Comprehensive Transcriptome and Mutational Profiling of Endemic Burkitt Lymphoma Reveals EBV Type-specific Differences

Yasin Kaymaz¹, Cliff I. Oduor²,³, Hongbo Yu⁴, Juliana A. Otieno⁵, John Michael Ong’echa², Ann M. Moormann⁶, Jeffrey A. Bailey¹,⁷*

Affiliations of authors:
¹ Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA
² Center for Global Health Research, Kenya Medical Research Institute, Kisumu, Kenya
³ Department of Biomedical Sciences and Technology, Maseno University, Maseno, Kenya
⁴ Department of Pathology, University of Massachusetts Medical School, Worcester, MA
⁵ Jaramogi Oginga Odinga Teaching and Referral Hospital, Ministry of Health, Kisumu, Kenya
⁶ Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA
⁷ Division of Transfusion Medicine, Department of Medicine, University of Massachusetts Medical School, Worcester, MA

*Corresponding Author:
Jeffrey A. Bailey, MD PhD
jeffrey.bailey@umassmed.edu
368 Plantation St. Albert Sherman Building 4-1077
Worcester, MA 01605
Phone: 508-856-8034
Fax: 508-856-0017

Running Title:
Expression and mutational landscape of eBL

Key Words:
Endemic Burkitt lymphoma, Epstein Barr virus, transcriptome, somatic mutations

Financial support:
This study was supported by the US National Institutes of Health, National Cancer Institute R01 CA134051, R01 CA189806 (AMM) and The Thrasher Research Fund 02833-7 (AMM), UMCC TS Pilot Project Program U1 LTR000161-04 (YK, JAB, and AMM), Turkish Ministry of National Education Graduate Study Abroad Program (YK).

Conflict of interest disclosure: The authors disclose no potential conflicts of interest.

Word Count: 6,261 (main text), 262 (abstract).

5 Figures, 2 Tables.
ABSTRACT

Endemic Burkitt lymphoma (eBL) is the most common pediatric cancer in malaria-endemic equatorial Africa and nearly always contains Epstein-Barr virus (EBV), unlike sporadic Burkitt Lymphoma (sBL) that occurs with a lower incidence in developed countries. Given these differences and the variable clinical presentation and outcomes, we sought to further understand pathogenesis by investigating transcriptomes using RNA sequencing (RNAseq) from multiple primary eBL tumors compared to sBL tumors. Within eBL tumors, minimal expression differences were found based on: anatomical presentation site, in-hospital survival rates, and EBV genome type; suggesting that eBL tumors are homogeneous without marked subtypes. The outstanding difference detected using surrogate variable analysis was the significantly decreased expression of key genes in the immunoproteasome complex (PSMB9/β1i, PSMB10/β2i, PSMB8/β5i, and PSME2/PA28β) in eBL tumors carrying type 2 EBV compared to type 1 EBV. Secondly, in comparison to previously published pediatric sBL specimens, the majority of the expression and pathway differences was related to the PTEN/PI3K/mTOR signaling pathway and was correlated most strongly with EBV status rather than geographic designation. Third, common mutations were observed significantly less frequently in eBL tumors harboring EBV type 1, with mutation frequencies similar between tumors with EBV type 2 and without EBV. In addition to the previously reported genes, a set of new genes mutated in BL including TFAP4, MSH6, PRRC2C, BCL7A, FOXO1, PLCG2, PRKDC, RAD50, and RPRD2 were identified. Overall, these data establish that EBV, particularly EBV type 1, supports BL oncogenesis alleviating the need for certain driver mutations in the human genome.

Implications:

Genomic and mutational analyses of Burkitt lymphoma tumors identify key differences based on viral content and clinical outcomes suggesting new avenues for the development of prognostic molecular biomarkers and therapeutic interventions.
INTRODUCTION

Burkitt lymphoma (BL) is a B-cell neoplasm composed of monomorphic, medium sized cells with basophilic cytoplasm and one of the highest proliferation rates known for human tumors (1). Its histological appearance is “sky” like with a background of homogeneous tumor cells punctuated by “stars” consisting of macrophages at apoptotic foci. The World Health Organization recognizes three clinical subtypes of BL: endemic BL (eBL), sporadic BL (sBL), and immunodeficiency (including HIV)-related BL (idBL) (2).

