30/30/10/10	16h at 30C	4	68	8.26	0.12
	1	TE	6	19,483	30/30/10/10	16h at 30C	0	55	0.21	0.06
	0	TE		11,551	30/30/10/10	16h at 30C		76		0.23
	10,000	TE	62,448	3,456	30/30/5/5	16h at 30C	4,847	2	1.12	0.01
	1,000	TE	10,038	13,911	30/30/5/5	16h at 30C	544	75	1.09	0.11
	100	TE	467	10,883	30/30/5/5	16h at 30C	72	65	3	0.12
	10	**TE**	**10**	**11,326**	**30/30/5/5**	**16h at 30C**	**3**	**65**	**5.97**	**0.11**
	1	TE	6	19,483	30/30/5/5	16h at 30C	0	81	0.22	0.09
	0	TE		11,551	30/30/5/5	16h at 30C		71		0.21
	10,000	TE	62,448	3,456	15/15/5/5	16h at 30C	3,483	16	0.81	0.07
	1,000	TE	10,038	13,911	15/15/5/5	16h at 30C	488	84	0.97	0.12
	100	TE	467	10,883	15/15/5/5	16h at 30C	73	109	3.04	0.19
	10	TE	10	11,326	15/15/5/5	16h at 30C	1	80	2.71	0.14
	1	TE	6	19,483	15/15/5/5	16h at 30C	0	103	0.25	0.11
	0	TE		11,551	15/15/5/5	16h at 30C		118		0.35
	10,000	TE	62,448	3,456	15/15/2/2	16h at 30C	3,324	12	0.77	0.05
	1,000	TE	10,038	13,911	15/15/2/2	16h at 30C	533	77	1.06	0.11
	100	TE	467	10,883	15/15/2/2	16h at 30C	78	105	3.24	0.19
	10	TE	10	11,326	15/15/2/2	16h at 30C	4	105	7.8	0.18

	1	TE	6	19,483	15/15/2/2	16h at 30C	0	110	0	0.12
	0	TE		11,551	15/15/2/2	16h at 30C		108		0.32
Denaturation Buffer, Incubation time and temperature	10	TE	10	11,326	30/30/5/5	8h at 30C	2	40	6.31	0.14
	10	TE	10	11,326	30/30/5/5	16h at 30C	2	40	8.74	0.14
	10	TE	10	11,326	30/30/5/5	8h at 35C	1	22	4.05	0.08
	10	TE	10	11,326	30/30/5/5	16h at 35C	2	22	6.14	0.08
	10	TE+Q sol	10	11,326	30/30/5/5	8h at 30C	2	41	9.44	0.14
	10	**TE+Q sol**	**10**	**11,326**	**30/30/5/5**	**16h at 30C**	**3**	**60**	**13.94**	**0.21**
	10	TE+Q sol	10	11,326	30/30/5/5	8h at 35C	3	28	10.72	0.1
	10	TE+Q sol	10	11,326	30/30/5/5	16h at 35C	4	49	**15.13**	0.17

One of the possible reasons for low amplification yield of EBV DNA is the GC content of the genome. Therefore, we put this fact into the center of our optimization efforts and tested various parameters accordingly. We hypothesized that if we change the ratio of deoxyribonucleotides used in reactions in favor of guanine and cytosine, the polymerase will get a higher chance to proceed when it reaches to high GC regions of DNA. To test this hypothesis, we conducted experiments with varying deoxyribonucleotide compositions such as 30/30/10/10 or 30/30/5/5 for dG/dC/dA/dT in the same order. We found that 30/30/5/5 composition yields relatively better EBV amplification which we measured as an increase in EBV DNA copy over the increase in human DNA copy (see **Figure 4.3** and **Table 4.1**). We also tested low copy template DNA with serially decreasing input EBV copies to determine sensitivity level of the amplification. Our results show that we were able to successfully amplify ten genomic copies per 20 ng total input DNA with almost six fold increase. Since we aimed to optimize amplification yield by adjusting conditions for GC rich contents, we tested Q-solution provided by Qiagen to determine if it improves overall yield. Compared to using TE buffer in template denaturation using Q-solution significantly increased the amplification yield for EBV genome. We also tested various incubation time and temperatures relying on the fact that elevated temperature might help to keep denatured DNA free of tertiary structures. We found that increasing isothermal incubation temperature to 35C from 30C, and 16 hours longer incubation yielded slightly better (see **Table 4.1**). However, we decided to proceed with the recommended temperature for Phi29 because thermostability of the enzyme needed to be monitored and further validation experiments required. Nevertheless, this suggests that the reason for under-representation of EBV genome DNA is most likely its high GC content when whole genome amplification method is used, and this further supports our hypothesis above.

Even though we improved the amplification yield by optimizing reaction conditions for EBV genome (see appendix for reaction conditions), this overall slight increase does not result with final virus DNA amount enough for library preparation and sequencing. At this point, we decided to test one approach involves conventional PCR primers designed for EBV genome. We named this approach as "Preamp" which stands for the preprocessing step right before whole genome amplification. In this pre-amplification, we utilized multiple pools of primers and conducted low cycle number PCR to boost up target DNA. We hypothesized that this preamp step would increase the template amount for the following specific whole genome amplification reaction. We borrowed primers sequences from (Hin Kwok et al. 2012a) and added a few new primers to be able to capture type 2 EBV strains (see appendix for primer pools). We observed that three separate pools of non-overlapping PCR primers work better compared to one single pool of all tiling primers. We tested varying cycle numbers 5, 10, 15, and 20 for the preamp. We found that increasing cycle number also increases hybrid capture efficiency and sequencing read coverage. However, increasing it further might also cause overly non-uniform read coverage distributions which produce shorter assembly amplicons (will discuss in the following parts). Therefore, we found 20 cycles as optimum for preamp reaction (see appendix for preamp reaction conditions). We also tested forward only or reverse only primers. Although the amount of yield was high with single primers compared to paired primers, the coverage distribution of those was extremely non-uniform which is not desired for genome assembly (see **Figure 4.4** and **Table 4.2**). In addition, we also tested three different polymerases, HotStar Taq, Long Range PCR polymerase, and Phusion. We found that the best working polymerase is Long Range PCR polymerase as expected because the average amplicon length is around 4.5Kb.

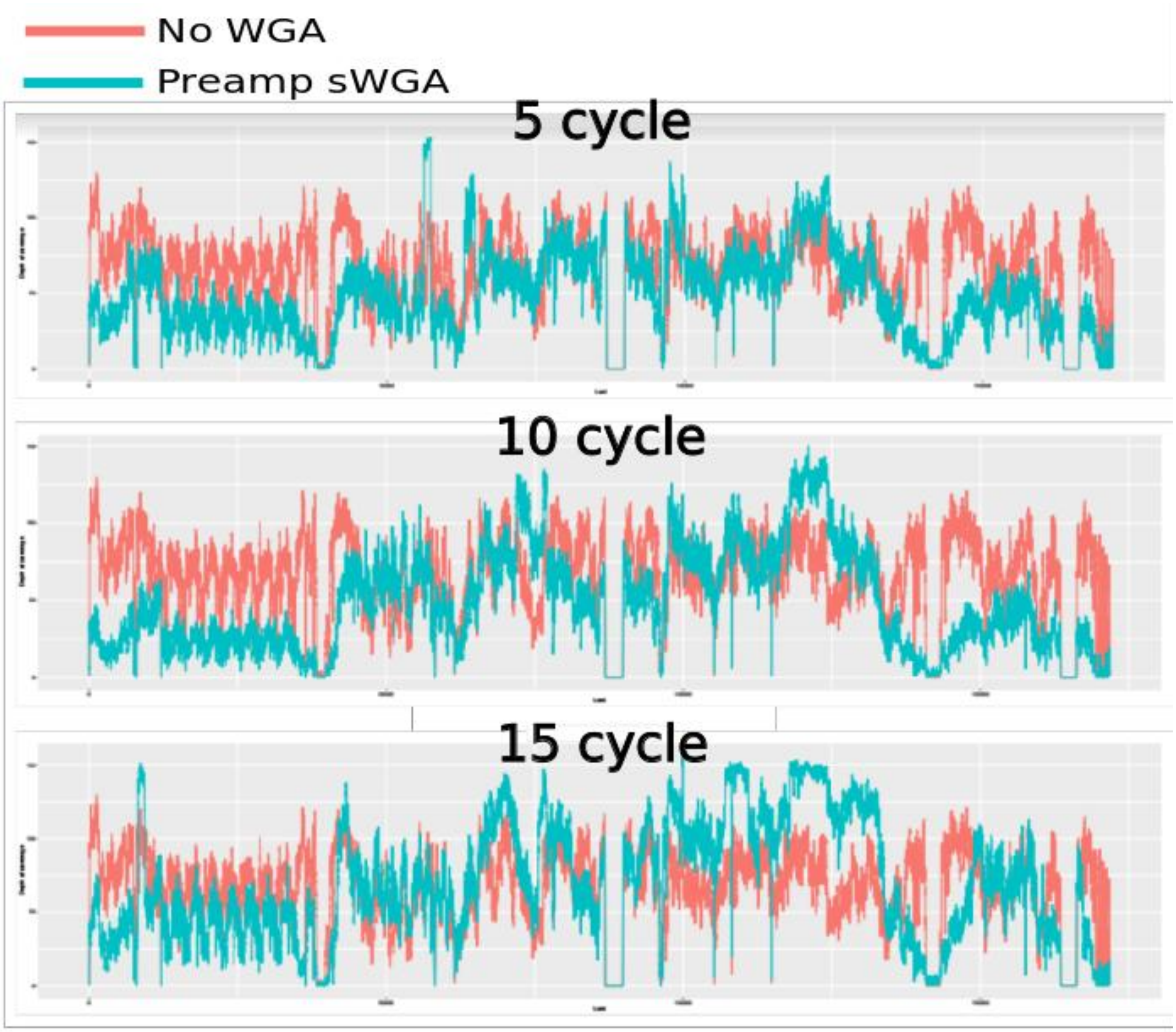


Figure 4.4 Sequencing coverage distribution over EBV genome comparing preamp-sWGA libraries with different cycle amplifications.

Table 4.2 Sequencing read statistics comparing preamp-sWGA libraries with different cycle amplifications.

		EBV	EBV	HUMAN	HUMAN
	READS	READS ALIGNED	% READS ALIGNED	READS ALIGNED	% READS ALIGNED
Raji	287,091	259,209	90.29%	24,592	8.57%
Raji_GP	1,270,227	1,040,137	81.89%	209,533	16.50%
Raji_GP_GC	530,698	417,933	78.75%	103,691	19.54%
Raji_Mix-1_sWGA	348,350	319,903	91.83%	22,757	6.53%
Raji_Mix-2_sWGA	376,570	356,406	94.65%	8,425	2.24%
Raji_5cyc_sWGA	2,451,526	250,133	10.20%	2,141,720	87.36%
Raji_10cyc_sWGA	642,930	194,795	30.30%	434,507	67.58%
Raji_15cyc_sWGA	1,014,156	506,977	49.99%	480,917	47.42%
Raji_Fw_sWGA	1,728,634	1,451,336	83.96%	237,268	13.73%
Raji_Rev_sWGA	6,245,778	5,081,132	81.35%	934,634	14.96%

4.5.3 Sequencing Library Preparation and Target enrichment

We prepared Illumina sequencing libraries using a custom protocol (see appendix B). Briefly, the protocol consists of steps for DNA shearing, blunt end repairing, 3'-adenylation and indexed sequencing adaptor ligation. After PCR amplification, the libraries for each sample were pooled at equal EBV genomic copy quantities. Pooled libraries were hybridized with custom design EBV sequence specific biotinylated RNA baits, produced by Mycroarray Inc. Hybridized DNA libraries were captured with streptavidin beads and purified for final PCR amplification using Kapa HiFi polymerase. During the sequencing library preparation, we checked DNA quality and fragment size distribution using Bioanalyzer Agilent 1000 kit. Final library qualities were confirmed with Bioanalyzer Agilent High Sensitivity DNA kit. Multiplexed DNA libraries were sequenced in multiple lanes of Illumina MiSeq/HiSeq2000/NextSeq 500 platform with 1x75bp, 2x100bp, and 2x150bp, respectively. Qualities of raw sequencing reads were assessed using FastQC (Andrews, n.d.). Pooled reads with unique sample barcode sequences were separated using Novobarcode (Novocraft Technologies, Malaysia). Residual Illumina adaptor sequences were trimmed using cutadapt (Martin, 2011). The overall capture efficiency was around 40%. This might be because of pooling many libraries.

4.6 Computational

4.6.1 Genome Assembly

Key points of a good genome assembly following re-sequencing are a good quality of DNA material, high base quality of sequencing reads, and uniform read coverage representation throughout the genome. Residual sequence tags or adapter sequences can easily interfere with assembly process and create false calls. Prior to de novo assembly of viral genomic sequences, we first aligned adaptor/tag cleaned reads to the human genome in order to prevent possible chimeric assemblies. In addition to base quality trimming starting from 3'-end of reads, we also observed that removing ultra-low complexity reads, which contain long homopolymer stretches possibly resulted from sequencing errors, would improve contig assembly. However, this should be applied with cautious since EBV genome might have such long homopolymer stretches naturally.

Contigs were generated with de novo sequence assembly tool Velvet (Zerbino and Birney 2008) with VelvetOptimiser (Consortium and Others 2012) with Kmer search ranging from 21 to 149, depending on read length. We used the Kmer that maximized N50, which is defined as the length of the shortest contig among the set of contigs whose sum makes up to 50% of all contig lengths. After initial assembly, we put contigs in genomic order according to the reference genome using ABACAS and extended their lengths when possible with additional read supports using IMAGE from PAGIT (Post assembly genome improvement toolkit) (Swain et al.

2012) (Swain et al., 2012). We further merged possibly overlapping regions in guidance with reference genomes using in-house scripts. By nature, Illumina short reads (50-250bp) fail to resolve long repetitive regions, and such regions create misassemblies. Therefore, we masked these regions to avoid possible errors. We then constructed the scaffolds out of contigs and aligned original sequencing reads back to these scaffolds separately for every sample. Read alignments back to assembled contigs allowed us to check the presence of alternative variants not represented in final assemblies. By considering the frequency of variant bases at each position, we incorporated the major variant nucleotide into final contigs in the case of multiple allele presences. As a result, this ensures that the assembled genomes are the representatives of major strains even if there is a mixture in the sample.

We deposited all analysis pipeline and custom scripts to GitHub under a repository named “EBV_Assembly_Pipeline”:

https://github.com/yasinkaymaz/EBV_Assembly_Pipeline.git

EBV_SequenceAnalysisPipeline.sh: A bash script for pipelining and processing for EBV sequencing datasets.

4.6.2 Estimation of Genome Assembly Error Rate.

The polymerase used during whole genome amplification, Phi29, has an error rate around 1 in 10^6-10^7 (Esteban, Salas, and Blanco 1993). Although a polymerase with this error rate is considered as high fidelity, the sequencing technique and assembly methods also contribute to misassignment of bases to the final result. Therefore, we designed a set of controlled experiments to investigate errors and estimate rates by comparing the same genomes sequenced using different amplification methods. One of the assessment was regarding the errors introduced during the whole genome amplification. For this purpose, we replicated 5 of the sequenced primary biopsy associated viral genome sequences with amplification and included both type 1 and type 2 strains as well as a known lab strain Raji. After data processing and assembly processes mentioned above, we compared genomes using multiple sequence alignment using Mafft (Katoh and Standley 2013). Then, we calculated the different bases between each other throughout genomes by excluding repetitive regions of EBV. In average, we found that the substitution rate between the whole genome amplified genomes using GenomiPhi v2 and their un-amplified counterparts was ~2.2E-05 (see **Table 4.3**). We also estimated the error rate associated with preamp step prior to sWGA as ~1.6E-04, roughly ten fold increased, compared to standard WGA. This was an expected outcome since we used a polymerase with relatively lower fidelity in the preamp reaction.

Table 4.3. Sequencing error rate estimations based on a number of substitutions between replicates. *Assemblies from Palser et al. 2015

Replicate 1	Replicate 2	Number of Substitutions	Number of Perfect Matches	Error Rate
DNA1_wga_EBV_type1	DNA1_EBV_type1	1	140069	7.14E-06
DNA2_wga_EBV_type2	DNA2_EBV_type2	4	142932	2.80E-05
DNA3_wga_EBV_type2	DNA3_EBV_type2	1	142154	7.03E-06
DNA4_wga_EBV_type1	DNA4_EBV_type1	5	141643	3.53E-05
DNA5_wga_EBV_type1	DNA5_EBV_type1	5	141643	3.53E-05
Daudi_CellLine	Daudi_D100_Preamp-sWGA	21	133445	1.57E-04
Raji_CellLine_longRead	Raji_CellLine_shortRead	3	136363	2.20E-05
Raji_CellLine_longRead	Raji_GenomiPhi	3	136454	2.20E-05
Raji_Rep1_Preamp-sWGA	Raji_Rep2_Preamp-sWGA	22	134478	1.64E-04
Raji_CellLine_longRead	Raji Assembly*	5	136145	3.67E-05
Daudi_CellLine	Daudi Assembly*	0	132782	0

4.6.3 Detection of Nucleotide Variation and Mixed Infection Analysis

With Illumina sequencing, haplotype phasing is limited to insert size which is the fragments of DNA sequenced (typically 300-400bp). Thus, informative haplotype calls are limited. However, individual variant bases can be precisely determined with a certain depth of coverage levels. Another essential component of variant detection is the presence of a reference genome. In the case of EBV, there are two existing reference genomes in NCBI database one for each subtype, NC_007605 and NC_009334, type 1 and type 2, respectively. Although, type 1 reference genome is a chimeric version of two EBV strains, B95-8 and Raji, type 2 reference is a complete with type stain from BL cell culture AG876 (Dolan et al. 2006). Given that both of the reference sequences are stains existed in the wild, the variant bases detected by comparison to references can also be interpreted as distances between strains.

In order to detect sequence variations between EBV genomes, we aligned reads against both reference genomes using bowtie2 (Langmead and Salzberg 2012) and marked PCR duplicate reads. After base recalibration and indel realigning, we called single nucleotide variations using GATK (McKenna et al. 2010c) with all alignment files simultaneously. We then filtered low quality and likely false calls due to repetitive sequences. For positions more than five non-PCR duplicate reads, if the variant frequency was >=95%, we flagged them as homogenous. If the variant frequency was between 20 and 94%, we considered it as heterogeneous loci (Hin Kwok et al. 2012b), and ambiguous otherwise.

Table 4.4 Sequencing read statistics of BL cell line mixture libraries.

Library ID	Viral subtype	Read Length	Total Reads	Note
J10D90_Rep1_PCRsWGA	Type1	300	496,533	
J10D90_Rep2_PCRsWGA	Type1	300	230,979	
J25D75_Rep1_PCRsWGA	Type1	300	529,976	
J25D75_Rep2_PCRsWGA	Type1	300	337,495	
J50D50_Rep1_PCRsWGA	Type2	300	920,063	
J50D50_Rep2_PCRsWGA	Type2	300	774,291	
J75D25_Rep1_PCRsWGA	Type2	300	1,303,011	
J75D25_Rep2_PCRsWGA	Type2	300	820,828	
J90D10_Rep1_PCRsWGA	Type2	300	840,718	
J90D10_Rep2_PCRsWGA	Type2	300	275,819	
Daudi_CellLine	Type1	300	114,210	
Daudi_D100_PCRsWGA	Type1	300	375,883	R-Preamp-sWGA
Jijoye_CellLine	Type2	300	15,094	
Jijoye_J100_PCRsWGA	Type2	300	917,470	R-Preamp-sWGA
Namalwa_CellLine	Type1	300	29,866	
Raji_CellLine_longRead	Type1	300	16,399	
Raji_Rep1_PCRsWGA	Type1	300	422,839	R-Preamp-sWGA
Raji_Rep2_PCRsWGA	Type1	300	1,250,938	R-Preamp-sWGA
Raji_GenomiPhi	Type1	150	1,277,696	
Raji_GenomiPhi+GC	Type1	150	535,253	
Raji_CellLine_shortRead	Type1	150	290,733	

One of the purposes of determining the variant positions in the sequenced strains is the detection of possible mixed infections. Infections with multiple different strains can be determined by comparing variant frequencies throughout the genome. However, this relies on the assumption that these strains are divergent enough to be distinguished with a few genomic loci. In fact, EBV genomes are generally stable except LMP1 genic region and it is quite hard to determine mixed infections unless strains are from two different types. In order to determine the sensitivity of our sequencing method for capturing multiple types of EBV, we conducted a controlled mixture sequencing experiment. We used Jijoye and Daudi as the representative genomes of type 1 and type 2 strains. We gradually increased the presence of one strain starting from 0%, 10%, 25%, 75%, and 100% while the other strain is decreasing accordingly (see **Table 4.4**). After sequencing the mixture DNAs, we aligned non-redundant reads back to assembled genomes. Then, we called the variants using read pile-ups with Samtools. This showed the alternative variant frequencies. We then checked frequencies of minor alleles which are kept out in the assembly. Since the assembly represents only major variants, checking the alternative variants is necessary to estimate the level of mixed infection. In addition, we constructed a phylogenetic tree with the genomic assembly of mixture sequences. This distance tree demonstrates that mixture genomes are correctly clustered according to their designated mixture ratios. This also ensures that our sequencing and assembly method correctly handles clinical specimens with possibly multiple strains.

EBV_SequenceAnalysisPipeline_SNPAnalysis.sh: This script is for viral genomic variant detection and several related downstream analysis. It can be download from https://github.com/yasinkaymaz/EBV_Assembly_Pipeline.git.

4.6.4 Phylogenetic Analysis

To determine the molecular epidemiological relationship between our EBV isolates and other published complete or partial EBV genome sequences from NCBI, we constructed phylogenetic trees. The whole genome sequences and gene coding sequences of these strains were then aligned using the Mafft. Phylogenetic analysis on the multiple sequence alignments generated for the whole genome sequences and selected genomic regions were constructed with Molecular Evolutionary Genetic Analysis program version 6 (MEGA v6.0) (Kumar, Stecher, and Tamura 2016). Trees have been built using the Neighbor-joining method and Jukes-Cantor substitution model, and as a measure of the robustness of each node, the bootstrap method with 1000 pseudo replicates was applied.

4.6.5 Recombination Analysis

Potential recombination events between divergent nucleotide sequences were explored using Recombination Detection Program (RDP v.4.35ß) software (D. P. Martin et al. 2015). RDP incorporates several published recombination detection methods in a single platform to detect potential recombination events and their origin from the group of accurately aligned DNA sequences. These methods include

RDP, GENECONV, Chimera, MaxChi, SiScan, BootScan, and 3Seq. In all cases, default parameters were applied with Bonferroni correction as multiple comparison corrections and only a P-value < 0.01 were considered as significant events. Only events predicted by more than four of the methods were considered as true recombinations. Recombination events were evaluated by determining possible break point locations as evidence of recombinations for every genome assembly.

