Decoding the B cell receptor in endemic Burkitt Lymphoma: Insights into pathogenesis and implications for disease detection

Katharine Lombardo

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Edus H. Warren, Chair

David Rawlings

Joseph Smith

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Decoding the B cell receptor in endemic Burkitt Lymphoma:
Insights into pathogenesis and implications for disease detection

Katharine Lombardo

Chair of the Supervisory Committee:
Adjunct Professor Edus H. Warren
Department of Pathology

The African endemic form of Burkitt Lymphoma (BL) is one of the most common pediatric malignancies in sub-Saharan Africa. BL is particularly prevalent in certain geographic regions, likely due to both environmental and genetic influences. BL is a malignancy of antigen-experienced, germinal center centroblasts. Normal B cells express a functional B cell receptor on their surface that is generated from ordered chromosomal rearrangements at the immunoglobulin (Ig) loci. We performed high-throughput sequencing on genomic DNA extracted from primary BL tumors. The sequencing was focused on the Ig loci from three independent BL patient cohorts to assess the complete repertoire of Ig rearrangements contained in BL tumors. This analysis demonstrated that 55 of 69 tumors harbored a clonal Ig heavy chain (*IGH*) repertoire. Amongst clonal tumors, a second rearranged *IGH* allele was only detected in 11 cases, suggesting

widespread monoallelic *IGH* rearrangements in BL tumors. Most tumors were characterized by extensive Ig sequence variation; hundreds of related, but unique, nucleotide sequences were detected in most tumors. These sequence families were associated with both complete VDJ rearrangements and incomplete DJ rearrangements, demonstrating the activity of active mutational processes at the Ig loci in BL tumors.

IGH sequences are highly specific and can serve as a molecular barcode for each B cell. Assessment of matched patient blood samples demonstrated that tumor-associated *IGH* sequences were regularly detected in circulation at diagnosis. Positive detection of circulating tumor DNA at diagnosis was associated with inferior patient survival than when tumor DNA was not detected, suggesting that tumor-specific *IGH* sequences may have prognostic value for BL patients. The data presented in this dissertation suggest an alternative model of BL pathogenesis and assess the utility of *IGH* sequences as a biomarker for BL detection. The goal of this work is to expand the BL research repertoire to provide insights that will contribute to disease prevention efforts and ultimately improve the outcome for BL patients.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my PI, Hootie, for his support and guidance over the course of my graduate studies. Thank you for providing a fun and enthusiastic environment for exploring and learning. Thank you for demonstrating how to be a great immunologist, scientist, and person. You impart a love for science in the lab that is so important. I look forward to seeing the wonderful things that the lab will accomplish next.

I would like to thank my thesis committee members, David Rawlings, Keith Jerome, Joe Smith and Michael Lagunoff, for their thoughtful critiques and insightful questions. Their support helped shape the path that my research took and helped me focus on important and relevant questions. In particular, thank you to Hootie, David, and Joe who served on my dissertation reading committee. I would also like to thank Chris Carlson and Erick Matsen for their productive discussions about our data.

To all members of the Warren Lab, past and present, thank you. I feel lucky to have worked amongst such great scientists and people over the last six years. Andrea Towlerton, thank you for everything; for your endless support and for keeping things running. David Coffey, thank you for sharing your R expertise and for your boundless ideas.

To Akiko Shimamura, thank you for training me as a research technician right out of college. You provided me with a solid foundation in research; I learned so much and will be forever grateful for the knowledge you shared with me.

And to my family. Thank you. To my parents, Marilyn and Jim – I don't have enough words to thank you sufficiently. Thank you for your unwavering support and endless encouragement. Thank you for telling me that girls could, in fact, be doctors when I doubted it as a child. To my brother Zach, thank you for always challenging and inspiring me.

To my husband, Jeff, thank you for your steadfast support, encouragement and laughs. You make every journey better. And to my son, Callen, thank you for showing me the world through your happy and curious eyes.

DEDICATION

This is for Callen.

Always remember to follow your dreams.

Chapter 1. INTRODUCTION

Sections of this chapter are adapted from a review in *Leukemia & Lymphoma* (Manuscript in progress).

Katharine A. Lombardo, David G. Coffey, and Edus H. Warren. Sporadic versus endemic: Comparing the pathogenesis of Burkitt Lymphoma subtypes. *Leuk Lymphoma*. (Manuscript in progress.)

1.1 Overview of Burkitt lymphoma

Burkitt lymphoma (BL) is an aggressive B cell neoplasm that is found worldwide. The endemic form of the disease (eBL) is one of the most common pediatric malignancies in sub-Saharan Africa. eBL occurs most commonly in children 2-13 years of age. Males have an increased risk of developing eBL, with an incidence as high as 4.7 cases per year, as compared to 3.0 cases per year for females (per 100,000 children <15 years of age).[1] eBL tumors grow extremely quickly, with a cell doubling time of only 24-48 hours.[2] eBL tumors frequently present as facial tumors, in the mandible or maxilla; or in the abdomen, particularly in the spleen, kidney, or ovary.

The development of eBL is largely attributed to co-infection with two distinct pathogens. eBL is highly associated with Epstein-Barr Virus (EBV; Human Herpes Virus 4) infection. EBV DNA is detected in >90% of endemic tumors. In fact, EBV was first discovered in an eBL tumor.[3] In addition to EBV, holoendemic *Plasmodium falciparum* malaria infection is associated with the development of eBL. eBL occurs primarily in equatorial Africa, and is particularly common in rural, lowland areas where the *Anopheles* mosquito, the malaria vector, is prevalent. Malaria eradication efforts have coincided with reduced rates of eBL.[4]

There are three epidemiological subtypes of BL. The endemic form is found most commonly in African children, as described above. The sporadic form of BL (sBL) occurs worldwide and the immunodeficiency-associated (id-BL) subtype occurs primarily in immunosuppressed or HIV-infected individuals. Both sBL and id-BL are also found in adults and are not highly associated with infection.

Survival disparities are reported for the endemic and sporadic subtypes. Long-term survival for BL patients in sub-Saharan Africa, where the endemic subtype predominates, is only 30-50%,[5] whereas long-term survival for BL patients in the United States and Europe is over 90%.[6,7] Additional co-factors, including disease stage, available treatment options, and patient genotype likely contribute to patient survival. However, this stark survival discrepancy demonstrates the potential for improving clinical outcomes across Africa. It is estimated that infection-related cancers account for 18% of the global cancer burden[8] and up to 33% of cancers in sub-Saharan Africa.[9] This subset of malignancies represents a significant, but uniquely actionable, global health threat.

1.2 TREATMENT LIMITATIONS

Children who present to tertiary African cancer centers with suspected BL undergo a tumor biopsy to confirm the diagnosis by histology. To assess disease dissemination, a chest x-ray and abdominal ultrasound are performed. A lumbar puncture and bone marrow biopsy, which are recommended staging procedures, are inconsistently performed due to the limited availability of needles. Even under optimal conditions, when all recommended tests are performed, these methods are only able to detect large deposits of disease. Thus, physicians are severely limited in their

ability to accurately stage BL and evaluate the efficacy of treatment. Without reliable methods to evaluate the extent of disease or the amount of residual disease, effective treatment is frequently impossible. More sensitive and precise methods are needed to adequately gauge BL disease dissemination; the development of a precise, molecular approach to BL detection would dramatically improve disease assessment and would likely improve patient outcome.

1.3 PATHOGENESIS

1.3.1 *Epstein-Barr Virus*

EBV is a gamma herpes virus with a tropism for B lymphocytes. More than 90% of adults worldwide are EBV-seropositive and are largely asymptomatic. EBV maintains a lifelong viral reservoir in a small subset of an individual's B cells; only 1-50 of every 1,000,000 circulating B cells contain EBV DNA.[10] However, EBV DNA is detected in 80-100% of eBL tumors, demonstrating a strong viral association with eBL (Table 1.1).[11-15] *In vitro*, EBV transforms B cells and is commonly used to generate rapidly proliferating lymphoblastoid cell lines in the laboratory. In addition to its role in BL pathogenesis, EBV is closely associated with nasopharyngeal carcinoma, Hodgkin's lymphoma, and several types of lymphoproliferative disorders.

Table 1.1. EBV status of eBL tumors in published studies

Study	EBV-positive	EBV-negative	Total tumors	EBV-positive tumors (%)
Bellan *et al.* 2005	12	3	15	80
Navari *et al.* 2015	7	1	8	87.5
Amato *et al.* 2016	26	0	26	100
Kaymaz *et al.* 2017	26	2	28	92.9
Lombardo *et al.* 2017	47	4	51	92.2

As a whole, African children acquire primary EBV infection at a very early age. In Eastern Uganda, virtually 100% of the population is EBV-positive by age three.[16] In regions of Kenya with holoendemic malaria infection, nearly 70% of children are infected with EBV by only 12 months of age.[17] In contrast, only 50% of children in the United States are infected with EBV by six to eight years of age.[18] This extremely early EBV acquisition may contribute to the role of EBV in the pathogenesis of eBL.

The majority of BL tumors demonstrate an EBV latency type I expression pattern, in which most of EBVs 85 genes are silenced. Under this latency program, BL tumors express two non-coding RNAs, EBER1 and EBER2, and the EBNA1 protein. EBNA1 is essential for the transformation efficiency of EBV[19] and indirectly destabilizes p53, leading to the inhibition of apoptosis in host cells.[20] Furthermore, EBNA1 has been demonstrated to have oncogenic potential in mice.[21,22] These data implicate EBNA1 in BL tumor cell survival.

There is increasing evidence that a significant proportion of BL tumors express more than the canonical three genes as defined by latency type I. This expanded EBV gene expression overlaps with other latency programs, and even includes evidence of lytic EBV reactivation.[14,23] Additional EBV genes that are reportedly expressed in BL tumors include LMP-1, -2A, EBNA-2, -3A, -3B, -3C, -LP and some lytic genes.[23,24] Additionally, another form of EBV latency has been described that occurs in approximately 15% of eBL cases, termed Wp-restricted latency, where EBNA2 is deleted and the EBNA3 and BHRF1 genes are expressed.[25,26] These studies demonstrate that a spectrum of EBV genes are expressed across BL tumors, and that BL-associated EBV gene expression profiles can be quite heterogeneous. A broad array of EBV gene products have been demonstrated to elicit an immune response,[27] which would target EBV-infected cells for destruction by the immune system. The continued expression of these gene products suggests that they may be essential for BL pathogenesis. For example, the latent EBV gene LMP2A is a B cell receptor (BCR) homologue that contains an immunoreceptor tyrosine activation motif (ITAM) on its N-terminus, which binds membrane-proximal BCR signaling molecules.[28] Expression of LMP2A has been linked to activation of BCR signaling through the PI3K pathway, likely contributing to tumor cell survival.[29-32] LMP2A-mediated signaling in B cells has also been shown to impair the humoral response by favoring the selection and differentiation of B cells with low-affinity BCRs.[33] This phenomenon has been demonstrated to lead to increased titers of autoreactive antibodies, which likely contribute to various pathologies.[33] It is likely that the broad array of EBV genes expressed in some BL tumors contribute to pathogenesis in additional ways that have yet to be elucidated.

From its latent B cell reservoir, EBV will periodically reactivate to produce infectious virions. The EBV-specific T cell response is essential for suppressing EBV replication. It has been

demonstrated that BL patients have a diminished EBNA1-specific T cell response and high levels of EBV viremia, suggesting that immunological control of EBV may be impaired in BL patients and further implicating EBV in the pathogenesis of BL.[17,34]

1.3.2 *Plasmodium falciparum*

eBL is also associated with holoendemic *P. falciparum* malaria infection. The CIDR1α domain of the malarial *P. falciparum* membrane protein 1 (PfEMP1), which is expressed on the surface of an infected red blood cell, is able to bind the BCR non-specifically, inducing B cell activation and hyperplasia in an infected individual.[35] Furthermore, malaria infection broadly suppresses the T cell response, possibly by interfering with dendritic cell maturation and antigen presentation, impeding the ability of dendritic cells to properly stimulate T cells.[36] This creates an environment of immunosuppression[37] and B cell proliferation that supports EBV reactivation. Children with acute malaria infection have a dampened EBV-specific immune response, higher EBV viral loads, and higher levels of circulating EBV-positive B cells.[38-41] This suggests that *P. falciparum* and EBV may cooperate to create a microenvironmental niche that supports the proliferation of EBV-positive B cells.

Hemozoin, a metabolic byproduct of hemoglobin by *P. falciparum*, has been demonstrated to induce Toll-like receptor 9 (TLR9) signaling, which can increase activation-induced cytidine deaminase (AID) expression.[42,43] It is hypothesized that deregulated AID expression could induce an increased frequency of mutations and even double-strand DNA breaks, contributing to BL pathogenesis.

1.3.3 *Chromosomal translocation*

All BL subtypes share a pathognomonic chromosomal translocation that contributes to pathogenesis. This translocation juxtaposes the *c-MYC* proto-oncogene at chromosome 8q24 with one of the immunoglobulin (Ig) loci. A t(8;14) translocation with the Ig heavy chain (*IGH*) locus on chromosome 14q32 occurs in 80% of BL tumors. A t(2;8) translocation with the Ig kappa light chain (*IGK)* locus on chromosome 2p12 or a t(8;22) translocation with the Ig lambda light chain (*IGL*) locus on chromosome 22q11 occurs in the remaining 20% of cases. These translocations put *c-MYC* under the control of the Ig regulatory elements, generating high levels of *c-MYC* expression in BL tumor cells. *c-MYC* encodes a transcription factor that binds to enhancer (E-box) domains to induce expression of a large number of genes involved in the cell cycle, differentiation, and apoptosis.

By southern blot, fluorescence in-situ hybridization, and long-distance PCR analysis, it has been demonstrated that the location of the double-strand DNA breakpoints, resulting in the reciprocal t(*c-MYC*;Ig) translocation, differ based on BL subtype. In sBL tumors, the *c-MYC* breakpoint regularly occurs adjacent to, or within, the first exon of *c-MYC*. In eBL tumors, the breakpoint generally occurs far upstream of the *c-MYC* locus and has been detected more than 300 kilobases away.[44-48] At the *IGH* locus, DNA breakpoints are reported to occur at roughly equal frequencies at the joining and isotype-specific constant and switch regions in both eBL and sBL.[46,47,49] It appears that *c-MYC* is deregulated by distinct mechanisms in each subtype of BL, perhaps indicating that the t(*c-MYC*;Ig) translocation occurs in unique cells of origin.

Despite its prevalence in BL tumors, the translocation alone is not sufficient for oncogenesis.[50] The accumulation of additional mutations, both within *c-MYC* and across the genome, is required for malignant transformation. In endemic tumors, point mutations cluster in

the 5' end of *c-MYC*.[44,46] In contrast, the 5' portion of *c-MYC* is frequently truncated in sporadic tumors due to the intragenic translocation breakpoint. The N-terminus of *c-MYC* contains negative regulatory elements and phosphorylation sites that induce proteasome-mediated mRNA degradation.[51] The disruption of these sites in BL leads to increased *c-MYC* transcription and mRNA stability, facilitating high *c-MYC* expression and promoting tumor cell proliferation.[52,53] This demonstrates that additional deregulation of *c-MYC* is required for malignant transformation.

1.4 THE B CELL RECEPTOR

1.4.1 *VDJ recombination*

BL is a malignancy of B cells. Healthy B cells are a critical component of the adaptive immune system. They respond to foreign antigens and initiate an immune response, which ultimately stimulates antibody production to combat infection. All normal B cells express a unique, functional BCR on their cell surface. The BCR is essentially a membrane-bound antibody; it is a heterotetrameric molecule that is comprised of two heavy chains and two light chains joined by disulfide bonds (Figure 1.1A).

Figure 1.1. The crystal structure of an immunoglobulin molecule.

A murine IgG1 monoclonal antibody is illustrated (PDB ID: 1IGY).[54] (A) The Ig heavy chains (*IGH)* are blue and Ig kappa light chains (*IGK)* are red. (B) The framework regions of the Ig molecule are green. (C) The individual complementarity-determining regions (CDRs) are indicated by color: the *IGH* CDR1 is red, *IGH* CDR2 is orange, and *IGH* CDR3 is yellow; the *IGK* CDR1 is green, *IGK* CDR2 is blue, and *IGK* CDR3 is purple. (Images were generated with the UCSF Chimera package.[55] Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311).)

The BCR is generated early in B cell development, while the cell resides in the bone marrow. BCRs are not encoded in germline DNA, but are created through a complex process of chromosomal rearrangements in each B cell. As a late pro-B cell, the cell will independently, but concurrently, rearrange the D and J gene segments on both immunoglobulin heavy chain (*IGH*) alleles (Figure 1.2B).[56] The cell will then rearrange the V to DJ gene segments on one allele (Figure 1.2C). If this is a productive rearrangement, meaning it is in frame and does not contain a stop codon, the cell will stop *IGH* rearrangement and the mature B cell will harbor a productive VDJ rearrangement on one allele and an incomplete DJ rearrangement on the other allele. If this initial VDJ rearrangement is unproductive, the cell will then rearrange the V to DJ gene segments on the second *IGH* allele (Figure 1.2D). If this rearrangement is productive, the mature B cell will contain one productively rearranged VDJ allele and a second, unproductively rearranged VDJ allele. Once a productive heavy chain has been generated, the cell undergoes a parallel process on the V and J gene segments, first at the immunoglobulin kappa (*IGK*) and then lambda (*IGL*) light chain loci. Successful rearrangement of the heavy and light chain loci permits expression of a functional BCR on the surface of the B cell.

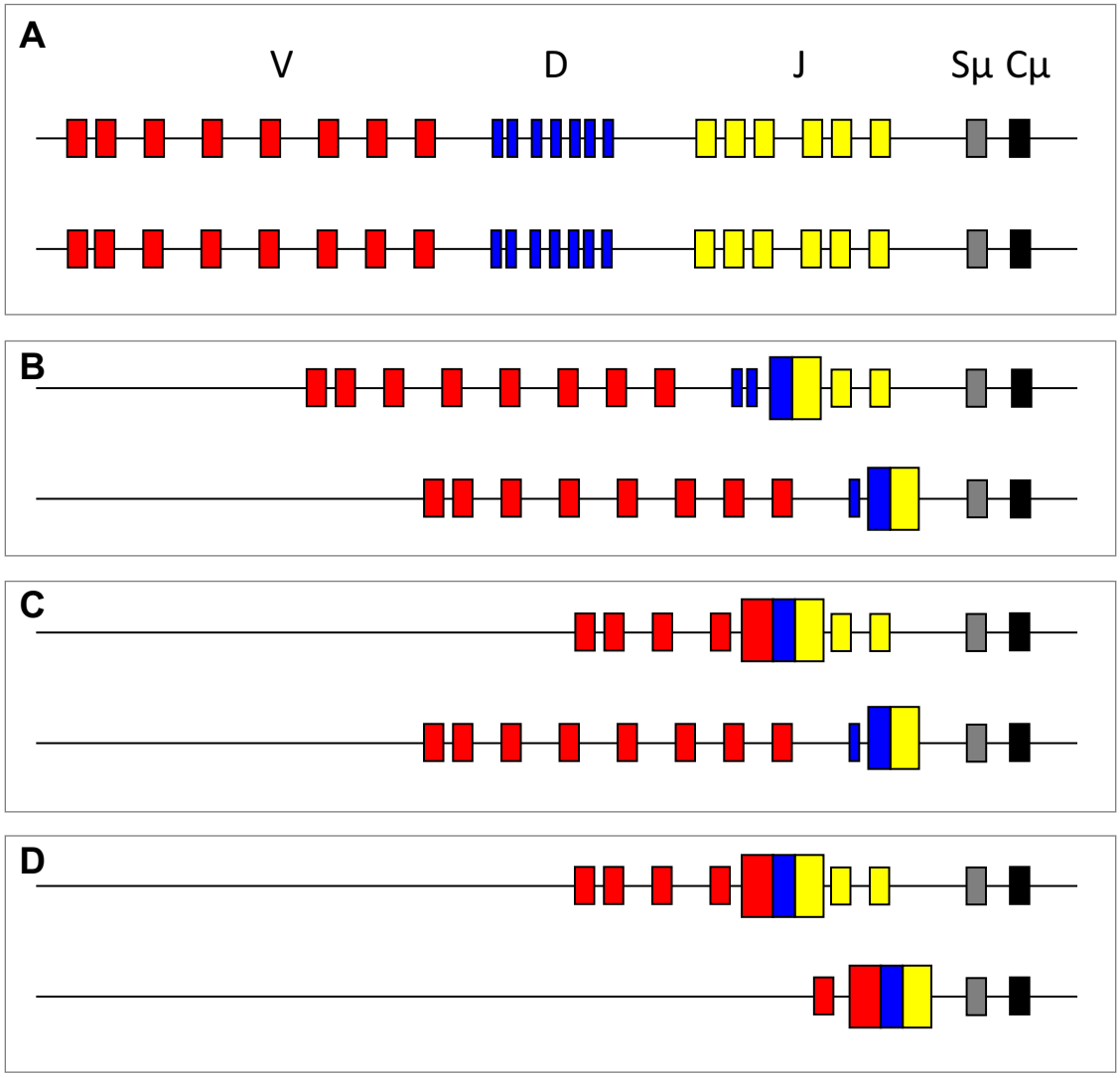


Figure 1.2. Progression of normal VDJ recombination on both *IGH* alleles.

IGHV gene segments are shown in red, *IGHD* gene segments in blue, and *IGHJ* gene segments in yellow. The first switch and constant regions are shown in grey and black, respectively. (A) The germline configuration of both *IGH* alleles. (B) At the pro-B cell stage, the D and J gene segments are rearranged on both *IGH* alleles (gene rearrangements are enlarged). (C) One *IGH* allele will proceed with VDJ rearrangement. If this is a productive rearrangement, the B cell will continue to mature and differentiate, leaving one *IGH* allele with only a DJ rearrangement. (D) If this initial VDJ rearrangement is unproductive, the cell will rearrange the VDJ on the other *IGH* allele. If productive, this cell will harbor two VDJ rearrangements, one productive and one unproductive.

These chromosomal rearrangements are mediated by the enzymes recombination-activating genes-1 and -2 (*RAG1* and *RAG2*). Unique BCRs are generated by the rearrangement of many distinct Ig gene segments. At the *IGH* locus, there are 40 functional V gene segments, 25 D gene segments, and 6 J gene segments that can be independently combined. The *IGH* rearrangement is independently paired with the *IGK*/*IGL* rearrangement. There are 40 V and 5 J gene segments at the *IGK* locus and 30 V and 4 J gene segments at the *IGL* locus.[57] In addition to the chromosomal recombination events, exonuclease activity and untemplated nucleotide insertions occur at each VD, DJ, and VJ junction, which are mediated by the enzyme terminal deoxytransferase. This generates additional sequence diversity within the BCR.

The BCR itself is comprised of framework regions (FRs), which provide structural support, and complementarity determining regions (CDRs), which are hypervariable, hypermutable, and create a pocket that interacts directly with antigen (Figures 1.1B,C).

1.4.2 *Somatic hypermutation*

Further Ig sequence diversity is generated in the germinal center (GC). The immature B cell, expressing its successfully rearranged heavy and light chains on the cell surface as a BCR, leaves the bone marrow and enters peripheral circulation. The B cell migrates to a secondary lymphoid organ, and upon binding antigen, the B cell proliferates and initiates GC formation. The GC is composed of a dark zone, where B cell proliferation occurs, and a light zone, where antigen presentation by follicular dendritic cells and co-stimulation by follicular T helper occurs. Also in the dark zone, activation-induced cytidine deaminase (AID) induces somatic hypermutation (SHM), a process by which random point mutations are introduced in Ig genes with the ultimate goal of increasing BCR affinity for its antigen. B cells expressing a BCR with high affinity for their cognate antigens are positively selected for within the GC and undergo class switch

recombination (CSR) to diversify their effector functions. A B cell can repeatedly traverse the GC to increase the binding affinity of its BCR to antigen. Upon completion of the GC reaction, the B cell will mature into a memory B cell, to rapidly respond if repeat antigen encounter occurs, or a plasma cell, to generate large amounts of antibody.

The paired processes of VDJ recombination and SHM generate a high degree of sequence diversity within the BCR and ensure that virtually every independently rearranged BCR is unique. Thus, the Ig loci are characterized by distinct nucleotide sequences that constitute a unique molecular barcode for every B cell.

1.4.3 *BCR signaling*

Low level, tonic signaling through the BCR is required for the survival of all normal B cells.[58] This signaling occurs through the PI3K/AKT signaling pathway,[59] which is involved in cell survival and cell cycle regulation. Binding of antigen to the BCR initiates signal transduction through several main pathways, including the NFκB, PI3K, MAPK, and NFAT pathways. Activation of these pathways serves to induce B cell proliferation and differentiation to initiate an adaptive immune response.

1.4.4 *The BCR in B cell malignancies*

Recurrent mutations in components of the BCR signaling cascade have been identified in a number of mature B cell malignancies. The mutations often serve to amplify the signal from the BCR and likely promote tumor cell proliferation. These gene variants have been reported in diffuse large B cell lymphoma,[60] follicular lymphoma,[61] and BL[13,15,23,62,63] tumors, suggesting that active BCR signaling contributes to oncogenesis in these malignancies.

The preferential utilization of particular *IGHV* genes on the productively rearranged *IGH* allele has been demonstrated in a number of mature B cell malignancies. This biased gene usage, termed BCR stereotypy, suggests a role for antigen in the pathogenesis of these cancers. It is believed that chronic antigenic stimulation drives recurrent GC reactions within particular subsets of B cells, increasing the chances of aberrant AID activity and malignant transformation. Preferential *IGHV* gene utilization has been described in mantle cell lymphoma,[64] chronic lymphocytic leukemia,[65-71] splenic and ocular adnexal marginal zone lymphomas,[72,73] and primary central nervous system lymphoma.[74]

The evidence for BCR affinity maturation, amplified BCR signaling, and BCR stereotypy in mature B cell cancers has led to the introduction of BCR signaling inhibitors as a treatment for several malignancies, including mantle-cell lymphoma and chronic lymphocytic leukemia. In particular, inhibitors of the PI3K and NFκB pathways have proven particularly effective.[75-77] The clinical efficacy of these inhibitors further demonstrates the fundamental role that BCR signaling can play in disease pathogenesis.

Studies indicate that activated BCR signaling may play a role in the development of BL as well. In many BL cell lines, attenuated BCR signaling through the PI3K pathway, either through siRNA knockdown or small molecule inhibition, is lethal.[78] Additionally, many BL cell lines demonstrate constitutive activation of the PI3K pathway.[78-80] In a mouse model, paired *c-MYC* overexpression and constitutive PI3K expression induce a BL-like malignancy.[81] The tumor-associated BCRs of some BL cases have been shown to bind to specific antigens, suggesting a possible role for antigen in BL.[82,83]

1.5 MOLECULAR CHARACTERISTICS OF BL

BL is a malignancy of antigen-experienced B cells. The Ig genes in BL tumor cells demonstrate evidence of SHM and affinity maturation that is characteristic of B cells that have traversed the GC. The frequency of non-synonymous nucleotide changes is higher, per nucleotide residue, in CDRs than in FRs, suggesting that positive selection for nucleotide changes within the antigen-binding region has occurred in tumor cells.[11,13] Additionally, BL tumor cells have markers on their cell surface that are characteristic of GC B cells, including CD19, CD20, and CD10. Gene expression studies suggest that BL tumors have a gene expression profile that resembles that of a GC B cell.[80] These data demonstrate that BL tumor cells have a GC phenotype.