Aside from geography differences in incidence, eBL tumors are commonly associated with Epstein-Barr virus (EBV) infection and have been linked to the protozoan parasite Plasmodium falciparum malaria, presenting in children between 5-9 years of age. It commonly presents in the jaw or facial bones as well as other extranodal sites such as the GI tract, kidneys, and breasts. In contrast, compared to eBL, pediatric sBL is found at a 10-fold lower incidence in developed countries where malaria is not endemic and only contains EBV in around 10-20% of cases. Pediatric sBL tends to afflict a higher proportion of males and adolescents and present in the abdomen often with disseminated disease (2). sBL incidence has a bimodal age distribution with peaks in children and older adults suggesting different etiologies. Adult sBL tends to have higher rates of EBV positivity, nodal presentation, along with poorer outcome, and often more variable pathologic features leading to designations of plasmacytoid or atypical-BL (3). These differences within sBL have raised the suggestion that adult sBL should be considered a separate entity (4) EBV positive and EBV negative tumors are defined separately (5).

The greatest difference in EBV prevalence in BL tumor classifications is seen between endemic (95%) and pediatric sporadic BL (10%) tumors. EBV positivity is intermediate in id-BL (2,3) and increases with age in adult sporadic cases (30-50%) (2). Unlike other lymphomas such as Hodgkin's (6) and DLBCL (7), EBV has not been associated with outcome (2). It does appear that EBV positive tumors may share a similar B cell origin compared to EBV negative tumors regardless of geographic origin (8). EBV positive eBL cases display a viral latency I expression profile, which includes EBNA1, EBERs, and the viral microRNAs within the BART region transcripts (9,10). Of note for eBL, the infecting EBV genome may be either of two divergent strains, type 1 or type 2, and comparative genomic studies have demonstrated type-specific divergence (11–13). While type 1 EBV is found globally, type 2 is more commonly found in Africa than other parts of the world.
world (14). Although, it has been reported that the transformation efficiency of EBV type 1 is higher compared to type 2 in lymphoblastoid cell line establishments (15), both strains are frequently found in African eBL cases and are prevalent within healthy populations in sub Saharan Africa (12,16). However, the expression and mutational profiles of EBV type 1 and type 2 within primary eBL tumors have not been compared and contrasted to determine if viral variation influences tumorigenesis.

Clinical features of eBL and response to conventional chemotherapy have not been examined with regards to expression or mutational profile of the tumor. Endemic BL shows distinctive presentation in either the jaw or the abdomen (17); among Kenyan children within our larger BL cohort, the tumor presentation sites were 43% jaw and 50% abdomen (18). In addition, during the study period between 2003 and 2011, 22% of the admitted patients died in-hospital and 78% completed the course of chemotherapy treatment (18). In this respect, there was a dramatic difference between the survival rates with 63% of patients with jaw tumors surviving compared to 33% for abdominal tumors. Attempts to associate antibody titers with tumor presentation site and prognosis have shown that Anti-Zta IgG levels were elevated in eBL patients with abdominal tumors compared to patients with jaw tumors (19). However, high throughput expression profiling and comparative assays applied here better address the question of distinct molecular features specific to tumor localization and/or survival outcome.

MYC oncogene deregulation and ectopic expression by chromosomal translocations is the key molecular driver and hallmark of BL. Even though deregulated expression and subsequent mutations of MYC gene severely alter the DNA binding efficiency of this transcription factor, these do not appear to be sufficient for tumorigenesis (20). The search for additional driver mutations in sBL has yielded several candidate tumor suppressors and oncogenes (21–23), however; eBL primary tumor biopsies have not been studied at a genome-wide level until recently with limited numbers of cases (24). The most common driver mutations in coding regions appear to occur in the transcription factor TCF3 (E2A) and its inhibitor ID3. Cell cycle regulator gene CCND3 which encodes for cyclin D3 is another gene frequently mutated, along with the TCF3 and ID3, in addition to MYC locus alterations especially in sBL cases. Here we investigate the transcriptome and mutational profiles of 28 eBL and 2 sBL primary tumors by deep sequencing and unlike previous studies, we
correlated our findings with clinical outcomes. We also explored the viral gene expression activity in EBV positive BL tumors comparing and contrasting type 1 and type 2 virus.