Chapter V. EBV Genome Sequences and Diversity

5.1 Summary

Complete genome sequence of Epstein Barr virus is now attainable directly from patient samples using targeted sequence enrichment methods combined with next generation sequencing. The goal of our study was to investigate viral genomes directly in primary clinical specimens collected from Kenyan children who are either diagnosed with eBL or at risk by living in the same malaria-endemic area. With this project we endeavor to reveal, first, whether there are particular substrains associated with the disease; second, whether the tumor cell associated virus is identical with the one circulating in the plasma. For these purposes, we sequenced EBV genomes from 41 primary eBL biopsies as well as 21 plasma EBV genomes from patients diagnosed with eBL in addition to 29 EBV genomes from healthy controls.

Analysis of individual EBV genomes consisted of de novo assemblies contigs, variation and recombination detection followed by comparative phylogenetic tree constructions. We extended the current study by comparing our sequences to available genomes up to date to provide more insight into regional variations. As a result, we observed that tumor-associated virus is identical to the circulating plasma virus. This supports the idea that the primary source of plasma viral load is the apoptotic tumor cells releasing intracellular content. We discovered a viral genome from an eBL tumor with one large 20 kb deletion resulting in a loss of multiple virions and lytic phase proteins. We detected three new inter-typic hybrid genomes from eBL patients, and all of them carried type 1 EBNA2 gene while their

EBNA3 genes were more similar to type 2 subtype. Regarding the subtype frequencies, we observed 75% type 1 association with eBL patients, regardless of the sample source such as plasma or tumor. Interestingly, type 1 and type 2 infection frequency among the healthy kids were at equal levels (50-50%). This resulted in a statistically significant association between type 1 virus and eBL cases as opposed to healthy controls (P=0.016, Chi-square test). Finally, our whole genome-based phylogenetic analysis revealed the high level of diversity even among African strains, and the major demarcation was correlated with the subtype classification.

Overall, our preliminary findings suggest that high-throughput sequencing will provide the means to fully unravel the complexity of EBV population structure Worldwide and provide insight into the viral variation that may influence oncogenesis and outcomes in eBL and other EBV-associated diseases.

5.2 Introduction

EBV infects more than 90% of world's population and its prevalence also geographically overlaps with various malignancies in different regions of the world (Lawrence S. Young and Rickinson 2004). EBV causes a high incidence of infectious mononucleosis in young adults of Western countries while the EBV-linked disease in Sub-Saharan Africa is endemic Burkitt Lymphoma, which is the most prevalent pediatric cancer, and it is nasopharyngeal carcinoma in some of the South Asian countries (Crawford 2001). It has been recently shown that the malaria *Plasmodium falciparum* infection causes polyclonal expansion of B cells, likely through metabolism sub-product of hemoglobin (hemozoin) (Torgbor et al. 2014b) or PfEMP1 antigen at the surface of infected red blood cells (Simone et al. 2011). The chronic malaria infection increases the AID expression which eventually enhances the likelihood of chromosomal translocation and somatic mutations (Robbiani et al. 2015). As a result, high mutagenic activity in B cells might result in increased random mutations in viral genomes and play a role in emergence of divergent strains.

The first publication of a type I EBV sequence from a long-term lymphoblastoid cell culture B95-8 was in 1982 along the first comprehensive viral transcriptome mapping has been done with this study (Baer et al. 1984). The second genome sequenced was a strain from a Chinese NPC culture cell, GD1 (Zeng et al. 2005). The first and only type 2 EBV genome sequenced was AG876 originated from a Ghanaian eBL case, which is still used as a reference for type 2 (Dolan et al.

2006). Recent advancements in high-throughput technologies accelerated the studies for whole genome sequencing. Even though the next generation sequencing helped to sequence whole genomes much more quickly compared to Sanger sequencing, the viral genome copy ratio to human DNA was still a huge challenge with metagenomic sequencing approach. The sequencing yield for the virus is extremely low as 0.014% relative to human (P. Liu et al. 2011). By increasing viral genome abundance through triggering lytic replication, commonly used culture strains, Akata (from a Japanese BL) and Mutu (from an African BL) have been sequenced (Z. Lin et al. 2013). As a solution to this abundance problem, an amplicon-based whole genome sequencing with multiple PCR products tiling across the genome has been used to sequence EBV from a primary NPC tumor biopsy (Hin Kwok et al. 2012a). However; the significant improvements were achieved by implementing the targeted capture techniques which selectively enriched the viral DNA from a mixture by sequence-specific hybridization (Depledge et al. 2011). This allowed to efficiently study with clinical samples in a better resolution in addition to reducing the sequencing cost. The major advantage of these developments is that the disease-related clinical information can easily be associated with molecular features and interpreted in the context of the natural host-pathogen environment. This also opens new opportunities to utilize primary clinical specimens and prevents limited research with only long-term cell lines. As a result, several studies provided multiple complete or partial viral genomic sequences from nasopharyngeal carcinoma tumors, EBV-associated gastric and lung cancers. Similar studies followed with BL tumors from Ghana, Brazil, and Argentina (H. Kwok et al. 2014; Y. Liu et al. 2016; S. Wang et al. 2016; Lei et al. 2015). Palser et al. conducted the largest study examining viral genome diversity Worldwide which was led by Wellcome Trust Sanger Institute, UK (Palser et al. 2015). This study revealed the first global EBV genome diversity and provided various genomes isolated from multiple tumors after establishing spontaneous lymphoblastoid cell lines (sLCL). Similar but smaller scale studies continued accumulating new genomes and proved more information about regional diversities (Simbiri et al. 2015). Although, new study designs have started to emerge investigating the associations between diseases and viral genomic diversity, these genome-wide examinations in case-control settings are still at very early stages (Chiara et al. 2016).

Here in this work, we attempted to conduct a genome-wide association study for the virus regarding eBL with a large case-control sample set from an endemic malaria area. Here, we test the hypothesis that a subset of EBV strains is associated significantly more frequently with eBL tumors, and viral genomes of healthy individuals from the same geographical location do not carry the same variant as frequent. For this purpose, we sequenced multiple primary clinical specimens from Kenyan children diagnosed with eBL and compared them to healthy control children from the same region.

5.3 Results

5.3.1 Study Set Demographics and Sequencing Data Quality

In this study, we sequenced EBV genomic DNA from various sources of primary clinical specimens. We sequenced viral genomes from 41 primary eBL tumors, 21 plasma of patients diagnosed with eBL, and 29 blood DNAs of healthy controls. In addition, we included EBV-positive BL culture genomes, Namalwa, Daudi, Jijoye, Raji, as well as viral genomes of newly established three eBL cell lines eBL_CL-1, eBL_CL-2, and eBL_CL-3. A total of 98 unique EBV libraries were sequenced with multiple sequencing lanes of Illumina. **Table 5.1** summarizes the sample set included in this study.

Table 5.1 The distribution of sample sources included in the whole genome EBV sequencing set.

DNA source	N (%)
Primary eBL Tumor	41 (41.8%)
Plasma DNA from eBL patients	21 (21.4%)
Blood DNA from healthy control kids	29 (29.6%)
New eBL cell lines	3 (3%)
Commonly used BL cell lines	4 (4%)

We isolated the DNA from clinical specimens and prepared sequencing libraries followed by viral genome enrichment. We amplified whole genome DNAs of some of these when necessary by following the methods explained in Chapter IV. We included a group of individuals as healthy controls with a certain level of viral loads (>1 EBV copy/ng blood DNA). Since the total viral copies in cell pellet DNA isolates from these healthy controls were not enough to directly sequence, we preprocess the DNA using the preamp-sWGA method we developed (explained in Chapter IV). We matched the age and geography of the control individuals with children diagnosed with eBL (case group) by collecting blood DNA from healthy children who lived in the same malaria endemic region. These kids did not demonstrate symptoms of any lymphoma before the time of collection, but they were possibly exposed to malaria multiple times.

Table 5.2 Samples in the sequencing set and their processing information in addition to basic sequencing statistics.

Sample ID	Preprocessing	Viral subtype	SampleSource	Read Length	Total Seq reads
eBL_CL-01	Direct Sequencing	Type1	New BL_CellLineDNA	300	27,728
eBL_CL-02	Direct Sequencing	Type2	New BL_CellLineDNA	300	37,293
eBL_CL-03	Direct Sequencing	Type1	New BL_CellLineDNA	300	130,447
Daudi	Direct Sequencing	Type1	BL_CellLineDNA	300	114,210
Jijoye	Direct Sequencing	Type2	BL_CellLineDNA	300	15,094
Namalwa	Direct Sequencing	Type1	BL_CellLineDNA	300	29,866
Raji	Direct Sequencing	Type1	BL_CellLineDNA	300	16,399
HC-0001	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	1,989,031
HC-0002	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	2,573,344
HC-0003	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	3,256,865
HC-0004	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	6,191,128
HC-0005	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	2,230,574
HC-0006	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	14,005,054
HC-0007	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	13,554,411
HC-0008	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	4,038,485
HC-0009	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	5,264,476
HC-0010	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	5,080,095
HC-0011	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	2,361,855
HC-0012	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	4,609,922
HC-0013	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	399,047
HC-0014	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	1,373,086
HC-0015	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	3,517,550
HC-0016	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	5,312,268
HC-0017	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	1,194,255
HC-0018	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	3,443,411
HC-0019	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	1,401,820
HC-0020	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	3,297,486
HC-0021	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	1,691,945
HC-0022	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	986,623
HC-0023	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	149,295
HC-0024	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	1,248,198
HC-0025	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	1,179,837
HC-0026	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	555,072
HC-0027	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	457,153
HC-0028	Preamp-sWGA	Type2	BloodDNA_fromHealthyKid	300	1,910,804
HC-0029	Preamp-sWGA	Type1	BloodDNA_fromHealthyKid	300	8,132,282
eBL-Tumor-0001	GenomiPhi-WGA	Type1	eBL_primaryTumorDNA	200/300	4,222,227
eBL-Tumor-0002	GenomiPhi-WGA	Type1	eBL_primaryTumorDNA	200/300	6,479,721
eBL-Tumor-0003	GenomiPhi-WGA	Type2	eBL_primaryTumorDNA	200/300	8,625,903
eBL-Tumor-0004	Preamp-sWGA	Type1	eBL_primaryTumorDNA	300	1,123,872
eBL-Tumor-0005	GenomiPhi-WGA	Type1	eBL_primaryTumorDNA	200/300	4,648,123
eBL-Tumor-0006	GenomiPhi-WGA	Type1	eBL_primaryTumorDNA	200/300	6,448,397
eBL-Tumor-0007	GenomiPhi-WGA	Type1	eBL_primaryTumorDNA	200/300	5,448,856
eBL-Tumor-0008	Preamp-sWGA	Type2	eBL_primaryTumorDNA	300	10,438,253
eBL-Tumor-0009	Preamp-sWGA	Type1	eBL_primaryTumorDNA	300	926,150
eBL-Tumor-0010	Preamp-sWGA	Type1	eBL_primaryTumorDNA	300	21,509,972
eBL-Tumor-0011	GenomiPhi-WGA	Type1	eBL_primaryTumorDNA	200/300	14,949,732
eBL-Tumor-0012	GenomiPhi-WGA	Type2	eBL_primaryTumorDNA	200/300	3,673,227
eBL-Tumor-0013	GenomiPhi-WGA	Type1	eBL_primaryTumorDNA	200/300	7,861,933
eBL-Tumor-0014	GenomiPhi-WGA	Type1	eBL_primaryTumorDNA	200/300	8,701,944

eBL-Tumor-0015	GenomiPhi-WGA	Type1	eBL_primaryTumorDNA	200/300	5,825,147
eBL-Tumor-0016	GenomiPhi-WGA	Type1	eBL_primaryTumorDNA	200/300	1,161,216
eBL-Tumor-0017	Direct Sequencing	Type2	eBL_primaryTumorDNA	200	10,699,217
eBL-Tumor-0018	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	11,258,545
eBL-Tumor-0019	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	16,192,472
eBL-Tumor-0020	Direct Sequencing	Type2	eBL_primaryTumorDNA	200	14,633,495
eBL-Tumor-0021	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	14,430,586
eBL-Tumor-0022	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	7,687,200
eBL-Tumor-0023	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	4,481,944
eBL-Tumor-0024	Direct Sequencing	Type2	eBL_primaryTumorDNA	200	2,585,251
eBL-Tumor-0025	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	9,775,107
eBL-Tumor-0026	Direct Sequencing	Type2	eBL_primaryTumorDNA	200	4,070,074
eBL-Tumor-0027	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	2,473,450
eBL-Tumor-0028	Direct Sequencing	Type2	eBL_primaryTumorDNA	200	2,785,996
eBL-Tumor-0029	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	7,602,074
eBL-Tumor-0030	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	6,079,669
eBL-Tumor-0031	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	3,972,963
eBL-Tumor-0032	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	7,118,274
eBL-Tumor-0033	Direct Sequencing	Type2	eBL_primaryTumorDNA	200	18,395,334
eBL-Tumor-0034	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	5,679,398
eBL-Tumor-0035	GenomiPhi-WGA	Type2	eBL_primaryTumorDNA	200/300	6,133,367
eBL-Tumor-0036	Preamp-sWGA	Type2	eBL_primaryTumorDNA	300	1,873,678
eBL-Tumor-0037	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	4,781,768
eBL-Tumor-0038	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	1,652,556
eBL-Tumor-0039	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	3,484,138
eBL-Tumor-0040	Direct Sequencing	Type2	eBL_primaryTumorDNA	200	3,758,609
eBL-Tumor-0041	Direct Sequencing	Type1	eBL_primaryTumorDNA	200	6,158,191
eBL-Plasma-0035	Preamp-sWGA	Type2	PlasmaDNA_fromBLpatient	300	385,303
eBL-Plasma-0036	Preamp-sWGA	Type2	PlasmaDNA_fromBLpatient	300	727,755
eBL-Plasma-0037	Preamp-sWGA	Type1	PlasmaDNA_fromBLpatient	300	52,615,528
eBL-Plasma-0038	Preamp-sWGA	Type1	PlasmaDNA_fromBLpatient	300	2,982,299
eBL-Plasma-0039	Preamp-sWGA	Type1	PlasmaDNA_fromBLpatient	300	2,834,853
eBL-Plasma-0040	Preamp-sWGA	Type2	PlasmaDNA_fromBLpatient	300	336,725
eBL-Plasma-0041	Preamp-sWGA	Type1	PlasmaDNA_fromBLpatient	300	1,055,728
eBL-Plasma-0042	Preamp-sWGA	Type1	PlasmaDNA_fromBLpatient	300	1,226,736
eBL-Plasma-0043	Preamp-sWGA	Type2	PlasmaDNA_fromBLpatient	300	643,566
eBL-Plasma-0044	Preamp-sWGA	Type1	PlasmaDNA_fromBLpatient	300	992,336
eBL-Plasma-0045	Preamp-sWGA	Type1	PlasmaDNA_fromBLpatient	300	9,600,176
eBL-Plasma-0046	Preamp-sWGA	Type1	PlasmaDNA_fromBLpatient	300	7,416,764
eBL-Plasma-0047	Direct Sequencing	Type1	PlasmaDNA_fromBLpatient	300	2,209,552
eBL-Plasma-0048	Direct Sequencing	Type2	PlasmaDNA_fromBLpatient	300	117,683
eBL-Plasma-0049	Direct Sequencing	Type2	PlasmaDNA_fromBLpatient	300	352,588
eBL-Plasma-0050	Direct Sequencing	Type1	PlasmaDNA_fromBLpatient	300	528,192
eBL-Plasma-0051	Direct Sequencing	Type1	PlasmaDNA_fromBLpatient	300	6,039,246
eBL-Plasma-0052	Preamp-sWGA	Type1	PlasmaDNA_fromBLpatient	300	1,682,856
eBL-Plasma-0053	Preamp-sWGA	Type1	PlasmaDNA_fromBLpatient	300	9,207,274
eBL-Plasma-0054	Direct Sequencing	Type2	PlasmaDNA_fromBLpatient	300	1,063,382
eBL-Plasma-0055	Direct Sequencing	Type1	PlasmaDNA_fromBLpatient	300	1,514,583

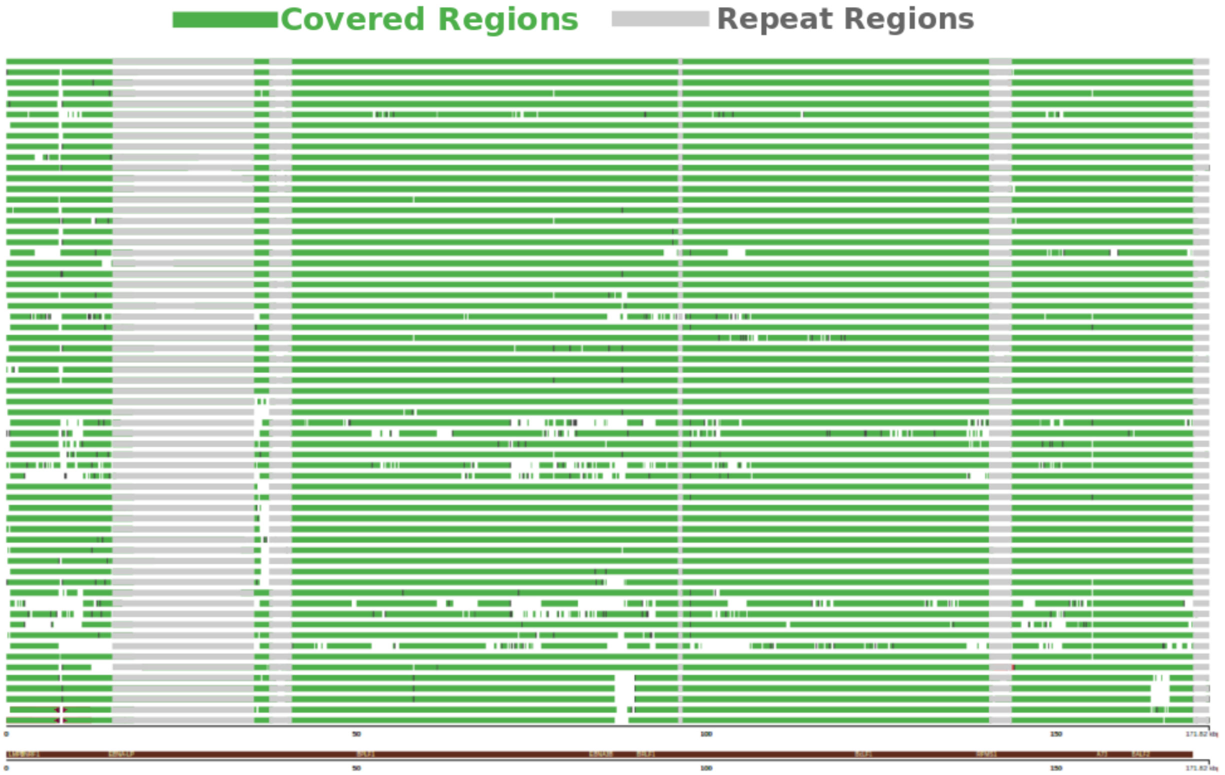


Figure 5.1 Genome coverage view of sequenced samples after De Novo assembly.

Following the sequencing with multiple lanes of Illumina, we *de novo* assembled the genomes. **Table 5.2** summarizes preprocessing information and sequencing statistics for each library. Analysis of individual EBV genomes consisted of de novo assemblies of genomes, variation detection followed by comparative phylogenetic tree constructions. **Figure 5.1** shows a representative genome-wide plot for the covered genomic regions after de novo assembly of a subset of samples. We extended the current study by comparing our sequences to publicly available genomes up to date to provide more insight into regional variations. Finally, we conducted an association test between the viral genome types and disease cases with a case-control design.

5.3.2 Genomic Comparison of Plasma EBV and Tumor EBV from patients

Circulating cell-free (CCF) DNA is defined as intact or mostly fragmented DNA found in the circulatory systems. The level of CCF EBV DNA in healthy carriers is often below the detection limits while it reaches to maximum levels when the person is diagnosed with EBV-associated carcinoma or lymphoma. For Burkitt patients, it has been shown that plasma EBV DNA levels and blood cellular fraction EBV levels were significantly correlated (Mulama et al. 2014). The source of CCF EBV also referred as plasma EBV DNA, while is not precisely known, can be tissue injury, pregnancy, or neoplasia. However, tumor cells are the usual suspect in cancer patients, and apoptotic or necrotic cells release cell-free DNA. Several studies showed that the plasma EBV DNA levels positively correlate with tumor burden and stage of NPC patients and removal of tumors led dramatic decrease in plasma EBV levels (J.-C. Lin et al. 2004; Ma et al. 2006; Chan et al. 2005). This indicates that the source of plasma EBV DNA is majorly tumor cells. Besides, studies showed that CCF EBV DNA is mostly naked rather than being protected by encapsidation using DNase treatment assays (Ryan et al. 2004).

Although the importance of plasma EBV levels of NPC patients has been boldly emphasized, the prognostic value for Burkitt lymphoma remains unexplored. Knowing that the plasma viral DNA and tumor-associated virus are same, at least same genome subtype, would open new avenues for utilizing plasma based prognostic approaches. Our initial question regarding the plasma viral genomes was whether they were identical genomes with the tumor virus or not.