Despite the extensive SHM observed in BL tumors, most tumor cells have not undergone class-switch recombination, but instead retain the IgM isotype. The signal that is generated from an IgM BCR is qualitatively different from that generated from an IgG BCR. The IgM signal promotes increased cell signaling and proliferation, whereas the signal from the IgG BCR promotes increased cell differentiation for subsequent antibody production.[84,85] *In vitro*, the IgM isotype demonstrates a growth advantage over the IgG1 isotype (though IgA had the greatest growth advantage).[85] These studies suggest an increased reliance on BCR signaling in BL.

Whole exome and RNA sequencing have identified particular genes that are recurrently mutated in BL tumors. The most commonly mutated gene is *c-MYC*, possibly due to its proximity to the Ig loci after translocation. Other recurrently mutated genes include *TP53*, *DDX3X*, *ID3*, *TCF3, SMARCA4* and *GNA13*.[13,15,23,62,63,78,86] Mutations in the transcription factor *ID3* and its negative regulator *TCF3* are common in both sBL (70%) and eBL (30-40%).[13,23,62,78] Wildtype *TCF3* increases BCR signaling in two ways: it increases expression of the Ig heavy and light chain genes and suppresses transcription of *PTEN*, a phosphatase that inhibits signaling through the

tonic, antigen-independent PI3K signaling pathway. *TCF3* mutations are commonly gain of function and result in increased protein expression.[78] *ID3* mutations are frequently biallelic, are found at AID hotspots (RGYW motifs), and are frequently deleterious, resulting in loss of protein expression.[62,78] The mutations in both *TCF3* and *ID3* serve to increase BCR signaling through the PI3K pathway in a large proportion of tumors.

Though the broad mutational landscape appears to be similar across BL subtypes, distinct mutation frequencies reportedly differ based on subtype. The genome-wide burden of somatic mutation is reportedly lower in endemic than sporadic BL.[14,23] This pattern suggests that other pathogenic forces, such as the transforming ability of EBV or recurrent antigenic stimulation, may play a role in the pathogenesis of the endemic subtype. In fact, eBL tumors infected with multiple herpes viruses, including Cytomegalovirus (HHV5) and Kaposi's sarcoma-associated herpesvirus (HHV8), tend to have even fewer somatic mutations.[23] Moreover, *TCF3*/*ID3* mutations and EBV infection are very rarely found in the same BL tumors, suggesting that both pathways may play analogous roles in malignant transformation.[23]

1.6 Previous research methods

Most previous studies of the Ig genes in BL tumor cells were performed using conventional, low-throughput Sanger sequencing. These experiments generated a small number of Ig sequence reads per tumor, limiting the ability to assess the full complement of Ig rearrangements within the tumor.[11,87,88] Additionally, many of the studies used a template of tumor RNA, preventing the detection of unproductive or incomplete Ig rearrangements.[89-91]

The advent of high-throughput sequencing (HTS) has profoundly enhanced our understanding of T and B cell malignancies by making the unique, clonal antigen receptor rearrangements carried by tumor cells readily identifiable and detectable with high sensitivity and specificity.[92,93] Detection of tumor-specific antigen receptor sequences by HTS is now being used to monitor treatment efficacy and disease relapse in many lymphoid malignancies,[94-96] demonstrating that antigen receptor sequences can serve as highly specific biomarkers. The studies presented in this dissertation utilize HTS on genomic DNA extracted from primary BL tumors and matched tissue samples to gain unprecedented resolution of the Ig repertoire, leading to novel insights into BL pathogenesis and disease dissemination.

Chapter 2. HIGH-THROUGHPUT SEQUENCING OF THE B CELL RECEPTOR IN AFRICAN BURKITT LYMPHOMA REVEALS CLUES TO PATHOGENESIS

This research was originally published in *Blood Advances*.

Katharine A. Lombardo, David G. Coffey, Alicia J. Morales, Christopher S. Carlson, Andrea M. H. Towlerton, Sarah E. Gerdts, Francis K Nkrumah, Janet Neequaye, Robert J. Biggar, Jackson Orem, Corey Casper, Sam M. Mbulaiteye, Kishor G. Bhatia, and Edus H. Warren. High-throughput sequencing of the B cell receptor in African Burkitt lymphoma reveals clues to pathogenesis. *Blood Adv*. 2017; 1(9):535-544.

2.1 ABSTRACT

Burkitt lymphoma (BL), the most common pediatric cancer in sub-Saharan Africa, is a malignancy of antigen-experienced B lymphocytes. High-throughput sequencing (HTS) of the immunoglobulin heavy (*IGH*) and light chain (*IGK*/*IGL*) loci was performed on genomic DNA from 51 primary BL tumors: 19 from Uganda and 32 from Ghana. RT-PCR analysis and sequencing of tumor RNA (RNAseq) was performed on the Ugandan tumors to confirm and extend the findings from HTS of tumor DNA. Clonal *IGH* and *IGK*/*IGL* rearrangements were identified in 41 and 46 tumors, respectively. Evidence for rearrangement of the second *IGH* allele was observed in only 6 of 41 tumor samples with a clonal *IGH* rearrangement, suggesting that the normal process of biallelic *IGHD* to *IGHJ* (DJ) rearrangement is often disrupted in BL progenitor cells. Most tumors, including those with a sole dominant non-expressed DJ rearrangement,

contained many *IGH* and *IGK*/*IGL* sequences that differed from the dominant rearrangement by <10 nucleotides, suggesting that the target of ongoing mutagenesis of these loci in BL tumor cells is not limited to expressed alleles. *IGHV* usage in both BL tumor cohorts revealed enrichment for *IGHV* genes that are infrequently utilized in memory B cells from healthy subjects. Analysis of publicly available DNA sequencing and RNAseq data revealed that these same *IGHV* genes were over-represented in dominant tumor-associated *IGH* rearrangements in several independent BL tumor cohorts. These data suggest that BL derives from an abnormal B cell progenitor and that aberrant mutational processes are active on the immunoglobulin loci in BL cells.

2.2 INTRODUCTION

Burkitt lymphoma (BL) is the most common pediatric malignancy in sub-Saharan Africa, with an incidence as high as 4.7 cases per year for males and 3.0 cases per year for females (per 100,000 children under the age of 15 years).[97] Long-term survival for BL patients in sub-Saharan Africa is only 30-50%.[5] African BL is associated with Epstein-Barr Virus (EBV)[98] and holoendemic *Plasmodium falciparum* infection.[99] BL tumors carry a *c-MYC*;immunoglobulin (Ig) chromosomal translocation, which is necessary, but not sufficient,[50] for malignant transformation.

BL tumor cells derive from germinal center centroblasts and express B cell receptors (BCRs) of the IgM and IgD isotype that demonstrate a high level of inferred somatic hypermutation (SHM).[11,87-91] Improved understanding of the molecular etiology of African BL, in particular the role of the BCR, could lead to better treatment and inform strategies for prevention. Most prior studies of Ig rearrangements in BL tumors used capillary sequencing of tumor RNA, which generated a limited number of sequence reads, and detected only productive, expressed Ig

rearrangements.[89-91] In this study, we utilized high-throughput sequencing (HTS) of tumor genomic DNA (gDNA) to analyze the complete ensemble of Ig gene rearrangements in primary BL tumor samples from two distinct patient cohorts. Our results comprehensively define the tumor-associated Ig rearrangements in these African BL patient cohorts with unprecedented resolution, and provide novel insights that have been missed by RNA sequencing, but have important implications for models of BL pathogenesis.

2.3 METHODS

Study populations: Tumor biopsies were acquired from two cohorts of African children with BL after obtaining verbal or written informed consent using IRB-approved protocols. One cohort comprised 19 children who presented to the Uganda Cancer Institute (UCI) in Kampala, Uganda between August 2013 and March 2014. All tumors underwent primary pathologic review in Uganda and secondary pathologic review and immunohistochemistry at a central U.S. site. Biopsies were obtained from facial tumors involving the mandible and/or maxilla. Histologically-confirmed, cryopreserved tumor biopsies from each patient were shipped from the UCI to the Fred Hutchinson Cancer Research Center (FHCRC) in Seattle, WA in the vapor phase of LN_2. A second cohort of 32 cryopreserved, archival BL tumor samples, collected between 1975 and 1992 at Korle Bu Hospital in Accra, Ghana, and stored long-term at the Frederick National Cancer Laboratory in Frederick, MD were shipped to FHCRC on dry ice. Samples from the Ghanaian cohort were primarily obtained from abdominal masses, most frequently from ovary, kidney, or spleen and were diagnosed based on clinical and cytological criteria; no histology was performed on these samples.

HTS of immunoglobulin heavy (*IGH)* and light (*IGK/)* chains: HTS of the *IGH* and *IGK*/*IGL* loci was performed on 1-3μg of gDNA from each tumor using the ImmunoSEQ platform at Adaptive Biotechnologies (Seattle, WA).[94,100] In brief, libraries of rearranged Ig loci were generated from tumor gDNA by multiplex PCR using sense primers specific for all V (and D) gene segments and antisense primers specific for all J gene segments for the *IGH* and *IGK*/*IGL* loci. The inclusion of primers specific for *IGHD* enabled the capture of incomplete *IGHD* to *IGHJ* (DJ) rearrangements. The PCR products were ligated to adapters, and 130-nucleotide (nt) sequence reads encompassing the 3' portion of framework region (FR) 3 and the entirety of complementarity-determining region (CDR) 3 of the *IGH* and *IGK*/*IGL* loci were generated on the Illumina MiSeq platform. Amplification bias from multiplexed primer sets was removed by direct sequencing of synthetic DNA templates.[101] Each molecule of amplified DNA was sequenced at least 10 times. Loading density on the MiSeq flow cell affected the total number of sequencing reads per sample. Adaptive Biotechnologies performed the initial analysis of raw sequence reads, including filtering and decomposition of reads into their component *V*, [*D*], and *J* segments,[102,103] and non-templated junctional nucleotide insertions. Tumors were defined as clonal if a unique Ig rearrangement that comprised ≥15% of the repertoire was detected. Sequencing data files are available on the Adaptive Biotechnologies website (https://clients.adaptivebiotech.com/pub/lombardo-2017-bloodadvances). Subsequent sequence analyses were performed using the LymphoSeq R package (http://www.bioconductor.org/packages/LymphoSeq) created by D.G.C.[104] Analysis of SHM was performed using IMGT/V-Quest[102,103] and IgBlast.[105]

HTS of T-cell receptor β (*TRB)*: HTS of the *TRB* locus was performed on tumor gDNA using the hsTCRB kit (Adaptive Biotechnologies). Sequencing and analysis of *TRB* sequencing data were performed in the same manner as for the *IGH* and *IGK/IGL* loci.

RNA sequencing (RNAseq): RNAseq was performed on the 18 Ugandan tumors with an RNA integrity number (RIN) >7 (no Ghanaian tumors were included due to low-quality RNA). RNAseq data is available in the Sequence Read Archive (accession number SRP099346). polyA-selected sequencing libraries were prepared with the Illumina TruSeq v2.0 kit. Paired-end, 50-nt sequencing was performed on the Illumina HiSeq platform at a depth of 100 million reads per sample. Variants were called by the Broad's Genome Analysis Tool Kit.[106] Normal tissue samples were not available to filter germline variants. Non-synonymous mutations were called if they were: 1) predicted to be deleterious by metaSVM,[107] 2) not present in a database of known single nucleotide polymorphisms defined by the 1000 Genomes Project and a panel of non-malignant samples from Sanger CGP sequencing, 3) observed in more than one tumor, and 4) previously reported in the COSMIC v77[108] database.

Statistical analyses: The Fisher Exact test was used to compare the frequency of *IGHV* utilization, with a Bonferroni correction to adjust for multiple comparisons. A Student's t-test was used to evaluate the frequency of SHM in *IGHV* gene regions and based on EBV status. The Wald test was used to assess differential gene expression, with a Benjamini & Hochberg adjustment for multiple comparisons. Clonal relatedness was defined as the fraction of unique sequences with an edit distance <10 from the most common sequence, where 1 indicates all sequences are related to the most frequent sequence and 0 indicates none of the sequences are related. Clonal relatedness considers both unproductive rearrangements and sequence relatedness, so was utilized preferentially over the clonality metric.

See Appendix A for additional methods.

2.4 RESULTS

2.4.1 *Study populations*

From August 2013 to March 2014, diagnostic tumor samples were obtained from 19 children with histologically-confirmed BL at the UCI in Kampala, Uganda. One of the patients was HIV-seropositive, and all four Ziegler disease stages were represented (Table 2.1). Only five patients (26%) had a sustained complete response to standard therapy (six cycles of cyclophosphamide, vincristine, and methotrexate), and were alive at one-year post study enrollment. One-year overall survival for the Ugandan cohort was 42% (median=253 days; Figure B1, Appendix B).

Table 2.1 Clinical information for the Ugandan and Ghanaian BL patient cohorts

Characteristic	Ugandan cohort (n = 19)	Ghanaian cohort (n = 29*)
Sex, no. (%)		
Female	7 (37)	13 (45)
Male	12 (63)	16 (55)
Age at enrollment, y		
Median	7	9
Range	4-12	4-13
HIV status, no. (%)		
Negative	18 (95)	29 (100)
Positive	1 (5)	0 (0)
Ziegler disease stage, no. (%)†		
A	7 (37)	2 (10)
B	3 (16)	0 (0)
C	3 (16)	15 (75)
D	6 (31)	3 (15)

*Clinical data not available for all 32 patients.
†Disease stage data only available for 20 of the Ghanaian patients.

Archival, cytologically-confirmed BL tumor samples from 32 children, collected between 1975 and 1992 at Korle Bu Hospital in Accra, Ghana, were received from the NCI sample repository. Clinical information was only available for a subset of patients. Seventy-five percent of the children in the Ghanaian cohort had Ziegler stage C disease (n=20; Table 2.1), and their one-year survival was 51% (median=458 days; n=29; Figure B1, Appendix B). The Ugandan and Ghanaian cohorts comprise a total of 51 BL patients, which form the focus of this study.

2.4.2 *Assessment of EBV DNA and RNA*

The EBV status of each BL tumor was determined by PCR-detection of EBV DNA sequences in tumor gDNA. EBV-encoded *EBER1, EBER2,* and *EBNA1* were detected in 47/51 (92%) of the tumors (Figure B2, Appendix B).

RT-PCR detection of *EBER* and *EBNA1* transcripts was used to evaluate the expression of EBV genes in a subset of tumors with quality RNA (RIN >5). *EBER1*, *EBER2*, and *EBNA1* RNA was detected in all tumors in which EBV DNA was detected (Figure B2, Appendix B). These results were confirmed by RNAseq analysis of EBV transcripts in the Ugandan cohort (n=18). All three of the Ugandan BL patients with EBV-negative tumors had expired one year after study enrollment.

2.4.3 *HTS of Ig gene rearrangements in BL tumors*

HTS was performed on gDNA from each of the 51 tumors to investigate the repertoire of *IGH* and *IGK*/*IGL* gene rearrangements. The sequencing strategy captured both incomplete, non-productive DJ rearrangements, as well as complete V to D to J (VDJ) rearrangements of the *IGH* locus, and V to J (VJ) rearrangements of the *IGK*/*IGL* loci, using 130-nt reads. (Figure B3, Appendix B; Table C1, Appendix C).

HTS of the *IGH* and *IGK*/*IGL* loci revealed a clonal rearrangement, in which the dominant rearrangement comprised ≥15% of the repertoire, in 41 and 46 of the 51 tumor samples, respectively (Figures 2.1A,B and 2.2A,B). An *IGH* VDJ rearrangement was detected in 31 of the 41 clonal tumors, and a sole DJ rearrangement was detected in the remaining 10 tumors. A clonal pattern of IGK/IGL VJ rearrangements was detected in all 41 of these cases. Five tumors carried a clonal *IGK/IGL* rearrangement but a polyclonal *IGH* repertoire, in which neither a clonal VDJ or DJ rearrangement was observed. In five tumors, all from the Ghanaian cohort, both the heavy and light chain repertoires were polyclonal. In contrast to the overwhelmingly clonal nature of the *IGH* and *IGK*/*IGL* repertoires in BL tumors, the repertoires of B cells from bone marrow and peripheral blood of healthy donors were highly polyclonal, with no sequences comprising ≥2.25% of the reads (Figure B4, Appendix B).

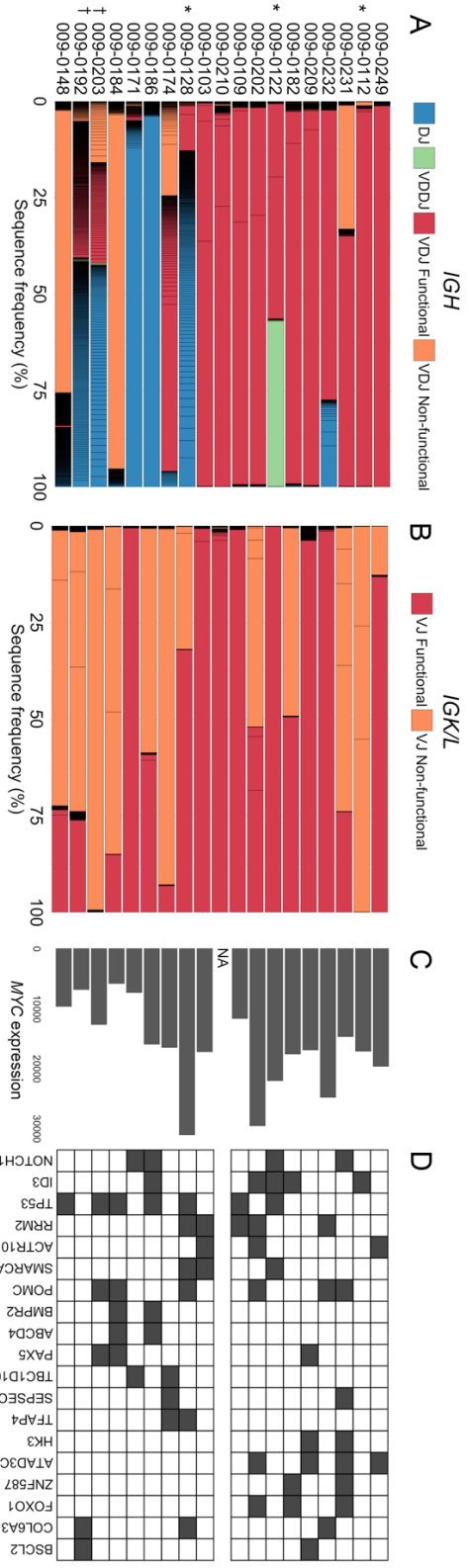


Figure 2.1. HTS of gDNA and RNA from Ugandan BL tumors identifies the repertoire of Ig rearrangements and sequence variants in tumor cells.

(A-B) Cumulative frequency plots of all unique *IGH* (A) or *IGK*/*IGL* (B) sequences identified by HTS of gDNA from the Ugandan BL cohort. Each segment in the bar plots represents a unique nucleotide sequence, and the color indicates the type of Ig rearrangement. Black lines separate the unique sequences, so highly polyclonal regions appear black in the figure. *Indicates that the tumor sample was EBV-negative; †indicates that the repertoire was classified as polyclonal. (C) Normalized *c-MYC* expression for each of the 18 Ugandan tumors on which RNAseq analysis was performed. RNAseq was not performed on sample 009-0210 due to poor quality RNA. (D) Genes with at least one single nucleotide variant (SNV) detected in RNAseq and predicted to be pathologic in each of the 18 Ugandan BL tumors. A black box indicates the presence of at least one SNV.

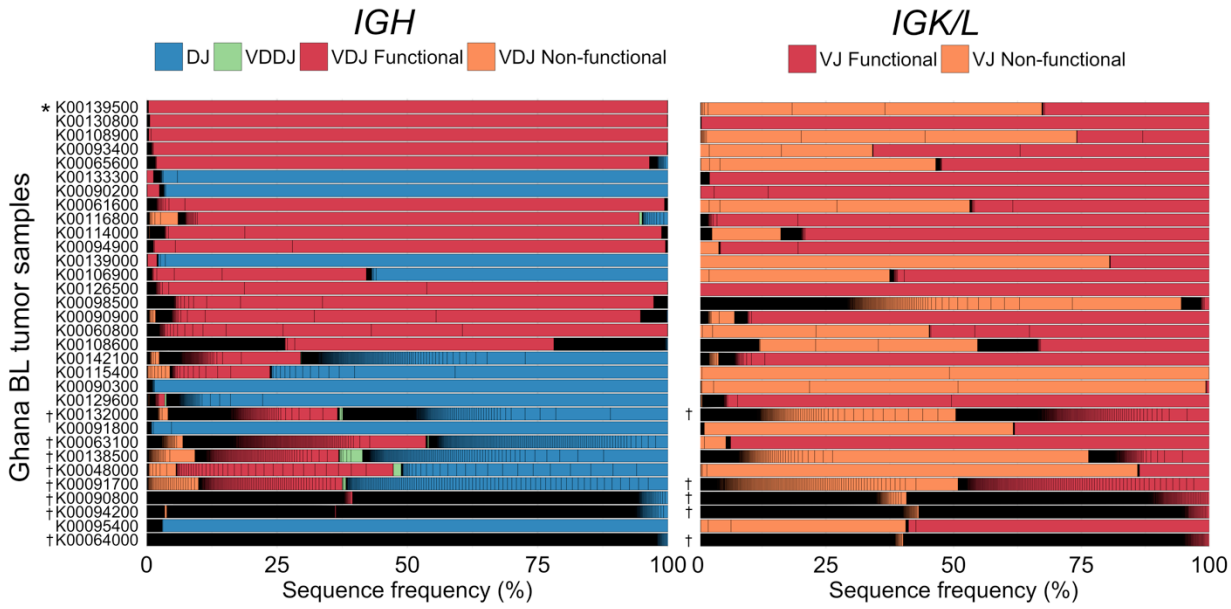


Figure 2.2. HTS of gDNA from Ghanaian BL tumors identifies the repertoire of Ig rearrangements in tumor cells.

(A-B) Cumulative frequency plots of all unique *IGH* (A) or *IGK*/*IGL* (B) sequences identified by HTS of gDNA from the Ghanaian BL cohort. Each segment in the bar plots represents a unique nucleotide sequence, and the color indicates the type of Ig rearrangement. Black lines separate the unique sequences, so highly polyclonal regions appear black in the figure. *Indicates that the tumor sample was EBV-negative; †indicates that the repertoire was classified as polyclonal.

2.4.4 *IGH and IGK/IGL sequence variation*

Analysis of the *IGH* and *IGK*/*IGL* repertoires in clonal tumors revealed, in most cases, a large number of Ig sequence variants. Using neighbor-joining tree estimation from edit distance matrices, phylogenetic trees were created for each tumor to visualize the relationships between all unique sequences (Figure 2.3A). Most of the unique sequences in each tumor are closely related to the dominant sequence, and cluster within the trees. A small fraction of highly dissimilar sequences in each tumor likely derive from non-malignant B cells captured in the tumor biopsy. We compiled density plots of the edit distance, defined as the number of nucleotide differences between the most frequent sequence and all other unique sequences in the tumor, and found that BL tumors are characterized by many related, low edit-distance (<10 nt) sequences (Figure 2.3B,C). In contrast, edit distance density plots from control PBMC[104] and bone marrow[109] samples contain very few unique sequences that are closely related to, and therefore have low edit distances from, the most frequent sequence (Figure 2.3E,F). Furthermore, HTS of the *TRB* locus in all 51 tumors revealed very few sequences with low edit distance, demonstrating that this phenomenon is restricted to the Ig loci in tumor cells (Figure 2.3D). The ensemble of unique sequences defined by the most frequent sequence, and all variants with an edit distance <10, likely define the malignant population.

In tumors with clonal Ig rearrangements, 3 to 2,091 unique *IGH* sequences with an edit distance <10 from the most frequent sequence were detected. Surprisingly, this rich sequence variation was observed in tumors with a dominant, non-productive DJ rearrangement, as well as in those with a dominant, productive VDJ rearrangement. The median number of variants with an edit distance <10 was 262 for VDJ sequence families, and 192 for DJ sequence families ($P = 0.54$). Sequence variation within VDJ and DJ families appears to be uniformly distributed throughout the

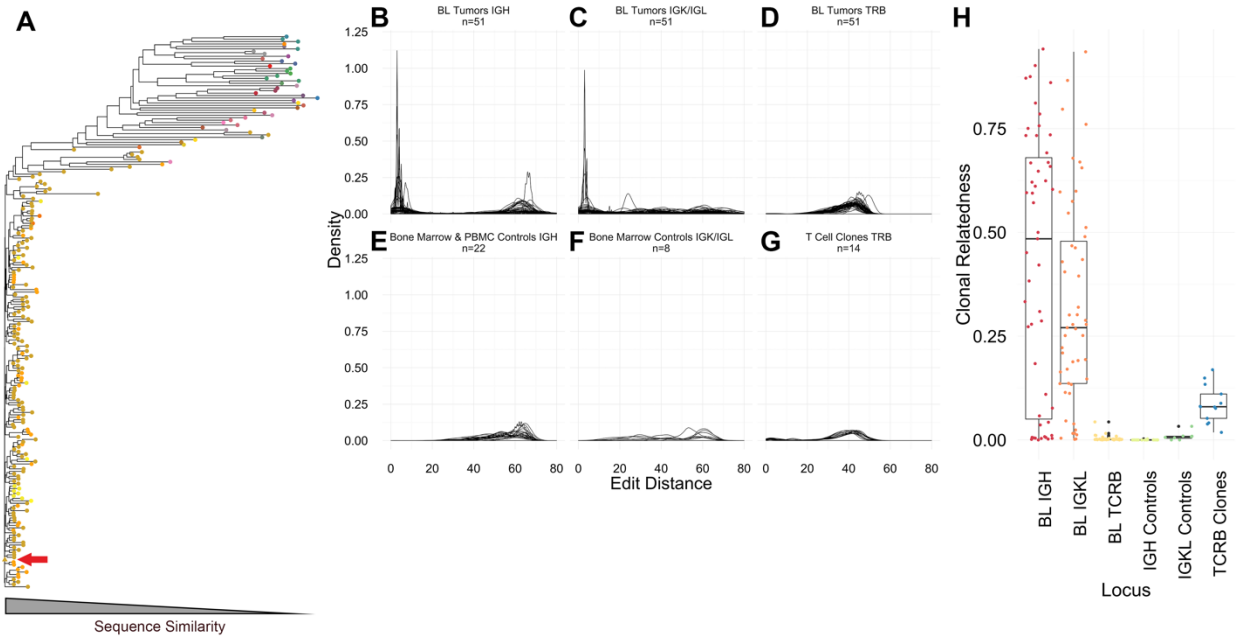


Figure 2.3. HTS of *IGH* in BL tumors reveals large families of closely related sequences.