MATERIALS AND METHODS

Ethical approval and sample collection

Fine needle aspirates (FNA) were prospectively obtained between 2009 and 2012 at time of diagnosis and prior to commencing chemotherapy at Jaramogi Oginga Odinga Teaching and Referral Hospital (JOOTRH), a regional referral hospital for pediatric cancer in western Kenya. Tumor FNAs were stained with Giemsa/May-Grünwald for morphologic diagnosis by microscopy. Morphology was assessed by two independent pathologists to verify the diagnosis. A second FNA was transferred into RNA later at the bedside and subsequently stored at -20°C. Written informed consent was obtained from a parent or legal guardian of the child before enrollment. Ethical approval was obtained from the Institutional Review Board at the University of Massachusetts Medical School and the Scientific and Ethics Review Unit at the Kenya Medical Research Institute.

In order to better compare BL subtypes, we also analyzed published RNAseq dataset of sBLs and cell lines (21). The sequences in fastq format were downloaded through the NCBI (SRP009316) for 28 sBL primary tumors and 13 long term BL cultures derived from sporadic and endemic cases. In addition, we also analyzed 89 mRNA sequencing from lymphoblastoid cell lines (LCLs) from healthy individuals involved in the 1000 genome project (Yoruba, YRI) (ERP001942), which we used to eliminate variant calls likely due to transcript assembly, mapping artifacts or RNA editing.

Sequencing Library Preparation

Briefly, starting with 1-5ug total RNA, we prepared strand-specific RNAseq libraries following the protocol from Zhang et al. (25) combined with mRNA enrichment with oligo-dT using Dynabeads mRNA purification kit (Life Technologies) (Supplementary methods). Final library qualities were confirmed with Bioanalyzer High sensitivity DNA kit (Agilent) and sequenced with paired end read (2x100bp) using multiple
lanes of Illumina HiSeq 2000 (Illumina Inc., San Diego, CA, USA). Data can be accessed at dbGAP with accession number (Pending).

**Differential Gene Expression Analysis**

After quality assessment and preprocessing the raw sequencing reads, we aligned read pairs to a transcriptome index built by RSEM (26) using Gencode v19 protein coding transcript annotations and hg19 genomic sequence. For EBV genes, we used GenBank gene annotations from both the type 1 and type 2 reference genomes (NC_007605 and NC_009334, respectively). To perform differential gene expression test, we used DESeq2 (27) in R computing environment. In order to be able to account for the batch variables and unknown factors while testing for the differential expression, we estimated the number of latent factors for every comparison separately using svaseq (28) while preserving the variation of interest. We then incorporated these surrogate variables into the testing model for DESeq2.

**Gene Set Enrichment Analysis**

We performed a standard gene set enrichment analysis (GSEA) using the GSEA module implemented by Broad Institute, Cambridge, MA (29). GSEA was performed on normalized expression data and on data after surrogate variable analysis. For a ranking metric, we used signal to noise value of each gene, and performed a permutation test for FDR by permuting sample phenotypes. The analysis included standard gene sets of hallmark and oncogenic signatures as well as the curated C2 gene sets from the Molecular Signatures Database (v5.0 MSigDB) (30).

**Single Nucleotide Variation Detection**

We mapped sequencing reads to human reference genome hg19 using the spliced aligner STAR (31) after quality trimming and removing the PCR duplicate reads. We followed the standard work flow by GATK (32) for calling variation within RNAseq data using the HaplotypeCaller module with additional stringency requirements (Supplementary methods). Variants observed in dbSNP v146 and low quality calls were excluded. We limited variant calling to translated sequences of protein coding genes in GenCode annotation v19.
RESULTS