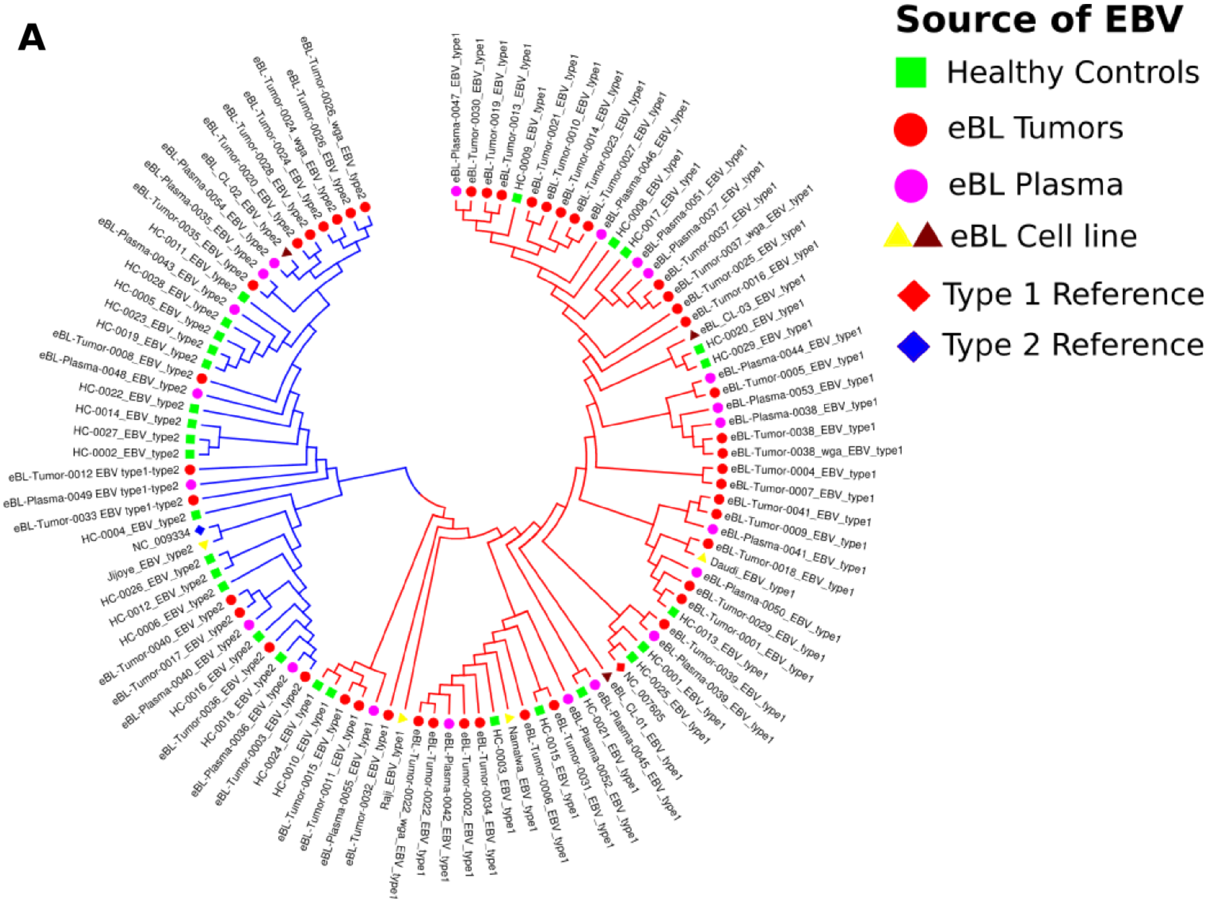
To answer this question, we sequenced seven eBL primary tumor virus genomes as well as their matching plasma associated virus DNAs. Following the sequencing and assemblies, we compared the pairs by multiple sequence alignment and looked at the differences between genomes. The phylogenetic tree demonstrates that almost all of plasma viral isolates are same as their tumor counterparts. This entirely agrees with the hypothesis that the primary source of plasma viral DNA is the apoptotic tumor cells which release cell-free DNA into the circulating blood stream. Only one of the plasma virus from the pairs showed slight differences from its tumor counterpart which suggest that there might be a secondary infection unique to that particular patient. Nevertheless, both of the viral isolates were the same subtype, type 1 EBV. This putative secondary infection case requires further validation with an alternative technique. In addition to 7 plasma-tumor pair viral genome sequences, we extended this investigation by genome-typing eight more patients using conventional PCR. **Table 5.3** shows that all of these tumor-plasma virus pairs also belong to matching virus subtypes. Overall, these findings strongly support the idea that plasma virus mainly originates from tumor cells.

Table 5.3. EBV genome types found in tumor-plasma pairs using PCR in addition to whole genome sequenced samples. Out of 8 randomly selected patients, all types were found to be same in plasmas and their tumor pairs.

Patient	Tumor DNA	Plasma DNA
eBL Patient 1	Type 1	Type 1
eBL Patient 2	Type 2	Type 2
eBL Patient 3	Type 2	Type 2
eBL Patient 4	Type 1	Type 1
eBL Patient 5	Type 1	Type 1
eBL Patient 6	Type 1	Type 1
eBL Patient 7	Type 1	Type 1
eBL Patient 8	Type 1	Type 1

5.3.3 Diversity and Phylogenetic analysis of EBV genomes Worldwide

We compared genomes in our sequencing set by constructing a phylogenetic tree following a multiple sequence alignment. We masked repetitive regions according to the reference genome. Pairwise distance calculations were based on Jukes-Cantor nucleotide substitution model, and the tree was constructed with the simple Neighbor-Joining method. **Figure 5.2** demonstrates the tree with all of the genomes in our sequencing set. The major demarcation is the separation of isolates based on their genomic subtypes, type 1 and type 2, regardless of their source. Although viral genomes, especially type 2, which were isolated from healthy children's blood tend to accumulate as subclusters, generally there is no clear sub-branching unique to healthy controls. Proper clustering of plasma tumor pair isolates can also be seen in this larger tree. The whole genome analysis also showed the segregation of EBV type 2 strains into two highly supported distinct clusters. Overall, this comparative whole genome analysis demonstrates the level of diversity in a relatively small geographic area, and yet it is hyperendemic regarding the malaria intensity.


A
Source of EBV
🟩 Healthy Controls
● eBL Tumors
● eBL Plasma
▲▲ eBL Cell line
◆ Type 1 Reference
◆ Type 2 Reference
eBL-Tumor-0026_wga_EBV_type2
eBL-Tumor-0026_EBV_type2
eBL-Tumor-0024_wga_EBV_type2
eBL-Tumor-0024_EBV_type2
eBL-Tumor-0028_EBV_type2
eBL-Tumor-0020_EBV_type2
eBL_CL-02_EBV_type2
eBL-Plasma-0054_EBV_type2
eBL-Plasma-0035_EBV_type2
eBL-Tumor-0035_EBV_type2
HC-0011_EBV_type2
eBL-Plasma-0043_EBV_type2
HC-0028_EBV_type2
HC-0005_EBV_type2
HC-0023_EBV_type2
HC-0019_EBV_type2
eBL-Tumor-0008_EBV_type2
eBL-Plasma-0048_EBV_type2
HC-0022_EBV_type2
HC-0014_EBV_type2
HC-0027_EBV_type2
HC-0002_EBV_type2
eBL-Tumor-0012 EBV type1-type2
eBL-Plasma-0049 EBV type1-type2
eBL-Tumor-0033 EBV type1-type2
HC-0004_EBV_type2
NC_009334
Jijoye_EBV_type2
HC-0026_EBV_type2
HC-0012_EBV_type2
HC-0006_EBV_type2
eBL-Tumor-0040_EBV_type2
eBL-Tumor-0017_EBV_type2
eBL-Plasma-0040_EBV_type2
HC-0016_EBV_type2
eBL-Tumor-0036_EBV_type2
HC-0018_EBV_type2
eBL-Plasma-0036_EBV_type2
eBL-Tumor-0003_EBV_type2
HC-0024_EBV_type1
HC-0010_EBV_type1
eBL-Tumor-0015_EBV_type1
eBL-Tumor-0011_EBV_type1
eBL-Plasma-0055_EBV_type1
eBL-Tumor-0032_EBV_type1
Raji_EBV_type1
eBL-Tumor-0022_wga_EBV_type1
eBL-Tumor-0022_EBV_type1
eBL-Plasma-0042_EBV_type1
eBL-Tumor-0002_EBV_type1
eBL-Tumor-0034_EBV_type1
HC-0003_EBV_type1
Namalwa_EBV_type1
eBL-Tumor-0006_EBV_type1
HC-0015_EBV_type1
eBL-Tumor-0031_EBV_type1
eBL-Plasma-0052_EBV_type1
HC-0021_EBV_type1
eBL-Plasma-0045_EBV_type1
eBL_CL-01_EBV_type1
NC_007605
HC-0025_EBV_type1
HC-0001_EBV_type1
eBL-Plasma-0039_EBV_type1
eBL-Tumor-0039_EBV_type1
HC-0013_EBV_type1
eBL-Tumor-0001_EBV_type1
eBL-Tumor-0029_EBV_type1
eBL-Plasma-0050_EBV_type1
Daudi_EBV_type1
eBL-Tumor-0018_EBV_type1
eBL-Plasma-0041_EBV_type1
eBL-Tumor-0009_EBV_type1
eBL-Tumor-0041_EBV_type1
eBL-Tumor-0007_EBV_type1
eBL-Tumor-0004_EBV_type1
eBL-Tumor-0038_wga_EBV_type1
eBL-Tumor-0038_EBV_type1
eBL-Plasma-0038_EBV_type1
eBL-Plasma-0053_EBV_type1
eBL-Tumor-0005_EBV_type1
eBL-Plasma-0044_EBV_type1
HC-0029_EBV_type1
HC-0020_EBV_type1
eBL_CL-03_EBV_type1
eBL-Tumor-0016_EBV_type1
eBL-Tumor-0025_EBV_type1
eBL-Tumor-0037_wga_EBV_type1
eBL-Tumor-0037_EBV_type1
eBL-Plasma-0051_EBV_type1
HC-0017_EBV_type1
HC-0008_EBV_type1
eBL-Plasma-0046_EBV_type1
eBL-Tumor-0027_EBV_type1
eBL-Tumor-0023_EBV_type1
eBL-Tumor-0014_EBV_type1
eBL-Tumor-0010_EBV_type1
eBL-Tumor-0021_EBV_type1
HC-0009_EBV_type1
eBL-Tumor-0013_EBV_type1
eBL-Tumor-0019_EBV_type1
eBL-Tumor-0030_EBV_type1
eBL-Plasma-0047_EBV_type1

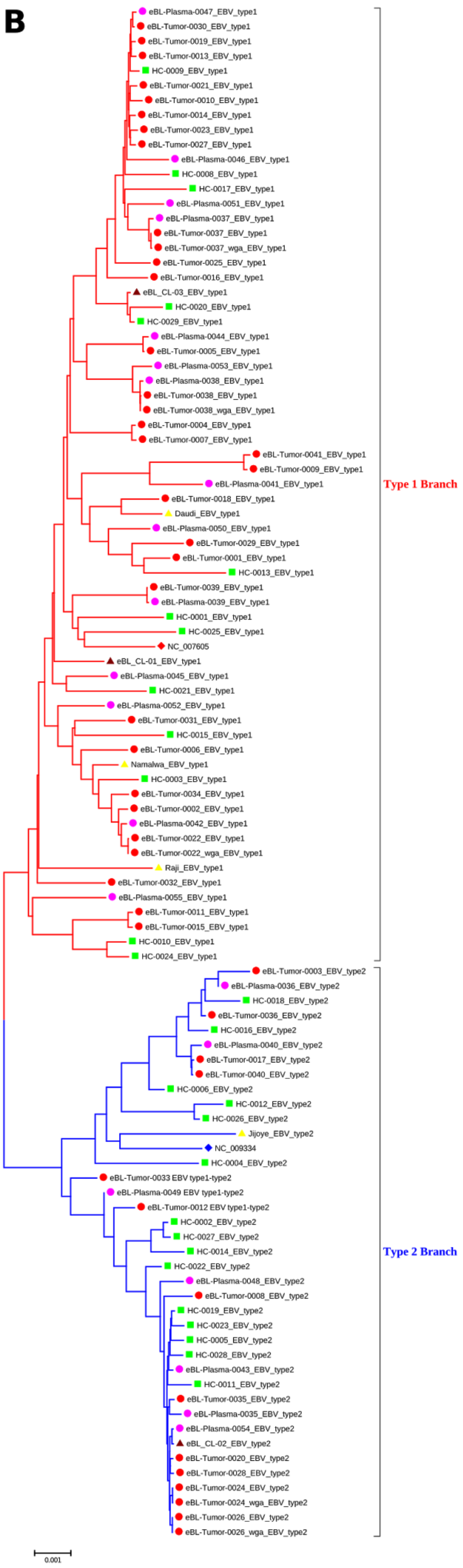
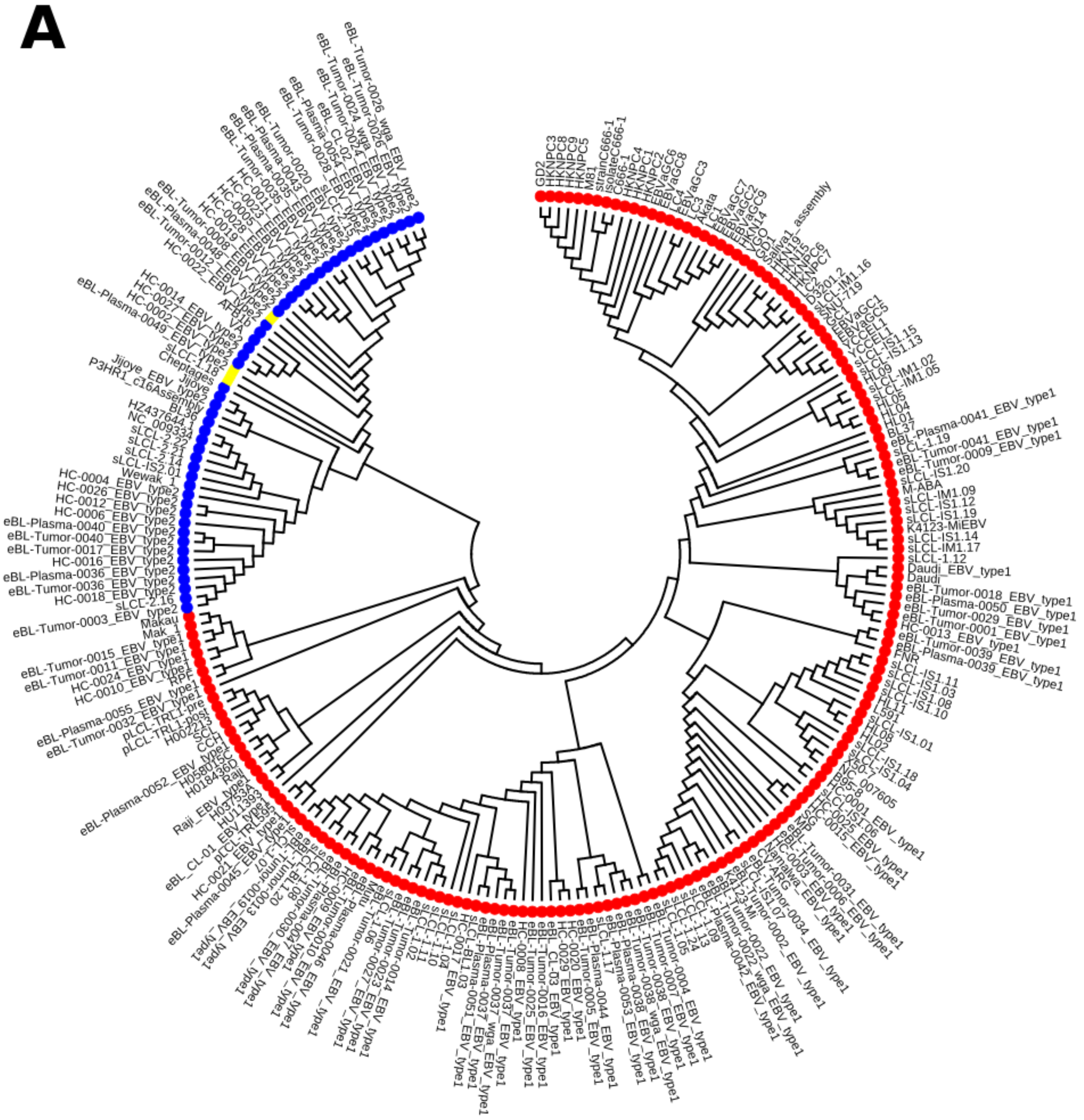

B
eBL-Plasma-0047_EBV_type1
eBL-Tumor-0030_EBV_type1
eBL-Tumor-0019_EBV_type1
eBL-Tumor-0013_EBV_type1
HC-0009_EBV_type1
eBL-Tumor-0021_EBV_type1
eBL-Tumor-0010_EBV_type1
eBL-Tumor-0014_EBV_type1
eBL-Tumor-0023_EBV_type1
eBL-Tumor-0027_EBV_type1
eBL-Plasma-0046_EBV_type1
HC-0008_EBV_type1
HC-0017_EBV_type1
eBL-Plasma-0051_EBV_type1
eBL-Plasma-0037_EBV_type1
eBL-Tumor-0037_EBV_type1
eBL-Tumor-0037_wga_EBV_type1
eBL-Tumor-0025_EBV_type1
eBL-Tumor-0016_EBV_type1
eBL_CL-03_EBV_type1
HC-0020_EBV_type1
HC-0029_EBV_type1
eBL-Plasma-0044_EBV_type1
eBL-Tumor-0005_EBV_type1
eBL-Plasma-0053_EBV_type1
eBL-Plasma-0038_EBV_type1
eBL-Tumor-0038_EBV_type1
eBL-Tumor-0038_wga_EBV_type1
eBL-Tumor-0004_EBV_type1
eBL-Tumor-0007_EBV_type1
eBL-Tumor-0041_EBV_type1
eBL-Tumor-0009_EBV_type1
eBL-Plasma-0041_EBV_type1
Type 1 Branch
eBL-Tumor-0018_EBV_type1
Daudi_EBV_type1
eBL-Plasma-0050_EBV_type1
eBL-Tumor-0029_EBV_type1
eBL-Tumor-0001_EBV_type1
HC-0013_EBV_type1
eBL-Tumor-0039_EBV_type1
eBL-Plasma-0039_EBV_type1
HC-0001_EBV_type1
HC-0025_EBV_type1
NC_007605
eBL_CL-01_EBV_type1
eBL-Plasma-0045_EBV_type1
HC-0021_EBV_type1
eBL-Plasma-0052_EBV_type1
eBL-Tumor-0031_EBV_type1
HC-0015_EBV_type1
eBL-Tumor-0006_EBV_type1
Namalwa_EBV_type1
HC-0003_EBV_type1
eBL-Tumor-0034_EBV_type1
eBL-Tumor-0002_EBV_type1
eBL-Plasma-0042_EBV_type1
eBL-Tumor-0022_EBV_type1
eBL-Tumor-0022_wga_EBV_type1
Raji_EBV_type1
eBL-Tumor-0032_EBV_type1
eBL-Plasma-0055_EBV_type1
eBL-Tumor-0011_EBV_type1
eBL-Tumor-0015_EBV_type1
HC-0010_EBV_type1
HC-0024_EBV_type1
eBL-Tumor-0003_EBV_type2
eBL-Plasma-0036_EBV_type2
HC-0018_EBV_type2
eBL-Tumor-0036_EBV_type2
HC-0016_EBV_type2
eBL-Plasma-0040_EBV_type2
eBL-Tumor-0017_EBV_type2
eBL-Tumor-0040_EBV_type2
HC-0006_EBV_type2
HC-0012_EBV_type2
HC-0026_EBV_type2
Jijoye_EBV_type2
NC_009334
HC-0004_EBV_type2
eBL-Tumor-0033 EBV type1-type2
eBL-Plasma-0049 EBV type1-type2
eBL-Tumor-0012 EBV type1-type2
HC-0002_EBV_type2
HC-0027_EBV_type2
HC-0014_EBV_type2
Type 2 Branch
HC-0022_EBV_type2
eBL-Plasma-0048_EBV_type2
eBL-Tumor-0008_EBV_type2
HC-0019_EBV_type2
HC-0023_EBV_type2
HC-0005_EBV_type2
HC-0028_EBV_type2
eBL-Plasma-0043_EBV_type2
HC-0011_EBV_type2
eBL-Tumor-0035_EBV_type2
eBL-Plasma-0035_EBV_type2
eBL-Plasma-0054_EBV_type2
eBL_CL-02_EBV_type2
eBL-Tumor-0020_EBV_type2
eBL-Tumor-0028_EBV_type2
eBL-Tumor-0024_EBV_type2
eBL-Tumor-0024_wga_EBV_type2
eBL-Tumor-0026_EBV_type2
eBL-Tumor-0026_wga_EBV_type2
0.001

Figure 5.2 A) Phylogenetic tree of whole EBV genomes in our sequencing set. Our sequencing set consists of 62 genomes from eBL patients, 29 genomes from healthy children, and cultured cells strain for sequencing control. EBV genomes from healthy kids are represented with green squares, primary eBL tumors with red circles, plasma EBV from eBL kids are pink circles, BL culture EBVs are yellow triangles, Type 1 reference genome (NC_007605) is red diamond, Type2 reference genome (NC_009334) is a blue diamond. Three new established BL culture EBV genomes are represented by brown triangles. Red color branches represent type 1 genomes, while type 2 is with blue branch color. **B)** Same tree in traditional view. The evolutionary history was inferred using the Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 0.07924741 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Jukes-Cantor method (Jukes, Cantor, and Others 1969) and are in the units of the number of base substitutions per site. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There was a total of 128380 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura et al. 2013).

To be able to estimate the effect of malaria on viral genome divergence, we further compared genomes in our sequencing set to publicly available strains collected from various countries all around the world. We added more than 120 genomes and constructed the tree again (**Figure 5.3 - A**). This tree of whole EBV genomes includes various tumor isolates such as BL, NPC, and EBVaGC, cell cultures, as well as sLCL EBV genomes from healthy individuals in addition to our direct isolates. The tree demonstrates the apparent divergence of genomes based on type as well as geographical locations. As previously observed, we show evidence for closely related phylogeny being found among sequences derived from different geographic regions (Palser et al. 2015). **Figure 5.3 - B** demonstrates the divergence based on whole genomes again after masking major type specific genes EBNA2 and EBNAs.