(A) Phylogenetic tree of all *IGH* sequences observed in a representative BL tumor sample. Unique *IGH* VDJ gene rearrangements are indicated by node color. The red arrow indicates the dominant unique sequence identified in the tumor. (B-G) Density plots of nucleotide edit distance (the number of nucleotide differences between a given unique sequence and the most frequent sequence in the tumor) for all sequences identified in all BL *IGH* samples (B), BL *IGK*/*IGL* samples (C), BL *TRB* samples, non-malignant bone marrow and PBMC *IGH* control samples (E), bone marrow control *IGK*/*IGL* samples (F), and for $CD8^+$ T cell clones (G). (H) Clonal relatedness scores for each of the sample populations listed above (defined as the total number of distinct unique sequences with an edit distance <10 from the most frequent sequence, divided by the total number of unique sequences).

length of the sequencing reads (Figure B5, Appendix B). The high degree of sequence diversity within both VDJ and DJ sequence families demonstrates that mutagenesis of *IGH* in BL tumor cells is not limited to productively rearranged, expressed alleles.

We calculated the clonal relatedness of each tumor, defined as the fraction of all unique sequences with an edit distance <10 from the most frequent sequence. Most tumors had high *IGH* clonal relatedness (median 0.48; n=51) (Figure 2.3H). To evaluate whether the sequence variation observed in BL tumors might be attributable to PCR or sequencing error, we compared the sequence diversity in BL tumors with that observed in clonal but non-malignant samples (Figure 2.3G). The median clonal relatedness of the *TRB* locus in 14 primary $CD8^+$ T-cell clones (generated using the same sequencing strategy and platform as used for the BL tumors) was 0.08. Thus, the high degree of BL-associated sequence variation is not likely due to sequencing error, but rather to active mutational processes within the tumor cells (Figure 2.3H). We propose that clonal relatedness has utility to evaluate ongoing hypermutation in BL tumors.

2.4.5 *Monoallelic IGH rearrangements in BL tumor cells*

At the pro-B cell stage of development, D to J rearrangements occur on both *IGH* alleles, followed by sequential (if necessary) V to DJ rearrangements on both *IGH* alleles (Figure 1.2).[56] If BL tumors develop from B cells that have followed this canonical developmental pathway, tumor cells should carry vestiges of two co-dominant *IGH* rearrangements, either one productive VDJ and one incomplete DJ rearrangement (Figure 2.2A, patient K00106900), or two VDJ rearrangements, one productive and one non-productive (Figure 2.1A, patient 009-0231). However, only 6/41 tumors with a clonal *IGH* repertoire appeared to carry two co-dominant *IGH* rearrangements. PCR, qRT-PCR, and droplet digital PCR were employed to assess the rearrangement status of the *IGH* alleles in BL tumors. However, the presence of non-malignant

cells in the tumor biopsies, such as infiltrating T cells and stromal cells, which carry germline *IGH* loci, prevented unambiguous resolution of this question. The VDJ:DJ sequence read ratios (Figure B6, Appendix B) clearly demonstrate that most tumors are characterized by one primary type of rearrangement.

2.4.6 *IGHV gene segment utilization*

To evaluate whether the BCRs expressed in African BL tumors exhibit stereotypy, manifested by biased Ig gene segment utilization, we analyzed the *IGHV*-gene segment usage of the dominant clone in 30 tumor samples that contained a clonal VDJ rearrangement, and for which an unambiguous gene segment assignment could be made. In addition, we examined the *IGHV* usage in 40 endemic and 33 sporadic tumor-associated BCRs of BL cases reported in the literature.[13,78,110] We also used a novel computational method to identify *IGHV* utilization in RNAseq data from 28 sporadic BL tumors.[78] Combining the 30 evaluable cases from our two BL cohorts with these independent tumor cohorts allowed us to examine *IGHV* usage in 131 BL cases: 70 endemic and 61 sporadic. BL-associated *IGHV* usage was compared with the *IGHV* usage in unfractionated PBMC, bone marrow, and purified B cell populations from 33 adult and 9 pediatric control subjects[104,109,111] (Figure 2.4A). *IGHV3-30*, *IGHV3-21*, *IGHV3-07*, *IGHV4-59*, and *IGHV4-34* were preferentially utilized in BL BCRs as compared to the control B cell populations (*P* values: $1.1\text{x}10^{-12}$, $1.5\text{x}10^{-6}$, $5.7\text{x}10^{-6}$, $8.9\text{x}10^{-6}$, and $1.5\text{x}10^{-2}$, respectively; Figure 2.4B). After correcting for multiple comparisons, no *IGHV* genes were differentially utilized in sporadic versus endemic BL BCRs.

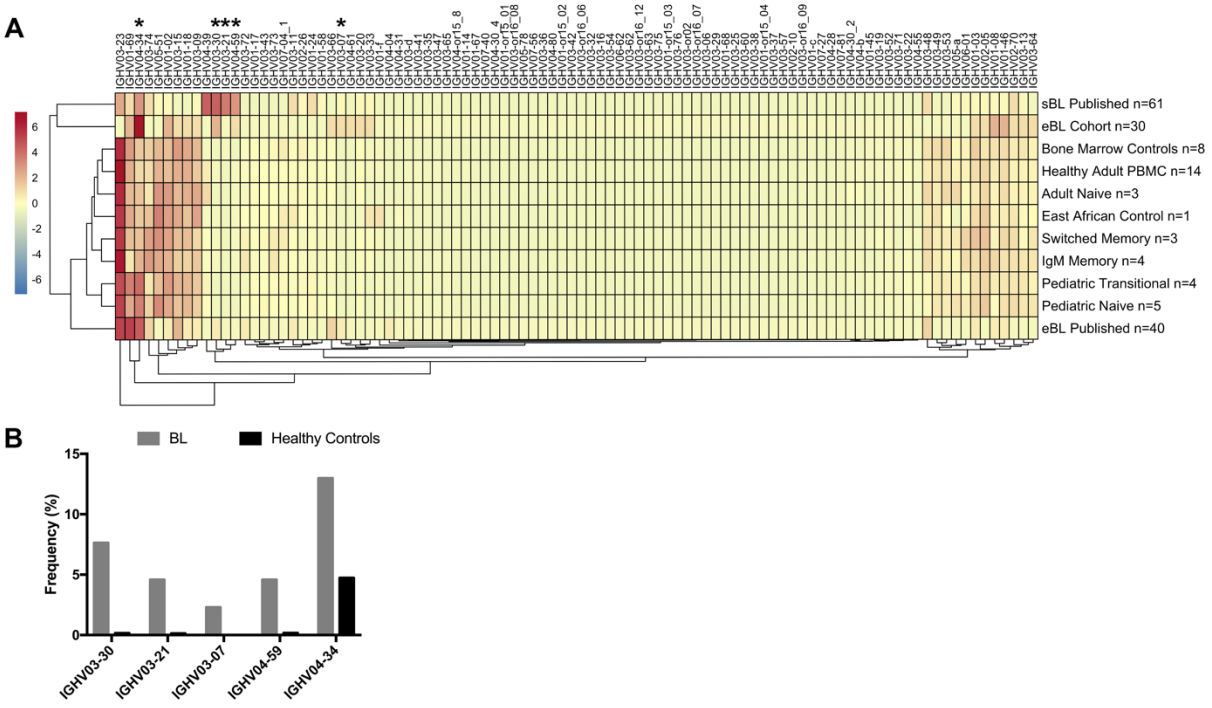


Figure 2.4. Biased *IGHV* gene utilization observed in BL tumors.

(A) Heat map illustrating *IGHV* gene usage in the 30 endemic BL samples from our cohorts, 40 published endemic BL samples, and 61 published sporadic BL samples, as compared with 42 healthy controls. Highly enriched *IGHV* genes are indicated in red, and genes with significant differential utilization ($P < .05$) in BL tumor and control B cells are indicated by an asterisk. Both *IGHV* gene utilization and all samples are clustered by similarity. (B) Frequency of *IGHV* gene utilization in the five significantly differentially used *IGHV* genes. *IGHV* gene frequencies in BL tumors are plotted in gray and *IGHV* gene frequencies of healthy controls are plotted in black.

2.4.7 *Transcription of BL tumor IGH and IGK/IGL rearrangements*

Analysis of BL tumor RNA by RT-PCR and RNAseq was performed to assess transcription of productive tumor-associated *IGH* and *IGK*/*IGL* rearrangements identified by HTS of tumor gDNA. This analysis focused on 18 Ugandan BL tumors from which RNA of sufficient quality was extracted. RNAseq was able to identify 7 of 12 clonal, productive *IGH* VDJ rearrangements

and 13 of 15 clonal, productive *IGK/IGL* rearrangements (Figure B7, Appendix B). Failure to detect transcription of dominant Ig rearrangements predicted by HTS of tumor gDNA was closely associated with the length of the CDR3. The average CDR3 length of tumor-associated *IGH* rearrangements not detected by RNAseq was 51 nt, which is longer than the length of the RNAseq reads (50 nt). In contrast, the average CDR3 length of *IGH* and *IGK/IGL* rearrangements that were detected by RNAseq was 43.5 and 31 nt, respectively.

RT-PCR was used to confirm transcription of dominant *IGH* and *IGK/IGL* rearrangements, and to determine the isotype of the putative tumor-associated rearrangement (Figure B3, Appendix B). Transcripts encoding the predicted heavy chain were identified in 13 Ugandan BL tumors with dominant, productive *IGH* rearrangements (Figure B8, Appendix B). Two tumors carried a dominant, non-productive *IGH* VDJ rearrangement, only one of which contained a detectable transcript. The inferred isotype of tumors with a dominant VDJ rearrangement was $IgM^{+}IgD^{+}$ in 12 cases and IgG^{+} in two cases (Figure B8, Appendix B). Moreover, 15 of the 16 evaluable tumors expressed a detectable *IGK/IGL* rearrangement, although two were predicted to be non-productive (Figure B8, Appendix B).

RNAseq differential gene expression analysis demonstrated two-fold higher *c-MYC* expression in tumors carrying a dominant productive *IGH* rearrangement, compared with those for which a productive rearrangement was not identified by HTS (P = 0.037, Figure 2.1C). No significant differences, however, were seen in the expression of BCR signaling pathway genes in tumors with and without a dominant, productive *IGH* rearrangement. RNAseq variant analysis revealed sequence variants in genes previously reported to be mutated in BL, including *ID3*, *TP53*, *SMARCA4*, *ZNF587* and *FOXO1*,[13,23,62,63,78] as well as 14 genes that are mutated in other cancers, including *NOTCH1*, *PAX5*, and *TFAP4* (Figure 2.1D; All sequence variants listed in Table C2,

Appendix C). Two tumors with a polyclonal *IGH* repertoire and a clonal *IGK/IGL* rearrangement by HTS of gDNA carried mutations in several genes mutated in other BL tumors, including *TP53*, *POMC*, *COL6A3,* and *BSCL2*, supporting the diagnosis of BL in these cases.

2.4.8 *SHM in BL BCRs*

A PCR-based strategy was used to sequence the complete variable region of productive *IGH* and *IGK*/*IGL* rearrangements not covered by HTS (Figure B3, Appendix B). Nucleotide-level analysis revealed a median of 23 candidate sites of SHM per *IGH* variable region (range, 4 to 43; n=29).[102,103,105] Non-synonymous SHM clustered in the CDRs as compared to the FRs (*P* values: FR1:CDR1 $7.1x10^{-6}$, CDR1:FR2 $1.0x10^{-4}$, FR2:CDR2 $5.2x10^{-7}$, CDR2:FR3 $7.6x10^{-5}$; Figures 2.5A,B). An analogous approach was utilized on the clonal *IGK/IGL* rearrangements, and a median of 13 candidate sites of SHM per *IGK/IGL* variable region were identified (range, 0 to 33; n=23).

The median number of SHM sites in EBV-positive and -negative tumors were 24 and 13, respectively ($P = 0.014$; Figure 2.5B). This disparity has been cited to support the hypothesis that EBV-positive and EBV-negative tumors arise from distinct germinal center sub-populations.[11] A higher number of candidate sites of SHM was found in patients with a favorable clinical outcome, but this difference was not statistically significant (Figure B9, Appendix B).

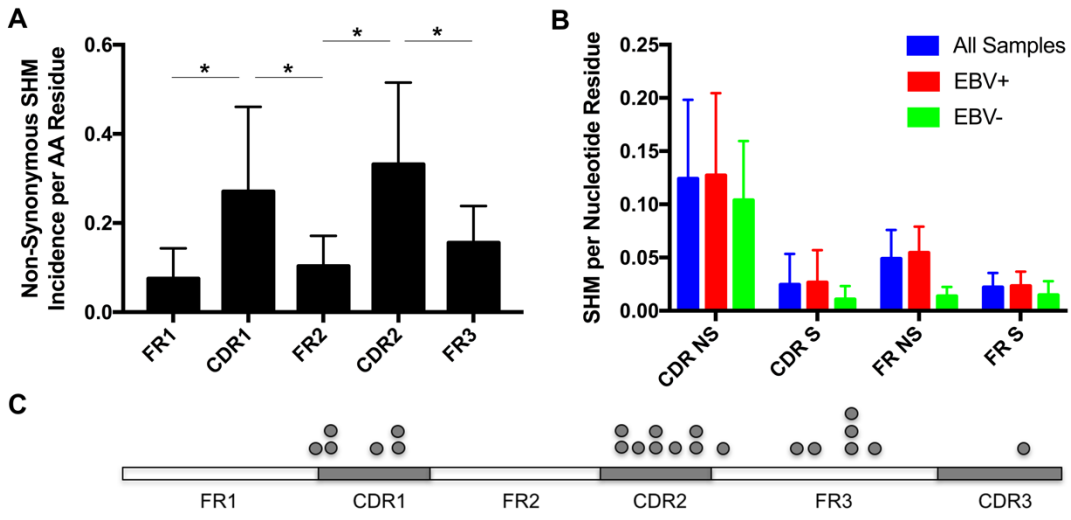


Figure 2.5. Non-synonymous sites of SHM and SHM-induced NLGS are enriched in CDRs.

Analyses were performed on all clonal BL tumors for which the complete V region of the most frequent sequence could be determined by capillary sequencing (n=29). (A) Sites of non-synonymous SHM plotted per amino acid residue in each *IGH* gene region. Statistically significant differences in SHM incidence are indicated by asterisks. (B) Sites of non-synonymous (NS) and synonymous (S) SHM plotted per nucleotide residue in EBV^+ and EBV^- tumors by CDR or FR. (C) A schematic of an *IGH* or *IGK*/*IGL* gene rearrangement is shown and the location of all NLGS (Asn-X-Ser/Thr amino acid motifs) identified within all complete, clonal BL tumor *IGH* and *IGK*/*IGL* sequences are indicated by grey circles.

2.4.9 *Enrichment for N-linked glycosylation sites (NLGS) in CDRs*

We examined the putative tumor-associated Ig sequences in the Ugandan and Ghanaian BL tumors for the canonical NLGS motif: Asn-X-Ser/Thr. Of 33 clonal BL tumors from which a complete *IGH* or *IGK*/*IGL* variable region sequence was captured, we identified 22 NLGS across 18 samples (Figure 2.5C). Nineteen of the sites were likely introduced by mutation, and three were present in the germline *IGHVH4-34* sequence.[102,103,105] The NLGS cluster within and around the

CDRs, which could enable interactions between BL tumor cells and stimulatory molecules in the tumor microenvironment.

2.5 DISCUSSION

This study demonstrates that Ig gene rearrangements in African BL tumors are characterized by far more molecular heterogeneity than suggested by previous studies, most[11,87-91,112] but not all[13] of which were performed using capillary sequencing. HTS of the Ig loci in gDNA from BL tumors identified at least one clonal locus in 90% (46/51) of tumors. Thirty-one tumors (61%) harbored a clonal *IGH* rearrangement, comparable to the clonality detection rate obtained by the BIOMED-2 FR3 primers (63%).[113,114] In five Ghanaian tumors cytologically-classified as BL, no dominant Ig rearrangements were identified by HTS. It is possible that these tumors were misdiagnosed as BL, particularly given the clinical limitations over 20 years ago when the samples were acquired. It is also possible that the 15% threshold used to define clonal tumors was set too high; however, this stringency seemed most appropriate due to the likelihood of reactive B cell expansion within the tumors. In the Ugandan cohort, RT-PCR and RNAseq confirmed transcription of one or both of the predicted productive *IGH* and *IGK*/*IGL* rearrangements in most tumors. We hypothesize that malignant transformation occurs in antigen-experienced B cells that expressed a functional heterotetrameric BCR, but that one or both of the BCR components may have been rendered non-productive by mutation during disease pathogenesis.

HTS demonstrates that most BL tumors carry many closely related *IGH* and *IGK*/*IGL* sequences with low edit distances that collectively define the malignant population. This rich sequence diversity is not likely due to PCR or sequencing error, as it was not observed within the

IGH repertoires of B cells from blood or marrow of healthy donors, nor in the *TRB* repertoire in any of the 51 tumors or in 14 non-malignant $CD8^+$ T-cell clones, all generated using the same sequencing strategy. Thus, the mutational mechanism that creates this diversity is not active on *IGH* in non-malignant B cells, nor on *TRB* in BL tumor-infiltrating T cells. Comparable *IGH* sequence diversity was observed within families of both VDJ and DJ rearrangements. Selective pressure from SHM-induced affinity maturation is only expected to occur on productive, actively transcribed Ig alleles.[115] These observations imply that a mutational mechanism that is not appropriately regulated by antigen-driven SHM and affinity maturation is active on both *IGH* and *IGK*/*IGL* in BL tumor cells. Although it is tempting to speculate that this mechanism involves activation-induced cytidine deaminase (AID), particularly given that AID transcripts were detected in all tumors by RNAseq, this remains to be proved. *IGH* sequence evolution has also been reported by HTS of pediatric acute lymphoblastic leukemia, a malignancy of an early B cell progenitor, in a strikingly similar magnitude as that discovered in BL.[116]

Although B cells undergo biallelic *IGH* DJ rearrangement during development, HTS of tumor gDNA identified only one *IGH* rearrangement in 35/41 (85%) of the clonal BL tumors. Identification of at least one rearrangement in 41/51 (80%) of tumors demonstrates that our HTS strategy could efficiently detect *IGH* rearrangements. Given the unique translocation involving the *IGH* locus that occurs in BL, one possible interpretation is that the DJ rearrangement on the second *IGH* allele may not have occurred in BL progenitors. DJ rearrangement may have been inhibited by the t(*c-MYC*;*IGH*) translocation that occurs in 80% of BL tumors (the remaining 20% of translocations are to the *IGK*/*IGL* loci).

Detection of only one *IGH* DJ rearrangement in BL tumors is consistent with a previous study that reported a high frequency of monoallelic DJ rearrangements in BL.[117] The consensus in

the BL literature is that t(*c-MYC;*Ig) translocations occur in mature B cells due to recurrent antigenic stimulation and aberrant AID activity in the germinal center.[118-121] However, our data raise the possibility that the *c-MYC* translocation may actually occur in a developing B cell, before that cell completes DJ rearrangement on both *IGH* alleles. Another group recently identified a pre-B cell population with concurrent expression of recombination activating genes (RAG) and AID, providing additional support for the possibility that the t(*c-MYC*;Ig) translocation could occur early in development.[122]

Biased, or stereotyped, usage of particular *IGHV* genes was observed in BL tumors, suggesting that BL progenitors carrying particular Ig genes preferentially differentiate into the memory B-cell compartment. Thus, BL progenitors are preferentially selected based on their BCR, and likely their antigenic-specificity. Chronic lymphocytic leukemia,[65,66,68,70,71] mantle cell lymphoma,[64] and splenic marginal zone lymphoma[72] all demonstrate BCR stereotypy, strongly suggesting a role for antigen in the pathogenesis of a substantial subset of mature B cell malignancies, including BL.

BL Ig sequences reportedly contain significantly more SHM-induced NLGS (82%) than do non-malignant antigen-experienced B cells (9%).[112] In our BL cohorts, NLGS were highly enriched and clustered within CDRs, which could allow for interactions with molecules in the tumor microenvironment. Lectins on tumor-resident macrophages and dendritic cells can bind to NLGS and induce signaling through the BCR,[123] suggesting a mechanism by which the BCR could be activated in BL and promote tumor survival. Furthermore, despite extensive SHM, IgM, rather than IgG, dominated the BL BCR repertoire, suggesting an increased dependence on BCR signaling in BL tumors.[84,124] BCR signaling has been implicated in the pathogenesis of a number of B cell malignancies,[60,66,68,69,125] against which BCR signaling pathway inhibitors have proven

effective.[75,76] These data suggest that these agents may also have utility for the treatment of African BL.

This study demonstrates that a distinct family of Ig rearrangements uniquely characterizes most African BL tumors, and suggests that these sequences may have utility as a biomarker to diagnose and monitor disease progression. Given the frequency of non-productive VDJ and DJ *IGH* rearrangements and the rich sequence variation detected in BL tumors, an RNA- or PCR-based approach would have limited utility. These findings support the potential use of a gDNA-based, HTS approach for disease monitoring. Indeed, we have used HTS to detect BL-associated *IGH* sequences in matched blood, serum, and cerebrospinal fluid samples from patients reported on in this study. Preliminary data suggest that detection of these sequences may have prognostic value for BL patients.

Chapter 3. THE BCR AS A UNIQUE BIOMARKER FOR BL

3.1 ABSTRACT

All normal B cells express a B cell receptor on their surface that is generated by the ordered rearrangement of immunoglobulin (Ig) gene segments. The African subtype of Burkitt lymphoma (BL) is a malignancy of mature, antigen-experienced B cells that have traversed the germinal center. The Ig genes in BL tumor cells are characterized by point mutations, indicating that they have undergone somatic hypermutation and affinity maturation in the germinal center. The combined processes of chromosomal rearrangements and somatic hypermutation generate a unique BCR on the surface of virtually every B cell. We utilized high-throughput sequencing to characterize the tumor-associated Ig rearrangements in 69 primary BL tumors from three independent cohorts. We then probed patient-matched peripheral blood mononuclear cells, plasma, and serum samples obtained at diagnosis for evidence of the tumor-associated Ig heavy chain (*IGH*) rearrangements. Tumor-associated *IGH* sequences were detected in circulation in a patient-specific manner. Furthermore, overall patient survival in the Ugandan cohort was inversely correlated with the detection of tumor-associated *IGH* sequences at diagnosis. This study demonstrates that circulating tumor DNA may have utility as a prognostic biomarker for BL patients.

3.2 INTRODUCTION

BL is the most common pediatric malignancy in sub-Saharan Africa. It is characterized by rapidly proliferating tumors of the jaw, orbit, and abdominopelvic region. BL occurs primarily in

children ages 2-13 and its incidence is as high as 21.5 cases per year (per 100,000 children under the age of 15) in highly endemic regions.[126] In these areas, BL is 6 times more frequent than pediatric ALL, the most common pediatric cancer in the United States.

Since the first report of African BL in 1958,[127] the standard treatment regimen has changed very little. Moreover, disease staging and treatment efficacy are most commonly assessed by tumor palpation and imprecise imaging techniques such as plain film radiographs and ultrasound. These relatively insensitive methods only detect large disease deposits, making disease severity and dissemination virtually impossible to assess. Diagnostic and treatment limitations likely contribute to poor patient outcomes; the long-term survival rate for BL patients within Africa is only 30-50%.[5] These limitations favor the development of a more sensitive, molecular approach to disease detection.

BL is a malignancy of antigen-experienced B cells. All normal, and most malignant, B cells express a functional B cell receptor (BCR) on the cell surface, which is generated by the ordered rearrangement of V, D, and J immunoglobulin (Ig) gene segments. In the bone marrow, the D and J gene segments on both Ig heavy chain (*IGH*) alleles are concurrently rearranged.[56] The B cell then performs V to DJ rearrangements independently on each allele, in search of a productive *IGH* rearrangement. Random nucleotide insertions and deletions occur at the gene segment junctions to increase BCR sequence diversity. Later in B cell development, upon antigen encounter, the Ig genes undergo AID-mediated somatic hypermutation (SHM), which generates point mutations to increase affinity for an antigen. These processes collectively generate two unique Ig rearrangements on both *IGH* alleles, each of which has the potential to be used as a cell-specific molecular barcode to identify a particular B cell or clonal B cell population.

The use of high-throughput sequencing (HTS) on lymphoid malignancies has made unique, clonal antigen receptor rearrangements carried by tumor cells readily identifiable and detectable with high sensitivity and specificity.[92,93] Detection of tumor-specific antigen receptor sequences by HTS is being used to monitor treatment efficacy and disease relapse in many lymphoid malignancies,[94-96,128,129] demonstrating that antigen receptor sequences can serve as highly specific biomarkers. The ability of HTS to detect a single tumor cell has been reported to be as sensitive as one in 1 x 10^6 genomes.[95,96]

Recent studies utilizing HTS of BL tumors have revealed novel tumor characteristics. Sequencing of tumor gDNA, as opposed to tumor RNA, has demonstrated that a large proportion of BL tumors harbor only a non-functional *IGH* allele (either a sole, incomplete DJ rearrangement, or an unproductive VDJ rearrangement), suggesting that the malignant cells in these tumors no longer express a functional BCR.[15] HTS of BL tumors has also revealed that seemingly clonal tumors are often characterized by a high degree of sequence variation at the *IGH* and *IGK*/*IGL* loci.[13,15] Hundreds of related Ig sequences are frequently detected in BL tumors, each containing small numbers of point mutations. The number of nucleotide changes in a given sequence, as compared to the most frequent sequence, is defined as the edit distance. BL tumors are characterized by a large number of unique sequences with low edit distance (<10 nucleotides), and this phenomenon is not observed in healthy B cell populations.[15] This extensive sequence variation is associated with both productive and unproductive *IGH* rearrangements, suggesting that the entirety of the low edit distance sequence repertoire arising from both *IGH* alleles may represent the malignant population and thus have utility as a biomarker for BL.

Several previous studies have demonstrated that general indicators of gross tumor burden, including serum lactate dehydrogenase (LDH) levels and clinical disease stage have prognostic

value for eBL patients. Patients with higher LDH levels and advanced stage disease have worse progression-free and overall survival[130,131] and have lower rates of complete remission.[132] Indicators of high EBV viremia, such as elevated levels of EBV DNA in patient plasma[133] and high EBV-specific antibody titers[132] were also found to be associated with eBL patient outcome.

We utilized HTS of the *IGH* locus in an effort to develop a more sensitive and specific measurement of BL tumor burden. HTS was performed on genomic DNA (gDNA) isolated from eBL patient tumors and matched peripheral blood mononuclear cells (PBMC), plasma, serum, and cerebrospinal fluid (CSF) samples from three independent BL cohorts to detect the tumor-associated family of *IGH* rearrangements. Our study demonstrates that unique tumor-associated *IGH* rearrangements are readily detected in circulation at diagnosis and have utility as a prognostic biomarker for BL.