Case information and sequencing summary

To survey the transcriptome of eBL, we sequenced 28 primary histologically-confirmed tumor FNA biopsies collected from Kenyan children with median age 8.2 years old (Table 1). We also sequenced two fresh frozen sBL tumors from diagnostic biopsies at the University of Massachusetts Medical School (UMMS). For the eBL patients, the tumor presenting site was 43% (12/28) jaw tumors and 57% (16/28) abdominal tumors. In terms of survival, the eBL samples included 3 patients who died prior to receiving any treatment, 5 patients who died during the course of treatment, and 16 patients who were able to complete the recommended chemotherapy treatment with resolution of their tumor and discharged from hospital (18). In-hospital survival for the children included in this study was 64% (18/28). For each of the samples, we performed strand-specific RNA sequencing generating on average 14M paired reads per library (range 8.9-53.7M reads; Supplementary Table S1). All 30 samples in the sequencing set showed high expression of associate BL markers, including traditional cell surface markers CD19, CD20, CD10 and CD79A/B, and intracellular markers of MYC and BCL6, consistent with the molecular phenotype of BL (3). All samples, including re-analyzed sBLs, showed high proportions of B cell specific expression (33) consistent with adequate aspirates of the tumor cells (Supplementary Fig. S1).

EBV positive eBL tumors are predominantly canonical latency I expression program

In concert with human transcriptome analysis, we first checked if EBV DNA was present in the tumor isolates using quantitative PCR. As expected, the vast majority of eBLs was positive (93%, 26/28) while only 2 eBLs and the 2 sBLs were negative (34). For EBV positive tumors, viral load assays indicated that tumor cells contained multiple copies of EBV DNA (mean 4,475 copies/ng tumor DNA; ~30 EBV/cell, median 1,542 copies/ng tumor DNA; ~10 EBV/cell). We also determined the virus type using distinguishing primers against viral gene EBNA3C. We found that 31% (N=8) of the EBV positive tumors were infected with type 2 EBV genomes whereas 69% (N=18) of them carried EBV type 1 genomes. We observed no mixed infection of both types within eBL tumors.
Given that these two types have divergent genomic sequences for several genes, we mapped the RNAseq reads to the appropriate viral transcriptome sequences. EBV positive tumors demonstrated significant viral gene expression, regardless of viral genome type. This expression from the EBV genome ranged across a continuum with the average around 200 RPM (reads per million) (ranging between 10 to 400 RPM; Supplementary Fig. S2A). The two EBV negative eBL tumors and two sBL tumors did not demonstrate any EBV specific reads supporting the absence of the virus based on qPCR. Interestingly, viral DNA copy numbers did not correlate significantly with overall viral transcriptome activity levels (Supplementary Fig. S2B). Along with BHRF1, BHLF1 and several EBV latent genes (EBERs, EBNA-1 and LMP2A/B) are weakly positively correlated with viral DNA levels in individual eBL tumors. On the other hand, BART transcripts, which are the most abundant transcripts including RPMS1, A73 and LF3, demonstrated no correlation with viral DNA levels (Supplementary Fig. S2C). This suggests that the observed viral expression levels within tumor cells are for the most part independent of viral load or lytic replication rates and may be dependent on other factors. In addition to the eBL tumors, four of the sBL primary tumors that were re-analyzed were also EBV positive (14%) and carried type 1 EBV genomes. Overall, viral genes in eBL tumors demonstrated a predominant expression pattern consistent with the latency I. However, hierarchical clustering of viral genes revealed several potential subgroups (Fig. 1). All sBLs clustered separately and showed latency I pattern but increased levels of BALF3 and BARF0, unlike eBLs. Within the eBLs there was also a substructure that appeared to be related to the relative amount of EBER1. The cell lines and three eBL samples showed higher levels of most genes suggestive of increased viral replication and lytic activity (Supplementary Fig. S2D). Among these three, two (eBL_02 and eBL_25) with elevated BHLF1 and lytic gene expression were patients who died in-hospital before receiving any treatment.