A
GD2
HKNPC3
HKNPC8
HKNPC9
HKNPC5
M81
strainC666-1
isolateC666-1
C666-1
HKNPC4
HKNPC1
HKNPC2
EBVaGC6
EBVaGC8
GC4
EBVaGC3
Akata
EBVaGC7
EBVaGC2
EBVaGC9
Mutu1_assembly
HKNPC6
HKNPC7
sLCL-IM1.16
SNU-719
EBVaGC1
EBVaGC5
sLCL-IS1.15
sLCL-IS1.13
sLCL-IM1.02
sLCL-IM1.05
HL05
HL04
HL01
BL37
eBL-Plasma-0041_EBV_type1
sLCL-1.19
eBL-Tumor-0041_EBV_type1
eBL-Tumor-0009_EBV_type1
sLCL-IS1.20
M-ABA
sLCL-IM1.09
sLCL-IS1.12
sLCL-IS1.19
K4123-MiEBV
sLCL-IS1.14
sLCL-IM1.17
sLCL-1.12
Daudi_EBV_type1
Daudi
eBL-Tumor-0018_EBV_type1
eBL-Plasma-0050_EBV_type1
eBL-Tumor-0029_EBV_type1
eBL-Tumor-0001_EBV_type1
HC-0013_EBV_type1
eBL-Tumor-0039_EBV_type1
eBL-Plasma-0039_EBV_type1
FNR
sLCL-IS1.11
sLCL-IS1.03
sLCL-IS1.08
sLCL-IS1.10
sLCL-IS1.01
sLCL-IS1.18
sLCL-IS1.04
007605
HC-0001_EBV_type1
HC-0025_EBV_type1
HC-0015_EBV_type1
eBL-Tumor-0031_EBV_type1
eBL-Tumor-0006_EBV_type1
HC-0003_EBV_type1
Namalwa_EBV_type1
eBL-Tumor-0034_EBV_type1
eBL-Tumor-0002_EBV_type1
eBL-Tumor-0022_EBV_type1
eBL-Tumor-0022_wga_EBV_type1
eBL-Plasma-0042_EBV_type1
sLCL-1.24
sLCL-1.05
sLCL-1.13
eBL-Tumor-0004_EBV_type1
eBL-Tumor-0007_EBV_type1
eBL-Tumor-0038_wga_EBV_type1
eBL-Tumor-0038_EBV_type1
eBL-Plasma-0053_EBV_type1
eBL-Plasma-0044_EBV_type1
eBL-Tumor-0005_EBV_type1
sLCL-1.17
HC-0029_EBV_type1
HC-0020_EBV_type1
eBL_CL-03_EBV_type1
eBL-Tumor-0016_EBV_type1
eBL-Tumor-0025_EBV_type1
HC-0008_EBV_type1
eBL-Tumor-0037_EBV_type1
eBL-Plasma-0037_wga_EBV_type1
eBL-Plasma-0051_EBV_type1
pLCL-BL1.03
HC-0017_EBV_type1
sLCL-1.04
sLCL-1.10
sLCL-1.11
eBL-Tumor-0014_EBV_type1
eBL-Tumor-0023_EBV_type1
eBL-Tumor-0027_EBV_type1
eBL-Tumor-0021_EBV_type1
eBL-Plasma-0046_EBV_type1
eBL-Tumor-0010_EBV_type1
HC-0009_EBV_type1
eBL-Tumor-0030_EBV_type1
eBL-Tumor-0013_EBV_type1
eBL-Tumor-0019_EBV_type1
eBL-Plasma-0045_EBV_type1
HC-0021_EBV_type1
eBL_CL-01_EBV_type1
Raji_EBV_type1
Raji
eBL-Plasma-0052_EBV_type1
pLCL-TRL1-pre
pLCL-TRL1-post
eBL-Tumor-0032_EBV_type1
eBL-Plasma-0055_EBV_type1
RPF
HC-0010_EBV_type1
HC-0024_EBV_type1
eBL-Tumor-0011_EBV_type1
eBL-Tumor-0015_EBV_type1
Mak_1
Makau
eBL-Tumor-0003_EBV_type2
sLCL-2.16
HC-0018_EBV_type2
eBL-Tumor-0036_EBV_type2
eBL-Plasma-0036_EBV_type2
HC-0016_EBV_type2
eBL-Tumor-0017_EBV_type2
eBL-Tumor-0040_EBV_type2
eBL-Plasma-0040_EBV_type2
HC-0006_EBV_type2
HC-0012_EBV_type2
HC-0026_EBV_type2
HC-0004_EBV_type2
Wewak_1
sLCL-IS2.01
sLCL-2.14
sLCL-2.21
sLCL-2.22
NC_009334
HZ437644
BL36
P3HR1_c16Assembly
Jijoye
Jijoye_EBV_type2
Cheptages
sLCL-1.18
eBL-Plasma-0049_EBV_type2
HC-0002_EBV_type2
HC-0027_EBV_type2
HC-0014_EBV_type2
HC-0022_EBV_type2
eBL-Tumor-0012_EBV_type2
eBL-Tumor-0048_EBV_type2
eBL-Plasma-0048_EBV_type2
eBL-Tumor-0008_EBV_type2
eBL-Plasma-0043_EBV_type2
eBL-Tumor-0020_EBV_type2
eBL-Plasma-0035_EBV_type2
eBL-Tumor-0028_EBV_type2
eBL-Plasma-0054_EBV_type2
eBL_CL-02_EBV_type2
eBL-Tumor-0024_EBV_type2
eBL-Tumor-0024_wga_EBV_type2
eBL-Tumor-0026_EBV_type2
eBL-Tumor-0026_wga_EBV_type2

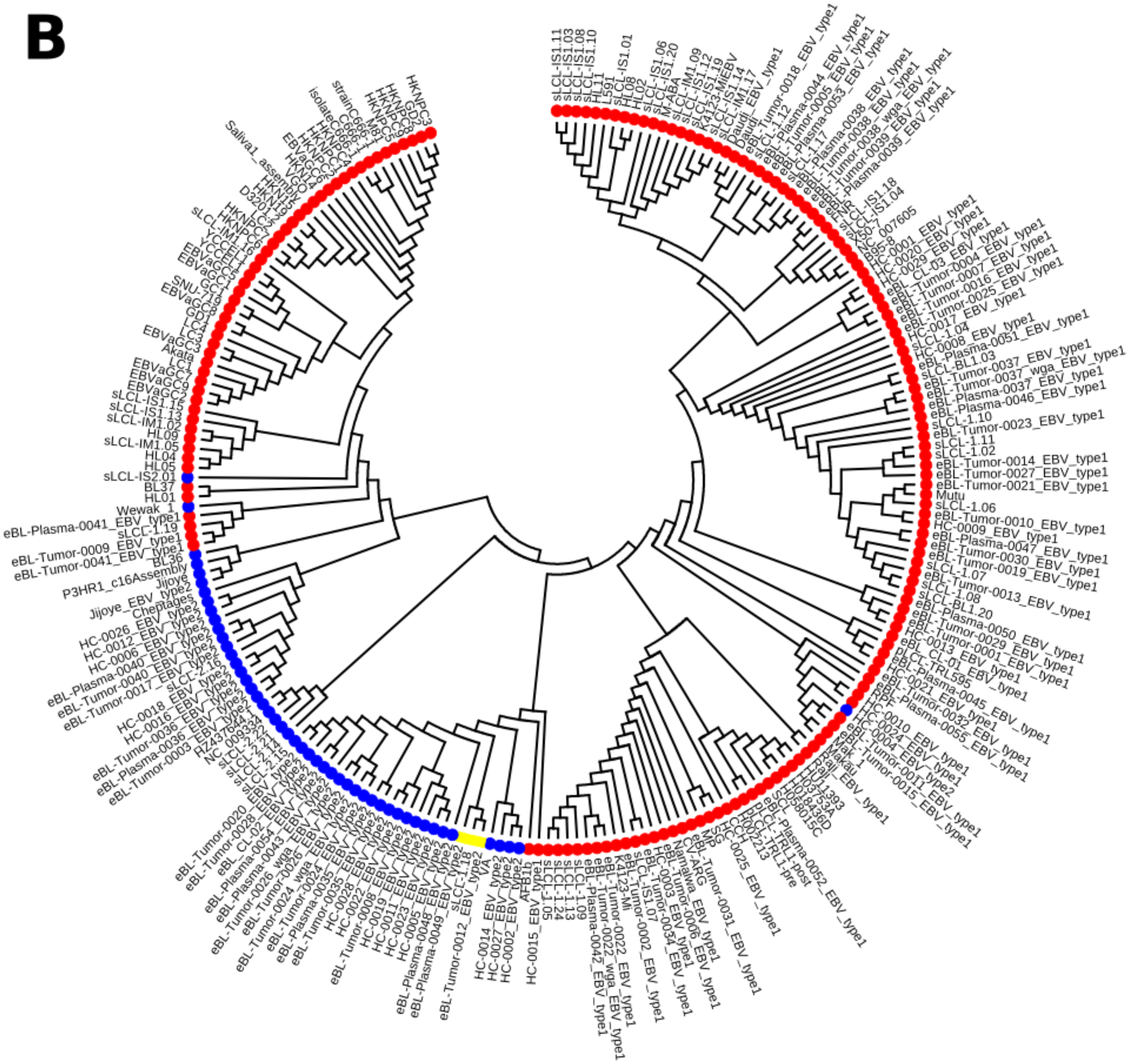


Figure 5.3 A) Phylogenetic tree of 226 complete genome sequences. Viral genomes in this tree are color coded based on their subtypes as type 1 genomes are represented by red circles while blue represents type 2. Yellow circles are used to label inter-typic hybrid genomes. **B)** A Same tree without EBNA2/EBNA3s regions. The evolutionary history was inferred using the Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 0.20057488 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Jukes-Cantor method (Jukes, Cantor, and Others 1969) and are in the units of the number of base substitutions per site. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position.

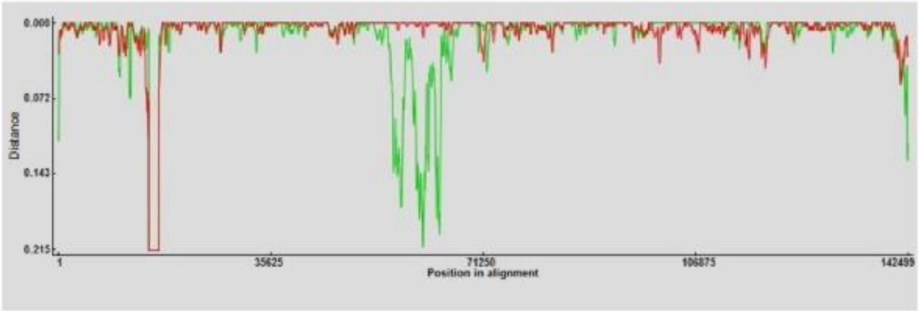
There was a total of 107023 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura et al. 2013).

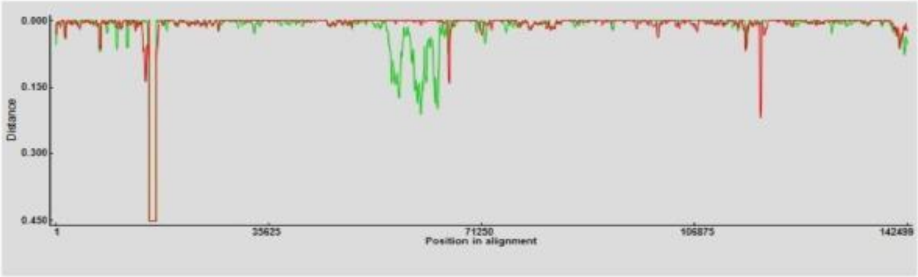
5.3.4 Novel Intertypic Hybrid Genomes

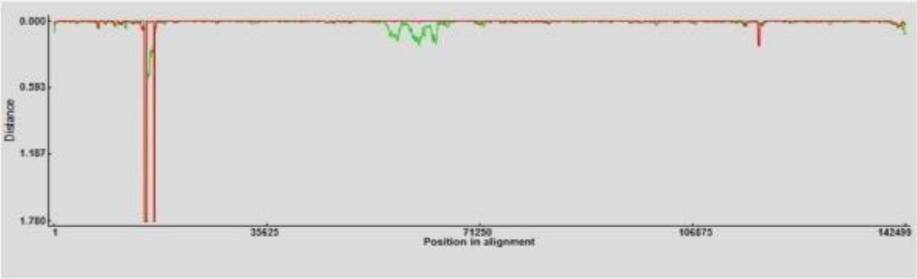
The first demonstration of a recombination between two types of EBV in the lab involved conducting a superinfection experiment with both types (Skare et al. 1985). The reports of naturally occurring recombinant viral genomes followed (Burrows et al. 1996; Yao, Tierney, Croom-Carter, Cooper, et al. 1996), and many more evidence have been provided for recombination events occurring between types of EBV. Thus, the continuance of type 1 and type 2 properties irrespective of the recombinations should be considered in clinical concepts. Although, it is hard to estimate the frequencies of recombinant genomes unless assays include all EBNA genes simultaneously, the observations from these individual studies emphasize on the underestimated functional importance of type 2 genomes in diseases. Intertypic recombinant virus genomes can be found in various cancer patients (Cho and Lee 2000) and lymphoma cells (Aguirre and Robertson 1999). These intertype viruses are replication competent and are still able to successfully transform cells to generate spontaneous LCLs (Kim, Kang, and Lee 2006). This supports the importance of EBNA2 gene of type 1 in efficiently transforming B cells compared to type 2 strains given intertypic recombinant virus genomes almost always contain type 1 EBNA2 and type 2 EBNA3s genes.

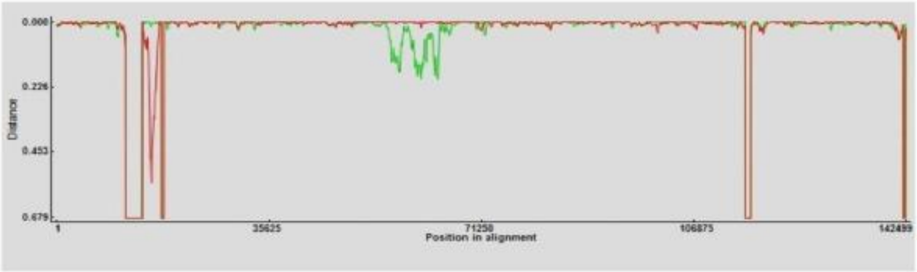
Recombinant genomes might be products of relatively recent events because the diversity of EBNA1 and LMP1 does not correlate with types and these regions still carry their geographic divergence patterns (Midgley et al. 2000). As an interpretation of this, one can infer that each geographic area has its unique hybrid genome as a result of more frequent recombination events. However, the necessity of both types to be present in the same cell at the same time via superinfection is the major limitation for the frequent occurrence of hybrid genomes. One of the suggested hypothesis for the mysterious emergence of EBV subtypes is that during human evolutionary history two human populations carrying same ancestor virus got separated from each other. The virus with them diverged into two distinct genomes at certain EBNA genes. Then, these two human populations converged again setting the stage for new recombination events between divergent subtypes (McGeoch and Gatherer 2007).

The biological question is whether these incidents are products of an early ancestral event or they occur more frequently as the different strains find a chance to recombine in a cell. To help to answer this question, we further investigated the genome isolates to determine whether there are any recombinant genomes. For this purpose, we compared the pairwise similarities of each genome against both type 1 and type 2 reference genomes. By applying a tiling window through the whole genome, we were able to determine which parts are more similar to any of the subtypes. As a result, we discovered three new hybrid EBV genomes which carry type 1 EBNA2 gene while their EBNA3 genes are more similar to type 2 subtype. **Figure 5.4** demonstrates genome-wide similarity tracks for each hybrid against both reference genomes. Also, we also included one previously known recombinant isolate from Palser et al. as a control for our detection approach. Two of these new hybrids were isolated from primary eBL tumors while the third was from a plasma again from a patient diagnosed with eBL. While each of these was from separate individuals, we did not observe any chimeric genomes isolated from healthy controls. However, the occurrence rate as 3 in 62 eBL isolates suggest that we might have missed the chance to capture one such genome with only sample size 29 healthy controls. This shows that it is hard to attribute significance to this event and to associate with the disease.


0.000
0.072
0.143
0.215
Distance
1
35625
71250
106875
142499
Position in alignment


0.000
0.150
0.300
0.450
Distance
1
35625
71250
106875
142499
Position in alignment


0.000
0.593
1.187
1.780
Distance
1
35625
71250
106875
142499
Position in alignment


0.000
0.226
0.453
0.679
Distance
1
35625
71250
106875
142499
Position in alignment

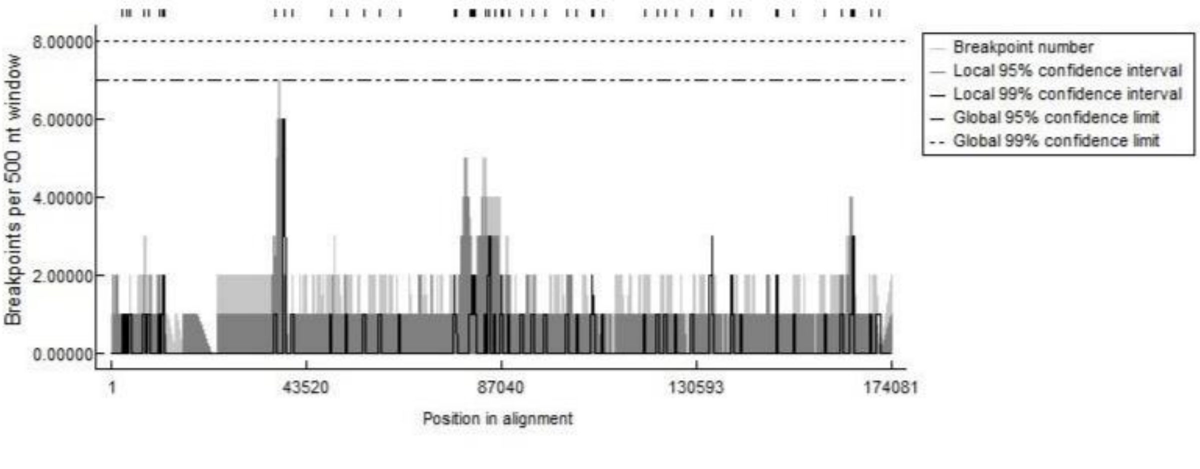

8.00000
6.00000
4.00000
2.00000
0.00000
Breakpoints per 500 nt window
1
43520
87040
130593
174081
Position in alignment
Breakpoint number
Local 95% confidence interval
Local 99% confidence interval
Global 95% confidence limit
Global 99% confidence limit

Figure 5.4. Similarity plots for identifying inter-typic hybrid genomes. Breakpoint distribution plot. These incidents in the genome sequences were supported by statistical evidence for the putative recombination breakpoints, and only breakpoints detected by all four algorithms were considered (RDP, BOOTSCAN, Chimaera, Sister-Scanning analysis).

5.3.5 Viral Genomic Variants and Associations with eBL

It has been suggested that the main driving force behind the sequence divergence among the viral genomes was HLA class II types of infected individuals. Since major HLA types significantly differ between different ethnic groups at various geographic locations, this also reflected viral genome variations (Tzellos and Farrell 2012). In addition to finding EBNA2 to be the key determinant of differential lymphocyte transformation efficiency (Cohen et al. 1989b; Lucchesi et al. 2008), LMP1 gene was also found to be essential for lymphocyte growth transformation (Kaye, Izumi, and Kieff 1993). Recently, EBNA3B has been determined to be playing a tumor suppressor role in lymphomagenesis (White et al. 2012). These findings ensure that genomic sequence variants of viral strains might lead to distinct outcomes of cell fate. In general, EBV sequence and disease association studies examined several cases and control sets for viral sequence and type differences between patient and healthy groups. The most commonly targeted viral genomic regions for their variants for association studies are EBNA genes (mainly for subtyping), other genes such as LMP1, and BZLF1 gene (different lytic replication) (reviewed in (Chang et al. 2009)). However, these studies fail to draw definitive conclusions regarding associations between diseases and any viral genomic variants including the subtypes type 1 and type 2. Although there is no clear significant association between these variants and diseases, these variants might differ in functions to induce several host signaling pathways. These studies were limited to only one or two viral gene sequences. In addition, there is no such association study examining differences between the virus in BL tumors and isolates from healthy individuals except one study focusing on LMP1 gene only (Wohlford et al. 2013). Other disease association studies also lack proper random sampling of individuals from same geographic and environmental conditions. We address these important factors in our design carefully by sequencing the whole genome of the virus instead of targeting certain regions and choosing patient and healthy controls from the same geographic area under the same malaria endemicity.

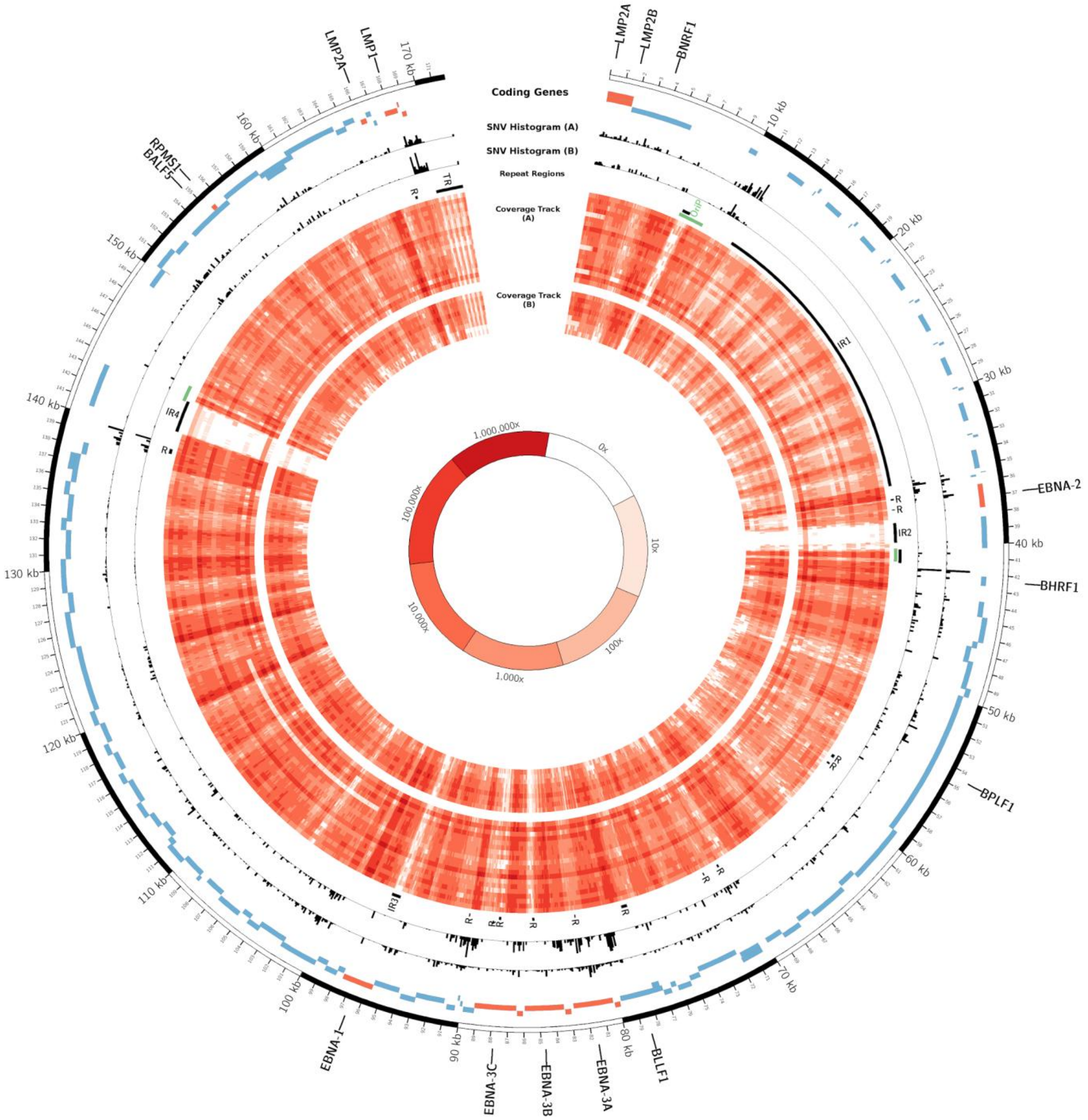


Figure 5.5 Coverage and variants histogram with the circos plot.

Previously, it has been reported that BL cell lines P3HR1 and Daudi have deleted regions spanning nuclear protein-coding EBNA2 gene (Kelly, Bell, and Rickinson 2002b). Additionally, loss of regions that encode for miRNA sequences in B95-8 and loss of EBNA3C in Raji strains have also been reported (B. D. Parker et al. 1990; Polack et al. 1984). We could not detect deletion of the mentioned genes in our isolates, however; one of the genomes in eBL tumors carries a large deletion between roughly 100kb and 120kb region spanning multiple virions related and lytic phase genes, such as BBRF1/2, BBLF1/3, BGLF1/2/3/4/5, and BDLF2/3/4 (see **Figure 5.5**). This circos plot also shows SNV distributions through the genome separate for type 1 and type 2. Highly divergent regions are EBNA3 genes and LMP1 gene.

We compare genomic sequences of viral DNA from eBL patients to viral genomes of 29 healthy kids from the same malaria endemic region with an age-geography matched case-control set. Our initial association test was regarding the genome subtype of the virus. To determine the type of the genomes, for each sample we compared the number of single nucleotide variations, against each reference genome, within EBNA2, EBNA3A-B-C genes. We also complemented this analysis by comparing overall sequencing read alignment rates of each sample genome against both references.