3.3 METHODS

Samples: Tumor biopsies from the Ugandan BL cohort (collected between August 2013 and March 2014) and the Ghanaian BL cohort (collected between 1975 and 1992) were obtained as previously described.[15] An additional 15 diagnostic, patient-matched PBMC samples from the Ugandan cohort and five diagnostic serum and nine diagnostic CSF samples from the archival Ghanaian cohort were obtained. gDNA from fine needle aspirates from an additional 18 BL tumors was obtained from the Jaramogi Oginga Odinga Teaching and Referral Hospital in western Kenya. The samples were obtained at diagnosis and a morphologic diagnosis of BL was confirmed in each case by two independent pathologists. Patient-matched plasma samples were also obtained from

the Kenyan cohort at diagnosis. The Kenyan samples were shipped to the Fred Hutchinson Cancer Research Center in Seattle, WA on dry ice.

gDNA isolation: gDNA was extracted from tumor samples using the QIAGEN DNeasy Blood and Tissue Kit according to the manufacturer's instructions. PBMC samples were washed two times with PBS containing 3.6mM EDTA, pH 8.0 and once with PBS alone before processing. gDNA was extracted from PBMC, serum, and CSF samples using the QIAGEN QIAamp Blood Mini Kit according to the manufacturer's instructions. gDNA was extracted from plasma samples using the QIAGEN QIAamp Blood Mini Kit or the QIAGEN QIAamp MinElute Virus Spin Kit according to the manufacturer's instructions.

HTS of *IGH*: Survey-level HTS of the *IGH* locus was performed on gDNA from each sample using the ImmunoSeq platform at Adaptive Biotechnologies.[94] As previously described,[15] PCR libraries of Ig rearrangements were generated using sense primers specific for all *IGHV* and *IGHD* gene segments and antisense primers specific for all *IGHJ* gene segments. The libraries were sequenced on the Illumina MiSeq platform and generated 130 nucleotide sequence reads encompassing the 3' portion of framework region 3 and the entirety of complementarity-determining region 3. The initial analysis of raw sequence reads, including filtering and decomposition of reads into their component *V*, *D*, and *J* segments, and non-templated junctional nucleotide insertions was performed at Adaptive Biotechnologies. Subsequent sequence analyses were performed using the LymphoSeq R package (http://www.bioconductor.org/packages/LymphoSeq) created by David G. Coffey. A tumor was classified as clonal if the most frequent rearrangement comprised >15% of the *IGH* repertoire.

Statistics: The Fisher Exact test was used to compare the frequency of *IGHV* utilization, with a Bonferroni correction to adjust for multiple comparisons. The number of sequence variants

detected for *IGH* VDJ versus *IGH* DJ rearrangements were compared using the Student's t-Test. BL patient survival based on detection of circulating tumor DNA (ct-DNA) was assessed using the Log-rank test.

3.4 RESULTS

3.4.1 *Study populations*

Clinical information on the BL patients in the Ugandan (n=19) and Ghanaian (n=32) cohorts has been previously reported.[15] From July 2009 through September 2012, BL patients were admitted to the Jaramogi Oginga Odinga Teaching and Referral Hospital in Kisumu, Kenya. Tumor biopsies were obtained from 18 pathologically-confirmed BL cases before the start of therapy. Most patients were male (72%) and the median age at hospital admission was eight years of age. All patients were HIV-negative (Table 3.1).

Table 3.1. Clinical information for the Kenyan BL patient cohort

Characteristic	Kenyan cohort (n = 18)
Sex, no. (%)	
Female	5 (28)
Male	13 (72)
Age at hospital admission	
Median	8
Range	2-13
HIV status, no (%)	
Negative	18 (100)
Positive	0 (0)

3.4.2 *HTS of BL tumors*

HTS of the *IGH* locus was performed on gDNA from all 18 BL tumors obtained in western Kenya. A median of 891,279 total *IGH* reads were obtained per tumor (range: 1,627-7,441,750) (Table D1, Appendix D). Sequencing revealed a clonal *IGH* rearrangement in 14 of 18 tumors (Figure 3.1). The remaining four tumors had a polyclonal *IGH* repertoire. Of the clonal tumors, only five had evidence of two rearranged *IGH* alleles; the remainder harbored only one detectable VDJ (n=5) or DJ (n=4) rearrangement. In addition to the Kenyan BL cohort, all subsequent analyses will include data from previously published HTS of BL tumors from Uganda and Ghana,[15] for a total BL cohort of 69 tumors.

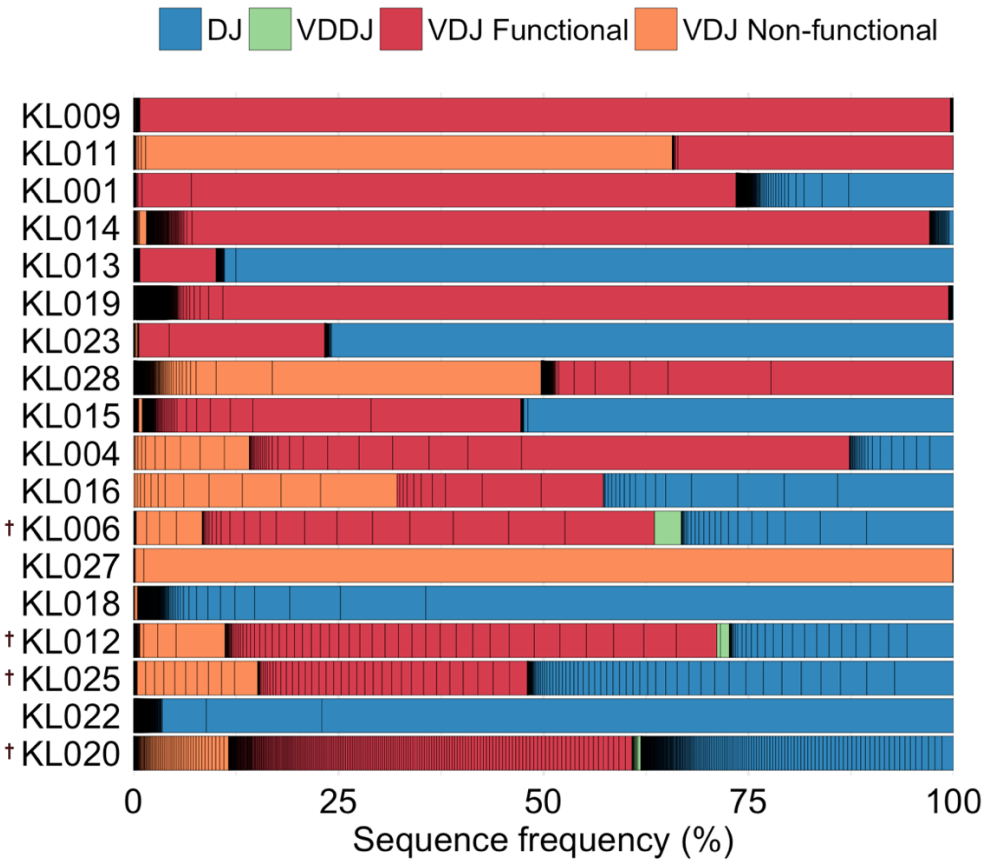


Figure 3.1. HTS of gDNA from Kenyan BL tumors identifies the repertoire of *IGH* rearrangements in tumor cells.

A cumulative frequency plot of all unique *IGH* sequences identified by HTS of gDNA from the Kenyan BL cohort. Each segment in the bar plots represents a unique nucleotide sequence, and the color indicates the type of Ig rearrangement. Black lines separate the unique sequences, so polyclonal regions appear black in the figures. The tumors are ordered vertically by their clonality score. †indicates that the repertoire was classified as polyclonal.

3.4.3 *IGHV utilization*

The *IGHV* gene segment utilized in the most frequent, productive *IGH* rearrangement detected in each clonal tumor was analyzed. Due to the limited HTS read length and the prevalence of dominant DJ rearrangements, not all tumors contained identifiable V gene segments. Three tumors harbored a single VDJ rearrangement with an unambiguous gene segment assignment. In those tumors, *IGHV4-39*, *IGHV3-15*, and *IGHV3-53* were utilized. Two tumors with identifiable V gene segments harbored evidence of two rearranged *IGH* alleles. One contained an incomplete DJ rearrangement, paired with a VDJ rearrangement that utilized *IGHV4-39*. The other contained an unproductive VDJ rearrangement that utilized *IGHV3-43* and a productive VDJ rearrangement that utilized *IGHV3-23*. None of these V gene segments were found to be enriched in the Ugandan or Ghanaian cohorts. Assessment of *IGHV* gene utilization in all three BL cohorts (n=35) revealed the addition of *IGHV4-39* as preferentially enriched in BL tumors ($P = 0.037$). *IGHV4-39* was utilized at a frequency of 6.6% in BL tumors, but only 1.6% in control B cell populations. *IGHV3-30*, *IGHV3-21*, *IGHV3-07*, *IGHV4-59*, and *IGHV4-34* were previously found to be enriched in the Ugandan and Ghanaian BL cohorts.[15]

3.4.4 *Sequence variation*

In addition to the dominant *IGH* rearrangement identified in clonal tumors from the Kenyan cohort, most tumors contained many, low-frequency sequence variants. These sequence variants were closely related to the putative tumor clone, as they differed from the dominant sequence by a low edit distance (<10 nucleotides) (Figure 3.2A). In clonal tumors, the number of variants per tumor ranged from 3 to 4509 unique, but related, sequences and were associated with both productive and unproductive *IGH* rearrangements. The degree of sequence variation was similar on *IGH* alleles harboring a VDJ or DJ rearrangement; there was a median of 135 sequence variants per tumor containing a clonal VDJ rearrangement, as compared to a median of 230 sequence variants per tumor containing a clonal DJ rearrangement (p=0.62). Thus, BL tumors are characterized by large families of closely related B cell populations with a high degree of sequence variation at the *IGH* locus. These low edit distance populations are not found in normal B cell populations[15] and likely define the malignant population. This phenomenon was also reported in the Ugandan and Ghanaian BL cohorts.[15] The edit distance analysis (Figure 3.2A) includes edit distance data from all three BL cohorts (n=69).

The clonal relatedness metric can be used to evaluate the degree of sequence variation within a given population.[15] Clonal relatedness is defined as the fraction of unique nucleotide sequences with an edit distance <10 from the most frequent sequence. Healthy B cell populations are characterized by low clonal relatedness scores. In contrast, BL tumors have high clonal relatedness scores, demonstrating that large malignant populations frequently occur in BL. BL tumors have a median clonal relatedness score of 0.38 at the *IGH* locus (n=69), 0.27 at the *IGK*/*IGL* loci (n=51), and 0.0013 at the *TRB* locus (n=51) (Figure 3.2B). Healthy PBMC and bone marrow controls demonstrate low clonal relatedness scores of 0.00009 at the *IGH* locus and 0.007 at the

IGK/*IGL* loci. These data suggest that mutational processes are active at the Ig loci in BL tumors and that large families of tumor-associated *IGH* sequences may be relevant as tumor biomarkers.

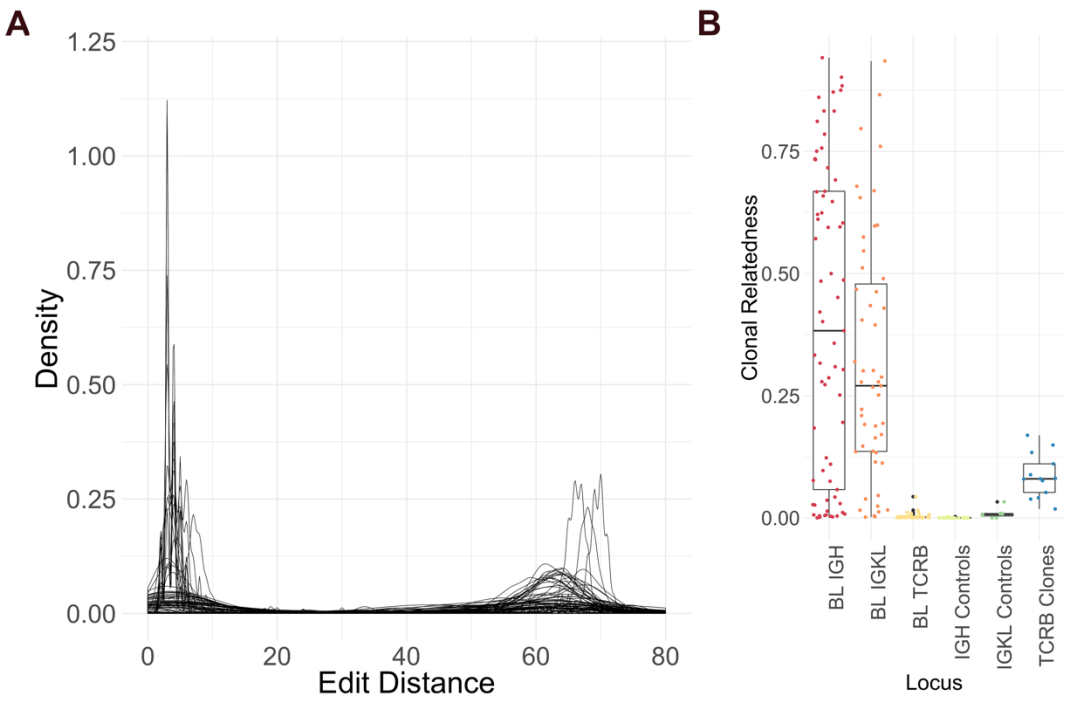


Figure 3.2. HTS of BL tumors reveals high degree of sequence variation.

(A) Density plot of sequence variation in BL tumors based on all *IGH* sequences identified in each tumor. Edit distance measures sequence variation and is defined as the number of nucleotide changes from the most frequent sequence. The plot includes sequencing data from 69 BL tumors, derived from three independent cohorts. (B) Clonal relatedness scores are plotted for each BL tumor at the *IGH* (n=69), *IGK*/*IGL* (n=51), and *TRB* (n=51) loci. Scores are plotted for control PBMC and bone marrow samples at the *IGH* (n=22) and *IGK*/*IGL* (n=8) loci. Scores are plotted for $CD8^+$ T cell clones at the *TRB* locus (n=14).

3.4.5 *The BCR as a biomarker*

3.4.5.1 Detection of circulating tumor-DNA in PBMCs

Due to the unique Ig rearrangements within each clonal B cell population, there is reason to believe that tumor-associated *IGH* sequences may have utility as a biomarker for BL. However, it has yet to be established whether circulating tumor-DNA (ct-DNA) can be detected in the blood of BL patients. To determine if tumor-associated DNA circulates in the peripheral blood of BL patients, gDNA from 15 diagnostic PBMC samples collected from the Ugandan cohort underwent sequencing of the *IGH* locus. (Two tumors were excluded from this analysis because they harbored a polyclonal *IGH* repertoire; and PBMC samples were unavailable for two of the patients with clonal tumors.) Each PBMC sample was probed for the unique nucleotide sequence that dominated the tumor *IGH* repertoire. 8 of 15 PBMC samples were positive for the most frequent *IGH* sequence detected in the tumor. This establishes that ct-DNA is frequently detectable in the blood of BL patients.

We then queried the PBMC samples for the complete family of low edit distance sequences that characterize the malignant clone. An edit distance of <8 nucleotides was used for optimal specificity. In 11 of 15 cases, the tumor-associated family of *IGH* sequences was successfully detected in circulation (Figure 3.3). Only tumor and blood samples from the same patient shared related *IGH* sequences, indicating that no cross-contamination occurred. The complete family of tumor-associated *IGH* sequences comprised 0.001% to 9% of the PBMC *IGH* repertoire.

The number of tumor-associated *IGH* sequences detected in each PBMC sample ranged from 1 to 23 unique nucleotide sequences (median: 1). In addition to the detection of productive VDJ rearrangements in peripheral blood, dominant tumor-associated unproductive VDJ and incomplete DJ rearrangements were also detected as ct-DNA. The median edit

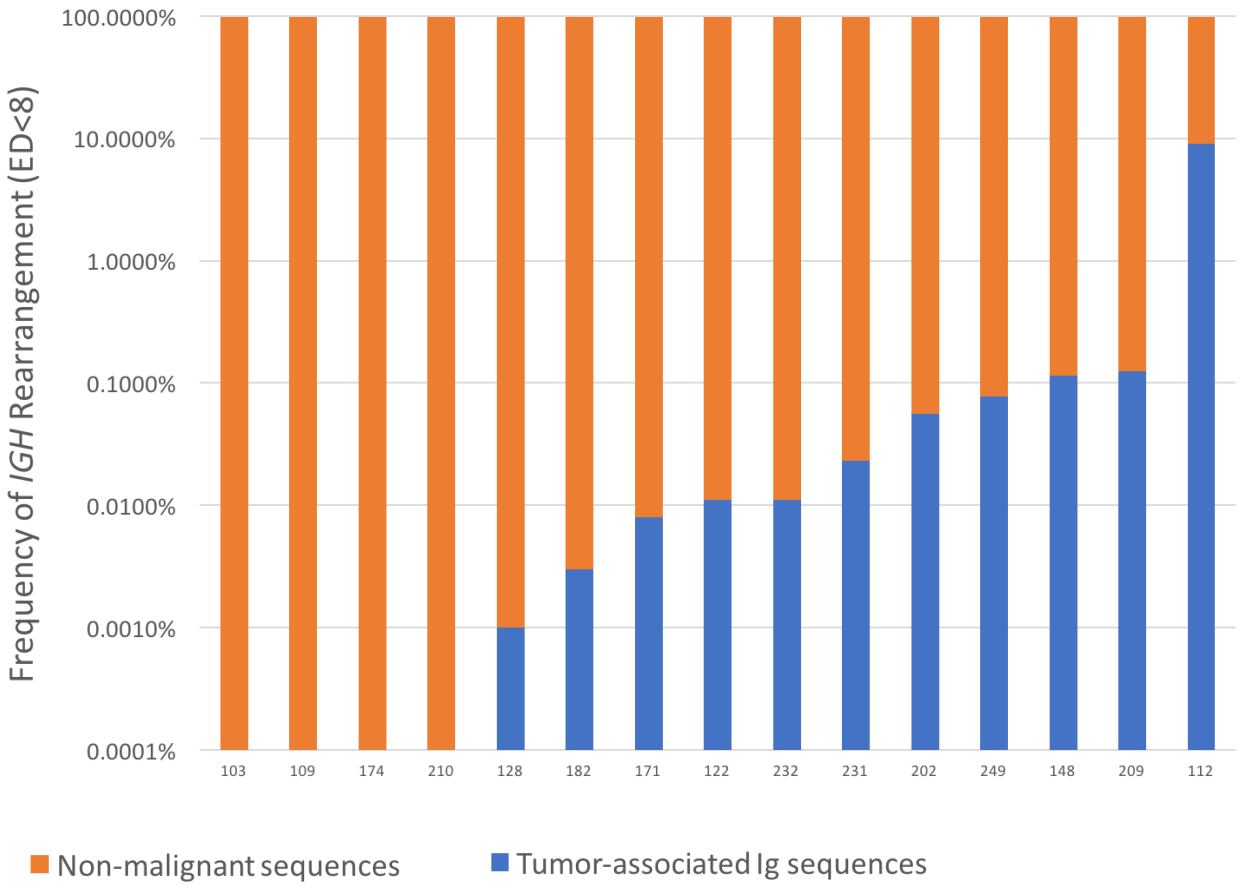


Figure 3.3. Detection of tumor-associated *IGH* rearrangements in the cellular component of blood at diagnosis.

The frequency of tumor-associated *IGH* rearrangements, including all sequences with an edit distance <8 nucleotides from the most frequent sequence identified in the tumor, is listed for each diagnostic PBMC sample from the Ugandan cohort. Frequency is listed on a log scale. All unique sequences detected in each PBMC sample are included in the plot. The putative tumor-associated *IGH* sequences are in blue and non-malignant *IGH* sequences are in orange.

distance of all of the *IGH* sequences detected in the PBMC samples was two nucleotides, demonstrating that sequences other than the dominant clone can be relevant markers of disease. Three PBMC samples harbored only low edit distance ct-DNA sequences, none of which were an exact match to the dominant tumor-associated sequence. Of the 65 total unique *IGH* sequences

from the tumor-associated family of rearrangements detected in all PBMC samples, 35 were not detected in the corresponding tumor sample. These novel, closely related sequence variants are likely just as relevant as the exact sequences identified in the tumor. Ct-DNA analysis that incorporates low edit distance sequence variants will be more likely to detect all tumor-associated sequences. The inclusion of all low edit distance sequences in the malignant population will also allow for detection of tumor evolution and will thus aid in the detection of minimal residual disease and disease relapse.

3.4.5.2 Detection of circulating tumor-DNA in serum

To confirm these results in a second, independent cohort, five patient-matched serum samples from clonal Ghanaian tumors were analyzed for the presence of ct-DNA. Two serum samples contained nucleotide sequences identical to those that dominated their matched tumor *IGH* repertoire. Three of the five samples, including the two above, contained *IGH* sequence variants related to the tumor-associated *IGH* rearrangement, all of which originated from a productive VDJ rearrangement (Figure 3.4). Ct-DNA sequence frequencies ranged from 0.45-54.5% of all sequences detected in the serum. Detection of tumor-associated Ig sequences in the serum demonstrates that tumor DNA is present in the cell-free component of blood in BL patients.

The number of unique, tumor-associated sequence variants detected in each serum sample ranged from 5 to 18 (median: 9) and the median edit distance of these sequences was four nucleotides. Of the 32 total tumor-associated *IGH* sequences identified in the serum samples, only three were detected in the original tumors. One serum sample harbored five unique tumor-associated *IGH* sequences that differed from the most frequent sequence detected in the tumor by one to five nucleotides. None of these exact sequences were detected in the corresponding tumor sample, demonstrating the relevance of circulating tumor-associated sequence variants.

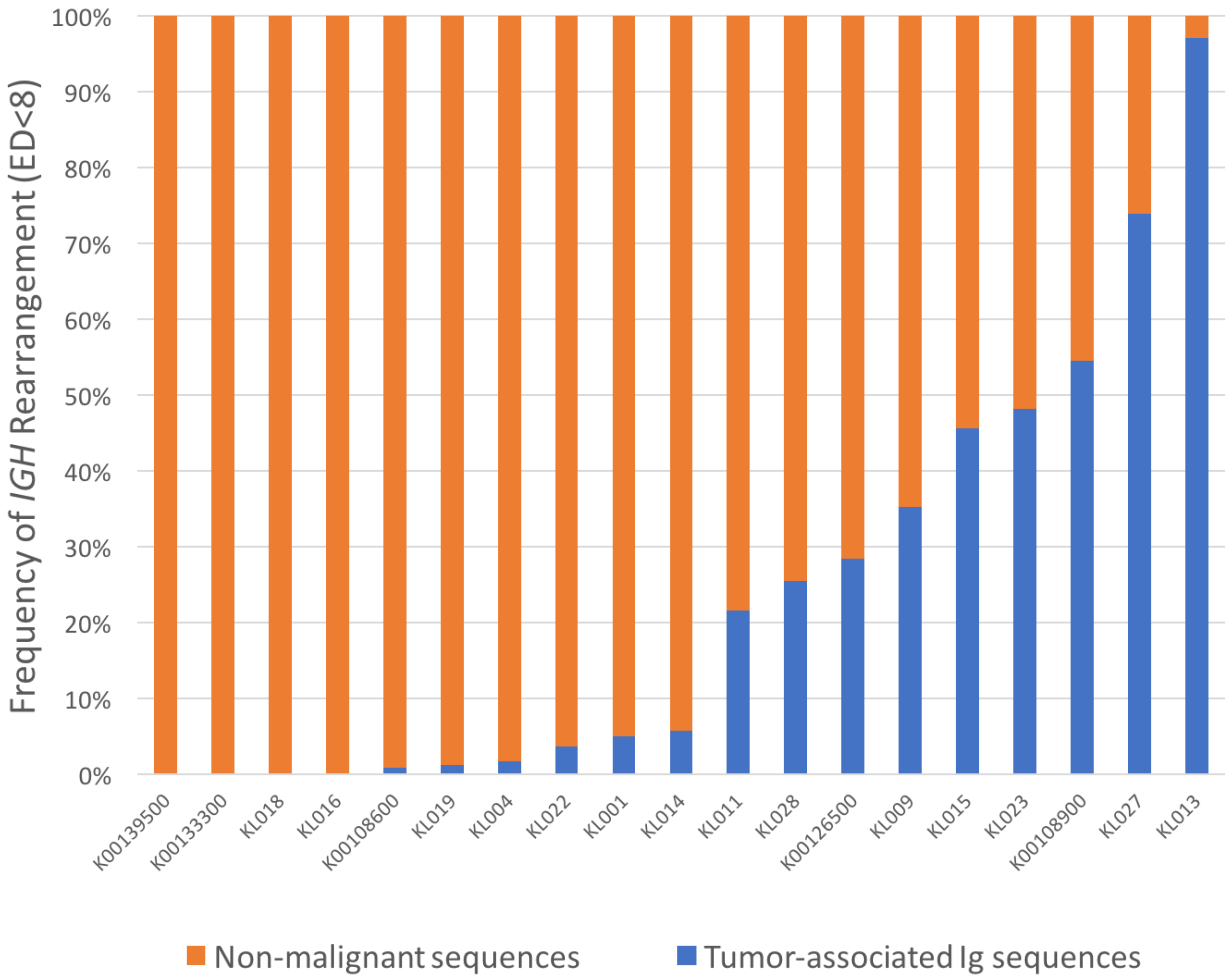


Figure 3.4. Detection of tumor-associated *IGH* rearrangements in the cell-free component of blood at diagnosis.

The frequency of tumor-associated *IGH* rearrangements, including all sequences with an edit distance <8 nucleotides from the most frequent sequence identified in the tumor, is listed for each diagnostic serum and plasma sample from the Ghanaian and Kenyan cohorts, respectively. All unique sequences detected in each serum or plasma sample are included in the plot. The putative tumor-associated *IGH* sequences are in blue and non-malignant *IGH* sequences are in orange.

3.4.5.3 Detection of circulating tumor-DNA in plasma

Finally, patient-matched plasma samples obtained from the Kenyan BL cohort were assessed for the presence of ct-DNA. Of 14 clonal tumor samples, 12 had detectable ct-DNA in the matched plasma sample (Figure 3.4). Tumor-associated *IGH* sequence frequencies ranged from 1.7-97% of all sequences detected in the plasma. The median number of unique sequence variants detected in each plasma sample was 3 (range: 1-70). Sequence families detected in the plasma originated from incomplete DJ rearrangements (n=4), unproductive VDJ rearrangements (n=3), and productive VDJ rearrangements (n=5). The median edit distance of the sequence variants detected in the plasma was two nucleotides. In one plasma sample, a low edit distance sequence variant was detected, but no exact nucleotide matches were found to the dominant tumor-associated *IGH* sequence. Of the 187 total tumor-associated *IGH* sequences identified in all of the plasma samples, 52 sequences were not detected in the matched tumor, representing novel, plasma-specific sequence variants.