In-depth assessment of correlated variation with clinical features and viral type

Our initial question related to the tumor transcriptome was if there were any major expression differences and if any major differences correlated with the features of anatomical tumor presentation site, in-hospital survival or EBV type. After normalization of expression, we performed unsupervised hierarchical clustering based on Pearson correlations on expressed genes with the greatest variation (Fig. 2). The overall correlations among eBLs were extremely high ($r>0.96$, average). The sporadic tumors which differed in the
biopsy collection procedure (surgical biopsy or fine-needle aspirate) and preservation methods (fresh frozen or RNAlater) still showed a high-degree of correlation ($r > 0.90$) with eBLs although they distinctly clustered away from eBLs. Similarly, the major principal components showed no discernible separation based on tumor presentation site, treatment outcome, and viral genome type (Supplementary Fig. S3). Overall, this suggests that eBL tumors are relatively homogeneous group without overt subtypes based on tumor presentation site, survival or EBV type.

We then checked individual genes for differential expression between eBL tumors with different clinical features. For tumor site, only $\text{NOS3}$ showed significantly higher expression in abdominal tumors. This gene encodes for nitric oxide synthase 3 (aka eNOS) and is known to be more highly expressed in abdominal endothelial (35). Given the molecular phenotype of eBL tumors appears relatively homogeneous, it may be that unaccounted variation, biases or stochastic noise may be obscuring the detection of true expression differences. Thus, we used surrogate variable analysis (SVA) to isolate and remove unaccounted variation while preserving the variation associated with the feature of interest (28). As a result, we still failed to determine any significantly differentially expressed genes or pathways between biopsies from two different clinical tumor presentation sites, jaw and abdomen. However, for in-hospital survival of those who commenced chemotherapy, we detected 10 significantly differentially expressed genes between tumors of patients who survived and those who died (Fig. 3A). $\text{AGPAT3}$, $\text{CTSL}$, $\text{ISCU}$, $\text{CTSD}$, and $\text{APOE}$ showed greater relative expression in tumors of patients who survived while $\text{TUBB6}$, $\text{SLC25A24}$, $\text{FAM127A}$, $\text{HOMER1}$, and $\text{SLC12A3}$ demonstrated greater expression in tumors of patients who died (Supplementary Table S2). Gene set enrichment and pathway analysis between expression profiles of these patient groups suggested several hallmark pathways including hypoxia, IL2/STAT5 signaling, MYC targets, and TNFα signaling via NFκβ (Supplementary Table S3). The leading edge genes (core of the enrichment signal) mutually shared by these hallmark gene sets are $\text{SERPINE1}$, $\text{CD44}$, $\text{ENO2}$, $\text{PLAUR}$, $\text{RHOB}$, and $\text{TNFAIP3}$. These genes represent potential prognostic biomarkers requiring further investigation.

Comparison between EBV type 1 and type 2 viral-containing eBL tumors revealed 13 significant, differentially expressed genes (Fig. 3B). Four out of 8 genes that have significantly higher expression in eBL tumors with type 1 EBV are coding for the required components of immunoproteasome complex formation;
PSMB9 (β1i), PSMB10 (β2i), PSMB8 (β5i), and PSME2 (PA28β) (Supplementary Table S2). In addition, all of the other proteasome gene transcripts showed increased expression on average in eBLs with type 1 EBV (Supplementary Fig. S4). Consistent with this, our gene set enrichment and pathway analysis revealed several significant differential gene sets involving MHC class I antigen presenting cascades, ubiquitination and proteasome degradation, and antigen cross presentation altered between type 1 and 2 (Supplementary Table S3). This difference in expression of IFN-gamma inducible immunoproteasome complex genes and enriched pathways suggest that type 1 and type 2 genomes of EBV might differ in the pathogenesis of infection as well as in their roles promoting oncogenesis.

**Human gene expression appears to be more differentiated based on EBV status rather than eBL and sBL geographic designation**