We first intended to assure that our sampling or inclusion criteria is completely random and does not depend on any factor. Since we relied on viral load levels of clinical specimens to prepare sequencing libraries, we checked to determine if there is any bias towards one of the types concerning viral load levels. **Figure 5.6 - A** shows the nonsignificant relationship between viral loads and viral genome types. This indicates that our sample inclusion criteria were not biased regarding the uniform sampling all viral subtypes. We also aimed to verify that the plasma virus is representative of the tumors from the patients and they can be included in the association test. As we demonstrated above, plasma viral types always matched with the tumor viral type from the same patients. In addition, the viral type frequencies in plasma samples from independent patients were entirely equivalent to type frequencies in BL tumors (see Figure 5.6 - B).

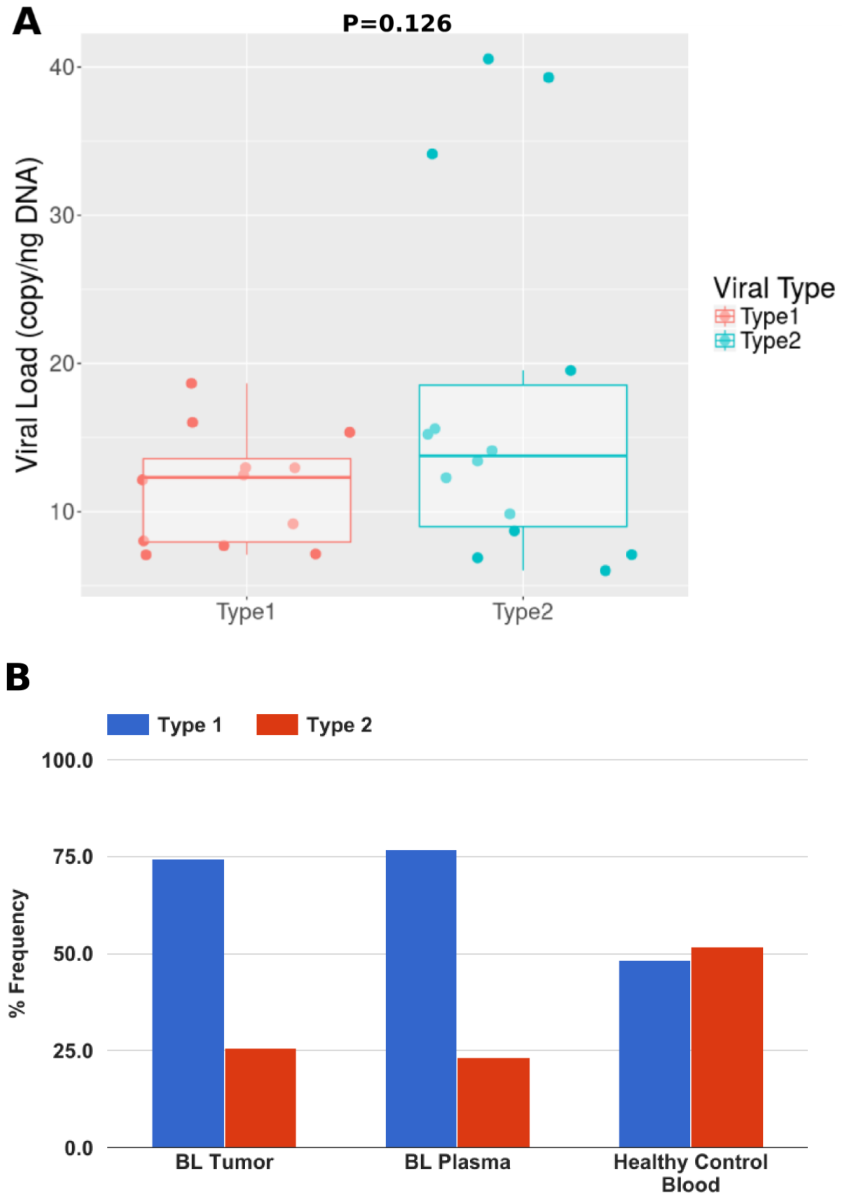


Figure 5.6 A) There is no significant relationship between viral genome type and viral load levels (P=0.126, t-test). This shows that our sampling method is not biased and skewed towards one of the genomes. **B)** Type 1 and Type 2 infection frequencies among the Kenyan children in malaria holoendemic region comparing eBL patients to healthy kids. The observed frequency of EBV types among healthy kids is 50-50% while it is skewed in favor of type 1 (75%) among eBL kids.

To be precise in our calculations, we excluded the three inter-typic hybrid genomes from this analysis as well as the plasma virus genomes paired with the tumor from same patients. As a result, we determined that the 74.5% of total 55 BL cases carried type 1 virus while only 25.5% carried type 2 infection. On the other hand, we found that the type 1 carrying healthy controls were at similar levels as type 2 carrying healthy children, 48.2% vs. 51.7%, respectively. This striking viral type frequency difference between BL cases and healthy controls is statistically significant with p=0.01605 and Chi-square statistics 5.79 (**Table 5.4**). To our knowledge, this is the first time to demonstrate the equal type frequencies in healthy carriers while type 1 is the predominant type in BL patients in Africa malaria hyper endemic region.

Table 5.4. Contingency table used for testing the significance of the association between EBV genome type and BL. Chi-square p-value is significant at 0.05 significance level, and the test statistics is 5.76.

	BL	**Healthy Controls**	*Marginal Row Totals*
Type 1 EBV	**41 (74.5%)**	**14 (48.2%)**	**55**
Type 2 EBV	**14 (25.5%)**	**15 (51.7%)**	**29**
Marginal Column Totals	**55**	**29**	**84 (Grand Total)**
The chi-square statistic is 5.7968. The p-value is 0.016056. **Count (% of marginal column totals)**			

5.4 Discussion

EBV is the first virus found within a tumor cell over 50 years ago. However, we are just starting to decipher its possible role in pathogenesis. Sequencing technologies help to answer questions regarding the host-pathogen interactions by diving into the principal source, DNA. Here in this study, our goal was to identify EBV genome sequence variants and correlate these with endemic Burkitt lymphoma incidences.

This study, for the first time, demonstrates by sequencing that circulating plasma virus of eBL patients are identical to their tumor virus. Although there were various attempts to utilize plasma virus as a biomarker in NPC patients, such studies were limited regarding the BL (J.-C. Lin et al. 2004; Chan et al. 2005; Ma et al. 2006). Knowing that plasma virus is identical to the tumor-associated virus, this can be utilized in various ways such as tumor diagnosis, treatment monitoring, or tumor evolution studies. Circulating cell-free viral DNA can be monitored over time and examined as a factor affecting survival outcome of the patient. Combining with the recent applications of monitoring plasma EBV of eBL patients, sequencing virus genomes in clinical settings can be used in a predictive model for foreseeing relapses (Westmoreland et al. 2017).

The first large-scale attempt to demonstrate the broad diversity among Worldwide strains was by Palser et al. This study proves that geography plays a significant role in viral variation and needs to be properly accounted for in comparative studies. However, our sequencing approach is fundamentally different than the one used in this study because we tried to avoid generating lab-cultured strains for enrichment via spontaneous lymphoblastoid cell lines. Directly sequencing the virus within the primary sources takes the snapshot of the population diversity unlike bottlenecking for replication competent strains with sLCL generation. Thus, the contribution of choosing the right sequence enrichment method is undeniable for capturing the real subtype frequencies. As a result, we discovered the dramatic difference between type 1 and type 2 infection rate in eBL patients compared to healthy control kids. By relying on the earliest genome typing studies, it is widely believed that the type 2 subtype is more prevalent in Africa compared to the rest of the World (L. S. Young et al. 1987a; Apolloni and Sculley 1994). However, there were reports also showing the contrary results and suggesting that type 2 might not be limited to Africa (Sixbey et al. 1989). Besides all of this controversy, the fundamental reasons for the existence of two types require more investigations.

Although it is hard, almost impossible, to infer linear ancestry from the phylogenetic tree because of possible extensive recombination levels, our comparative analysis demonstrated the level of diversity even in smaller geographies. Both of the parent strains have to be present within the same cell in order to recombine and give birth to a third strain. The likelihood of this event to happen would vary depending on the geographic area. The equivalent levels of both subtypes among healthy children might explain observing inter-typic hybrid genomes frequently in the same region. However, whether these genomes are products of early ancestral events or often occur more frequently is not known. Interestingly, all hybrid genomes carried same recombination pattern which had the same EBNA2 and EBNA3s combinations. This entirely agrees with previous functional experiments demonstrating that EBNA2 gene is the key transformative factor in creating lymphoblastoid cell lines and EBNA3B playing a tumor suppressor role in lymphomagenesis (White et al. 2012; Lucchesi et al. 2008). What kind of other functional advantages, such as better expanding host range or increasing virulence, a recombination gives to the virus needs further investigations in the light of hybrid genomes. On the other hand, finding hybrid viral genomes only in eBL patients does not indicate that these are unique to the disease because it is likely to capture such genomes among healthy controls with higher sample sizes.

Finally, we detected a virus in an eBL tumor with a substantially large deletion causing a loss of various lytic and late genes. Increasing the examples of such incidences with more sequencing might help to understand the actual role of EBV in pathogenesis. As a conclusion, whole genome sequence differences between type 1 and type 2 EBV need to be investigated comprehensively with larger cohorts in order to pinpoint this fundamental divergence. Such investigation will not only decipher the puzzling pathogenic differences but also will help to understand how these two EBV types persist in the population at the same time.

Chapter VI. Conclusions

General Conclusions and Future Directions

In this research, we examined the expression and mutational spectrum vis-a-vis clinical and molecular features of Kenyan children diagnosed with eBL and compared to publicly available data for sBL. We observed relative homogeneity of expression within tumors collected from our patient population suggesting no overt subtypes within eBL. Our initial expectation regarding the gene expression was to determine clinical subtypes with distinct expression profiles. We found minimal differences between tumors presenting in the jaw or abdomen but observed differences in expression that correlated with survival, viral presence, and type of EBV. We tested the hypothesis that eBL tumor cells carry mutated genes different from sBL tumors. Our comparative analysis confirmed the frequently mutated gene sets in eBLs to be almost same as sBLs; however, mutation frequencies were significantly lower in eBLs. We also detected previously undescribed somatically mutated genes and showed that the BL mutational spectrum appears to most greatly differ based on the type of EBV infecting the tumor rather than the geographic origin of the patient. Furthermore, we found that tumors harboring EBV type 1 display a significantly different host mutational profile compared to BL tumors with EBV type 2 and without EBV. This distinct mutational frequency profiles of tumors with different viral subtypes perfectly inlines with the known fact regarding the transformation ability differences of viral subtypes.

We also found that the differences in tumor mutational spectrum were more striking when categorized by viral presence or absence rather than geographic origin. This was also supported by stronger differences in expression profiles in critical pathways as well as involving likely downregulation of *PTEN* affecting the central pathway of *PI3K-Akt* signaling followed by mTORC1 activation. It has been previously shown that EBV can modulate *mTOR* pathway by LMP2A (Moody et al. 2005). Given the limited viral gene expression in latency I, EBV could be interacting with major host regulators especially through viral miRNAs (H.-J. Yang et al. 2013) or by regulating cellular miRNAs expressions (Forte et al. 2012). It has been recently demonstrated that EBV microRNA Bart6-3p can inhibit *PTEN* translation (Cai et al. 2015). Even though this could be the mechanism by which the virus interferes with critical cellular pathways, the consequences of viral miRNA interactions with *PTEN* and whether translational inhibition or mRNA degradation in EBV-positive BL needs further clarification. Our data shows that the mRNA transcription of *PTEN* itself is not significantly differentially expressed between EBV-positive and negative BLs. However, Ambrosio et al. found that protein levels of *PTEN* are significantly lower in EBV-positive BLs compared to negatives (Ambrosio et al. 2014). This suggests a mechanism in which viral miRNAs interact with *PTEN* causing a translational inhibition. Combined, this suggests that the virus plays a key role in oncogenesis beyond the possible role in potentiating the translocation. However, further functional assays are needed in order to validate viral compensation for the less frequent *ID3*, *TCF3* or *CCND3* mutations found in EBV-positive eBL tumors.

Third, we discovered new mutations occurring in genes previously not reported in other BL studies. While many of these additional genes are mutated at low frequencies (<10%), they support the key roles of previously identified pathways (e.g. *BCL7A* as another member of SWI/SNF). These genes also implicate DNA repair regarding oncogenesis where we identify three previously undescribed genes (*MSH6*, *RAD50*, and *PRKDC*) involved in double strand repair and nonhomologous end joining.

Fourth, we found that not only does tumor gene mutation rates and distribution vary based on the presence of EBV but that tumors have different patterns of mutation based on EBV type. We observed that BLs with type 2 has a significantly higher average number of mutated genes relative to type 1 and that this rate is on par with viral negative tumors. The only observed consistent mutational difference was a lower rate of mutations in *ID3* and *TCF3*, which supports the idea that the virus may play a critical role regulating these pathways during oncogenesis by alternatively driving AKT/mTOR signaling. The overall lower mutational rates in BLs with type 1 virus, suggest that type 1 virus may be providing survival advantages in other ways. This is consistent with the known ability of type 1 virus to better transform peripheral B cells to create lymphoblastoid cell lines. Given that previous studies have not seen significant differences in viral types in tumors relative to population controls, this suggests that many of these driver mutations while offering relative advantages in tumor growth are not in and of themselves necessary regarding oncogenesis. Further studies with greater number of tumors and population controls can help to better understand the distribution of EBV within the general population in contrast to their role in eBL pathogenesis.

Finally, we observed that the viral expression pattern is consistent with viral latency I, as we expected, where EBV is substantially quiescent and maintained with EBNA1 expression. However, increased detection of lytic gene expression was suggestive of poor prognosis. It has been argued that this observed expression pattern may primarily be due to various levels of lytic reactivation (Fujita et al. 2004). Consistent with the previous reports (Arvey et al. 2015), our results demonstrated similar heterogeneous viral gene expression in BL suggesting that tumor cells could be targeted by antiviral immunotherapies. Dysfunctional T cell immunity has been reported for children diagnosed with eBL who were defective for EBNA1 specific IFN-gamma T cell responses (Moormann et al. 2009b). This defect putatively allows latency I tumors to escape from immune surveillance. Interestingly, we found that one-third of the eBL tumors which carried type 2 EBV genome had significantly suppressed immunoproteasome complex gene transcriptions compared to eBLs with type 1 EBV. One explanation for this novel observation could be that type 2 EBV more readily infects immunocompromised individuals. Baarle et al. reported an increased prevalence of type 2 EBV among HIV patients (van Baarle et al. 2000). However, a larger cohort of HIV patients showed that T-cell impairment does not sensitize individuals for type 2 EBV infections (Yao et al. 1998b). Thus, an alternate explanation is that type 2 EBV directly interacts with host regulatory components in order to interfere with immunoproteasome complex formation. Reduced transcriptional expression of these genes in EBV type 2 eBL tumors implies a mechanism in which viral components (coding or non-coding) suppress *JAK/STAT1* mediated transcription. This could be an additional mechanism by which type 2 EBV is able to escape from immune surveillance by preventing T-cell responses (Sijts et al. 2002). Since we did not observe any viral transcriptional pattern differences compared to type 1 genes, this phenomenon requires further investigation to confirm.

As a summary of our mutational investigation, we have illustrated the key pathways implicated in BL oncogenesis integrating our analytical results with current literature (Figure 2.5 C). This expanded view of BL oncogenesis more clearly defines a role for EBV. Driving proliferation through *PTEN/PI3K/AKT* and *CCND3* pathways may be an essential step toward bypassing the lack of mutations in the *TCF3/ID3/CCND3* axis that may include SWI/SNF interactions of *SMARCA4* as well. *MYC* translocation provides the pivotal accelerant while the gain of function mutations in *CCND3* strengthens the pressure. Although the mutated genes which function in B cell development and chromatin remodeling complexes might contribute to this signaling, our results regarding gene expression and pathway differences suggest a role for viral microRNAs which can inhibit *PTEN* function and cause activated BCR signaling via AKT. Other genes frequently mutated in BL play roles in distinct but relevant pathways such as DNA repair and focal adhesion. Overall, this combined model demonstrates pathogenic mechanistic routes to tumorigenesis BL and introduces a defined role for EBV that warrants further interrogation.

Although RNAseq based expression profiling technique gives clues about viral genome subtypes in tumors, we pursued to extend the investigation further by sequencing the EBV genomes in primary eBL tumors and of infected Kenyan children. Our initial expectation regarding the circulating plasma virus and tumor virus of the same patients was that they both originate from the same source and, thus, carry the same genome. As a result of our comparative analysis, we concluded that plasma viral genomes of eBL kids are identical with their tumor viruses as we expected. This opens up new opportunities for cell-free circulating plasma DNA to be utilized as a diagnostic or a prognostic tool. Besides, knowing that eBL tumor cells carry the same virus in plasma, better association studies can be conducted with larger cohorts without waiting to reach enough tumor tissue samples. Secondly, we intended to measure the association frequencies of viral subtypes and eBL with a case control study. In the light of previous studies and literature, we expected to observe a similar subtype association rate even among healthy children control group. However, we found that infection rates of two EBV subtypes were at equal levels among healthy children. Strikingly, on the other hand, we determined a significant association between the type 1 EBV and eBL cases in a case-control study which involves virus genomes from healthy age-geography matched children. The significant difference between the frequencies of two subtypes suggests that children carrying type 1 infection have a higher likelihood of developing lymphoma.

In conjunction with our findings regarding the mutated host gene rate differences between type 1 carrying tumors and type 2 carrying tumors, it becomes more evident that the viral genome typing assays should be incorporated into clinical diagnostic assays to guide local pediatric oncologists in Kenya.

Distinct EBV strains might contribute to different diseases. To test this, one can compare viral genomes of NPC cases in East Africa to NPC-associated viruses from southern-China. This might shed light on some common viral sequence features shared by both cases. As a matter of fact, an association study that involves multiple virus-associated diseases worldwide can be much more informative. In such study, various geographic regions and proper healthy controls should be included. However, there are still many unanswered questions in the EBV research area; do viral genome variations contribute to the viral preferences for infecting epithelial cells or B cells? Why there are two types of EBV genomes is still an unexplored phenomenon. Both types should have functional properties which would allow them to persist in the population without getting wiped-out. Are the intertypic hybrid genomes the products of early ancestral events and should they be considered as the third subtype, or they occur more frequently than we can think?

In the near future, it is possible that researchers will sequence many more viral genomes to investigate disease associations and utilize sequence information for biomarker or vaccine development. The increasing sequence samples will require accurate handling with standardized bioinformatics analysis pipelines. Therefore, an online public database storing all sequenced EBV genomes and their relevant clinical information would serve as a base for investigators in the process of hypothesis generation.

Here in this dissertation, we developed novel bioinformatic approaches and workflows to analyze and investigate expression profiles of tumors as well as viral genomic variations. We identified important differences based on viral content and clinical outcomes by genomic and mutational analyses of Burkitt lymphoma tumors. Overall, these suggested new avenues for the development of prognostic molecular biomarkers and therapeutic interventions.

REFERENCES

Abate, Francesco, Maria Raffaella Ambrosio, Lucia Mundo, Maria Antonella Laginestra, Fabio Fuligni, Maura Rossi, Sakellarios Zairis, et al. 2015a. “Distinct Viral and Mutational Spectrum of Endemic Burkitt Lymphoma.” *PLoS Pathogens* 11 (10): e1005158.

Abbas, A. R., D. Baldwin, Y. Ma, W. Ouyang, A. Gurney, F. Martin, S. Fong, et al. 2005. “Immune Response in Silico (IRIS): Immune-Specific Genes Identified from a Compendium of Microarray Expression Data.” *Genes and Immunity* 6 (4): 319–31.

Aguirre, Andrew J., and Erle S. Robertson. 1999. “Characterization of Intertypic Recombinants of the Epstein–Barr Virus from the Body-Cavity-Based Lymphomas Cell Lines BC-1 and BC-2.” *Virology* 264 (2): 359–69.

Ambrosio, Maria Raffaella, Mohsen Navari, Lorena Di Lisio, Eduardo Andres Leon, Anna Onnis, Sara Gazaneo, Lucia Mundo, et al. 2014. “The Epstein Barr-Encoded BART-6-3p microRNA Affects Regulation of Cell Growth and Immuno Response in Burkitt Lymphoma.” *Infectious Agents and Cancer* 9 (April): 12.

Andrews, S. 2010. “FastQC.” *A Quality Control Tool for High Throughput Sequence Data*, 13.

Apolloni, A., and T. B. Sculley. 1994. “Detection of A-Type and B-Type Epstein-Bart Virus in Throat Washings and Lymphocytes.” *Virology* 202 (2): 978–81.

Ariumi, Yasuo, Misao Kuroki, Ken-Ichi Abe, Hiromichi Dansako, Masanori Ikeda, Takaji Wakita, and Nobuyuki Kato. 2007. “DDX3 DEAD-Box RNA Helicase Is Required for Hepatitis C Virus RNA Replication.” *Journal of Virology* 81 (24): 13922–26.

Armitage, J. O., and D. D. Weisenburger. 1998. “New Approach to Classifying Non-Hodgkin’s Lymphomas: Clinical Features of the Major Histologic Subtypes. Non-Hodgkin's Lymphoma Classification Project.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 16 (8): 2780–95.

Arvey, Aaron, Akinyemi I. Ojesina, Chandra Sekhar Pedamallu, Gianna Ballon, Joonil Jung, Fujiko Duke, Lorenzo Leoncini, et al. 2015. “The Tumor Virus Landscape of AIDS-Related Lymphomas.” *Blood* 125 (20): e14–22.

Asito, Amolo S., Erwan Piriou, Peter Sumba Odada, Nancy Fiore, Jaap M. Middeldorp, Carole Long, Sheetij Dutta, et al. 2010a. “Elevated Anti-Zta IgG Levels and EBV Viral Load Are Associated with Site of Tumor Presentation in Endemic Burkitt’s Lymphoma Patients: A Case Control Study.” *Infectious Agents and Cancer* 5 (July): 13.

Baarle, D. van, E. Hovenkamp, N. H. Dukers, N. Renwick, M. J. Kersten, J. Goudsmit, R. A. Coutinho, F. Miedema, and M. H. van Oers. 2000. “High Prevalence of Epstein-Barr Virus Type 2 among Homosexual Men Is Caused by Sexual Transmission.” *The Journal of Infectious Diseases* 181 (6): 2045–49.

Baarle, D. van, E. Hovenkamp, M. J. Kersten, M. R. Klein, F. Miedema, and M. H. van Oers. 1999. “Direct Epstein-Barr Virus (EBV) Typing on Peripheral Blood Mononuclear Cells: No Association between EBV Type 2 Infection or Superinfection and the Development of Acquired Immunodeficiency Syndrome-Related Non-Hodgkin’s Lymphoma.” *Blood* 93 (11): 3949–55.

Baer, R., A. T. Bankier, M. D. Biggin, P. L. Deininger, P. J. Farrell, T. J. Gibson, G. Hatfull, et al. 1984. “DNA Sequence and Expression of the B95-8 Epstein—Barr Virus Genome.” *Nature* 310 (5974). Nature Publishing Group: 207–11.

Bell, Melissa J., Rebekah Brennan, John J. Miles, Denis J. Moss, Jacqueline M. Burrows, and Scott R. Burrows. 2008. “Widespread Sequence Variation in Epstein-Barr Virus Nuclear Antigen 1 Influences the Antiviral T Cell Response.” *The Journal of Infectious Diseases* 197 (11): 1594–97.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1). [Royal Statistical Society, Wiley]: 289–300.