Two tumors from the Kenyan cohort were characterized by the same clonal *IGH* sequence. This sequence was present at a frequency of 14% in tumor KL016 and 77% in tumor KL022. The sequence originated from a DJ rearrangement that utilized *IGHD03-22* and *IGHJ05-01*, two relatively uncommonly utilized Ig gene segments, joined by an 18 base pair untemplated nucleotide insertion. In tumor KL022, there were an estimated 16,866 diploid genomes derived from this sequence, but only two estimated genomes in tumor KL016. The two tumors shared nine unique nucleotide sequences, which comprised 22% of the *IGH* repertoire detected in tumor KL016. These two tumors were processed for gDNA on the same day. Due to these factors, it is most likely that sample cross-contamination occurred, rather than the independent generation of these unique sequence rearrangements.

3.4.6 *The second IGH allele as a biomarker*

Eleven of the tumors across the three BL cohorts demonstrated evidence of two rearranged *IGH* alleles. Of those, four had matched PBMC samples and five had matched plasma samples available for analysis of ct-DNA. All nine dual-allelic tumor samples were positive for tumor-associated *IGH* detection in the blood based on the analysis of the most frequent rearrangement. We then probed the blood samples for the secondary, co-dominant *IGH* rearrangement to determine if that sequence could be utilized as a tumor-specific indicator as well. In seven tumors, the secondary *IGH* rearrangement was specifically detected in the patient-matched blood sample. Most secondary tumor sequences (6/7) originated from VDJ rearrangements and only one originated from a DJ rearrangement. The number of unique, low edit distance sequences detected in the blood that originated from the secondary *IGH* rearrangement ranged from 1 to 77 (median: 8). The remaining two tumors contained a secondary DJ rearrangement that was detected in blood samples from other patients. This demonstrates that in many cases, both *IGH* alleles may be relevant markers of disease.

3.4.7 *Detection of low-frequency sequences*

One tumor from the Ghanaian BL cohort (patient K00138500) contained a dominant DJ rearrangement that comprised 12% of the *IGH* repertoire. This rearrangement was just below the 15% threshold that was used to define a clonal tumor, so it was excluded from our initial biomarker analysis. However, it was noted that this tumor-associated rearrangement was also uniquely detected in the patient-matched serum sample. Nine unique tumor-associated *IGH* sequences were detected that comprised 48% of the serum *IGH* repertoire. Patient-matched CSF samples were also obtained from nine of the Ghanaian patients. Most CSF samples were not found to contain

detectable tumor-associated *IGH* sequences. However, the same DJ sequence family detected in the serum of patient K00138500 was also uniquely detected in that patient's CSF sample. Three unique, tumor-associated sequences were detected in the CSF, one of which matched the dominant tumor-associated rearrangement exactly and two that each differed by two nucleotides. These three sequences comprised 46% of the *IGH* repertoire in the CSF. Although this clonal B cell population was present at a lower frequency, it may nonetheless have important implications for disease monitoring in various tissue compartments.

3.4.8 *Patient survival as a function of ct-DNA detection*

Ziegler Staging, the most commonly used method for staging BL in Africa, was first described in 1974.[134] Disease stage is scored on a spectrum from A to D. Stage A signifies early, localized disease and stage D indicates advanced, disseminated disease. Detailed clinical information was available for the Ugandan patient cohort, so patient survival was assessed. Ziegler disease stage did not appear to correlate with overall patient survival in this BL patient population (Figure 3.5A).

Patient survival was then assessed based on the detection of ct-DNA at diagnosis. A strong inverse correlation was observed between tumor-associated DNA detection and clinical outcome. When ct-DNA was undetectable, patient survival was 75% (n=4), but when tumor-associated sequences were detected in the blood at diagnosis, patient survival fell to just 9% (n=11) (P = 0.0489) (Figure 3.5B). These data suggest that detection of the tumor-associated molecular signature in the blood of BL patients is a strong prognostic indicator. Importantly, five patients diagnosed with Ziegler stage A disease had detectable ct-DNA. Though they were diagnosed and treated for early stage disease, the detection of tumor DNA in circulation suggests they likely have

a worse prognosis. Our findings demonstrate that unique tumor-associated *IGH* sequences may serve as a more effective staging method for BL than the current approach.

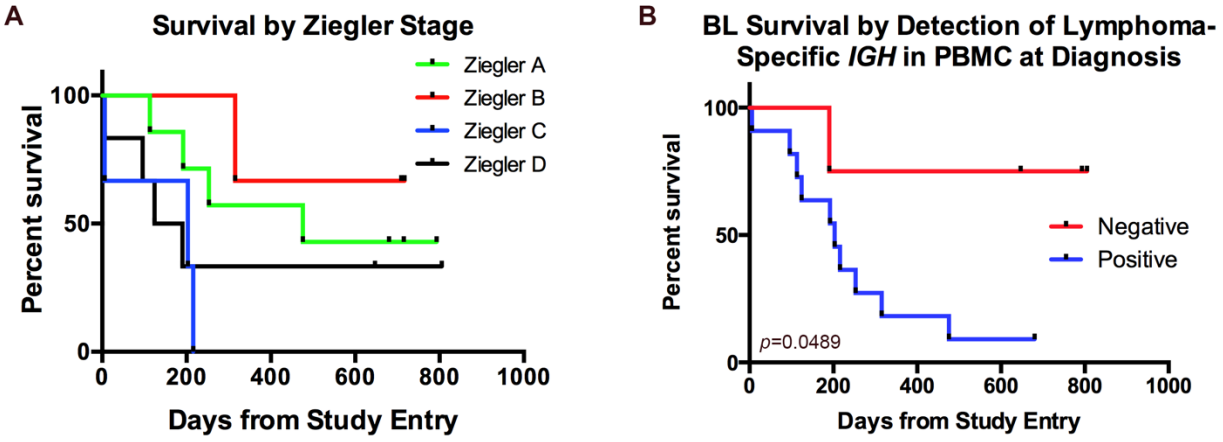


Figure 3.5. Detection of tumor-specific *IGH* rearrangements in the blood at diagnosis as a prognostic indicator.

Kaplan-Meier survival curves are shown for (A) the Ugandan BL cohort by patient Ziegler disease stage (n=19) and (B) by detection of ct-DNA at diagnosis (n=15).

3.5 DISCUSSION

HTS of the *IGH* locus in an independent BL cohort of 18 pediatric patients in western Kenya demonstrated that molecular characteristics of BL tumors are consistent across geographically distinct BL cohorts. Many of the novel molecular patterns discovered in the Ugandan and Ghanaian BL tumor cohorts[15] were also observed in this Kenyan cohort. This includes the detection of only one *IGH* allele in most tumors, a clonal *IGH* rearrangement in most, but not all BL tumors, and a high degree of *IGH* sequence variation in both VDJ- and DJ-dominant tumors.

The lack of two detectable, rearranged *IGH* alleles in most BL tumors confirms earlier reports of monoallelic germline *IGH* rearrangements in BL tumors.[15,117] Data compiled from all three cohorts demonstrates that only one dominant *IGH* allele was detected in 44 of 55 (80%) clonal BL tumors. Virtually all BL tumor cells harbor a t(*c-MYC*;Ig) translocation; in 80% of tumors this translocation involves the *IGH* locus and in the remaining 20% it involves either the *IGK* or *IGL* locus. This study provides additional support to the hypothesis that the *c-MYC* translocation may occur early in B cell development. The existence of the *c-MYC* translocation on one *IGH* allele may interfere with normal D and J gene segment rearrangement and thus preclude its detection in a large proportion of BL tumors.

Most tumors in the Kenyan cohort (14/18) harbored a clonal *IGH* rearrangement. Four of the tumors contained a polyclonal *IGH* repertoire comprised of many, low-frequency sequences. The absence of a clonal rearrangement may be biologically real, or may be a sampling or technical artifact. It is possible that a small number of tumors either lost the distal segments of the long arm of chromosome 14, where *IGH* is encoded, or did not carry a rearranged *IGH* region. It is also possible that a small number of tumors were misdiagnosed as BL, or that the DNA that was sequenced was actually derived from normal adjacent tissue. It is also possible that the sequencing primers were unable to bind and amplify the *IGH* region due to extensive SHM or novel *IGHV* genes utilized within the tumor.

The presence of ct-DNA in the peripheral blood of most patients at diagnosis suggests that the BCR might have utility as a biomarker for BL. Given the significant limitations on disease detection in many resource-limited settings, the ability to assess the presence of disease by a simple blood test would be extremely valuable. Serum LDH levels are used as an indicator of gross tumor burden, but LDH is a general marker of tissue damage and is associated with a variety of

conditions. Tumor-associated *IGH* sequences may prove more useful as a disease-specific indicator of tumor burden.

In addition to the productively arranged VDJ allele, our study demonstrates that non-functional *IGH* rearrangements, including unproductive VDJ and incomplete DJ rearrangements, can function as biomarkers as well. This finding highlights the importance of using gDNA as opposed to RNA for the assessment of ct-DNA. Furthermore, in the fraction of tumors with two detectable *IGH* alleles, both unique rearrangements can frequently be used as a marker for the presence of disease.

This study demonstrates that the clonal *IGH* sequence that characterizes the tumor can be uniquely detected in the matched patient blood sample. In addition to the exact *IGH* sequence, low edit distance sequence variants were also detected in circulation. In fact, five blood samples harbored only variants of the dominant tumor-associated sequence, and two of those contained variants that were not detected in the tumors themselves. The high degree of *IGH* sequence diversity in BL tumors provides important information for the assessment of ct-DNA. Molecule-specific approaches, such as PCR, may not capture all relevant tumor-associated sequences. A broad, unbiased approach for sequence detection will be required in order to capture all possible tumor-associated sequences. Importantly, these unbiased approaches will still detect disease even after tumor evolution, allowing for assessment of treatment efficacy over time.

Metastatic progression of BL to the central nervous system (CNS) is associated with a very poor clinical outcome and is essentially incurable in many settings.[135] A sensitive molecular method to detect disease in this compartment would have great clinical value. The use of *IGH* sequences as a biomarker for BL progression to the CNS would allow for assessment of disease progression and support appropriate therapeutic intervention.

The detection of the same low-frequency BL-associated *IGH* sequence family in patient-matched serum and CSF samples has important implications for ct-DNA detection. Primarily, it demonstrates the utility of the BCR as a tumor-specific biomarker in both blood and CSF. It also suggests that the 15% threshold used to define tumor clonality may be too high and that lower frequency sequences are likely biologically relevant. These low frequency sequences may also play an important role in disease monitoring.

In the Ugandan cohort, patient survival was inversely correlated with tumor *IGH* detection in the blood at the time of diagnosis, demonstrating the role of the BCR as a prognostic indicator. Five of six patients diagnosed with early stage disease (Ziegler stage A) were found to have detectable ct-DNA. With the current diagnostic tools, early stage patients are believed to have a limited tumor burden and are treated accordingly. Our data suggest that even patients with early stage disease may require more aggressive treatment to realize better outcomes. Detection of ct-DNA in the blood of BL patients may thus allow physicians to more appropriately gauge therapy on an individual basis and potentially improve overall patient survival.

Chapter 4. DISCUSSION

4.1 SUMMARY OF RESEARCH

This body of research includes experiments and insights that are novel to the study of BL. The use of HTS to analyze gDNA extracted from primary tumors allowed us to assess the complete repertoire of *IGH* rearrangements in BL tumors. Most of the tumors (55/69) from the three independent BL cohorts contained a clonal *IGH* rearrangement. However, a substantial fraction of BL tumors (20%) were characterized by a polyclonal *IGH* repertoire. Furthermore, a large proportion of clonal tumors did not appear to contain a productively rearranged *IGH* allele. Sixteen of 55 clonal tumors harbored only a dominant DJ or an unproductive VDJ *IGH* rearrangement. Together, these findings demonstrate that a substantial fraction of BL cases (30/69; 43%) do not appear to express a clonal, functional BCR. Though contrary to the BL dogma, our analysis of publicly available RNAseq data[23,78] supported this finding and demonstrated a small subset of broadly polyclonal BL tumors in two additional cohorts. If validated in additional study populations, this novel finding may revise the model of BL pathogenesis in a large proportion of tumors.

The finding that 14 of 69 BL tumors harbored a polyclonal *IGH* repertoire was unexpected. The method of diagnosis and the corresponding rigor of each approach varied for the different cohorts in this study. The most rigorous diagnostic approach was utilized for the Ugandan cohort. Tumor biopsies were initially diagnosed in Uganda and were each confirmed by an experienced hematopathologist in the United States. Immunohistochemistry was performed to confirm each diagnosis. Only two of 19 (10%) of the Ugandan tumors were polyclonal at the *IGH* locus. Both

of these contained clonal light chain rearrangements and carried BL-associated mutations, suggesting that they were accurately diagnosed as BL. For the archival Ghanaian cohort, diagnoses were made based on clinical presentation and cytological appearance, the best methods available at the time. Eight of 32 (25%) tumors in the Ghanaian cohort were polyclonal at the *IGH* locus and five were polyclonal at the *IGK*/*IGL* loci. For the Kenyan cohort, each tumor was diagnosed by two independent pathologists based on tumor morphology. Four of 18 (22%) of the Kenyan tumors were polyclonal at the *IGH* locus. (Sequencing of the *IGK*/*IGL* loci was not performed on the Kenyan tumors due to limited sample availability.) The trend towards less rigorous diagnostic methods and an increased number of polyclonal samples amongst the cohorts suggests that a small fraction of the cases studied may have been misdiagnosed. The actual percentage of polyclonal BL tumors may be slightly lower than that represented in this study.

In the majority of clonal tumors (44/55; 80%), only one *IGH* allele was detected by HTS. Due to the organized chromosomal rearrangements at the Ig loci early in B cell development, two rearranged *IGH* alleles should be detectable by HTS in any mature B cell. Because BL tumor cells are characterized by a recurrent chromosomal translocation involving the *IGH* locus in 80% of tumors, we hypothesize that the absence of a detectable allele in the same proportion of tumors suggests that the pathognomonic translocation occurs before the ordered rearrangement of both *IGH* alleles. The finding of monoallelic *IGH* rearrangements in BL tumors has previously been reported in the literature,[117] but would have been missed in previous studies using Sanger sequencing or tumor RNA.

The current consensus in the BL literature is that the translocation occurs in the germinal center due to aberrant AID activity.[118-121,136,137] Most of the support for this theory is derived from the association between AID and translocations involving the Ig and *c-MYC* loci. AID sequence

motifs are reportedly enriched at *IGH*/*c-MYC* junctions in BL tumors.[120] A proportion of the breakpoints occur in *IGH* switch regions, where dsDNA breaks regularly occur due to AID-mediated class-switch recombination in the germinal center. In mouse models, AID expression is closely associated with t(*c-MYC*;*IGH*) translocations, in a dose-dependent manner.[121,137] However, AID expression may not be confined to the germinal center; it was recently reported that AID is expressed in a pre-B cell population in the bone marrow. Furthermore, *P. falciparum* gametocytes develop in erythrocytes in the bone marrow,[138] where they produce hemozoin, a by-product of hemoglobin metabolism. Hemozoin has been demonstrated to induce AID expression via activation of TLR9.[42,43] In light of these data, our experiments challenge the notion that the translocation always occurs in the germinal center and suggests that translocations could be mediated by AID in the bone marrow.

Another explanation for the detection of only one rearranged *IGH* allele is that the other allele was lost due to genomic instability within the tumor. Several experimental approaches were utilized to detect evidence of both *IGH* alleles in BL tumors. First, PCR was performed on tumor gDNA to amplify intronic sequences between the D and J gene segments. Second, quantitative PCR was utilized to evaluate the presence of both alleles at various positions along the *IGH* locus. Third, PCR was performed to amplify a microsatellite region located at the telomeric end of the *IGH* locus on both alleles of chromosome 14. And finally, droplet digital PCR was utilized to quantify the number of copies of *IGH* at various positions along the locus. In all of these experiments, the presence of non-malignant cells in the tumor, which harbored germline *IGH* loci, confounded our results. Sequencing of the T cell receptor β locus demonstrated a clonally diverse T cell infiltrate in the BL tumors studied. The presence of these, and other non-malignant cells, in the tumor prevented the detection of a tumor-specific signal. In general, the assessment of bulk

tumor material will be blurred by the heterogeneous nature of the tissue. Moving forward, experiments utilizing single-cell or sorted tumor populations will be needed to address this question definitively.

The utilization of HTS to assess the Ig rearrangements in BL tumors allowed for a more comprehensive analysis of the Ig repertoire than has previously been performed. In addition to detection of the clonal Ig rearrangement, the full spectrum of Ig sequences was captured, providing a more accurate snapshot of both the normal and malignant B cells contained in BL tumors. This analysis revealed that BL tumors are characterized by large families of Ig rearrangements. The family members are all related to the dominant clonal rearrangement, but differ at 1-10 nucleotide positions. The identification of these populations broadens the malignant population identified in BL tumors and demonstrates that hyperactive mutational processes are targeted to the Ig loci in tumors.

Analysis of *IGHV* gene usage in the clonal *IGH* rearrangement detected in each tumor demonstrated that BL tumors exhibit biased usage of particular *IGHV* gene segments, as compared to normal B cell populations. This finding has been demonstrated in a number of mature B cell malignancies and suggests that antigenic stimulation may play a role in the pathogenesis of BL. The enrichment of particular groups of naïve B cells, utilizing specific *IGHV* gene segments, in the tumor population suggests that antigen plays a contributing role in the multi-step development of BL.

The detection of tumor-associated *IGH* sequences in the blood and CSF of BL patients and its association with patient survival has important implications for BL disease monitoring. This analysis demonstrates that ct-DNA is regularly detected in the blood of BL patients at diagnosis. Importantly, both VDJ and DJ rearrangements were uniquely detected in circulation, suggesting

that both productive and unproductive *IGH* rearrangements can be used as disease biomarkers. Furthermore, tumor-associated *IGH* sequence variants were also detected in circulation, demonstrating that the entire repertoire of tumor-associated *IGH* rearrangements is important to consider for biomarker detection.

The presence of tumor-associated *IGH* sequences in gDNA extracted from PBMCs was associated with inferior survival, suggesting that tumor DNA detection may be an indicator of overall tumor burden. This analysis was performed on blood cells, suggesting that the tumor DNA was derived from circulating tumor cells. This could suggest disease dissemination. The parallel studies that were performed on the cell-free component of blood, either serum or plasma, suggest that tumor-associated, cell-free DNA is also present in circulation. This likely derives from apoptotic or necrotic tumor cells. Future studies are planned to determine if the detection of cell-free ct-DNA is indicative of tumor burden or dissemination and is associated with patient survival as well.

4.2 FUTURE DIRECTIONS

An increasing body of evidence suggests that signaling through the BCR may be important for driving the pathogenesis of BL. Analysis of gDNA revealed that 43% of tumors (30/69) do not express a clonal, functional BCR. If BCR signaling is essential for cell survival and proliferation, an analysis of the mutational landscape of tumors harboring a functional versus nonfunctional BCR would be informative. Our RNAseq analysis was performed on a small number of tumors; an expanded study designed for the purpose of detecting sequence variants may be beneficial. Perhaps tumors with non-functional BCRs are enriched for mutations that amplify signaling

through the BCR signaling cascade. Additionally, if tumor cells do depend on BCR signaling, the introduction of signaling inhibitors as a therapeutic intervention may be effective.

The biased utilization of *IGHV* gene segments in BL tumors suggests that antigenic-stimulation contributes to the development of BL. If true, the tumor-associated BCR would specifically recognize that particular antigen, providing novel insights into the emergence of BL. To determine the antigenic specificity of tumor-associated BCRs, we have generated soluble, recombinant BCRs derived from complete BL tumor heavy and light chain Ig sequences (Figure 4.1A). The heavy and light chains are properly linked by disulfide bonds that are dissociated under reducing conditions (Figures 4.1B, C). Furthermore, the BCRs contain the appropriate post-translational modifications; they are properly glycosylated, as illustrated by a shift in size after incubation with enzymes that cleave N-linked glycoproteins (Figure 4.1D). We plan to use these proteins to assess the binding specificity of tumor-associated BCRs. We will use protein microarrays tiled with epitopes derived from EBV, *P. falciparum*, and autoantigens to evaluate BCR recognition. The discovery of a particular antigen contributing to the development of BL would provide important information in the effort to prevent BL.

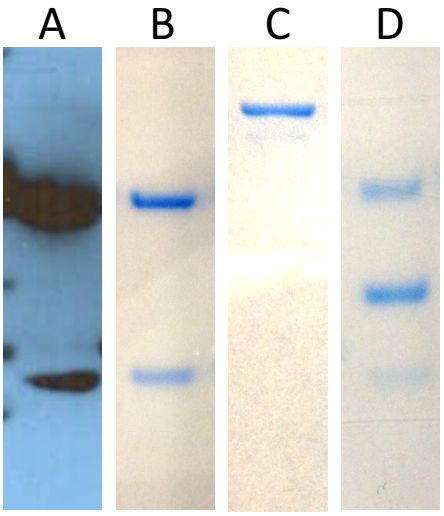


Figure 4.1. Characterization of a soluble, recombinant BL tumor-associated BCR. All panels include characterization of a BCR derived from patient 009-0103 from the Ugandan BL cohort as imaged by denaturing SDS-PAGE. (A) A western blot of a BL BCR probed with IgG1- and IgK-specific primary antibodies. (B) An image of a reducing, coomassie-stained gel. (C) An image of a non-reducing, coomassie-stained gel. (D) An image of a coomassie-stained gel after BCR incubation with PNGase F, an enzyme that cleaves N-linked glycans.

The data presented in the third chapter of this dissertation demonstrate that the detection of tumor DNA at diagnosis is associated with poor patient survival. This finding suggests that assessment of patient blood for tumor DNA may have prognostic value in the clinic. A simple blood test could identify patients with more aggressive disease who could benefit from more intensive therapy. This finding introduces many follow-up questions, including: can the detection of tumor-associated *IGH* rearrangements be used to assess the efficacy of treatment, if detected at the completion of therapy? Can BL patients' blood be monitored for the presence of tumor-associated sequences to predict disease relapse? If the t(*c-MYC*; Ig) translocation is an initiating

event in disease development and tumor DNA is present in circulation, can the translocation be detected in the blood? And if so, could detection of the translocation in African children be used as a screening method to identify high-risk individuals who might be more likely to develop BL? We plan to begin addressing these questions with hundreds of newly acquired BL patient tumor, PBMC, plasma, and serum samples from the Uganda Cancer Institute. The collection of high-quality, well-annotated clinical specimens are invaluable to the study of BL and to the ability to ultimately improve treatment for BL patients.

4.3 PROPOSED MODEL OF BL PATHOGENESIS

The data compiled in this dissertation support a multi-step model of BL pathogenesis (Figure 4.2). A combination of genetic and environmental factors are likely required for malignant transformation. The t(*c-MYC*; Ig) translocation may occur in the bone marrow, early in B cell development. This is the initiating step to oncogenesis. However, the t(8;14) translocation has been detected in the blood of healthy individuals,[50] suggesting that additional genetic hits are required for full transformation to BL.

Once the B cell has matured and left the bone marrow, it will enter peripheral circulation and acquire EBV infection. This will lead to a transient increase in host cell proliferation, as EBV has growth transforming ability and can inhibit *c-MYC*-induced apoptosis. This will expand the pool of EBV-infected B cells harboring a translocation. With the appropriate T cell responses, EBV infection is properly controlled and the expanded population of B cells will shrink. However, recurrent *P. falciparum* malaria infection induces broad T cell immunosuppression, allowing EBV-infected B cells to proliferate unchecked. *P. falciparum* can also potentiate EBV reactivation,

leading to more infectious EBV virions in the microenvironment. The CIDR1 domain of the malarial PfEMP1 protein can non-specifically bind the BCR, stimulating broad B cell hyperplasia. This creates an environment where B cells harboring EBV infection and carrying a chromosomal translocation can rapidly accumulate.

EBV-positive, translocation-positive B cells will then encounter antigen in a secondary lymphoid organ. This will stimulate another round of cell proliferation, again leading to an enlarged pool of primed B cells. Chronic antigenic stimulation likely leads to the BCR stereotypy observed in BL tumors. The particular stimulating antigen may be EBV, *P. falciparum*, or another pathogen or autoantigen that has yet to be associated with BL.

Upon antigenic stimulation, the germinal center reaction is initiated. The B cell will undergo recurrent rounds of affinity maturation as it recycles from dark zone to light zone. AID-mediated point mutations are induced in Ig loci. Numerous non-Ig genes have also been reported to be targets of AID and to contain AID-induced hypermutation.[139-142] Chronic antigenic stimulation likely increases the chance of deleterious mutations across the genome. These off-target mutations in various oncogenes likely contribute to malignant transformation.

In addition to the chromosomal translocation, further deregulation of *c-MYC* is required for transformation. In endemic tumors, point mutations are frequently found in the first exon of *c-MYC* and may be attributed to off-target AID activity in the germinal center. *c-MYC* is one of the most heavily mutated genes in BL tumors. The N-terminus of *c-MYC* contains important regulatory elements that control *c-MYC* transcription and proteasome-mediated mRNA degradation.[51] The disruption of this region in BL leads to increased *c-MYC* expression and mRNA stability, facilitating high *c-MYC* levels and promoting tumor cell proliferation.[52,53] Fully

dysregulated *c-MYC* expression likely occurs after EBV infection to overcome *c-MYC*-induced apoptosis.

Recurrent mutations in genes commonly mutated in a number of cancers have also been detected in BL tumors. In addition to *c-MYC*, other commonly mutated genes include *TP53*, *DDX3X*, *TCF3*, *SMARCA4*, *ID3*, and *GNA13*.[13,15,23,62,63,78,86] This array of mutations likely serve to amplify the BCR signaling cascade and other contributing pathways. The accumulation of these mutations and subsequent dysregulation of these genes likely occurs in the germinal center and contributes to BL pathogenesis.

In our model of BL pathogenesis, the accumulation of somatic mutations in the germinal center provides the catalyst for full malignant transformation. Upon acquisition of growth-transforming mutations, EBV will enter viral latency and restrict its gene expression in an effort to avoid detection by the immune system. The malignant B cell will exhibit extremely high *c-MYC* expression, rapid cell proliferation, and a block to terminal differentiation. Therefore, the transformed cell will retain a germinal center phenotype as it proliferates unchecked.

It is well established that normal B cells require tonic signaling through the BCR for survival and development. This proposed model of BL pathogenesis relies on the generation of a functional BCR in the bone marrow to drive cell survival, differentiation, and maturation. However, in 14 tumors, an incomplete DJ rearrangement was the only clonal *IGH* rearrangement detected. Unproductively rearranged Ig alleles are believed to be transcriptionally silenced. However, high *c-MYC* expression in BL tumors demonstrates that the translocated allele is able to overcome this negative regulation. *c-MYC* demonstrates strong growth-promoting effects and it is possible that even moderate levels of these signals, early in B cell development, could provide sufficient survival signals to bypass the requirement for BCR signaling. *c-MYC* overexpression

has been demonstrated to alter B cell tolerance by promoting the survival of autoreactive B cells.[143] In this way, a B cell harboring an incomplete DJ rearrangement on one allele and a t(*c-MYC*;*IGH*) translocation on the other allele may be able to leave the bone marrow and continue to develop. This model does not encompass every BL tumor, but it serves to expand the possible paths to pathogenesis for a substantial portion of BL cases. Additional studies are needed to address these possibilities.