We next investigated whether sBLs differ from eBLs in terms of human gene expression profiles. While it is inherently challenging to compare samples that have been collected and experimentally processed with non-identical methods, we attempted to control for collection and processing differences by accounting for them in the comparison sets using SVA. We included 7 BL cell cultures as well as two sBL primary biopsies from our sequencing set and observed proper clustering according to their eBL and sBL designations (Supplementary Fig. S5A). Differential gene expression analysis comparing only eBL and sBL primary biopsy samples resulted in 504 genes with significantly different expression profiles based on geographic BL subtype classification (Supplementary Table S2 and Supplementary Table S3). Leading edge analysis following the GSEA of differentially expressed genes reoccurring in multiple Gene Ontology (GO) gene sets demonstrated that the genes involving biological processes such as vasculature or blood vessel development (BH $P_{adj} = 6.0 \times 10^{-24}$) and angiogenesis (BH $P_{adj} = 4.0 \times 10^{-23}$) are the major variation source between our eBL biopsy collections with FNA and sBLs with FFB. These dominant enrichment sets are likely associated with the different biopsy collection techniques rather than pathological distinctions. On the other hand, the differentially expressed genes between eBL tumors and sBL tumors also resulted in significant enrichments in Hallmark gene sets including apoptosis, IL2/STAT5 signaling, Notch signaling, KRAS signaling, and TNFα signaling via NFκβ. Leading edge genes in these hallmark sets point to strong differential expression of the
PI3K-Akt signaling pathway \( (BH \ P_{adj} = 3.0 \times 10^{-23}) \) which plays a central role in BL pathogenesis/oncogenesis (36,37).

Given EBV presence is highly correlated with eBL tumors, we hypothesized that EBV may be a major determinant affecting differential expression between BL tumor subtypes. Therefore, we stratified our samples sets by their EBV content. Hierarchical clustering of the sample correlations demonstrates that we successfully preserved the variation associated with only BL tumors' EBV status and removed other unwanted covariates (Supplementary Fig. S5B). As a sign of this, three EBV positive BL cell cultures as well as 4 EBV positive sBL tumors clustered with the rest of the EBV positive eBL tumors. Confirming this stratification, two EBV negative eBLs and two sBLs from our sequencing set properly clustered with the rest of the negative sBLs. We then performed differential gene expression and pathway enrichment analysis between the primary BL tumors, excluding the cell lines. This resulted in 1658 significantly differentially expressed genes between EBV positive and negative BL tumors (Supplementary Table S2). Increased number of significantly differentially expressed genes suggests that EBV presence in BL tumor affects host expression profile more dramatically than subtype designations based on geography. These differentially expressed genes highlighted functions in biological processes involving DNA replication, mismatch repair as well as cell cycle regulation pathways (Supplementary Table S3). Interestingly, gene set enrichment of the differentially expressed genes between EBV positive and negative BL tumors resulted a significant enrichment in one of the oncogenic signature gene sets which consists of genes down-regulated when PTEN was experimentally knock-down (FDR q=0.096). Fig. 4A shows the genes that have higher expression in EBV negative BLs compared to EBV positive BLs. This suggests that EBV positive BLs, regardless of their geographic origin, share a common mechanism in which PTEN is suppressed. Supporting this, enrichment of another gene set in which genes are upregulated through activation of mTORC1 (mTOR complex 1) suggests the loss of regulatory role of PTEN on this signaling pathway in EBV positive BL tumors (Fig. 4B). Combined these suggest increased activity of PI3K and subsequently AKT/mTOR pathway driving cell cycle and proliferation.

**Examination of transcript mutations**

We next explored the transcriptome for somatic mutations in eBL and compared it to previously sequenced sBL in order to investigate whether gene mutation frequencies diverge as well as gene expression
profiles. After excluding the known genomic variants (SNPs), a total of 2728 putative somatic mutations were determined across the 56 tumor samples (Supplementary Table S4 and Supplementary Fig. S6). Our carefully controlled variant detection allowed us to compare the gene mutation frequencies between eBL and sBL and clinical correlates (Supplementary methods). This resulted a total of 21 genes mutated in 4 or more (>7%) of the sporadic and endemic tumors. Interestingly, the number of mutations did not differ significantly between sBL and eBL with an average of per tumor 3.6 vs 4.1 genes mutated in eBL and sBL respectively ($P = 0.24$, t-test, 2-tailed). However, for the top ten most commonly mutated genes the difference was significant with 2.5 and 3.5 genes mutated per tumor in eBL and sBL, respectively ($P = 0.017$, t-test, 2-tailed). Two of the top 3 genes were equally mutated genes including MYC and DDX3X (Fig. 5A, pink-cyan bars). The mutation rates of the genes ID3, CCND3, TCF3, and SMARCA4 were notably less frequently mutated in eBL tumors accounting for the difference between eBL and sBL. ID3 was significantly different, mutated in 32% of the eBL compared to 67% of the sBL ($P = 0.007$, Fisher's Exact). Also, 36% of the sBLs carried mutations in CCND3 compared to 14% of eBLs ($P = 0.061$, Fisher's Exact).