Blenk, S., J. Engelmann, M. Weniger, J. Schultz, M. Dittrich, A. Rosenwald, H. K. Müller-Hermelink, T. Müller, and T. Dandekar. 2007. “Germinal Center B Cell-like (GCB) and Activated B Cell-like (ABC) Type of Diffuse Large B Cell Lymphoma (DLBCL): Analysis of Molecular Predictors, Signatures, Cell Cycle State and Patient Survival.” *Cancer Informatics* 3 (December): 399–420.

Boerma, E. G., Gustaaf W. van Imhoff, Inge M. Appel, Nic Jgm Veeger, Ph M. Kluin, and J. C. Kluin-Nelemans. 2004a. “Gender and Age-Related Differences in Burkitt Lymphoma--Epidemiological and Clinical Data from The Netherlands.” *European Journal of Cancer* 40 (18). Elsevier: 2781–87.

Boerma, E. G., G. W. van Imhoff, I. M. Appel, N. J. G. M. Veeger, Ph M. Kluin, and J. C. Kluin-Nelemans. 2004b. “Gender and Age-Related Differences in Burkitt Lymphoma – Epidemiological and Clinical Data from The Netherlands.” *European Journal of Cancer* 40 (18): 2781–87.

Boxer, L. M., and C. V. Dang. 2001. “Translocations Involving c-Myc and c-Myc Function.” Oncogene 20 (40): 5595–5610.

Boyle, M. J., W. A. Sewell, T. B. Sculley, A. Apolloni, J. J. Turner, C. E. Swanson, R. Penny, and D. A. Cooper. 1991. “Subtypes of Epstein-Barr Virus in Human Immunodeficiency Virus-Associated Non-Hodgkin Lymphoma.” *Blood* 78 (11): 3004–11.

Boyle, M. J., E. Vasak, M. Tschuchnigg, J. J. Turner, T. Sculley, R. Penny, D. A. Cooper, B. Tindall, and W. A. Sewell. 1993. “Subtypes of Epstein-Barr Virus (EBV) in Hodgkin’s Disease: Association between B-Type EBV and Immunocompromise.” *Blood* 81 (2): 468–74.

Buckle, Geoffrey, Louise Maranda, Jodi Skiles, John Michael Ong’echa, Joslyn Foley, Mara Epstein, Terry A. Vik, et al. 2016a. “Factors Influencing Survival among Kenyan Children Diagnosed with Endemic Burkitt Lymphoma between 2003 and 2011: A Historical Cohort Study.” *International Journal of Cancer. Journal International Du Cancer*, May. doi:10.1002/ijc.30170.

Buck, M., S. Cross, K. Krauer, N. Kienzle, and T. B. Sculley. 1999. “A-Type and B-Type Epstein-Barr Virus Differ in Their Ability to Spontaneously Enter the Lytic Cycle.” *The Journal of General Virology* 80 (Pt 2) (February): 441–45.

Burkitt, Denis, and Burkitt Denis. 1961. “MALIGNANT LYMPHOMA IN AFRICAN CHILDREN.” *The Lancet* 277 (7191): 1410–11.

Burrows, J. M., R. Khanna, T. B. Sculley, M. P. Alpers, D. J. Moss, and S. R. Burrows. 1996. “Identification of a Naturally Occurring Recombinant Epstein-Barr Virus Isolate from New Guinea That Encodes Both Type 1 and Type 2 Nuclear Antigen Sequences.” *Journal of Virology* 70 (7): 4829–33.

Cai, Longmei, Jinbang Li, Xiaona Zhang, Yaoyong Lu, Jianguo Wang, Xiaoming Lyu, Yuxiang Chen, et al. 2015. “Gold Nano-Particles (AuNPs) Carrying Anti-EBV-miR-BART7-3p Inhibit Growth of EBV-Positive Nasopharyngeal Carcinoma.” *Oncotarget* 6

(10): 7838–50.

Chang, Cindy M., Kelly J. Yu, Sam M. Mbulaiteye, Allan Hildesheim, and Kishor Bhatia. 2009. “The Extent of Genetic Diversity of Epstein-Barr Virus and Its Geographic and Disease Patterns: A Need for Reappraisal.” *Virus Research* 143 (2): 209–21.

Chan, K. C. Allen, Anthony T. C. Chan, Sing-Fai Leung, Jesse C. S. Pang, Angela Y. M. Wang, Joanna H. M. Tong, Ka-Fai To, et al. 2005. “Investigation into the Origin and Tumoral Mass Correlation of Plasma Epstein-Barr Virus DNA in Nasopharyngeal Carcinoma.” *Clinical Chemistry* 51 (11): 2192–95.

Chen, Bo-Jung, Sheng-Tsung Chang, Shih-Feng Weng, Wan-Ting Huang, Pei-Yi Chu, Pin-Pen Hsieh, Yun-Chih Jung, Chun-Chi Kuo, Yu-Ting Chuang, and Shih-Sung Chuang. 2016a. “EBV-Associated Burkitt Lymphoma in Taiwan Is Not Age-Related.” *Leukemia & Lymphoma* 57 (3): 644–53.

Chêne, Arnaud, Daria Donati, André Ortlieb Guerreiro-Cacais, Victor Levitsky, Qijun Chen, Kerstin I. Falk, Jackson Orem, Fred Kironde, Mats Wahlgren, and Maria Teresa Bejarano. 2007. “A Molecular Link between Malaria and Epstein–Barr Virus Reactivation.” *PLoS Pathogens* 3 (6). Public Library of Science: e80.

Chen, W. G., Y. Y. Chen, M. M. Bacchi, C. E. Bacchi, M. Alvarenga, and L. M. Weiss. 1996. “Genotyping of Epstein-Barr Virus in Brazilian Burkitt’s Lymphoma and Reactive Lymphoid Tissue. Type A with a High Prevalence of Deletions within the Latent Membrane Protein Gene.” *The American Journal of Pathology* 148 (1): 17–23.

Cheung, S. T., S. F. Leung, K. W. Lo, K. W. Chiu, J. S. Tam, T. F. Fok, P. J. Johnson, J. C. Lee, and D. P. Huang. 1998. “Specific Latent Membrane Protein 1 Gene Sequences in Type 1 and Type 2 Epstein-Barr Virus from Nasopharyngeal Carcinoma in Hong Kong.” *International Journal of Cancer. Journal International Du Cancer* 76 (3): 399–406.

Chhangawala, Sagar, Gabe Rudy, Christopher E. Mason, and Jeffrey A. Rosenfeld. 2015. “The Impact of Read Length on Quantification of Differentially Expressed Genes and Splice Junction Detection.” *Genome Biology* 16 (June): 131.

Chiara, Matteo, Caterina Manzari, Claudia Lionetti, Rosella Mechelli, Eleni Anastasiadou, Maria Chiara Buscarinu, Giovanni Ristori, et al. 2016. “Geographic Population Structure in Epstein-Barr Virus Revealed by Comparative Genomics.” *Genome Biology and Evolution* 8 (11): 3284–91.

Cho, Sung-Gyu, and Won-Keun Lee. 2000. “Analysis of Genetic Polymorphisms of Epstein-Barr Virus Isolates from Cancer Patients and Healthy Carriers.” *Journal of Microbiology and Biotechnology* 10 (5). The Korean Society for Applied Microbiology and Biotechnology: 620–27.

Cingolani, Pablo, Cingolani Pablo, Platts Adrian, Le Lily Wang, Coon Melissa, Nguyen Tung, Wang Luan, Susan J. Land, Lu Xiangyi, and Douglas M. Ruden. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff.” *Fly* 6 (2): 80–92.

Cohen, J. I., F. Wang, J. Mannick, and E. Kieff. 1989a. “Epstein-Barr Virus Nuclear Protein 2 Is a Key Determinant of Lymphocyte Transformation.” *Proceedings of the National Academy of Sciences* 86 (23): 9558–62.

Coleman, Carrie B., Eric M. Wohlford, Nicholas A. Smith, Christine A. King, Julie A. Ritchie, Paul C. Baresel, Hiroshi Kimura, and Rosemary Rochford. 2015. “Epstein-Barr Virus Type 2 Latently Infects T Cells, Inducing an Atypical Activation Characterized by Expression of Lymphotactic Cytokines.” *Journal of Virology* 89 (4): 2301–12.

Consortium, Victorian Bioinformatics, and Others. 2012. “Velvetoptimiser.” *Available: Bioinformatics. Net. Au/software. Velvetoptimiser. Shtml. Accessed* 22.

Correa, Rita Mariel, María Dolores Fellner, Lidia Virginia Alonio, Karina Durand, Angélica R. Teyssié, and María Alejandra Picconi. 2004. “Epstein-Barr Virus (EBV) in Healthy Carriers: Distribution of Genotypes and 30 Bp Deletion in Latent Membrane Protein-1 (LMP-1) Oncogene.” *Journal of Medical Virology* 73 (4): 583–88.

Corvalan, Alejandro, Shan Ding, Chihaya Koriyama, Edwin Carrascal, Gabriel Carrasquilla, Claudia Backhouse, Luz Urzua, et al. 2006. “Association of a Distinctive Strain of Epstein-Barr Virus with Gastric Cancer.” *International Journal of Cancer. Journal International Du Cancer* 118 (7): 1736–42.

Crawford, D. H. 2001. “Biology and Disease Associations of Epstein-Barr Virus.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 356 (1408): 461–73.

Cui, Ying, Yun Wang, Xia Liu, Yan Chao, Xiaoming Xing, Chengquan Zhao, Chengyu Liu, and Bing Luo. 2011. “Genotypic Analysis of Epstein-Barr Virus Isolates Associated with Nasopharyngeal Carcinoma in Northern China.” *Intervirology* 54 (3): 131–38.

Dambaugh, T., K. Hennessy, L. Chamnankit, and E. Kieff. 1984a. “U2 Region of Epstein-Barr Virus DNA May Encode Epstein-Barr Nuclear Antigen 2.” *Proceedings of the National Academy of Sciences of the United States of America* 81 (23): 7632–36.

Dave, Sandeep S., Kai Fu, George W. Wright, Lloyd T. Lam, Philip Kluin, Evert-Jan Boerma, Timothy C. Greiner, et al. 2006. “Molecular Diagnosis of Burkitt’s Lymphoma.” *The New England Journal of Medicine* 354 (23): 2431–42.

Depledge, Daniel P., Anne L. Palser, Simon J. Watson, Imogen Yi-Chun Lai, Eleanor R. Gray, Paul Grant, Ravinder K. Kanda, Emily Leproust, Paul Kellam, and Judith Breuer. 2011. “Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples.” *PloS One* 6 (11): e27805.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013a. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

Dolan, Aidan, Clare Addison, Derek Gatherer, Andrew J. Davison, and Duncan J. McGeoch. 2006. “The Genome of Epstein-Barr Virus Type 2 Strain AG876.” *Virology* 350 (1): 164–70.

Donati, Daria, Eva Espmark, Fred Kironde, Edward Katongole Mbidde, Moses Kamya, Ake Lundkvist, Mats Wahlgren, Maria Teresa Bejarano, and Kerstin I. Falk. 2006. “Clearance of Circulating Epstein-Barr Virus DNA in Children with Acute Malaria after Antimalaria Treatment.” *The Journal of Infectious Diseases* 193 (7): 971–77.

Donati, Daria, Li Ping Zhang, Arnaud Chêne, Qijun Chen, Kirsten Flick, Maja Nyström, Mats Wahlgren, and Maria Teresa Bejarano. 2004. “Identification of a Polyclonal B-Cell Activator in Plasmodium Falciparum.” *Infection and Immunity* 72 (9): 5412–18.

Esteban, J. A., M. Salas, and L. Blanco. 1993. “Fidelity of Phi 29 DNA Polymerase. Comparison between Protein-Primed Initiation and DNA Polymerization.” *The Journal of Biological Chemistry* 268 (4): 2719–26.

Feng, Fu-Tuo, Qian Cui, Wen-Sheng Liu, Yun-Miao Guo, Qi-Sheng Feng, Li-Zhen Chen, Miao Xu, et al. 2015. “A Single Nucleotide Polymorphism in the Epstein-Barr Virus Genome Is Strongly Associated with a High Risk of Nasopharyngeal Carcinoma.” *Chinese Journal of Cancer* 34 (12): 563–72.

Ferry, Judith A. 2006a. “Burkitt’s Lymphoma: Clinicopathologic Features and Differential

Diagnosis." *The Oncologist* 11 (4): 375–83.

Forte, Eleonora, Raul E. Salinas, Christina Chang, Ting Zhou, Sarah D. Linnstaedt, Eva Gottwein, Cassandra Jacobs, et al. 2012. "The Epstein-Barr Virus (EBV)-Induced Tumor Suppressor microRNA MiR-34a Is Growth Promoting in EBV-Infected B Cells." *Journal of Virology* 86 (12): 6889–98.

Fujita, Shuichi, Nathan Buziba, Atsushi Kumatori, Masachika Senba, Akira Yamaguchi, and Kan Toriyama. 2004. "Early Stage of Epstein-Barr Virus Lytic Infection Leading to the 'Starry Sky' Pattern Formation in Endemic Burkitt Lymphoma." *Archives of Pathology & Laboratory Medicine* 128 (5): 549–52.

Garber, Manuel, Manfred G. Grabherr, Mitchell Guttman, and Cole Trapnell. 2011. "Computational Methods for Transcriptome Annotation and Quantification Using RNA-Seq." *Nature Methods* 8 (6). Nature Publishing Group: 469–77.

Gatto, Francesca, Giulia Cassina, Francesco Broccolo, Giuseppe Morreale, Edoardo Lanino, Eddi Di Marco, Efthiya Vardas, et al. 2011. "A Multiplex Calibrated Real-Time PCR Assay for Quantitation of DNA of EBV-1 and 2." *Journal of Virological Methods* 178 (1-2): 98–105.

Gire, Stephen K., Augustine Goba, Kristian G. Andersen, Rachel S. G. Sealfon, Daniel J. Park, Lansana Kanneh, Simbirie Jalloh, et al. 2014. "Genomic Surveillance Elucidates Ebola Virus Origin and Transmission during the 2014 Outbreak." *Science* 345 (6202): 1369–72.

Hansen, Kasper Daniel, and Zhijin Wu. 2012. "CQN (Conditional Quantile Normalization)." *Dim (montgomery. Subset)* 1 (23552): 10.

Hansen, Kasper D., Rafael A. Irizarry, and Zhijin Wu. 2012. "Removing Technical Variability in RNA-Seq Data Using Conditional Quantile Normalization." *Biostatistics* 13 (2): 204–16.

Hassan, Rocio, Claudete Esteves Klumb, Fabricio E. Felisbino, Deisy M. Guiretti, Lídia R. White, Claudio Gustavo Stefanoff, Mario Henrique M. Barros, Héctor N. Seuánez, and Ilana R. Zalcberg. 2008. "Clinical and Demographic Characteristics of Epstein-Barr Virus-Associated Childhood Burkitt's Lymphoma in Southeastern Brazil: Epidemiological Insights from an Intermediate Risk Region." *Haematologica* 93 (5): 780–83.

Hecht, J. L., and J. C. Aster. 2000. "Molecular Biology of Burkitt's Lymphoma." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 18 (21): 3707–21.

Heikkinen, K., S-M Karppinen, Y. Soini, M. Mäkinen, and R. Winqvist. 2003. "Mutation Screening of Mre11 Complex Genes: Indication of RAD50 Involvement in Breast and Ovarian Cancer Susceptibility." *Journal of Medical Genetics* 40 (12): e131.

Heyer, Erin E., Hakan Ozadam, Emiliano P. Ricci, Can Cenik, and Melissa J. Moore. 2015. "An Optimized Kit-Free Method for Making Strand-Specific Deep Sequencing Libraries from RNA Fragments." *Nucleic Acids Research* 43 (1): e2.

Hu, L. F., F. Chen, X. Zheng, I. Ernberg, S. L. Cao, B. Christensson, G. Klein, and G. Winberg. 1993. "Clonability and Tumorigenicity of Human Epithelial Cells Expressing the EBV Encoded Membrane Protein LMP1." *Oncogene* 8 (6): 1575–83.

Hu, L. F., E. R. Zabarovsky, F. Chen, S. L. Cao, I. Ernberg, G. Klein, and G. Winberg. 1991. "Isolation and Sequencing of the Epstein-Barr Virus BNLF-1 Gene (LMP1) from a Chinese Nasopharyngeal Carcinoma." *The Journal of General Virology* 72 (Pt 10) (October): 2399–2409.

Hummel, Michael, Stefan Bentink, Hilmar Berger, Wolfram Klapper, Swen Wessendorf, Thomas F. E. Barth, Heinz-Wolfram Bernd, et al. 2006. “A Biologic Definition of Burkitt’s Lymphoma from Transcriptional and Genomic Profiling.” *The New England Journal of Medicine* 354 (23): 2419–30.

Ikuta, K., Y. Satoh, Y. Hoshikawa, and T. Sairenji. 2000. “Detection of Epstein-Barr Virus in Salivas and Throat Washings in Healthy Adults and Children.” *Microbes and Infection / Institut Pasteur* 2 (2): 115–20.

Janz, Siegfried, Michael Potter, and Charles S. Rabkin. 2003. “Lymphoma- and Leukemia-Associated Chromosomal Translocations in Healthy Individuals.” *Genes, Chromosomes & Cancer* 36 (3): 211–23.

Jarrett, Ruth F., Gail L. Stark, Jo White, Brian Angus, Freda E. Alexander, Andrew S. Krajewski, June Freeland, G. Malcolm Taylor, Penelope R. A. Taylor, and Scotland and Newcastle Epidemiology of Hodgkin Disease Study Group. 2005. “Impact of Tumor Epstein-Barr Virus Status on Presenting Features and Outcome in Age-Defined Subgroups of Patients with Classic Hodgkin Lymphoma: A Population-Based Study.” *Blood* 106 (7): 2444–51.

Jeng, K. C., C. Y. Hsu, M. T. Liu, T. T. Chung, and S. T. Liu. 1994. “Prevalence of Taiwan Variant of Epstein-Barr Virus in Throat Washings from Patients with Head and Neck Tumors in Taiwan.” *Journal of Clinical Microbiology* 32 (1): 28–31.

Jia, Yuping, Yun Wang, Yan Chao, Yongzheng Jing, Zhifu Sun, and Bing Luo. 2010. “Sequence Analysis of the Epstein-Barr Virus (EBV) BRLF1 Gene in Nasopharyngeal and Gastric Carcinomas.” *Virology Journal* 7 (November): 341.

Jukes, Thomas H., Charles R. Cantor, and Others. 1969. “Evolution of Protein Molecules.” *Mammalian Protein Metabolism* 3 (21). New York: 132.

Jung, Peter, and Heiko Hermeking. 2009. “The c-MYC-AP4-p21 Cascade.” *Cell Cycle* 8 (7): 982–89.

Kadoch, Cigall, Diana C. Hargreaves, Courtney Hodges, Laura Elias, Lena Ho, Jeff Ranish, and Gerald R. Crabtree. 2013. “Proteomic and Bioinformatic Analysis of Mammalian SWI/SNF Complexes Identifies Extensive Roles in Human Malignancy.” *Nature Genetics* 45 (6): 592–601.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80.

Kawauchi, Kiyotaka, Toshie Ogasawara, Masako Yasuyama, Kuniaki Otsuka, and Osamu Yamada. 2009. “The PI3K/Akt Pathway as a Target in the Treatment of Hematologic Malignancies.” *Anti-Cancer Agents in Medicinal Chemistry* 9 (5): 550–59.

Kaye, K. M., K. M. Izumi, and E. Kieff. 1993. “Epstein-Barr Virus Latent Membrane Protein 1 Is Essential for B-Lymphocyte Growth Transformation.” *Proceedings of the National Academy of Sciences of the United States of America* 90 (19): 9150–54.

Kazembe, P., P. B. Hesseling, B. E. Griffin, I. Lampert, and G. Wessels. 2003. “Long Term Survival of Children with Burkitt Lymphoma in Malawi after Cyclophosphamide Monotherapy.” *Medical and Pediatric Oncology* 40 (1): 23–25.

Kelly, Gemma, Andrew Bell, and Alan Rickinson. 2002a. “Epstein-Barr Virus-Associated Burkitt Lymphomagenesis Selects for Downregulation of the Nuclear Antigen EBNA2.” *Nature Medicine* 8 (10). Nature Publishing Group: 1098–1104.

Khan, G., E. M. Miyashita, B. Yang, G. J. Babcock, and D. A. Thorley-Lawson. 1996. “Is EBV Persistence in Vivo a Model for B Cell Homeostasis?” *Immunity* 5 (2): 173–79.

Khanim, F., Q. Y. Yao, G. Niedobitek, S. Sihota, A. B. Rickinson, and L. S. Young. 1996. "Analysis of Epstein-Barr Virus Gene Polymorphisms in Normal Donors and in Virus-Associated Tumors from Different Geographic Locations." *Blood* 88 (9): 3491–3501.

Kim, Sung-Min, So-Hee Kang, and Won-Keun Lee. 2006. "Identification of Two Types of Naturally-Occurring Intertypic Recombinants of Epstein-Barr Virus." *Molecules and Cells* 21 (2): 302–7.

Kimura, H., M. Morita, Y. Yabuta, K. Kuzushima, K. Kato, S. Kojima, T. Matsuyama, and T. Morishima. 1999. "Quantitative Analysis of Epstein-Barr Virus Load by Using a Real-Time PCR Assay." *Journal of Clinical Microbiology* 37 (1): 132–36.

Krieg, A. M. 2000. "The Role of CpG Motifs in Innate Immunity." *Current Opinion in Immunology* 12 (1): 35–43.

Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. 2016. "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets." *Molecular Biology and Evolution* 33 (7): 1870–74.

Kwok, Hin, Amy H. Y. Tong, Chi Ho Lin, Si Lok, Paul J. Farrell, Dora L. W. Kwong, and Alan K. S. Chiang. 2012a. "Genomic Sequencing and Comparative Analysis of Epstein-Barr Virus Genome Isolated from Primary Nasopharyngeal Carcinoma Biopsy." *PloS One* 7 (5). Public Library of Science: e36939.

Kwok, H., C. W. Wu, A. L. Palser, P. Kellam, P. C. Sham, D. L. W. Kwong, and A. K. S. Chiang. 2014. "Genomic Diversity of Epstein-Barr Virus Genomes Isolated from Primary Nasopharyngeal Carcinoma Biopsy Samples." *Journal of Virology* 88 (18): 10662–72.

Kyaw, M. T., L. Hurren, L. Evans, D. J. Moss, D. A. Cooper, E. Benson, D. Esmore, and T. B. Sculley. 1992. "Expression of B-Type Epstein-Barr Virus in HIV-Infected Patients and Cardiac Transplant Recipients." *AIDS Research and Human Retroviruses* 8 (11): 1869–74.

Langmead, Ben, Langmead Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.

Lay, Meav-Lang J., Robyn M. Lucas, Cheryl Toi, Mala Ratnamohan, Anne-Louise Ponsonby, and Dominic E. Dwyer. 2012. "Epstein-Barr Virus Genotypes and Strains in Central Nervous System Demyelinating Disease and Epstein-Barr Virus-Related Illnesses in Australia." *Intervirology* 55 (5): 372–79.