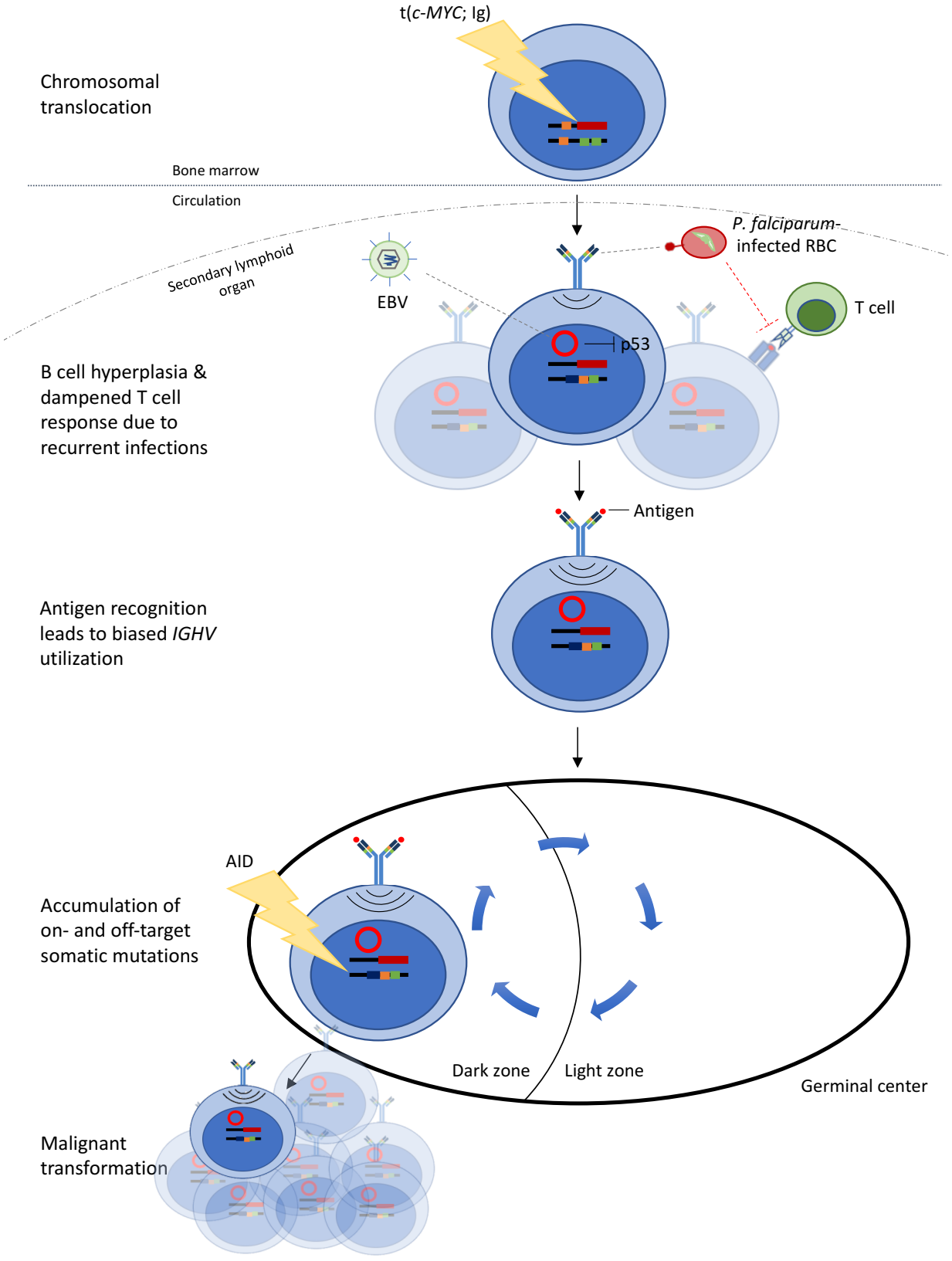


Figure 4.2. Proposed model of BL pathogenesis.

4.4 CONCLUDING REMARKS

I think that this dissertation introduces more questions about BL than it answers. It challenges much of the dogma about BL tumors themselves and about how pathogenesis occurs. It suggests that the family of *IGH* rearrangements that characterize the tumor may be used as a biomarker with clinical utility, but raises questions about the circumstances under which it may be utilized and implemented. The research on BL in the African setting is limited. I hope that this dissertation can provide a roadmap for navigating an improved understanding of the etiology of BL. Only through improved knowledge of the biological origin of BL can we better treat, and ultimately prevent, this malignancy.

APPENDIX A: SUPPLEMENTAL METHODS FOR CHAPTER 2

HIV Status: HIV serostatus of the children in the Ugandan cohort was evaluated at Makerere University, by the Walter Reed Project, using a Bio-Rad Genetic Systems HIV-1/HIV-2 PLUS O Enzyme Immunoassay. Children who were HIV seropositive on initial testing underwent repeat assessment, and those with second positive tests underwent confirmatory testing by western blot using the Bio-Rad Genetic Systems HIV-1 Western Blot Kit. HIV serostatus of the children in the archival BL cohort from Ghana was not tested.

gDNA Isolation: Small pieces of each BL tumor were digested to completion in buffer containing Proteinase K for 2-4 hours at 56°C. The DNeasy Blood and Tissue Kit (QIAGEN) was used to extract gDNA from the digested tumor samples, according to the manufacturer's instructions.

RNA Isolation: Tumor samples were dissociated by pipetting and homogenized through a QIAshredder column (QIAGEN). The RNeasy Plus Mini Kit (QIAGEN) was used to extract total RNA from each sample, according to the manufacturer's instructions. First-strand complementary DNA was generated from total tumor RNA using the Transcriptor First Strand cDNA Synthesis Kit (Roche). RNA quality was evaluated with an Agilent 2200 TapeStation.

EBV Status: The EBV status of each tumor was determined by PCR and Reverse Transcription-PCR (RT-PCR) using EBV gene product-specific primers. Previously published[144] *EBNA1*-specific primers were used, with a modification in the reverse primer to increase specificity for EBV strains common to Africa (5'-CAGACAATGGACTCCCTTAGTG-3'). The PCR products

were amplified as follows: 94°C for 5 minutes; 94°C for 30 seconds, 54°C (*EBER*) or 51°C (*EBNA1*) for 45 seconds, and 72°C for 1 minute, for 40 cycles; 72°C for 5 minutes.

Sequencing of Complete Ig Variable Regions in Tumor-associated BCRs: Analysis of HTS data from BL tumors enabled identification of the 3' portion of the presumptive *IGH* and *IGK*/*IGL* rearrangements carried in BL tumor cells. PCR was used to determine the sequence of the 5' portion of tumor-associated Ig rearrangements, with V family-specific sense primers complementary to the leader peptide sequences of the inferred *IGH*[145] or *IGK*/*IGL*[146] V gene segment and antisense primers complementary to the CDR3 region of the inferred tumor-associated Ig rearrangements. The amplification programs used were: 94°C for 5 minutes; 94°C for 30 seconds, 50-58°C for 45 seconds, and 72°C for 1 minute, for 40 cycles; 72°C for 5 minutes. PCR products of the predicted size were cloned and sequenced by capillary methods; at least 5 independent *IGH* and *IGK*/*IGL* clones were sequenced for each tumor.

Ig Expression Analysis and Isotype: RT-PCR was used to determine whether the predicted tumor-associated *IGH* and *IGK*/*IGL* chains were expressed in tumor cells, and to identify the heavy chain isotype of tumor-associated BCRs. For this analysis, sense primers specific for the CDR3 of the most frequent *IGH* or *IGK*/*IGL* rearrangement and antisense primers complementary to the *IGH* [147] and *IGK*/*IGL* [146] constant regions were used to amplify an interval spanning the 3' portion of the Ig variable region and the 5' portion of the Ig constant region from first-strand BL tumor cDNA. The amplification programs used were: 94°C for 5 minutes; 94°C for 30 seconds, 51-57°C for 45 seconds, and 72°C for 1 minute, for 40 cycles; 72°C for 5 minutes.

Ig Gene Segment Utilization and Analysis of Somatic Hypermutation (SHM): Initial identification of the gene segments utilized in Ig rearrangements carried in BL tumor cells was performed at Adaptive Biotechnologies, and was independently confirmed using IgBLAST[105]

(http://www.ncbi.nlm.nih.gov/igblast/) and IMGT V-QUEST[102,103] (http://www.imgt.org/IMGT_vquest/share/textes/). These tools were also used to identify putative sites of SHM in tumor-associated Ig sequences. Identification of the *IGHV* gene segments used in BL tumor cells was also performed in an independent BL tumor cohort with 28 sporadic[78] BL cases for which RNA sequencing (RNAseq) data is publicly available in the NCBI Sequence Read Archive (SRA048058). For this analysis, *IGH* CDR3 sequences were extracted from RNAseq fastq files using MiXCR v1.7 set to default parameters.[148]

RNAseq: Reads were aligned to the human reference genome (Hg19) using the Spliced Transcripts Alignment to a Reference (STAR) aligner v2.5, 2-pass method.[149] Reads that did not align to the human genome were aligned separately to the EBV genome, AG876 strain (NC_009334.1) using STAR. Gene annotation files used for EBV alignment were provided by the Erik Flemington Laboratory at Tulane University (http://flemingtonlab.com/rnaseq.html). EBV positive tumors were defined as those with greater than 100 mapped EBV reads per million human reads. Read counts were generated for each gene using the Python package HTSeq v0.5.4 with default settings. Differential expression of normalized gene counts was performed using DESeq2 Bioconductor package.[150] Variants were called by the Broad's Genome Analysis Tool Kit[106] according to the most up-to-date best practices workflow for RNAseq (version 2015-12-07).

APPENDIX B: SUPPLEMENTAL DATA FOR CHAPTER 2

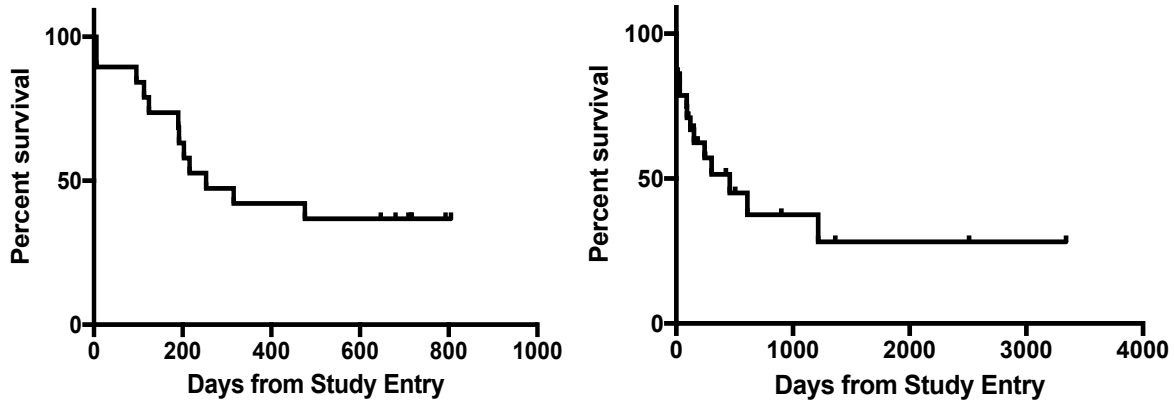


Figure B1. Survival of BL patients from Ugandan and Ghanaian cohorts is poor, and has not improved significantly over the past 40 years. Kaplan-Meier survival curves are shown for (A) the Ugandan BL cohort (n=19) and for (B) the Ghanaian BL cohort (n=29). Survival was plotted by days from study entry. The median overall survival was 253 days in the Ugandan cohort and 458 days in the Ghanaian cohort. Patients in the Ugandan cohort were followed from 2013-2014 and patients in the Ghanaian cohort were followed from 1975-1992.

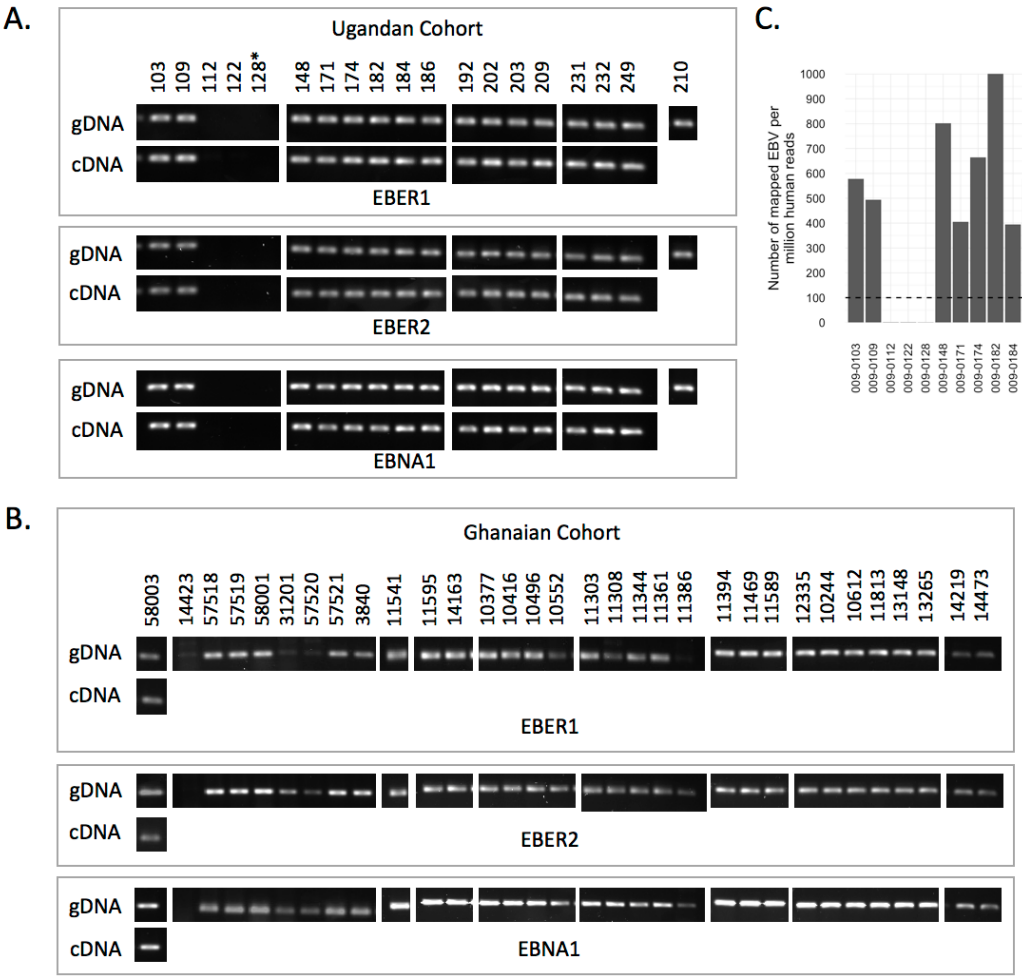


Figure B2. Most BL tumor samples from the Ugandan and Ghanaian cohorts are EBV-positive by PCR analysis of EBER1, EBER2, and EBNA1 and by RNAseq. PCR was performed to amplify DNA from the EBV gene products EBER1, EBER2, and EBNA1 on all BL tumor samples and the resulting agarose gels are shown. RT-PCR was only performed on those samples with RIN > 5. (A) RT-PCR and PCR analyses on tumors from the Ugandan cohort for each of the three EBV gene products (n=19). An asterisk indicates that the patient was HIV-positive. (B) RT-PCR and PCR analyses on tumors from the Ghanaian cohort for each of the three EBV gene products (n=32). 47/51 BL samples were EBV positive. (C) Number of mapped EBV reads per million human reads by RNAseq (n=18). EBV-positive tumors were defined as those with greater than 100 mapped EBV reads per million human reads. RNAseq was not performed on sample 009-0210 due to poor quality RNA.

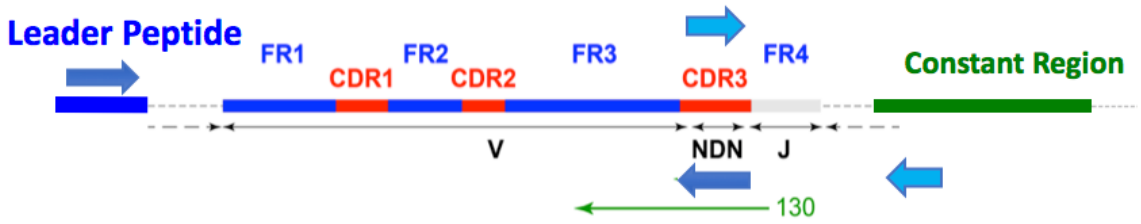


Figure B3. PCR and RT-PCR experimental strategies for *IGHV* sequencing and expression analysis. Schematic diagram of an idealized *IGH* rearrangement demonstrates the strategy used to sequence *IGH* rearrangements carried in BL tumor cells. HTS reads of 130 nucleotides (thin green arrow) encompassed the 3' portion of Framework Region 3 (FR3) and the entirety of Complementarity-Determining Region 3 (CDR3). Thick dark blue arrows indicate the binding sites of the leader peptide- and CDR3-specific primers used for PCR amplification of the 5' portion of the *IGH* variable region, including FR1, CDR1, FR2, CDR2, and FR3. Thick light blue arrows indicate the binding sites of CDR3- and constant region-specific primers used for RT-PCR to evaluate Ig expression and isotype.

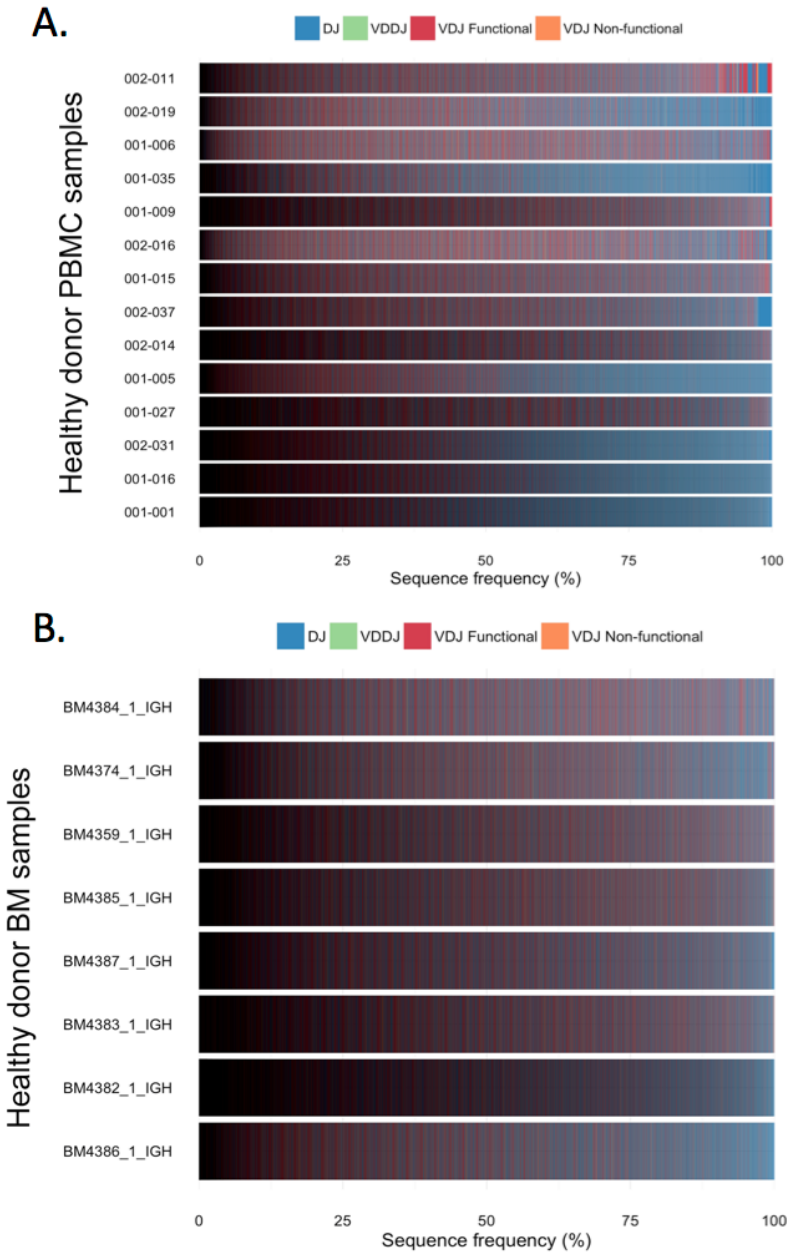


Figure B4. HTS on gDNA from healthy donor bone marrow and PBMC samples demonstrates a polyclonal *IGH* repertoire. (A-B) Cumulative frequency plots of all unique *IGH* sequences identified by HTS on gDNA from healthy donor PBMC samples (A) and healthy donor bone marrow samples (B). Each panel in the bar plots represents a unique nucleotide sequence and the color indicates the type of *IGH* rearrangement. Black lines separate the unique sequences, so increasingly polyclonal regions appear black in the figure.

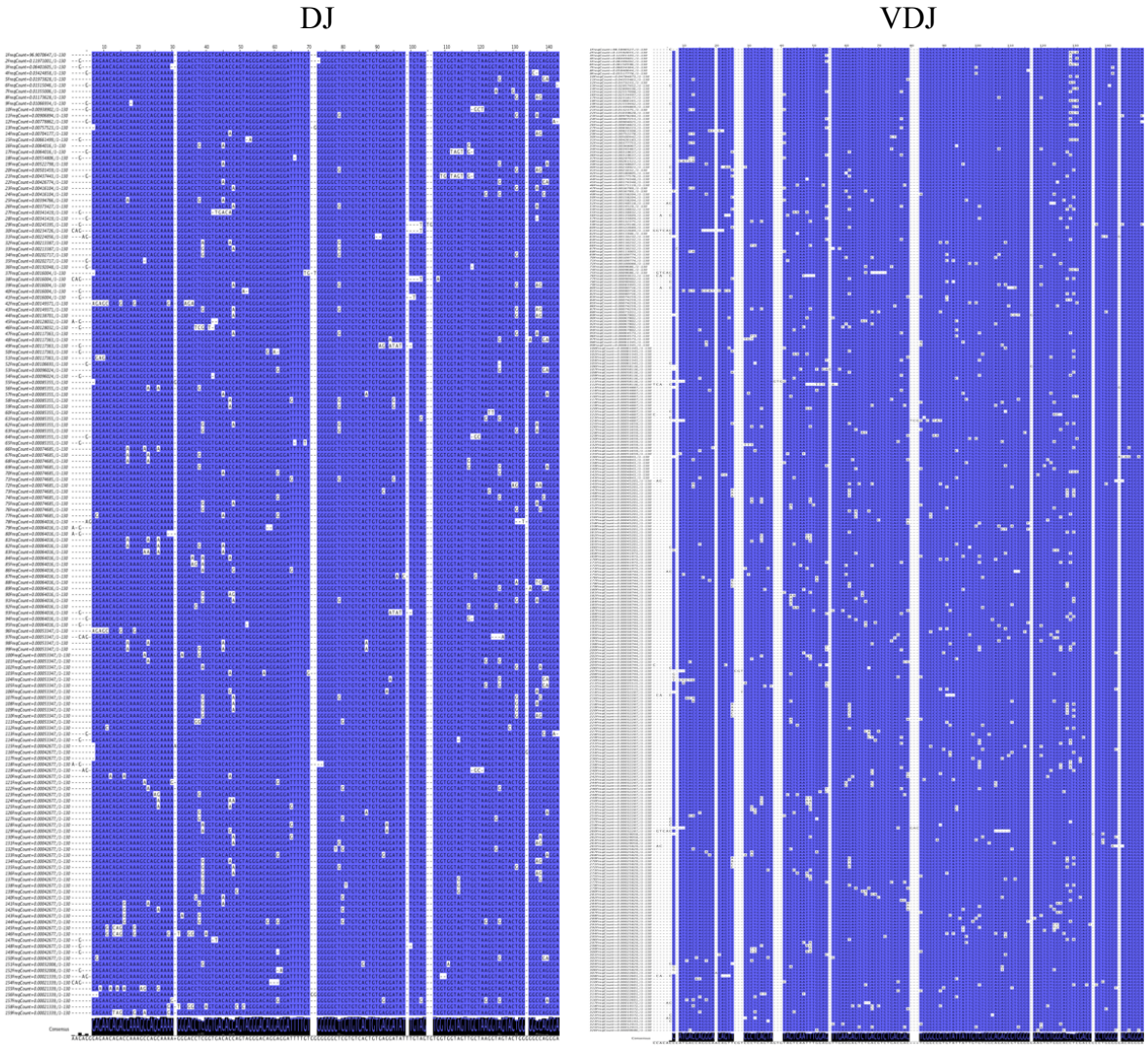


Figure B5. Sequence variation is evenly distributed along *IGH* rearrangements for tumors with dominant DJ and VDJ rearrangements. Sequence alignments are shown for a representative DJ (patient K00095400) and a VDJ (patient K00093400) tumor. All sequences shown are within an edit distance of 10 nucleotides from the most frequent sequence. Point mutations and nucleotide insertions/deletions are indicated in white and sequence alignment is indicated in blue.

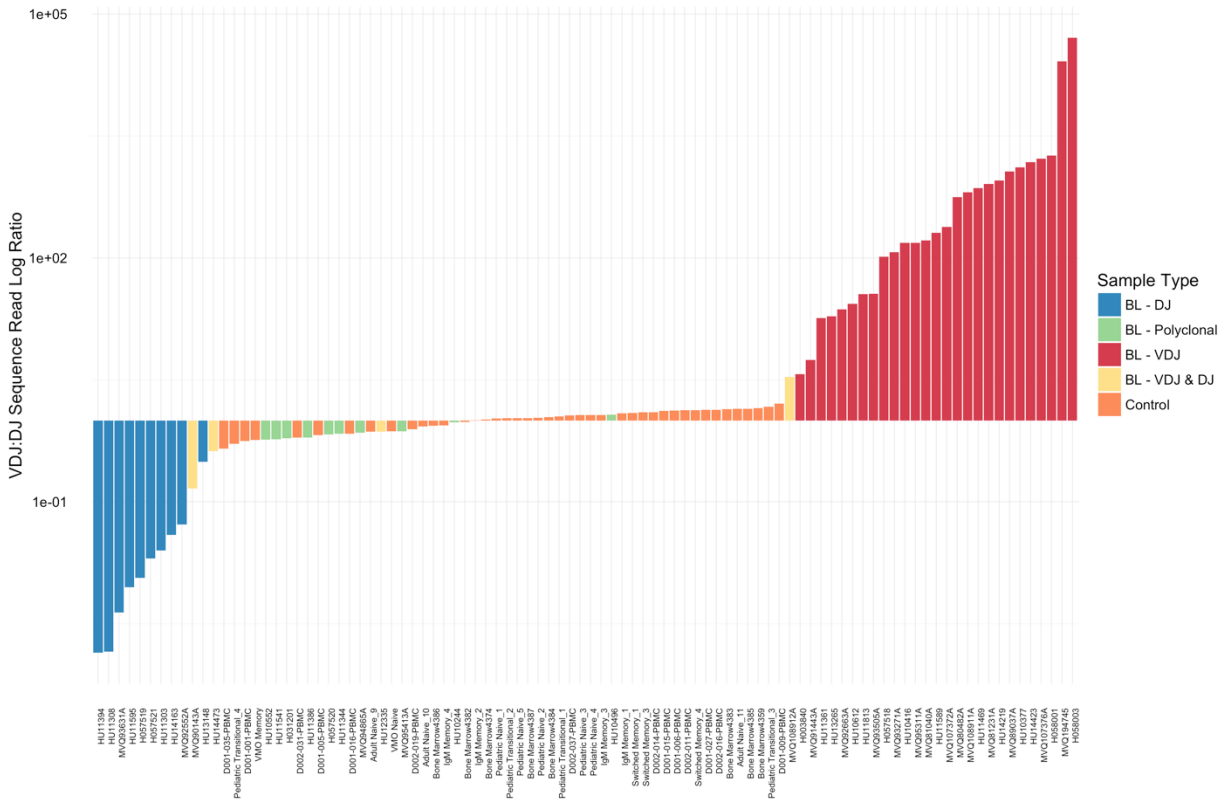


Figure B6. BL tumors characterized by one dominant *IGH* rearrangement. The ratio of total VDJ or DJ sequence reads are shown for each tumor and controls. Most clonal tumors are characterized by either VDJ or DJ *IGH* sequence reads that likely represent the malignant population. Tumors with a polyclonal *IGH* repertoire and control B cell populations have VDJ to DJ sequence ratios around 1.

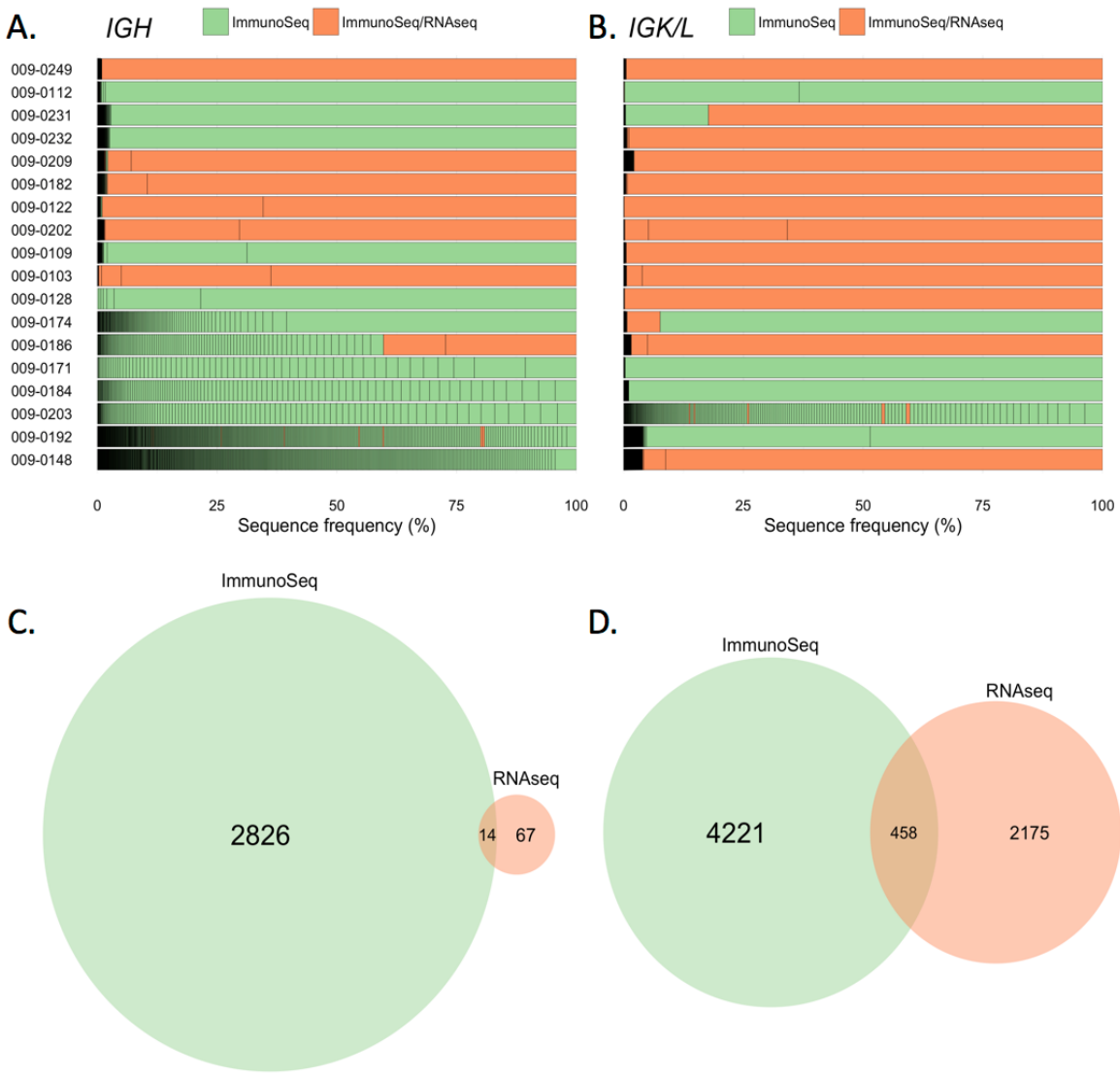
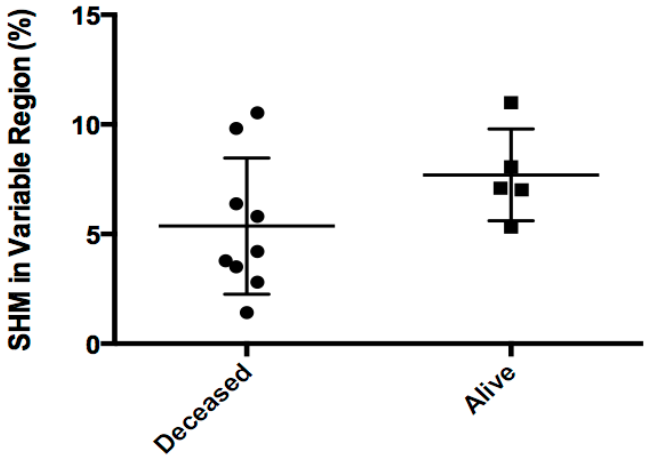


Figure B7. High frequency *IGH* and *IGK/IGL* sequences identified by HTS of gDNA are frequently detected by MiXCR analysis of RNAseq data. The frequency of each unique, *productive* sequence identified by HTS of tumor gDNA is illustrated for the *IGH* locus (A) and the *IGK/IGL* loci (B). Sequences that were only identified by HTS of gDNA are indicated in green and sequences that were identified both by HTS of gDNA and by RNAseq are indicated in orange. (C-D) Unique, productive sequences identified by ImmunoSeq, RNAseq, or both are indicated for *IGH* (C) and *IGK/IGL* (D).

IGH Expression	
Isotype	**BL Sample Number (N=15)**
IgM^+ IgD^+	12
IgG^+	2
No Expression	1

IGκ/λ Expression	
Expression	**BL Sample Number (N=16)**
$IGκ^+$	12
$IGλ^+$	3
No Expression	1

Figure B8. Most clonal BL tumors in Ugandan cohort express IgM^+IgD^+ and IGK^+ BCRs. *IGH* and *IGK/IGL* expression and isotype for clonal tumors are described.



Supplemental Figure S9. Higher levels of SHM associated with slightly improved clinical outcome in BL. The percentage of SHM present in the dominant *IGH* variable region is plotted based on clinical outcome in the Ugandan cohort (n=14). This difference was not statistically significant.

APPENDIX C: SUPPLEMENTAL TABLES FOR CHAPTER 2

Table C1. BL tumor high-throughput sequencing read, clonality and clonal relatedness data

Sample ID	IGH VDJ Reads	IGH DJ Reads	Total IGH Reads	IGK/IGL VJ Reads	IGH Clonality	IGK/IGL Clonality	IGH Clonal Relatedness	IGK/IGL Clonal Relatedness	TRB Clonal Relatedness
MVQ80482A	1,174,087	2107	1,176,194	3,433,066	0.77	0.97	0.658823529	0.866055046	0.015873016
MVQ81040A	2,041,567	12446	2,054,013	5,146,610	0.87	0.99	0.421393841	0.429230769	0.002742732
MVQ81231A	3,653,478	4488	3,657,966	22,580,837	0.98	0.89	0.621037464	0.209554832	0.002568493
MVQ89037A	4,485,673	3860	4,489,533	18,605,319	0.87	1	0.286743516	0.760299625	0.002228826
MVQ90143A	1,694	11579	13,273	7,492,926	0.66	1	0.006097561	0.511820331	0.000489237
MVQ91443A	155,187	27961	183,148	5,205,184	0.06	0.91	0.109919571	0.133679055	0.001773399
MVQ92552A	3,042	57752	60,794	2,886,937	0.1	0.99	0.272727273	0.655483871	0.003997716
MVQ92663A	15,830	678	16,508	6,675,933	0.52	0.96	0.733333333	0.267772512	0.000970874
MVQ93271A	177,026	1502	178,528	2,455,649	0.9	0.99	0.5	0.251582278	0.010869565
MVQ93505A	143,344	3969	147,313	5,520,456	0.08	0.98	0.383027523	0.278409091	0.000594177
MVQ93631A	4,143	951354	955,497	2,311,763	0.25	0.95	0.756828194	0.193654267	0.000637146
MVQ94865A	21,857	30798	52,655	2,558,577	0.06	0.85	0.005012531	0.038729198	0.000180002
MVQ95311A	562,423	3665	566,088	3,324,710	0.87	0.86	0.647368421	0.163833076	0.00203666
MVQ95413A	9,293	12636	21,929	1,246,313	0.08	0.09	0.003636364	0.222222222	0.006779661
MVQ107372A	659,998	2736	662,734	534,332	0.94	0.96	0.734969325	0.251386322	0.000651466
MVQ107376A	867,871	522	868,393	585,041	0.86	0.96	0.941763727	0.796721311	0.002478315
MVQ108911A	1,531,909	2387	1,534,296	4,134,693	0.96	0.92	0.594509804	0.288228155	0.003947368
MVQ108912A	629,507	183322	812,829	1,738,413	0.96	0.98	0.691622103	0.599243856	0.001451379
MVQ194745A	3,124,392	119	3,124,511	464,254	0.98	0.99	0.81152648	0.678756477	0.000895656
HU14423	1,073,890	714	1,074,604	8,199,248	1	0.98	0.624242424	0.170565302	0.00268524
H057518	590,367	5702	596,069	3,662,217	0.88	0.92	0.668449198	0.574768519	0.008050089
H057519	15,869	1371304.5	1,387,173	2,335,848	0.97	0.98	0.901869159	0.489726027	0.00121102
H058001	2,955,394	1629	2,957,023	2,666,436	0.79	1	0.875216638	0.546448087	0.001422688
H031201	259,203	429969	689,172	423,429	0.27	0.24	0.00617856	0.016289894	8.60E-05
H057520	18,780	27791	46,571	295,014	0.16	0.46	0.007894737	0.045914397	0.001292643
H057521	27,076	1350162	1,377,238	1,473,192	0.86	0.89	0.484536082	0.319767442	0.00132626
H003840	1,048,606	283539	1,332,145	858,086	0.62	0.67	0.004320786	0.012414244	0.000157183
H058003	2,614,747	51	2,614,798	5,919,146	0.99	0.87	0.733333333	0.278079009	0.000795545
HU11541	708,402	1204190	1,912,592	1,032,253	0.04	0.1	0.001467501	0.02405271	0.000710508
HU11595	8,292	928973	937,265	5,792,264	0.03	0.97	0.076699029	0.114443932	0.00128123
HU14163	6,074	155307	161,381	757,785	0.28	0.83	0.451219512	0.669527897	0.001113586
HU10377	5,831,418	4498	5,835,916	4,460,275	0.74	0.83	0.785215606	0.301610542	0.00108313
HU10416	3,903,815	25458	3,929,273	9,622,617	0.93	0.92	0.33315508	0.112225938	0.001294498
HU10496	62,922	53235	116,157	6,115,072	0.25	0.98	0.309409888	0.191304348	0.001232202
HU10552	915,698	1581187	2,496,885	4,994,294	0.03	0.1	9.42E-05	0.002433977	0.001116773
HU11303	77,583	3094295	3,171,878	1,655,026	0.94	0.92	0.860773481	0.597402597	0.002417405
HU11308	5,604	3906237	3,911,841	4,855,607	0.28	0.7	0.871462264	0.135752688	0.000700771
HU11344	304,596	443623	748,219	689,073	0.04	0.11	0.002197363	0.001571421	0.001533978
HU11361	300,290	16453	316,743	2,763,809	0.75	0.95	0.042672656	0.301081389	0.001272669
HU11386	2,243	3630	5,873	14,010	0.06	0.11	0.003875969	0.004291845	0.011389522
HU11394	3,038	2194392	2,197,430	122,025	0.26	0.98	0.750491159	0.188284519	0.009189641
HU11469	3,092,988	4273	3,097,261	7,886,910	0.98	0.81	0.667355372	0.43454039	0.002155689
HU11589	1,764,466	8683	1,773,149	4,005,064	0.87	0.93	0.59563543	0.270471464	0.002097169
HU12335	1,144,037	1568856	2,712,893	6,495,462	0.81	0.95	0.278810409	0.39483675	0.001208581
HU10244	6,663	6987	13,650	188,902	0.11	0.94	0.010989011	0.467741935	0.002212389
HU10612	267,826	9829	277,655	228,926	0.97	0.96	0.183908046	0.146892655	0.001607028
HU11813	1,425,281	39824	1,465,105	489,161	0.78	0.21	0.611180905	0.015790418	0.000178221
HU13148	5,117	16536	21,653	731,451	0.36	0.27	0.057971014	0.462809917	0.001126761
HU13265	81,982	4295	86,277	329,541	0.93	0.88	0.603550296	0.93499044	0.001484414
HU14219	2,259,407	2524	2,261,931	350,401	0.99	0.99	0.571428571	0.404907975	0.001104972
HU14473	57,461	136129	193,590	153,996	0.43	0.88	0.036166365	0.136363636	0.043478261

Table C2. Recurrent, single nucleotide variants detected in BL tumors that are predicted to be deleterious and are present in the Cosmic v77 database (Hg19)

SampleID	GeneSymbol	Chr	CytoBand	Start	End	Ref	Alt	MetaSVMscore	snp138	Cosmic77
009-0103	RRM2	chr2	2p25.1	10263566	10263566	A	C	0.955		COSM5045669
009-0103	ACTR10	chr14	14q23.1	58666970	58666970	A	G	1.095	rs76669080	COSM4593637
009-0103	SMARCA4	chr19	19p13.2	11143994	11143994	G	A	1.025		COSM1266237
009-0109	RRM2	chr2	2p25.1	10263566	10263566	A	C	0.955		COSM5045669
009-0109	TP53	chr17	17p13.1	7577120	7577120	C	T	0.939	rs28934576	COSM1645335
009-0112	ID3	chr1	1p36.12	23885728	23885728	G	A	1.094		COSM1159767
009-0122	ID3	chr1	1p36.12	23885728	23885728	G	A	1.094		COSM1159767
009-0122	NOTCH1	chr9	9q34.3	139413097	139413097	T	G	0.972	rs200520088	COSM4163567
009-0122	TP53	chr17	17p13.1	7577124	7577124	C	T	0.953		COSM99950
009-0122	SMARCA4	chr19	19p13.2	11143993	11143993	C	T	0.996		COSM2813735
009-0128	RRM2	chr2	2p25.1	10263566	10263566	A	C	0.955		COSM5045669
009-0128	POMC	chr2	2p23.3	25384477	25384477	T	C	0.17	rs200380417	COSM4141002
009-0128	COL6A3	chr2	2q37.3	238261167	238261167	G	A	0.901	rs116690555	COSM1406570
009-0128	TFAP4	chr16	16p13.3	4312613	4312613	C	T	0.968		COSM241880
009-0128	TP53	chr17	17p13.1	7577121	7577121	G	C	0.964		COSM3719992
009-0128	SMARCA4	chr19	19p13.2	11143993	11143993	C	T	0.996		COSM2813735
009-0148	TP53	chr17	17p13.1	7579356	7579356	G	T	1.097		COSM43790
009-0171	NOTCH1	chr9	9q34.3	139413097	139413097	T	G	0.972	rs200520088	COSM4163567
009-0171	TBC1D10B	chr16	16p11.2	30369645	30369645	C	T	0.014	rs188813224	COSM3387373
009-0174	SEPSECS	chr4	4p15.2	25161954	25161954	A	C	0.354	rs201339389	COSM4591819
009-0174	TFAP4	chr16	16p13.3	4312620	4312620	G	A	1.002		COSM1289282
009-0174	TBC1D10B	chr16	16p11.2	30369645	30369645	C	T	0.014	rs188813224	COSM3387373
009-0182	ID3	chr1	1p36.12	23885728	23885728	G	A	1.094		COSM1159767
009-0182	FOXO1	chr13	13q14.11	41240285	41240285	G	C	1.095		COSM4805886
009-0182	ZNF587	chr19	19q13.43	58370766	58370766	G	A	0.208	rs77577775	COSM3363049
009-0184	POMC	chr2	2p23.3	25384477	25384477	T	C	0.17	rs200380417	COSM4141002
009-0184	BMPR2	chr2	2q33.2	203419995	203419995	G	A	0.36	rs201440272	COSM1404595
009-0184	PAX5	chr9	9p13.2	37006494	37006494	C	T	0.21	rs115889954	COSM303899
009-0184	ABCD4	chr14	14q24.3	74764652	74764652	G	A	0.893	rs145141432	COSM195954
009-0184	TP53	chr17	17p13.1	7577581	7577581	A	T	0.961		COSM220768
009-0186	ID3	chr1	1p36.12	23885751	23885751	G	A	1.054		COSM1159769
009-0186	BMPR2	chr2	2q33.2	203419995	203419995	G	A	0.36	rs201440272	COSM1404595
009-0186	NOTCH1	chr9	9q34.3	139413097	139413097	T	G	0.972	rs200520088	COSM4163567
009-0186	ABCD4	chr14	14q24.3	74764652	74764652	G	A	0.893	rs145141432	COSM195954
009-0186	TP53	chr17	17p13.1	7578406	7578406	C	T	0.92	rs28934578	COSM99023
009-0192	COL6A3	chr2	2q37.3	238280477	238280477	G	A	0.24	rs73998894	COSM1406581
009-0192	BSCL2	chr11	11q12.3	62459867	62459867	C	T	0.875	rs190842600	COSM1580620
009-0202	ATAD3C	chr1	1p36.33	1391672	1391672	G	A	0.695	rs77225021	COSM3930324
009-0202	ID3	chr1	1p36.12	23885728	23885728	G	A	1.094		COSM1159767
009-0202	RRM2	chr2	2p25.1	10263566	10263566	A	C	0.955		COSM5045669
009-0202	POMC	chr2	2p23.3	25384477	25384477	T	C	0.17	rs200380417	COSM4141002
009-0202	FOXO1	chr13	13q14.11	41240279	41240279	G	A	1.076		COSM220647
009-0202	ACTR10	chr14	14q23.1	58667012	58667012	A	G	0.631	rs140960545	COSM238565
009-0203	POMC	chr2	2p23.3	25384477	25384477	T	C	0.17	rs200380417	COSM4141002
009-0203	PAX5	chr9	9p13.2	37006494	37006494	C	T	0.21	rs115889954	COSM303899
009-0203	TP53	chr17	17p13.1	7577566	7577566	T	C	1.014		COSM1717141
009-0209	ATAD3C	chr1	1p36.33	1391672	1391672	G	A	0.695	rs77225021	COSM3930324
009-0209	HK3	chr5	5q35.2	176317899	176317899	G	A	0.559	rs113978418	COSM1543168
009-0209	PAX5	chr9	9p13.2	37006494	37006494	C	T	0.21	rs115889954	COSM303899
009-0209	BSCL2	chr11	11q12.3	62459867	62459867	C	T	0.875	rs190842600	COSM1580620
009-0231	ATAD3C	chr1	1p36.33	1391672	1391672	G	A	0.695	rs77225021	COSM3930324
009-0231	POMC	chr2	2p23.3	25384477	25384477	T	C	0.17	rs200380417	COSM4141002
009-0231	SEPSECS	chr4	4p15.2	25161954	25161954	A	C	0.354	rs201339389	COSM4591819
009-0231	HK3	chr5	5q35.2	176314610	176314610	C	T	0.752	rs61741552	COSM737627
009-0231	NOTCH1	chr9	9q34.3	139413097	139413097	T	G	0.972	rs200520088	COSM4163567
009-0231	FOXO1	chr13	13q14.11	41240349	41240349	T	C	0.003		COSM220645
009-0231	ZNF587	chr19	19q13.43	58370766	58370766	G	A	0.208	rs77577775	COSM3363049
009-0232	RRM2	chr2	2p25.1	10263566	10263566	A	C	0.955		COSM5045669
009-0232	POMC	chr2	2p23.3	25384477	25384477	T	C	0.17	rs200380417	COSM4141002
009-0232	COL6A3	chr2	2q37.3	238303364	238303364	G	A	0.741		COSM5008252
009-0249	ATAD3C	chr1	1p36.33	1391672	1391672	G	A	0.695	rs77225021	COSM3930324
009-0249	ACTR10	chr14	14q23.1	58666970	58666970	A	G	1.095	rs76669080	COSM4593637

APPENDIX D: SUPPLEMENTAL TABLE FOR CHAPTER 3

Table D1. BL tumor high-throughput sequencing read, clonality and clonal relatedness data

Sample ID	IGH VDJ Reads	IGH DJ Reads	Total IGH Reads	IGH Clonality	IGH Clonal Relatedness
KL001_Tumor	543,882	196,473	740,355	0.91	0.0969
KL004_Tumor	6,143	892	7,035	0.45	0.3167
KL006_Tumor	10,961	5,443	16,404	0.22	0.0294
KL009_Tumor	4,659,740	16,264	4,676,004	0.99	0.3036
KL011_Tumor	7,440,790	960	7,441,750	0.97	0.3575
KL012_Tumor	25,061	9,422	34,483	0.16	0.0272
KL013_Tumor	727,989	6,492,028	7,220,017	0.91	0.1956
KL014_Tumor	912,651	27,412	940,063	0.91	0.2514
KL015_Tumor	3,349,888	3,736,487	7,086,375	0.71	0.1230
KL016_Tumor	932	695	1,627	0.22	0.0750
KL018_Tumor	1,624	277,678	279,302	0.17	0.8842
KL019_Tumor	3,564,585	19,773	3,584,358	0.89	0.8327
KL020_Tumor	136,880	84,329	221,209	0.07	0.0139
KL022_Tumor	894	890,385	891,279	0.12	0.8326
KL023_Tumor	1,127,437	3,708,340	4,835,777	0.88	0.4018
KL024_Tumor	1,310	706	2,016	0.14	0.0217
KL025_Tumor	28,792	31,175	59,967	0.13	0.0265
KL027_Tumor	2,502,663	1,198	2,503,861	0.18	0.7166
KL028_Tumor	902,918	609	903,527	0.73	0.4868

REFERENCES

1. Chene A, Donati D, Orem J, et al. Endemic Burkitt's lymphoma as a polymicrobial disease: new insights on the interaction between Plasmodium falciparum and Epstein-Barr virus. *Semin Cancer Biol.* 2009;19(6):411-420.
2. Ferry JA. Burkitt's lymphoma: clinicopathologic features and differential diagnosis. *Oncologist.* 2006;11(4):375-383.
3. Epstein MA, Achong BG, Barr YM. VIRUS PARTICLES IN CULTURED LYMPHOBLASTS FROM BURKITT'S LYMPHOMA. *Lancet.* 1964;1(7335):702-703.
4. Geser A, Brubaker G, Draper CC. Effect of a malaria suppression program on the incidence of African Burkitt's lymphoma. *Am J Epidemiol.* 1989;129(4):740-752.
5. Schmitz R, Ceribelli M, Pittaluga S, Wright G, Staudt LM. Oncogenic mechanisms in Burkitt lymphoma. *Cold Spring Harb Perspect Med.* 2014;4(2).
6. Patte C, Auperin A, Michon J, et al. The Societe Francaise d'Oncologie Pediatrique LMB89 protocol: highly effective multiagent chemotherapy tailored to the tumor burden and initial response in 561 unselected children with B-cell lymphomas and L3 leukemia. *Blood.* 2001;97(11):3370-3379.
7. Todeschini G, Bonifacio M, Tecchio C, et al. Intensive short-term chemotherapy regimen induces high remission rate (over 90%) and event-free survival both in children and adult patients with advanced sporadic Burkitt lymphoma/leukemia. *Am J Hematol.* 2012;87(1):22-25.
8. Parkin DM. The global health burden of infection-associated cancers in the year 2002. *Int J Cancer.* 2006;118(12):3030-3044.
9. de Martel C, Ferlay J, Franceschi S, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol.* 2012;13(6):607-615.
10. Khan G, Miyashita EM, Yang B, Babcock GJ, Thorley-Lawson DA. Is EBV persistence in vivo a model for B cell homeostasis? *Immunity.* 1996;5(2):173-179.
11. Bellan C, Lazzi S, Hummel M, et al. Immunoglobulin gene analysis reveals 2 distinct cells of origin for EBV-positive and EBV-negative Burkitt lymphomas. *Blood.* 2005;106(3):1031-1036.
12. Navari M, Etebari M, De Falco G, et al. The presence of Epstein-Barr virus significantly impacts the transcriptional profile in immunodeficiency-associated Burkitt lymphoma. *Front Microbiol.* 2015;6:556.
13. Amato T, Abate F, Piccaluga P, et al. Clonality Analysis of Immunoglobulin Gene Rearrangement by Next-Generation Sequencing in Endemic Burkitt Lymphoma Suggests Antigen Drive Activation of BCR as Opposed to Sporadic Burkitt Lymphoma. *Am J Clin Pathol.* 2016;145(1):116-127.
14. Kaymaz Y, Oduor CI, Yu H, et al. Comprehensive Transcriptome and Mutational Profiling of Endemic Burkitt Lymphoma Reveals EBV Type-specific Differences. *Mol Cancer Res.* 2017.