Lazzi, S., F. Ferrari, A. Nyongo, N. Palummo, A. de Milito, M. Zazzi, L. Leoncini, P. Luzi, and P. Tosi. 1998. "HIV-Associated Malignant Lymphomas in Kenya (Equatorial Africa)." *Human Pathology* 29 (11): 1285–89.

Leek, Jeffrey T. 2014a. "Svaseq: Removing Batch Effects and Other Unwanted Noise from Sequencing Data." *Nucleic Acids Research* 42 (21). doi:10.1093/nar/gku864.

Leichty, Aaron R., and Dustin Brisson. 2014. "Selective Whole Genome Amplification for Resequencing Target Microbial Species from Complex Natural Samples." *Genetics* 198 (2): 473–81.

Lei, Haiyan, Tianwei Li, Bingjie Li, Shien Tsai, Robert J. Biggar, Francis Nkrumah, Janet Neequaye, et al. 2015. "Epstein-Barr Virus from Burkitt Lymphoma Biopsies from Africa and South America Share Novel LMP-1 Promoter and Gene Variations." *Scientific Reports* 5 (November): 16706.

Levin, Joshua Z., Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev. 2010. "Comprehensive Comparative Analysis of Strand-Specific RNA Sequencing Methods." *Nature Methods* 7 (9): 709–15.

Liberzon, Arthur, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. 2011. "Molecular Signatures Database (MSigDB) 3.0." *Bioinformatics* 27 (12): 1739–40.

Li, Bo, Li Bo, and Colin N. Dewey. 2011a. "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome." *BMC Bioinformatics* 12 (1): 323.

Lin, De-Chen, Xuan Meng, Masaharu Hazawa, Yasunobu Nagata, Ana Maria Varela, Liang Xu, Yusuke Sato, et al. 2014. "The Genomic Landscape of Nasopharyngeal Carcinoma." *Nature Genetics* 46 (8): 866–71.

Lin, Jin-Ching, Wen-Yi Wang, Kuang Y. Chen, Yau-Huei Wei, Wen-Miin Liang, Jian-Sheng Jan, and Rong-San Jiang. 2004. "Quantification of Plasma Epstein-Barr Virus DNA in Patients with Advanced Nasopharyngeal Carcinoma." *The New England Journal of Medicine* 350 (24): 2461–70.

Lin, Zhen, Xia Wang, Michael J. Strong, Monica Concha, Melody Baddoo, Guorong Xu, Carl Baribault, et al. 2013. "Whole-Genome Sequencing of the Akata and Mutu Epstein-Barr Virus Strains." *Journal of Virology* 87 (2): 1172–82.

Liu, Pan, Xiaodong Fang, Zizhen Feng, Yun-Miao Guo, Rou-Jun Peng, Tengfei Liu, Zhiyong Huang, et al. 2011. "Direct Sequencing and Characterization of a Clinical Isolate of Epstein-Barr Virus from Nasopharyngeal Carcinoma Tissue by Using next-Generation Sequencing Technology." *Journal of Virology* 85 (21): 11291–99.

Liu, Ying, Wenjun Yang, Yaqi Pan, Jiafu Ji, Zheming Lu, and Yang Ke. 2016. "Genome-Wide Analysis of Epstein-Barr Virus (EBV) Isolated from EBV-Associated Gastric Carcinoma (EBVaGC)." *Oncotarget* 7 (4): 4903–14.

Lo, Kwok-Wai, Grace Tin-Yun Chung, and Ka-Fai To. 2012. "Deciphering the Molecular Genetic Basis of NPC through Molecular, Cytogenetic, and Epigenetic Approaches." *Seminars in Cancer Biology* 22 (2): 79–86.

Love, Cassandra, Love Cassandra, Sun Zhen, Jima Dereje, Li Guojie, Zhang Jenny, Miles Rodney, et al. 2012. "The Genetic Landscape of Mutations in Burkitt Lymphoma." *Nature Genetics* 44 (12): 1321–25.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014a. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550.

Lucchesi, Walter, Gareth Brady, Oliver Dittrich-Breiholz, Michael Kracht, Rainer Russ, and Paul J. Farrell. 2008. "Differential Gene Regulation by Epstein-Barr Virus Type 1 and Type 2 EBNA2." *Journal of Virology* 82 (15): 7456–66.

Ma, Brigette B. Y., Ann King, Y. M. Dennis Lo, Y. Y. Yau, Benny Zee, Edwin P. Hui, Sing F. Leung, et al. 2006. "Relationship between Pretreatment Level of Plasma Epstein-Barr Virus DNA, Tumor Burden, and Metabolic Activity in Advanced Nasopharyngeal Carcinoma." *International Journal of Radiation Oncology, Biology, Physics* 66 (3): 714–20.

Magrath, I. 1990. "The Pathogenesis of Burkitt's Lymphoma." *Advances in Cancer Research* 55: 133–270.

Magrath, Ian T. 1991. "African Burkitt's Lymphoma: History, Biology, Clinical Features, and Treatment." *Journal of Pediatric Hematology/oncology* 13 (2): 222.

Marriott, Susan J., and Oliver John Semmes. 2005. "Impact of HTLV-I Tax on Cell Cycle Progression and the Cellular DNA Damage Repair Response." *Oncogene* 24 (39): 5986–

95.

Martin, Darren P., Ben Murrell, Michael Golden, Arjun Khoosal, and Brejnev Muhire. 2015. “RDP4: Detection and Analysis of Recombination Patterns in Virus Genomes.” *Virus Evolution* 1 (1): vev003.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal* 17 (1): 10–12.

Martin, Marcel, and Martin Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal* 17 (1): 10.

Marzluff, William F., Eric J. Wagner, and Robert J. Duronio. 2008. “Metabolism and Regulation of Canonical Histone mRNAs: Life without a poly (A) Tail.” *Nature Reviews. Genetics* 9 (11): 843–54.

Mathieu, Anne-Laure, Estelle Verronese, Gillian I. Rice, Fanny Fouyssac, Yves Bertrand, Capucine Picard, Marie Chansel, et al. 2015. “PRKDC Mutations Associated with Immunodeficiency, Granuloma, and Autoimmune Regulator-Dependent Autoimmunity.” *The Journal of Allergy and Clinical Immunology* 135 (6): 1578–88.e5.

Mbulaiteye, Sam M., Robert J. Biggar, Kishor Bhatia, Martha S. Linet, and Susan S. Devesa. 2009. “Sporadic Childhood Burkitt Lymphoma Incidence in the United States during 1992-2005.” *Pediatric Blood & Cancer* 53 (3): 366–70.

McGeoch, Duncan J., and Derek Gatherer. 2007. “Lineage Structures in the Genome Sequences of Three Epstein-Barr Virus Strains.” *Virology* 359 (1): 1–5.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010a. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303.

McLaughlin-Drubin, Margaret E., and Karl Münger. 2009. “Oncogenic Activities of Human Papillomaviruses.” *Virus Research* 143 (2): 195–208.

Midgley, R. S., N. W. Blake, Q. Y. Yao, D. Croom-Carter, S. T. Cheung, S. F. Leung, A. T. Chan, et al. 2000. “Novel Intertypic Recombinants of Epstein-Barr Virus in the Chinese Population.” *Journal of Virology* 74 (3): 1544–48.

Moody, Cary A., Rona S. Scott, Nazanin Amirghahari, Cherie-Ann Nathan, Lawrence S. Young, Chris W. Dawson, and John W. Sixbey. 2005. “Modulation of the Cell Growth Regulator mTOR by Epstein-Barr Virus-Encoded LMP2A.” *Journal of Virology* 79 (9): 5499–5506.

Moormann, Ann M., Kiprotich Chelimo, Odada P. Sumba, Mary L. Lutzke, Robert Ploutz-Snyder, Duane Newton, James Kazura, and Rosemary Rochford. 2005. “Exposure to Holoendemic Malaria Results in Elevated Epstein-Barr Virus Loads in Children.” *The Journal of Infectious Diseases* 191 (8): 1233–38.

Moormann, Ann M., Kevin N. Heller, Kiprotich Chelimo, Paula Embury, Robert Ploutz-Snyder, Juliana A. Otieno, Margaret Oduor, Christian Münz, and Rosemary Rochford. 2009a. “Children with Endemic Burkitt Lymphoma Are Deficient in EBNA1-Specific IFN-Gamma T Cell Responses.” *International Journal of Cancer. Journal International Du Cancer* 124 (7): 1721–26.

Morin, Ryan D., Karen Mungall, Erin Pleasance, Andrew J. Mungall, Rodrigo Goya, Ryan D. Huff, David W. Scott, et al. 2013. “Mutational and Structural Analysis of Diffuse Large B-Cell Lymphoma Using Whole-Genome Sequencing.” *Blood* 122 (7): 1256–65.

Morrow, R. H., Jr. 1985. “Epidemiological Evidence for the Role of Falciparum Malaria in the Pathogenesis of Burkitt’s Lymphoma.” *IARC Scientific Publications*, no. 60: 177–86.

Mulama, David H., Jeffrey A. Bailey, Joslyn Foley, Kiprotich Chelimo, Collins Ouma, Walter Gzo Jura, Juliana Otieno, John Vulule, and Ann M. Moormann. 2014. "Sickle Cell Trait Is Not Associated with Endemic Burkitt Lymphoma: An Ethnicity and Malaria Endemicity-Matched Case--Control Study Suggests Factors Controlling EBV May Serve as a Predictive Biomarker for This Pediatric Cancer." *International Journal of Cancer* 134 (3). Wiley Online Library: 645–53.

Mwanda, O. W. 2004. "Clinical Characteristics of Burkitt's Lymphoma Seen in Kenyan Patients." *East African Medical Journal*, no. 8 Suppl (August): S78–89.

Navari, Mohsen, Maryam Etebari, Giulia De Falco, Maria R. Ambrosio, Davide Gibellini, Lorenzo Leoncini, and Pier Paolo Piccaluga. 2015. "The Presence of Epstein-Barr Virus Significantly Impacts the Transcriptional Profile in Immunodeficiency-Associated Burkitt Lymphoma." *Frontiers in Microbiology* 6 (June): 556.

Neves, Marco, Joana Marinho-Dias, Joana Ribeiro, and Hugo Sousa. 2017. "Epstein-Barr Virus Strains and Variations: Geographic or Disease-Specific Variants?" *Journal of Medical Virology* 89 (3): 373–87.

Ogwang, Martin D., Kishor Bhatia, Robert J. Biggar, and Sam M. Mbulaiteye. 2008. "Incidence and Geographic Distribution of Endemic Burkitt Lymphoma in Northern Uganda Revisited." *International Journal of Cancer. Journal International Du Cancer* 123 (11): 2658–63.

Ojesina, Akinyemi I., Lee Lichtenstein, Samuel S. Freeman, Chandra Sekhar Pedamallu, Ivan Imaz-Rosshandler, Trevor J. Pugh, Andrew D. Cherniack, et al. 2014. "Landscape of Genomic Alterations in Cervical Carcinomas." *Nature* 506 (7488): 371–75.

Okkels, Henrik, Karen Lindorff-Larsen, Ole Thorlasius-Ussing, Mogens Vyberg, Jan Lindebjerg, Lone Sunde, Inge Bernstein, Louise Klarskov, Susanne Holck, and Henrik Bygum Krarup. 2012. "MSH6 Mutations Are Frequent in Hereditary Nonpolyposis Colorectal Cancer Families with Normal pMSH6 Expression as Detected by Immunohistochemistry." *Applied Immunohistochemistry & Molecular Morphology: AIMM / Official Publication of the Society for Applied Immunohistochemistry* 20 (5): 470–77.

Palma, Icela, Ana Elena Sánchez, Elva Jiménez-Hernández, Francisco Alvarez-Rodríguez, Margarita Nava-Frias, Pedro Valencia-Mayoral, Citlatepet Salinas-Lara, et al. 2013. "Detection of Epstein-Barr Virus and Genotyping Based on EBNA2 Protein in Mexican Patients with Hodgkin Lymphoma: A Comparative Study in Children and Adults." *Clinical Lymphoma, Myeloma & Leukemia* 13 (3): 266–72.

Palser, Anne L., Nicholas E. Grayson, Robert E. White, Craig Corton, Samantha Correia, Mohammed M. Ba Abdullah, Simon J. Watson, et al. 2015. "Genome Diversity of Epstein-Barr Virus from Multiple Tumor Types and Normal Infection." *Journal of Virology* 89 (10): 5222–37.

Parker, B. D., A. Bankier, S. Satchwell, B. Barrell, and P. J. Farrell. 1990. "Sequence and Transcription of Raji Epstein-Barr Virus DNA Spanning the B95-8 Deletion Region." *Virology* 179 (1): 339–46.

Parker, H. S., J. T. Leek, A. V. Favorov, M. Considine, X. Xia, S. Chavan, C. H. Chung, and E. J. Fertig. 2014. "Preserving Biological Heterogeneity with a Permuted Surrogate Variable Analysis for Genomics Batch Correction." *Bioinformatics* 30 (19): 2757–63.

Parkhomchuk, Dmitri, Tatiana Borodina, Vyacheslav Amstislavskiy, Maria Banaru, Linda Hallen, Sylvia Krobitsch, Hans Lehrach, and Alexey Soldatov. 2009. "Transcriptome Analysis by Strand-Specific Sequencing of Complementary DNA." *Nucleic Acids*

Research 37 (18): e123.

Park, Sarah, Jeeyun Lee, Young Hyeh Ko, Arum Han, Hyun Jung Jun, Sang Chul Lee, In Gyu Hwang, et al. 2007. "The Impact of Epstein-Barr Virus Status on Clinical Outcome in Diffuse Large B-Cell Lymphoma." *Blood* 110 (3): 972–78.

Peh, Suat-Cheng, Lian-Hua Kim, and Sibrand Poppema. 2002. "Frequent Presence of Subtype A Virus in Epstein-Barr Virus-Associated Malignancies." *Pathology* 34 (5): 446–50.

Pelicci, P. G., D. M. Knowles 2nd, I. Magrath, and R. Dalla-Favera. 1986. "Chromosomal Breakpoints and Structural Alterations of the c-Myc Locus Differ in Endemic and Sporadic Forms of Burkitt Lymphoma." *Proceedings of the National Academy of Sciences of the United States of America* 83 (9): 2984–88.

Peng, Stanford L. 2005. "Signaling in B Cells via Toll-like Receptors." *Current Opinion in Immunology* 17 (3): 230–36.

Piccaluga, Pier Paolo, Giulia De Falco, Manjunath Kustagi, Anna Gazzola, Claudio Agostinelli, Claudio Tripodo, Eleonora Leucci, et al. 2011a. "Gene Expression Analysis Uncovers Similarity and Differences among Burkitt Lymphoma Subtypes." *Blood* 117 (13): 3596–3608.

Polack, A., H. Delius, U. Zimber, and G. W. Bornkamm. 1984. "Two Deletions in the Epstein-Barr Virus Genome of the Burkitt Lymphoma Nonproducer Line Raji." *Virology* 133 (1): 146–57.

Price, Alexander M., and Micah A. Luftig. 2015. "To Be or Not IIb: A Multi-Step Process for Epstein-Barr Virus Latency Establishment and Consequences for B Cell Tumorigenesis." PLoS Pathogens 11 (3): e1004656.

Rainey, Jeanette Jane. 2005. *Epidemiological and Environmental Co-Factors Linked to Endemic Burkitt's Lymphoma in Kenya*. University of Michigan.

Rapisuwon, Suthee, Eveline E. Vietsch, and Anton Wellstein. 2016. "Circulating Biomarkers to Monitor Cancer Progression and Treatment." *Computational and Structural Biotechnology Journal* 14 (June): 211–22.

Richter, Julia, Matthias Schlesner, Steve Hoffmann, Markus Kreuz, Ellen Leich, Birgit Burkhardt, Maciej Rosolowski, et al. 2012. "Recurrent Mutation of the ID3 Gene in Burkitt Lymphoma Identified by Integrated Genome, Exome and Transcriptome Sequencing." *Nature Genetics* 44 (12): 1316–20.

Rickert, Robert C. 2013. "New Insights into Pre-BCR and BCR Signalling with Relevance to B Cell Malignancies." *Nature Reviews. Immunology* 13 (8): 578–91.

Rickinson, A. B., L. S. Young, and M. Rowe. 1987a. "Influence of the Epstein-Barr Virus Nuclear Antigen EBNA 2 on the Growth Phenotype of Virus-Transformed B Cells." *Journal of Virology* 61 (5): 1310–17.

Robbiani, Davide F., Stephanie Deroubaix, Niklas Feldhahn, Thiago Y. Oliveira, Elsa Callen, Qiao Wang, Mila Jankovic, et al. 2015. "Plasmodium Infection Promotes Genomic Instability and AID-Dependent B Cell Lymphoma." *Cell* 162 (4): 727–37.

Rochford, Rosemary, Martin J. Cannon, and Ann M. Moormann. 2005. "Endemic Burkitt's Lymphoma: A Polymicrobial Disease?" *Nature Reviews. Microbiology* 3 (2): 182–87.

Rowe, M., D. T. Rowe, C. D. Gregory, L. S. Young, P. J. Farrell, H. Rupani, and A. B. Rickinson. 1987. "Differences in B Cell Growth Phenotype Reflect Novel Patterns of Epstein-Barr Virus Latent Gene Expression in Burkitt's Lymphoma Cells." *The EMBO Journal* 6 (9): 2743–51.

Rowe, M., L. S. Young, K. Cadwallader, L. Petti, E. Kieff, and A. B. Rickinson. 1989.

"Distinction between Epstein-Barr Virus Type A (EBNA 2A) and Type B (EBNA 2B) Isolates Extends to the EBNA 3 Family of Nuclear Proteins." *Journal of Virology* 63 (3): 1031–39.

Ruprecht, Claudia R., and Antonio Lanzavecchia. 2006. "Toll-like Receptor Stimulation as a Third Signal Required for Activation of Human Naive B Cells." *European Journal of Immunology* 36 (4): 810–16.

Ryan, Julie L., Hongxin Fan, Lode J. Swinnen, Steven A. Schichman, Nancy Raab-Traub, Mary Covington, Sandra Elmore, and Margaret L. Gulley. 2004. "Epstein-Barr Virus (EBV) DNA in Plasma Is Not Encapsidated in Patients with EBV-Related Malignancies." *Diagnostic Molecular Pathology: The American Journal of Surgical Pathology, Part B* 13 (2): 61–68.

Saitou, N., and M. Nei. 1987. "The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees." *Molecular Biology and Evolution* 4 (4): 406–25.

Sandvej, K., J. W. Gratama, M. Munch, X. G. Zhou, R. L. Bolhuis, B. S. Andresen, N. Gregersen, and S. Hamilton-Dutoit. 1997. "Sequence Analysis of the Epstein-Barr Virus (EBV) Latent Membrane Protein-1 Gene and Promoter Region: Identification of Four Variants among Wild-Type EBV Isolates." *Blood* 90 (1): 323–30.

Satou, Akira, Satou Akira, Asano Naoko, Nakazawa Atsuko, Osumi Tomoo, Tsurusawa Masahito, Ishiguro Atsushi, et al. 2015. "Epstein-Barr Virus (EBV)-Positive Sporadic Burkitt Lymphoma." *The American Journal of Surgical Pathology* 39 (2): 227–35.

Schmitz, Roland, Schmitz Roland, Ryan M. Young, Ceribelli Michele, Jhavar Sameer, Xiao Wenming, Zhang Meili, et al. 2012a. "Burkitt Lymphoma Pathogenesis and Therapeutic Targets from Structural and Functional Genomics." *Nature* 490 (7418): 116–20.

Sculley, T. B., A. Apolloni, L. Hurren, D. J. Moss, and D. A. Cooper. 1990. "Coinfection with A- and B-Type Epstein-Barr Virus in Human Immunodeficiency Virus-Positive Subjects." *The Journal of Infectious Diseases* 162 (3): 643–48.

Shepard, Peter J., Eun-A Choi, Jente Lu, Lisa A. Flanagan, Klemens J. Hertel, and Yongsheng Shi. 2011. "Complex and Dynamic Landscape of RNA Polyadenylation Revealed by PAS-Seq." *RNA* 17 (4): 761–72.

Sheppard, Sarah, Nathan D. Lawson, and Lihua Julie Zhu. 2013. "Accurate Identification of Polyadenylation Sites from 3' End Deep Sequencing Using a Naive Bayes Classifier." *Bioinformatics* 29 (20): 2564–71.

Shu, C. H., Y. S. Chang, C. L. Liang, S. T. Liu, C. Z. Lin, and P. Chang. 1992. "Distribution of Type A and Type B EBV in Normal Individuals and Patients with Head and Neck Carcinomas in Taiwan." *Journal of Virological Methods* 38 (1): 123–30.

Sijts, Alice, Yuancheng Sun, Katarina Janek, Sylvie Kral, Annettte Paschen, Dirk Schadendorf, and Peter-M Kloetzel. 2002. "The Role of the Proteasome Activator PA28 in MHC Class I Antigen Processing." *Molecular Immunology* 39 (3-4): 165–69.

Simbiri, Kenneth O., Nicholas A. Smith, Richard Otieno, Eric E. M. Wohlford, Ibrahim I. Daud, Sumba P. Odada, Frank Middleton, and Rosemary Rochford. 2015. "Epstein-Barr Virus Genetic Variation in Lymphoblastoid Cell Lines Derived from Kenyan Pediatric Population." *PloS One* 10 (5): e0125420.

Simone, Olivia, Maria Teresa Bejarano, Susan K. Pierce, Salvatore Antonaci, Mats Wahlgren, Marita Troye-Blomberg, and Daria Donati. 2011. "TLRs Innate Immunereceptors and Plasmodium Falciparum Erythrocyte Membrane Protein 1 (PfEMP1) CIDR1α-Driven Human Polyclonal B-Cell Activation." *Acta Tropica* 119 (2-

3): 144–50.

Simon, K. C., X. Yang, K. L. Munger, and A. Ascherio. 2011. "EBNA1 and LMP1 Variants in Multiple Sclerosis Cases and Controls." *Acta Neurologica Scandinavica* 124 (1): 53–58.

Sims, David, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, and Chris P. Ponting. 2014. "Sequencing Depth and Coverage: Key Considerations in Genomic Analyses." *Nature Reviews. Genetics* 15 (2): 121–32.

Sixbey, J. W., P. Shirley, P. J. Chesney, D. M. Buntin, and L. Resnick. 1989. "Detection of a Second Widespread Strain of Epstein-Barr Virus." *The Lancet* 2 (8666): 761–65.

Skare, J., J. Farley, J. L. Strominger, K. O. Fresen, M. S. Cho, and H. zur Hausen. 1985. "Transformation by Epstein-Barr Virus Requires DNA Sequences in the Region of BamHI Fragments Y and H." *Journal of Virology* 55 (2): 286–97.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50.

Swain, Martin T., Isheng J. Tsai, Samual A. Assefa, Chris Newbold, Matthew Berriman, and Thomas D. Otto. 2012. "A Post-Assembly Genome-Improvement Toolkit (PAGIT) to Obtain Annotated Genomes from Contigs." *Nature Protocols* 7 (7): 1260–84.

Tamura, Koichiro, Glen Stecher, Daniel Peterson, Alan Filipski, and Sudhir Kumar. 2013. "MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0." *Molecular Biology and Evolution* 30 (12): 2725–29.

Teng, B., K. S. Murthy, J. F. Kuemmerle, J. R. Grider, K. Sase, T. Michel, and G. M. Makhlouf. 1998. "Expression of Endothelial Nitric Oxide Synthase in Human and Rabbit Gastrointestinal Smooth Muscle Cells." *The American Journal of Physiology* 275 (2 Pt 1): G342–51.

Thorley-Lawson, David A., and Andrew Gross. 2004. "Persistence of the Epstein-Barr Virus and the Origins of Associated Lymphomas." The New England Journal of Medicine 350 (13): 1328–37.

Tiwawech, Danai, Petcharin Srivatanakul, Anant Karalak, and Takafumi Ishida. 2008. "Association between EBNA2 and LMP1 Subtypes of Epstein-Barr Virus and Nasopharyngeal Carcinoma in Thais." *Journal of Clinical Virology: The Official Publication of the Pan American Society for Clinical Virology* 42 (1): 1–6.

Torgbor, Charles, Peter Awuah, Kirk Deitsch, Parisa Kalantari, Karen A. Duca, and David A. Thorley-Lawson. 2014a. "A Multifactorial Role for P. Falciparum Malaria in Endemic Burkitt's Lymphoma Pathogenesis." *PLoS Pathogens* 10 (5): e1004170.

Tsujimoto, Kyoko, Takeshi Ono, Masaki Sato, Takashi Nishida, Takemi Oguma, and Takushi Tadakuma. 2005. "Regulation of the Expression of Caspase-9 by the Transcription Factor Activator Protein-4 in Glucocorticoid-Induced Apoptosis." *The Journal of Biological Chemistry* 280 (30): 27638–44.

Tzellos, Stelios, and Paul J. Farrell. 2012. "Epstein-Barr Virus Sequence Variation—Biology and Disease." *Pathogens* 1 (2). Multidisciplinary Digital Publishing Institute: 156–74.

Wang, Shanshan, Hongchao Xiong, Shi Yan, Nan Wu, and Zheming Lu. 2016. "Identification and Characterization of Epstein-Barr Virus Genomes in Lung Carcinoma Biopsy Samples by Next-Generation Sequencing Technology." *Scientific*

Reports 6 (May): 26156.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews. Genetics* 10 (1): 57–63.

Weiner, George J. 2009. "CpG Oligodeoxynucleotide-Based Therapy of Lymphoid Malignancies." *Advanced Drug Delivery Reviews* 61 (3): 263–67.

Westmoreland, Katherine D., Nathan D. Montgomery, Christopher C. Stanley, Nader Kim El-Mallawany, Peter Wasswa, Toon van der Gronde, Idah Mtete, et al. 2017. "Plasma Epstein-Barr Virus DNA for Pediatric Burkitt Lymphoma Diagnosis, Prognosis and Response Assessment in Malawi." *International Journal of Cancer. Journal International Du Cancer*, March. doi:10.1002/ijc.30682.

White, Robert E., Patrick C. Rämer, Kikkeri N. Naresh, Sonja Meixlsperger, Laurie Pinaud, Cliona Rooney, Barbara Savoldo, et al. 2012. "EBNA3B-Deficient EBV Promotes B Cell Lymphomagenesis in Humanized Mice and Is Found in Human Tumors." *The Journal of Clinical Investigation* 122 (4): 1487–1502.

Wilmore, Joel R., Amolo S. Asito, Chungwen Wei, Erwan Piriou, P. Odada Sumba, Iñaki Sanz, and Rosemary Rochford. 2015. "AID Expression in Peripheral Blood of Children Living in a Malaria Holoendemic Region Is Associated with Changes in B Cell Subsets and Epstein-Barr Virus." *International Journal of Cancer* 136 (6). Wiley Online Library: 1371–80.

Wohlford, Eric M., Amolo S. Asito, Kiprotich Chelimo, Peter O. Sumba, Paul C. Baresel, Rebecca A. Oot, Ann M. Moormann, and Rosemary Rochford. 2013. "Identification of a Novel Variant of LMP-1 of EBV in Patients with Endemic Burkitt Lymphoma in Western Kenya." *Infectious Agents and Cancer* 8 (1): 34.

Wysoker, A., K. Tibbetts, and T. Fennell. 2013. "Picard Tools Version 1.90."

Xue, Shao-An, Louise G. Labrecque, Qi-Long Lu, S. Kate Ong, Irvin A. Lampert, Peter Kazembe, Elizabeth Molyneux, Robin L. Broadhead, Eric Borgstein, and Beverly E. Griffin. 2002. "Promiscuous Expression of Epstein-Barr Virus Genes in Burkitt's Lymphoma from the Central African Country Malawi." *International Journal of Cancer* 99 (5). Wiley Online Library: 635–43.

Yang, Dongmei, Wuguo Chen, Jie Xiong, Carly J. Sherrod, David H. Henry, and Dirk P. Dittmer. 2014. "Interleukin 1 Receptor-Associated Kinase 1 (IRAK1) Mutation Is a Common, Essential Driver for Kaposi Sarcoma Herpesvirus Lymphoma." *Proceedings of the National Academy of Sciences of the United States of America* 111 (44): E4762–68.

Yang, Hong-Jie, Yang Hong-Jie, Huang Tie-Jun, Yang Chang-Fu, Peng Li-Xie, Liu Ran-Yi, Yang Guang-Da, et al. 2013. "Comprehensive Profiling of Epstein-Barr Virus-Encoded miRNA Species Associated with Specific Latency Types in Tumor Cells." *Virology Journal* 10 (1): 314.

Yao, Q. Y., D. S. Croom-Carter, R. J. Tierney, G. Habeshaw, J. T. Wilde, F. G. Hill, C. Conlon, and A. B. Rickinson. 1998a. "Epidemiology of Infection with Epstein-Barr Virus Types 1 and 2: Lessons from the Study of a T-Cell-Immunocompromised Hemophilic Cohort." *Journal of Virology* 72 (5): 4352–63.

Yao, Q. Y., R. J. Tierney, D. Croom-Carter, G. M. Cooper, C. J. Ellis, M. Rowe, and A. B. Rickinson. 1996. "Isolation of Intertypic Recombinants of Epstein-Barr Virus from T-Cell-Immunocompromised Individuals." *Journal of Virology* 70 (8): 4895–4903.

Yao, Q. Y., R. J. Tierney, D. Croom-Carter, D. Dukers, G. M. Cooper, C. J. Ellis, M. Rowe, and A. B. Rickinson. 1996. "Frequency of Multiple Epstein-Barr Virus Infections in T-

Cell-Immunocompromised Individuals." *Journal of Virology* 70 (8): 4884–94.

Young, Lawrence S., and Alan B. Rickinson. 2004. "Epstein–Barr Virus: 40 Years on." *Nature Reviews. Cancer* 4 (10). Nature Publishing Group: 757–68.

Young, L. S., Q. Y. Yao, C. M. Rooney, T. B. Sculley, D. J. Moss, H. Rupani, G. Laux, G. W. Bornkamm, and A. B. Rickinson. 1987a. "New Type B Isolates of Epstein-Barr Virus from Burkitt's Lymphoma and from Normal Individuals in Endemic Areas." *The Journal of General Virology* 68 (Pt 11) (November): 2853–62.

Zani, V. J., N. Asou, D. Jadayel, J. M. Heward, J. Shipley, E. Nacheva, K. Takasuki, D. Catovsky, and M. J. Dyer. 1996. "Molecular Cloning of Complex Chromosomal Translocation t (8;14;12) (q24.1;q32.3;q24.1) in a Burkitt Lymphoma Cell Line Defines a New Gene (BCL7A) with Homology to Caldesmon." *Blood* 87 (8): 3124–34.

Zeng, Mu-Sheng, Da-Jiang Li, Qing-Lun Liu, Li-Bing Song, Man-Zhi Li, Ru-Hua Zhang, Xing-Juan Yu, Hui-Min Wang, Ingemar Ernberg, and Yi-Xin Zeng. 2005. "Genomic Sequence Analysis of Epstein-Barr Virus Strain GD1 from a Nasopharyngeal Carcinoma Patient." *Journal of Virology* 79 (24): 15323–30.

Zerbino, Daniel R., and Ewan Birney. 2008. "Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs." *Genome Research* 18 (5): 821–29.

Zhang, Zhao, William E. Theurkauf, Zhiping Weng, and Phillip D. Zamore. 2012. "Strand-Specific Libraries for High Throughput RNA Sequencing (RNA-Seq) Prepared without poly (A) Selection." *Silence* 3 (1): 9.

Zimber, U., H. K. Adldinger, G. M. Lenoir, M. Vuillaume, M. V. Knebel-Doeberitz, G. Laux, C. Desgranges, P. Wittmann, U. K. Freese, and U. Schneider. 1986. "Geographical Prevalence of Two Types of Epstein-Barr Virus." *Virology* 154 (1): 56–66.

APPENDIX

A. EBV Genome Typing and Viral Load Assays

The ratio of viral genomic copy number to human genomic DNA was determined with bi-plex qPCR using EBV specific primers against BALF5 gene (Kimura et al. 1999) and human beta-actin gene. Primer sequences for BALF5;
fw: 5`-CGGAAGCCCTCTGGACTTC-3`,
rw: 5`-CCCTGTTTATCCGATGGAATG-3`,
for b-actin;
fw: 5`-TCACCCACACTGTGCCCATCTACGA-3`,
rw: 5`-CAGCGGAACCGCTCATTGCCAATGG-3`
(Moormann et al. 2005).

Genotyping the EBV DNA was performed using conventional primers spanning EBNA3C gene producing 153bp and 246bp products for type I and type II genomes, respectively.
Primer sequences for EBNA3C;
fw: 5`-AGAAGGGGAGCGTGTGTTG-3`,
rw: 5`-GGCTCGTTTTTGACGTCGG-3`.

B. EBV Genome Sequencing Complete Protocol

Input DNA preparation:

1. Extraction/Isolation/Purification of DNA from clinical specimen (Biopsy, plasma, cell line)
2. Quantification of Input DNA with PicoGreen
3. Whole Genome Amplification with GenomiPhi v2 Phi29 polymerase
4. Alternatively, PCR-sWGA
5. Cleaning/purification after WGA using XP Ampure beads
6. Quantify the Amplified DNA with PicoGreen
7. Check the quality of DNA with NanoDrop for OD 260/280 ratio
8. Determine viral/human DNA before and after WGA

Sequencing Library Preparation:

9. Shearing DNA with Covaris ultrasonicator
10. Assess DNA quality
11. Blunt-end repair
12. Cleaning/purification/"Size selection" using XP Ampure beads
13. 3'-end Adenylation of DNA fragments
14. Cleaning/purification using XP Ampure beads – 1.8x
15. Y-shaped linker ligation
16. Cleaning/purification using XP Ampure beads – 1.8x
17. Barcode incorporation with PCR using indexed primers
18. Cleaning/purification using XP Ampure beads – 1.8x
19. Quantify library using Illumina adapter primers with Sybr-green qPCR
20. Determine viral/human DNA in the library

EBV Hybrid Selection Enrichment:

21. Pool the libraries by balancing based on EBV DNA fragment contents.
22. Hybrid pull-down with RNA baits
23. After-enrichment PCR
24. Cleaning/purification using XP Ampure beads – 1.8x
25. Assess DNA quality
26. Sequence on the same flow cell.

Input DNA preparation

Step 1. Extraction/Isolation/Purification of DNA from clinical specimen (Biopsy, plasma, cell line)

Step 2. Quantification of Input DNA for Sequencing Library Preparation using PicoGreen

Step 3. Whole Genome Amplification with preamp

For the preamplification reaction, we used regular PCR primers tiling across the virus genome in two pools with 20 primer pairs. Most of the primer sequences and the reaction conditions were obtained from Kwok et al. (Hin Kwok et al. 2012b) except additional primers that are unique to type 2 genomes. Following the 20 cycle preamplification LD-PCR polymerase, reaction solutions were mixed with phi29 reaction buffer containing 1.5 U phi29 polymerase and EBV genome specific 3'-protected oligos (Leichty and Brisson 2014). Then, we incubated at 30C for 16h.

For whole genome amplification with GenomiPhi v2 kit, please follow the kit guidelines. Preamp-sWGA is a method for pre-amplification template DNA boostup with tiling PCR primers followed by whole genome amplification using Phi29 and EBV specific protected oligos.

This reaction condition has been tested for 100 EBV copies/20ng Template DNA. So, input DNA should have at least this level of viral copy as a starting material (Please adjust the H2O volume depending on template DNA vol).

Preamp PCR reactions with two pools;

	1x 25uL Reaction			1x 25uL Reaction
LongRange PCR Buffer with Mg2+, 10x	2.5		LongRange PCR Buffer with Mg2+, 10x	2.5
dNTP mix (10 mM of each)	1.25		dNTP mix (10 mM of each)	1.25
Primer Pool 1	2		Primer Pool 2	2
LongRange PCR Enzyme Mix	0.15		LongRange PCR Enzyme Mix	0.15
5x Q-Solution	5		5x Q-Solution	5
MgCl2 (25mM)	0		MgCl2 (25mM)	0
H2O	0		H2O	0
Total	11		Total	11
Template DNA	14		Template DNA	14

20 cycle amplification:

95C for 3min
REPEAT 20 times
Denature at 95C for 30sec
Gradient Annealing 58C to 49C for 15sec each
Extension at 72C 7min
Final Extension at 68 for 10min
4C hold

Following the preamp reactions, mix the two reaction solutions well and proceed to next steps. Denature the mixture at 95C 3min, keep on ice immediately.

Important: this denaturation allows sWGA oligos to anneal on ssDNA template at isothermal conditions. So, please do not skip denaturation step!

Post-PCR sWGA:

Oligo sequence (5' to 3') used for preparing “EBV oligo mix” for the sWGA reaction:

GCCGCOG
CCGCCEC
GGTCTOG
GCGGGOC
CGCCAOC
CCGCCFC
GTGGCOG
GGGCCET
CGGGGZC
GTCCGEG

Prepare the following mixture for every sample:

	For 50uL preamp mix:
phi29 Reaction Buffer 10X	**7**
dNTP (Mix-2)*	**3**
EBV oligo mix	**7**
phi29 Polymerase (10U/uL)	**2**
BSA (0.1ug/uL) 100x	**0.7**
H2O	**0.3**
total	**20**

*dNTP Mix-2 (GTP, CTP, ATP,TTP; with 30, 30, 5, 5 mM, respectively)

When 20uL reaction solution is ready, directly add to denatured solution preamp mix.

Incubate at 30C for 16h followed by 65C for 15min.

Step 4. Cleaning/purification after WGA using 1.8X XP Ampure beads elute in 70uL H2O

Quantify the Amplified DNA with PicoGreen

Check the quality of DNA with NanoDrop for OD 260/280 ratio

Determine viral/human DNA before and after WGA qPCR

Sequencing Library Preparation

Step 5. Shear DNA

Make sure genomic DNA samples are of high quality with an OD 260/280 ratio ranging
from 1.8 to 2.0.

Covaris recommends:
<u>DNA input:</u> from 100 ng to 5 µ g purified DNA
<u>Buffer:</u> Tris EDTA, pH 8.0
<u>DNA quality:</u> Genomic DNA (> 10 kb). For lower quality DNA, Covaris recommends setting up a time dose response experiment for determining appropriate treatment times.
<u>Sample volume:</u> 130ul (+/-5ul) for microTUBE AFA Fiber Snap-cap (Covaris p/n 520045)

For each DNA sample to be sequenced, prepare 1 library.

1. Dilute ***3 µg*** of high-quality gDNA with 1X Low TE Buffer in a 1.5-mL LoBind tube to a total volume of 130 µL.

2. Set up the Covaris E-series or S-series instrument.

a. Check that the water in the Covaris tank is filled with fresh deionized water to the appropriate fill line level according to the manufacturer's recommendations for the particular instrument model and sample tube or plate in use.

b. Check that the water covers the visible glass part of the tube.

c. On the device control panel, push the Degas button. Degas the instrument for least 2 hours before use, or according to the manufacturer's recommendations.

d. Set the chiller temperature to between 2°C to 5°C to ensure that the temperature reading in the water bath displays 5°C.

e. Optional. Supplement the circulated water chiller with ethylene glycol to 20% volume to prevent freezing.

Refer to the Covaris instrument user guide for more details.

3. Put a Covaris microTube into the loading and unloading station.
Keep the cap on the tube.

4. Use a tapered pipette tip to slowly transfer the 130-µL DNA sample through the pre-split septa.

Be careful not to introduce a bubble into the bottom of the tube.

5. Secure the microTube in the tube holder and shear the DNA with the settings in the table below, depending on the Covaris instrument SonoLab software version used.

The target DNA fragment size is 300 to 500bp (peak at 400bp).

Shear settings for Covaris instruments using SonoLab software prior to version 7:

Instrument E220:

Setting	Value
Duty Factor	10%
Peak Incident Power (W)	140
Cycles per Burst	200
Time	55
Temperature	4° to 7°C

6. Put the Covaris microTube back into the loading and unloading station.

7. While keeping the snap-cap on, insert a pipette tip through the pre-split septa, then slowly remove the sheared DNA.
8. Transfer each 130-µL sheared DNA sample to a 96 well plate.

!!! KEEP 65 ul of Sheared DNA as back up.

”Size selection” using XP Ampure beads. 0.3X (20uL) discard beads. Add 0.9X (60uL) more keep the beads. Elute DNA in 17uL H20.

Assess quality with the 2100 Bioanalyzer DNA 1000 kit (optional)

Step 6. Blunt End Repair

Quick Blunting Kit (NEB: E1201L)

Mix the following components with Sheared and Purified DNA (up to 5 µg, 17 µl) in a sterile well of 96 well plate:

Using Quick Blunting Kit (NEB: E1201L)	**µl Per reaction**
Sheared Clean DNA	**17ul**
10X Blunting Buffer	**2.5**
10mM dNTP Mix	**2.5**
Blunt Enzyme Mix	**1**
H2O	**2**
Total	**25**

Place the reaction into a thermal cycler and set the cycler for 20min at 12C followed by 15min at 37C. Immediately inactivate enzyme in the blunting reaction by heating at 70°C for 10 minutes.

Cycler:
20 min at 12°C (Do Not heat the lid)
15 min at 37°C (Do Not heat the lid)
10 min at 70°C

Step 7. Cleaning/purification using 1.2X XP Ampure beads. Elution volume is ~30uL.

Step 8. Adenylate the 3' end of the DNA fragments

Klenow Fragment (3'-->5' exo-) (NEB: M0212S)
dATP (NEB: N0440S)

Mix the following components in a sterile well of 96 well plate:

End repaired Clean DNA	30uL
second strand buffer/10× NEB Buffer 2	**5**
dATP (10 mM)	**2**
Klenow Fragment 3' to 5' exo– (5 U/µl) M0212L	**3**
H2O	**10**
Total	**50**

Incubate 60min at room temperature (25°C). Immediately inactivate enzyme in the adenylation reaction by heating at 75°C for 20 minutes.

Cycler:
60 min at 25°C (Do Not heat the lid)
20 min at 70°C

Step 9. Cleaning/purification using XP Ampure beads – 1.8x. Elution volume is ~20uL.

Step 10. Y-shaped Adapter Ligation

Add the Y-shaped adapters (oligonucleotide sequences © 2007–2011 Illumina, Inc. All rights reserved.). **To prepare the Y-shaped adapter, mix 25 µl adapter oligo 1 and oligo 2 (each at 50 µM stock concentration)**. ONLY ONCE STEP!

Heat at 95°C for 2 minutes, then ramp down slowly to room temperature. We usually heat the oligo mixture in an aluminum heat block for 2 minutes. Then remove the block from the heater and **let it cool down to room temperature, for approximately 30 minutes.**

Using Quick Ligation™ Kit (M2200L)

Quick Adapter Ligation:	**µl Per reaction**
A-tailed Clean DNA	**20uL**
2X Quick Ligation Buffer	**25**
Y shaped adapter (10 µM)	**2**
dATP (100 mM)	**1**
Quick Ligase	**2**
Total	**50**

Incubate at **room temperature for 30 minutes**.

Step 11. Cleaning/purification using 0.9X XP Ampure beads. Elution volume is ~30uL.

Step 12. Barcode incorporation with PCR using indexed primers

PCR amplify the Library:	µl Per reaction
Adapter ligated Clean DNA	30uL
2X KAPA HiFi HotStart ReadyMix	25
dNTP (10 mM each)	2
10 µM PCR Primer 1 (barcoded-optional)	1.5
H2O	0
Total	58.5
10 µM PCR Primer 2 (barcoded)	1.5

Incubate the tube at 98°C for 40 seconds, 65°C for 30 seconds and 72°C for 30 seconds. After the incubation, pause the PCR machine and then add 1.5 µl 10 µM PCR Primer 1. Continue the PCR with **10 cycles** of 98°C for 10 seconds, 65°C for 30 seconds, 72°C for 30 seconds, followed by incubation at 72°C for 3 minutes.

Step 13. Cleaning/purification using 0.9X XP Ampure beads. Elution volume is ~30uL.

Step 14. Quantify library qPCR using Illumina adapter primers.

KAPA Library Quantification Kit Illumina® platforms - KK4824. Follow the protocol.

Quantify Library:	µl Per reaction
2x Power SyberGreen PCR Mix	**12.5**
qPCR Primer Mix (2 uM) P7-3, P5-2	**2.5**
H2O	**0**
Total	**15**

Step 15. Determine viral/human DNA in the library with Sybr-green qPCR.
Pool the libraries by balancing based on EBV DNA fragment contents.

Step 16. Hybrid pull-down with RNA baits. >36h
Follow MyBait protocol !!!

Step 17. After-enrichment PCR

qPCR Primer 1 - HPLC Purified - 5’AATGATACGGCGACCACCGA -3’
qPCR Primer 2 - HPLC Purified - 5’CAAGCAGAAGACGGCATACGA -3’

Post-Capture amplify the Library:	**µl Per reaction**
2X KAPA HiFi HotStart ReadyMix	**25**
dNTP (10 mM each)	**2**
(10 µM each) PCR Primer Mix – P5-2/P7-3	**1.5**
H2O	**5.5**
Total	**34**

Step 18. Cleaning/purification using XP Ampure beads – 1.8x

Assess quality with the 2100 Bioanalyzer High Sensitivity DNA kit

Step 19. EBV qPCR→ Pool to one tube.

Step 20. Sequence on the same flow cell.

Copyright Information

Several experimental steps in the protocol provided here are copied partially or as a whole from

SureSelect XT2 Target Enrichment System for Illumina Multiplexed Sequencing Featuring Pre-Capture Indexing Reagents and Protocols Version C.1, December 2012, G9630-90000.

Zhang, Zhao, William E. Theurkauf, Zhiping Weng, and Phillip D. Zamore. 2012. “Strand-Specific Libraries for High Throughput RNA Sequencing (RNA-Seq) Prepared without poly (A) Selection.” *Silence* 3 (1): 9.