15. Lombardo KA, Coffey DG, Morales AJ, et al. High-throughput sequencing of the B-cell receptor in African Burkitt lymphoma reveals clues to pathogenesis. *Blood Advances.* 2017;1(9):535-544.
16. de-The G, Day NE, Geser A, et al. Sero-epidemiology of the Epstein-Barr virus: preliminary analysis of an international study - a review. *IARC Sci Publ.* 1975(11 Pt 2):3-16.
17. Piriou E, Asito AS, Sumba PO, et al. Early age at time of primary Epstein-Barr virus infection results in poorly controlled viral infection in infants from Western Kenya: clues to the etiology of endemic Burkitt lymphoma. *J Infect Dis.* 2012;205(6):906-913.
18. Balfour HH, Jr., Sifakis F, Sliman JA, Knight JA, Schmeling DO, Thomas W. Age-specific prevalence of Epstein-Barr virus infection among individuals aged 6-19 years in the United States and factors affecting its acquisition. *J Infect Dis.* 2013;208(8):1286-1293.
19. Humme S, Reisbach G, Feederle R, et al. The EBV nuclear antigen 1 (EBNA1) enhances B cell immortalization several thousandfold. *Proc Natl Acad Sci U S A.* 2003;100(19):10989-10994.
20. Saridakis V, Sheng Y, Sarkari F, et al. Structure of the p53 binding domain of HAUSP/USP7 bound to Epstein-Barr nuclear antigen 1 implications for EBV-mediated immortalization. *Mol Cell.* 2005;18(1):25-36.
21. Wilson JB, Bell JL, Levine AJ. Expression of Epstein-Barr virus nuclear antigen-1 induces B cell neoplasia in transgenic mice. *EMBO J.* 1996;15(12):3117-3126.
22. Kube D, Vockerodt M, Weber O, et al. Expression of epstein-barr virus nuclear antigen 1 is associated with enhanced expression of CD25 in the Hodgkin cell line L428. *J Virol.* 1999;73(2):1630-1636.
23. Abate F, Ambrosio MR, Mundo L, et al. Distinct Viral and Mutational Spectrum of Endemic Burkitt Lymphoma. *PLoS Pathog.* 2015;11(10):e1005158.
24. Xue SA, Labrecque LG, Lu QL, et al. Promiscuous expression of Epstein-Barr virus genes in Burkitt's lymphoma from the central African country Malawi. *Int J Cancer.* 2002;99(5):635-643.
25. Kelly G, Bell A, Rickinson A. Epstein-Barr virus-associated Burkitt lymphomagenesis selects for downregulation of the nuclear antigen EBNA2. *Nat Med.* 2002;8(10):1098-1104.
26. Kelly GL, Stylianou J, Rasaiyaah J, et al. Different patterns of Epstein-Barr virus latency in endemic Burkitt lymphoma (BL) lead to distinct variants within the BL-associated gene expression signature. *J Virol.* 2013;87(5):2882-2894.
27. Hislop AD, Taylor GS, Sauce D, Rickinson AB. Cellular responses to viral infection in humans: lessons from Epstein-Barr virus. *Annu Rev Immunol.* 2007;25:587-617.
28. Longnecker R, Druker B, Roberts TM, Kieff E. An Epstein-Barr virus protein associated with cell growth transformation interacts with a tyrosine kinase. *J Virol.* 1991;65(7):3681-3692.
29. Portis T, Longnecker R. Epstein-Barr virus (EBV) LMP2A mediates B-lymphocyte survival through constitutive activation of the Ras/PI3K/Akt pathway. *Oncogene.* 2004;23(53):8619-8628.
30. Scholle F, Bendt KM, Raab-Traub N. Epstein-Barr virus LMP2A transforms epithelial cells, inhibits cell differentiation, and activates Akt. *J Virol.* 2000;74(22):10681-10689.

31. Swart R, Ruf IK, Sample J, Longnecker R. Latent membrane protein 2A-mediated effects on the phosphatidylinositol 3-Kinase/Akt pathway. *J Virol.* 2000;74(22):10838-10845.
32. Fukuda M, Longnecker R. Latent membrane protein 2A inhibits transforming growth factor-beta 1-induced apoptosis through the phosphatidylinositol 3-kinase/Akt pathway. *J Virol.* 2004;78(4):1697-1705.
33. Minamitani T, Yasui T, Ma Y, et al. Evasion of affinity-based selection in germinal centers by Epstein-Barr virus LMP2A. *Proc Natl Acad Sci U S A.* 2015.
34. Moormann AM, Heller KN, Chelimo K, et al. Children with endemic Burkitt lymphoma are deficient in EBNA1-specific IFN-gamma T cell responses. *Int J Cancer.* 2009;124(7):1721-1726.
35. Donati D, Zhang LP, Chene A, et al. Identification of a polyclonal B-cell activator in Plasmodium falciparum. *Infect Immun.* 2004;72(9):5412-5418.
36. Urban BC, Ferguson DJ, Pain A, et al. Plasmodium falciparum-infected erythrocytes modulate the maturation of dendritic cells. *Nature.* 1999;400(6739):73-77.
37. Snider CJ, Cole SR, Chelimo K, et al. Recurrent Plasmodium falciparum malaria infections in Kenyan children diminish T-cell immunity to Epstein Barr virus lytic but not latent antigens. *PLoS One.* 2012;7(3):e31753.
38. Njie R, Bell AI, Jia H, et al. The effects of acute malaria on Epstein-Barr virus (EBV) load and EBV-specific T cell immunity in Gambian children. *J Infect Dis.* 2009;199(1):31-38.
39. Chattopadhyay PK, Chelimo K, Embury PB, et al. Holoendemic malaria exposure is associated with altered Epstein-Barr virus-specific CD8(+) T-cell differentiation. *J Virol.* 2013;87(3):1779-1788.
40. Whittle HC, Brown J, Marsh K, et al. T-cell control of Epstein-Barr virus-infected B cells is lost during P. falciparum malaria. *Nature.* 1984;312(5993):449-450.
41. Lam KM, Syed N, Whittle H, Crawford DH. Circulating Epstein-Barr virus-carrying B cells in acute malaria. *Lancet.* 1991;337(8746):876-878.
42. Coban C, Igari Y, Yagi M, et al. Immunogenicity of whole-parasite vaccines against Plasmodium falciparum involves malarial hemozoin and host TLR9. *Cell Host Microbe.* 2010;7(1):50-61.
43. Torgbor C, Awuah P, Deitsch K, Kalantari P, Duca KA, Thorley-Lawson DA. A multifactorial role for P. falciparum malaria in endemic Burkitt's lymphoma pathogenesis. *PLoS Pathog.* 2014;10(5):e1004170.
44. Pelicci PG, Knowles DM, 2nd, Magrath I, Dalla-Favera R. Chromosomal breakpoints and structural alterations of the c-myc locus differ in endemic and sporadic forms of Burkitt lymphoma. *Proc Natl Acad Sci U S A.* 1986;83(9):2984-2988.
45. Joos S, Falk MH, Lichter P, et al. Variable breakpoints in Burkitt lymphoma cells with chromosomal t(8;14) translocation separate c-myc and the IgH locus up to several hundred kb. *Hum Mol Genet.* 1992;1(8):625-632.
46. Shiramizu B, Barriga F, Neequaye J, et al. Patterns of chromosomal breakpoint locations in Burkitt's lymphoma: relevance to geography and Epstein-Barr virus association. *Blood.* 1991;77(7):1516-1526.
47. Busch K, Keller T, Fuchs U, et al. Identification of two distinct MYC breakpoint clusters and their association with various IGH breakpoint regions in the t(8;14) translocations in sporadic Burkitt-lymphoma. *Leukemia.* 2007;21(8):1739-1751.

48. Barriga F, Kiwanuka J, Alvarez-Mon M, et al. Significance of chromosome 8 breakpoint location in Burkitt's lymphoma: correlation with geographical origin and association with Epstein-Barr virus. *Curr Top Microbiol Immunol.* 1988;141:128-137.
49. Basso K, Frascella E, Zanesco L, Rosolen A. Improved long-distance polymerase chain reaction for the detection of t(8;14)(q24;q32) in Burkitt's lymphomas. *Am J Pathol.* 1999;155(5):1479-1485.
50. Muller JR, Janz S, Goedert JJ, Potter M, Rabkin CS. Persistence of immunoglobulin heavy chain/c-myc recombination-positive lymphocyte clones in the blood of human immunodeficiency virus-infected homosexual men. *Proc Natl Acad Sci U S A.* 1995;92(14):6577-6581.
51. Rabbitts PH, Forster A, Stinson MA, Rabbitts TH. Truncation of exon 1 from the c-myc gene results in prolonged c-myc mRNA stability. *EMBO J.* 1985;4(13b):3727-3733.
52. Bahram F, von der Lehr N, Cetinkaya C, Larsson LG. c-Myc hot spot mutations in lymphomas result in inefficient ubiquitination and decreased proteasome-mediated turnover. *Blood.* 2000;95(6):2104-2110.
53. Cesarman E, Dalla-Favera R, Bentley D, Groudine M. Mutations in the first exon are associated with altered transcription of c-myc in Burkitt lymphoma. *Science.* 1987;238(4831):1272-1275.
54. Harris LJ, Skaletsky E, McPherson A. Crystallographic structure of an intact IgG1 monoclonal antibody. *J Mol Biol.* 1998;275(5):861-872.
55. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem.* 2004;25(13):1605-1612.
56. Alt FW, Yancopoulos GD, Blackwell TK, et al. Ordered rearrangement of immunoglobulin heavy chain variable region segments. *EMBO J.* 1984;3(6):1209-1219.
57. Murphy K TP, Walport M. *Janeway's Immunobiology.* 8th ed. New York, NY: Garland Science; 2012.
58. Lam KP, Kuhn R, Rajewsky K. In vivo ablation of surface immunoglobulin on mature B cells by inducible gene targeting results in rapid cell death. *Cell.* 1997;90(6):1073-1083.
59. Srinivasan L, Sasaki Y, Calado DP, et al. PI3 kinase signals BCR-dependent mature B cell survival. *Cell.* 2009;139(3):573-586.
60. Davis RE, Ngo VN, Lenz G, et al. Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. *Nature.* 2010;463(7277):88-92.
61. Krysiak K, Gomez F, White BS, et al. Recurrent somatic mutations affecting B-cell receptor signaling pathway genes in follicular lymphoma. *Blood.* 2017;129(4):473-483.
62. Richter J, Schlesner M, Hoffmann S, et al. Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat Genet.* 2012;44(12):1316-1320.
63. Love C, Sun Z, Jima D, et al. The genetic landscape of mutations in Burkitt lymphoma. *Nat Genet.* 2012;44(12):1321-1325.
64. Hadzidimitriou A, Agathangelidis A, Darzentas N, et al. Is there a role for antigen selection in mantle cell lymphoma? Immunogenetic support from a series of 807 cases. *Blood.* 2011;118(11):3088-3095.
65. Agathangelidis A, Darzentas N, Hadzidimitriou A, et al. Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. *Blood.* 2012;119(19):4467-4475.

66. Duhren-von Minden M, Ubelhart R, Schneider D, et al. Chronic lymphocytic leukaemia is driven by antigen-independent cell-autonomous signalling. *Nature.* 2012;489(7415):309-312.
67. Catera R, Silverman GJ, Hatzi K, et al. Chronic lymphocytic leukemia cells recognize conserved epitopes associated with apoptosis and oxidation. *Mol Med.* 2008;14(11-12):665-674.
68. Zwick C, Fadle N, Regitz E, et al. Autoantigenic targets of B-cell receptors derived from chronic lymphocytic leukemias bind to and induce proliferation of leukemic cells. *Blood.* 2013;121(23):4708-4717.
69. Hoogeboom R, van Kessel KP, Hochstenbach F, et al. A mutated B cell chronic lymphocytic leukemia subset that recognizes and responds to fungi. *J Exp Med.* 2013;210(1):59-70.
70. Seiler T, Woelfle M, Yancopoulos S, et al. Characterization of structurally defined epitopes recognized by monoclonal antibodies produced by chronic lymphocytic leukemia B cells. *Blood.* 2009;114(17):3615-3624.
71. Chu CC, Catera R, Hatzi K, et al. Chronic lymphocytic leukemia antibodies with a common stereotypic rearrangement recognize nonmuscle myosin heavy chain IIA. *Blood.* 2008;112(13):5122-5129.
72. Zibellini S, Capello D, Forconi F, et al. Stereotyped patterns of B-cell receptor in splenic marginal zone lymphoma. *Haematologica.* 2010;95(10):1792-1796.
73. Zhu D, Bhatt S, Lu X, et al. Chlamydophila psittaci-negative ocular adnexal marginal zone lymphomas express self polyreactive B-cell receptors. *Leukemia.* 2015;29(7):1587-1599.
74. Montesinos-Rongen M, Purschke FG, Brunn A, et al. Primary Central Nervous System (CNS) Lymphoma B Cell Receptors Recognize CNS Proteins. *J Immunol.* 2015;195(3):1312-1319.
75. Wang ML, Rule S, Martin P, et al. Targeting BTK with ibrutinib in relapsed or refractory mantle-cell lymphoma. *N Engl J Med.* 2013;369(6):507-516.
76. Gopal AK, Kahl BS, de Vos S, et al. PI3Kdelta inhibition by idelalisib in patients with relapsed indolent lymphoma. *N Engl J Med.* 2014;370(11):1008-1018.
77. Byrd JC, Furman RR, Coutre SE, et al. Targeting BTK with ibrutinib in relapsed chronic lymphocytic leukemia. *N Engl J Med.* 2013;369(1):32-42.
78. Schmitz R, Young RM, Ceribelli M, et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature.* 2012;490(7418):116-120.
79. Walter R, Pan KT, Doebele C, et al. HSP90 promotes Burkitt lymphoma cell survival by maintaining tonic B-cell receptor signaling. *Blood.* 2017;129(5):598-608.
80. Piccaluga PP, De Falco G, Kustagi M, et al. Gene expression analysis uncovers similarity and differences among Burkitt lymphoma subtypes. *Blood.* 2011;117(13):3596-3608.
81. Sander S, Calado DP, Srinivasan L, et al. Synergy between PI3K signaling and MYC in Burkitt lymphomagenesis. *Cancer Cell.* 2012;22(2):167-179.
82. Riboldi P, Gaidano G, Schettino EW, et al. Two acquired immunodeficiency syndrome-associated Burkitt's lymphomas produce specific anti-i IgM cold agglutinins using somatically mutated VH4-21 segments. *Blood.* 1994;83(10):2952-2961.

83. Roncella S, Cutrona G, Favre A, et al. Apoptosis of Burkitt's lymphoma cells induced by specific interaction of surface IgM with a self-antigen: implications for lymphomagenesis in acquired immunodeficiency syndrome. *Blood.* 1996;88(2):599-608.
84. Horikawa K, Martin SW, Pogue SL, et al. Enhancement and suppression of signaling by the conserved tail of IgG memory-type B cell antigen receptors. *J Exp Med.* 2007;204(4):759-769.
85. Cheong TC, Compagno M, Chiarle R. Editing of mouse and human immunoglobulin genes by CRISPR-Cas9 system. *Nat Commun.* 2016;7:10934.
86. Havelange V, Pepermans X, Ameye G, et al. Genetic differences between paediatric and adult Burkitt lymphomas. *Br J Haematol.* 2016;173(1):137-144.
87. Chapman CJ, Wright D, Stevenson FK. Insight into Burkitt's lymphoma from immunoglobulin variable region gene analysis. *Leuk Lymphoma.* 1998;30(3-4):257-267.
88. Tamaru J, Hummel M, Marafioti T, et al. Burkitt's lymphomas express VH genes with a moderate number of antigen-selected somatic mutations. *Am J Pathol.* 1995;147(5):1398-1407.
89. Chapman CJ, Mockridge CI, Rowe M, Rickinson AB, Stevenson FK. Analysis of VH genes used by neoplastic B cells in endemic Burkitt's lymphoma shows somatic hypermutation and intraclonal heterogeneity. *Blood.* 1995;85(8):2176-2181.
90. Klein U, Klein G, Ehlin-Henriksson B, Rajewsky K, Kuppers R. Burkitt's lymphoma is a malignancy of mature B cells expressing somatically mutated V region genes. *Mol Med.* 1995;1(5):495-505.
91. Riboldi P, Ikematsu W, Brambilla B, Caprani C, Gerosa M, Casali P. Diversity and somatic hypermutation of the Ig VHDJH, V kappa J kappa, and V lambda J lambda gene segments in lymphoma B cells: relevance to the origin of the neoplastic B cell clone. *Hum Immunol.* 2003;64(1):69-81.
92. Boyd SD, Marshall EL, Merker JD, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med.* 2009;1(12):12ra23.
93. Wu D, Sherwood A, Fromm JR, et al. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci Transl Med.* 2012;4(134):134ra163.
94. Wu D, Emerson RO, Sherwood A, et al. Detection of minimal residual disease in B lymphoblastic leukemia by high-throughput sequencing of IGH. *Clin Cancer Res.* 2014;20(17):4540-4548.
95. Kurtz DM, Green MR, Bratman SV, et al. Noninvasive monitoring of diffuse large B-cell lymphoma by immunoglobulin high-throughput sequencing. *Blood.* 2015;125(24):3679-3687.
96. Roschewski M, Dunleavy K, Pittaluga S, et al. Circulating tumour DNA and CT monitoring in patients with untreated diffuse large B-cell lymphoma: a correlative biomarker study. *Lancet Oncol.* 2015;16(5):541-549.
97. Orem J, Mbidde EK, Lambert B, de Sanjose S, Weiderpass E. Burkitt's lymphoma in Africa, a review of the epidemiology and etiology. *Afr Health Sci.* 2007;7(3):166-175.
98. Magrath I. Epidemiology: clues to the pathogenesis of Burkitt lymphoma. *Br J Haematol.* 2012;156(6):744-756.
99. Kafuko GW, Burkitt DP. Burkitt's lymphoma and malaria. *Int J Cancer.* 1970;6(1):1-9.

100. DeWitt WS, Lindau P, Snyder TM, et al. A Public Database of Memory and Naive B-Cell Receptor Sequences. *PLoS One.* 2016;11(8):e0160853.
101. Carlson CS, Emerson RO, Sherwood AM, et al. Using synthetic templates to design an unbiased multiplex PCR assay. *Nat Commun.* 2013;4:2680.
102. Brochet X, Lefranc MP, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 2008;36(Web Server issue):W503-508.
103. Giudicelli V, Brochet X, Lefranc MP. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc.* 2011;2011(6):695-715.
104. Kanakry CG, Coffey DG, Towlerton AM, et al. Origin and evolution of the T cell repertoire after posttransplantation cyclophosphamide. *JCI Insight.* 2016;1(5).
105. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 2013;41(Web Server issue):W34-40.
106. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics.* 2013;43:11.10.11-33.
107. Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet.* 2015;24(8):2125-2137.
108. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* 2015;43(Database issue):D805-811.
109. Bertucci C RM. Bone Marrow From Healthy Adults. 2015.
110. Baptista MJ, Calpe E, Fernandez E, et al. Analysis of the IGHV region in Burkitt's lymphomas supports a germinal center origin and a role for superantigens in lymphomagenesis. *Leuk Res.* 2014;38(4):509-515.
111. Kolhatkar NS, Brahmandam A, Thouvenel CD, et al. Altered BCR and TLR signals promote enhanced positive selection of autoreactive transitional B cells in Wiskott-Aldrich syndrome. *J Exp Med.* 2015;212(10):1663-1677.
112. Zhu D, Ottensmeier CH, Du MQ, McCarthy H, Stevenson FK. Incidence of potential glycosylation sites in immunoglobulin variable regions distinguishes between subsets of Burkitt's lymphoma and mucosa-associated lymphoid tissue lymphoma. *Br J Haematol.* 2003;120(2):217-222.
113. van Krieken JH, Langerak AW, Macintyre EA, et al. Improved reliability of lymphoma diagnostics via PCR-based clonality testing: report of the BIOMED-2 Concerted Action BHM4-CT98-3936. *Leukemia.* 2007;21(2):201-206.
114. EuroClonality.
115. Fukita Y, Jacobs H, Rajewsky K. Somatic hypermutation in the heavy chain locus correlates with transcription. *Immunity.* 1998;9(1):105-114.
116. Gawad C, Pepin F, Carlton VE, et al. Massive evolution of the immunoglobulin heavy chain locus in children with B precursor acute lymphoblastic leukemia. *Blood.* 2012;120(22):4407-4417.
117. Bhatia K, Gutierrez M, Magrath IT. Burkitt's lymphoma cells frequently carry monoallelic DJ rearrangements. *Curr Top Microbiol Immunol.* 1992;182:319-324.

118. Robbiani DF, Deroubaix S, Feldhahn N, et al. Plasmodium Infection Promotes Genomic Instability and AID-Dependent B Cell Lymphoma. *Cell.* 2015;162(4):727-737.
119. Robbiani DF, Bothmer A, Callen E, et al. AID is required for the chromosomal breaks in c-myc that lead to c-myc/IgH translocations. *Cell.* 2008;135(6):1028-1038.
120. Greisman HA, Lu Z, Tsai AG, Greiner TC, Yi HS, Lieber MR. IgH partner breakpoint sequences provide evidence that AID initiates t(11;14) and t(8;14) chromosomal breaks in mantle cell and Burkitt lymphomas. *Blood.* 2012;120(14):2864-2867.
121. Takizawa M, Tolarova H, Li Z, et al. AID expression levels determine the extent of cMyc oncogenic translocations and the incidence of B cell tumor development. *J Exp Med.* 2008;205(9):1949-1957.
122. Swaminathan S, Klemm L, Park E, et al. Mechanisms of clonal evolution in childhood acute lymphoblastic leukemia. *Nat Immunol.* 2015.
123. Coelho V, Krysov S, Ghaemmaghami AM, et al. Glycosylation of surface Ig creates a functional bridge between human follicular lymphoma and microenvironmental lectins. *Proc Natl Acad Sci U S A.* 2010;107(43):18587-18592.
124. Waisman A, Kraus M, Seagal J, et al. IgG1 B cell receptor signaling is inhibited by CD22 and promotes the development of B cells whose survival is less dependent on Ig alpha/beta. *J Exp Med.* 2007;204(4):747-758.
125. Sachen KL, Strohman MJ, Singletary J, et al. Self-antigen recognition by follicular lymphoma B-cell receptors. *Blood.* 2012;120(20):4182-4190.
126. Wright NJ, Hesseling PB, McCormick P, Tchintseme F. The incidence, clustering and characteristics of Burkitt lymphoma in the Northwest province of Cameroon. *Trop Doct.* 2009;39(4):228-230.
127. Burkitt D. A sarcoma involving the jaws in African children. *Br J Surg.* 1958;46(197):218-223.
128. Sala Torra O, Othus M, Williamson DW, et al. Next-Generation Sequencing in Adult B Cell Acute Lymphoblastic Leukemia Patients. *Biol Blood Marrow Transplant.* 2017;23(4):691-696.
129. Martinez-Lopez J, Lahuerta JJ, Pepin F, et al. Prognostic value of deep sequencing method for minimal residual disease detection in multiple myeloma. *Blood.* 2014;123(20):3073-3079.
130. Arthur FK, Owusu L, Yeboah FA, Rettig T, Osei-Akoto A. Prognostic significance of biochemical markers in African Burkitt's lymphoma. *Clin Transl Oncol.* 2011;13(10):731-736.
131. Buckle G, Maranda L, Skiles J, et al. Factors influencing survival among Kenyan children diagnosed with endemic Burkitt lymphoma between 2003 and 2011: A historical cohort study. *Int J Cancer.* 2016.
132. Magrath I, Lee YJ, Anderson T, et al. Prognostic factors in Burkitt's lymphoma: importance of total tumor burden. *Cancer.* 1980;45(6):1507-1515.
133. Westmoreland KD, Montgomery ND, Stanley CC, et al. Plasma Epstein-Barr virus DNA for pediatric Burkitt lymphoma diagnosis, prognosis and response assessment in Malawi. *Int J Cancer.* 2017;140(11):2509-2516.
134. Magrath IT, Ziegler JL, Templeton AC. A comparison of clinical and histopathologic features of childhood malignant lymphoma in Uganda. *Cancer.* 1974;33(1):285-294.
135. Ziegler JL, Bluming AZ, Morrow RH, Fass L, Carbone PP. Central nervous system involvement in Burkitt's lymphoma. *Blood.* 1970;36(6):718-728.

136. Robbiani DF, Bunting S, Feldhahn N, et al. AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. *Mol Cell.* 2009;36(4):631-641.
137. Ramiro AR, Jankovic M, Callen E, et al. Role of genomic instability and p53 in AID-induced c-myc-Igh translocations. *Nature.* 2006;440(7080):105-109.
138. Joice R, Nilsson SK, Montgomery J, et al. Plasmodium falciparum transmission stages accumulate in the human bone marrow. *Sci Transl Med.* 2014;6(244):244re245.
139. Shen HM, Peters A, Baron B, Zhu X, Storb U. Mutation of BCL-6 gene in normal B cells by the process of somatic hypermutation of Ig genes. *Science.* 1998;280(5370):1750-1752.
140. Pasqualucci L, Migliazza A, Fracchiolla N, et al. BCL-6 mutations in normal germinal center B cells: evidence of somatic hypermutation acting outside Ig loci. *Proc Natl Acad Sci U S A.* 1998;95(20):11816-11821.
141. Gordon MS, Kanegai CM, Doerr JR, Wall R. Somatic hypermutation of the B cell receptor genes B29 (Igbeta, CD79b) and mb1 (Igalpha, CD79a). *Proc Natl Acad Sci U S A.* 2003;100(7):4126-4131.
142. Muschen M, Re D, Jungnickel B, Diehl V, Rajewsky K, Kuppers R. Somatic mutation of the CD95 gene in human B cells as a side-effect of the germinal center reaction. *J Exp Med.* 2000;192(12):1833-1840.
143. Refaeli Y, Field KA, Turner BC, Trumpp A, Bishop JM. The protooncogene MYC can break B cell tolerance. *Proc Natl Acad Sci U S A.* 2005;102(11):4097-4102.
144. Hon GM, Hassan MS, van Rensburg SJ, Erasmus RT, Matsha TE. Assessment of Epstein-Barr virus in blood from patients with multiple sclerosis. *Metab Brain Dis.* 2012;27(3):311-318.
145. Scheid JF, Mouquet H, Ueberheide B, et al. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science.* 2011;333(6049):1633-1637.
146. Wardemann H, Kofer J. Expression cloning of human B cell immunoglobulins. *Methods Mol Biol.* 2013;971:93-111.
147. Jiang N, He J, Weinstein JA, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci Transl Med.* 2013;5(171):171ra119.
148. Bolotin DA, Poslavsky S, Mitrophanov I, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods.* 2015;12(5):380-381.
149. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15-21.
150. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.

VITA

Katie Lombardo was born and raised in Seattle, WA. Her interest in science was piqued in middle school and high school due to great teachers who inspired an understanding of the biological world. Katie graduated *cum laude* from the University of Southern California, in Los Angeles, CA, in 2007 with a Bachelor of Science in biological sciences and a minor in cultural anthropology. Between her junior and senior years of college, Katie worked as an undergraduate researcher in the lab of Antonio Bedalov, M.D. at the Fred Hutchinson Cancer Research Center (FHCRC). After college, Katie worked in the lab of Akiko Shimamura, M.D., Ph.D. at FHCRC, studying the bone marrow failure disease, Shwachman Diamond Syndrome. Katie started in the Molecular and Cellular Biology graduate program in 2011 and joined the lab of Edus H. Warren, M.D., Ph.D. in 2012. Katie has gained experience teaching undergraduates as a teaching assistant at the University of Washington and mentoring a high school science teacher through the Science Education Partnership program. Katie was a recipient of the Chromosome, Metabolism and Cancer Training Grant. During her graduate studies, Katie presented at scientific conferences in Colorado, San Francisco, and Morocco. Her scientific interests are focused on cancer, the immune system, and infectious diseases.