Given that ID3, TCF3 and CCND3 are thought to be key drivers for sBL oncogenesis and eBLs do not carry mutations as frequent as sBLs, we hypothesized again that EBV may be already affecting these pathways abrogating the need for additional mutation. While EBV status is strongly correlated with endemic versus sporadic status, we re-analyzed the tumors based on the viral content to see the effect on the mutational spectrum. Interestingly, the difference in the frequency of the 10 most common genes differentiated further to 2.4 to 3.7 for EBV positive and negative tumors, respectively. Only 196 out of 10,000 permutations, in which 2 eBLs and 4 sBLs were randomly assigned as virus negative and positive, respectively at each iteration, equaled or exceed the difference observed between EBV positive and negative tumors ($P = 0.0198$, t-test, 2-tailed). While individual genes in the simulation did not reach significance, CCND3, SMARCA4, and TCF3 gene mutation frequencies showed further differentiation and reached significance by Fisher’s exact comparing EBV positive and negative tumors ($P < 0.05$ for each) (Fig. 5A, yellow-green bars). Overall, these findings suggest that the molecular feature of gene mutations correlates better with the presence or absence of EBV within the tumor. Further supporting this, EBV negative BL tumors carried significantly more TCF3 or
ID3 or CCND3 mutations ($P = 0.0021$) and there was only a single occurrence of more than one of these genes being mutated in EBV positive tumor compared with 11 EBV negative tumors.

Apart from ID3, TCF3 and CCND3, two key genes SMARCA4 and ARID1A involved in chromatin and nucleosome remodeling as part of the SWI/SNF complex are highly mutated. Interestingly, ARID1A was mutated in roughly equivalent levels in either categorization. In contrast, SMARCA4 demonstrates decreased gene mutation when comparing either eBL and sBL or EBV positive and negative. Furthermore, Ras homolog family member A, RHOA, gene was mutated in 14% of eBL tumors in a mutually exclusive manner with TCF3/ID3 mutations. In addition, when we examined the mutations of two small GTPases, GNA13 and GNA12, which are also recurrently mutated in DLBCL (diffuse large B-cell lymphoma) (38), this mutually exclusive mutation pattern among the BL tumors could be further extended to include mutations in GNA12 ($P = 0.005$, Fisher’s Exact) but not GNA13. While this potentially suggests an alternative path for tumor drives other than deregulated ID3/TCF3, CCND3 and RHOA were not mutually exclusive potentially suggesting that the effects of CCND3 may be independent of TCF3’s drive.

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

We identified several new genes that appear somatically mutated (Table 2). TFAP4, transcription factor AP-4, whose down regulation can protect against glucocorticoid-induced cell death (39) and appears to be involved in p21-myc regulation of the cell cycle (40). TFAP4 carried 5 mutations in its HLH (helix-loop-helix) domain of likely deleterious effect in addition to stop gain and start loss mutations (Supplementary Fig. S7). A new category of mutations seen in our analysis are genes involved in DNA repair including RAD50, PRKDC and MSH6. While such mutations are common in other cancers such as colorectal cancer, breast and ovarian cancer, and several autoimmune diseases (41–43), they have not been reported previously in BL. Another previously undocumented gene showing somatic mutation was BCL7A. It has only been previously implicated in lymphoma through the observation of complex rearrangements (44) and its overexpression was associated with Germinal Center (GC) phenotype in DLBCL (45). Interestingly, it is also a member of the SWI/SNF complex further highlighting the complex’s importance (46). Mutations in BCL7A, SMARCA4, and ARID1A were not observed to co-occur in either eBL or sBL. In addition, we detected mutations in the PRRC2C,
RPRD2, FOXO1, PLCG2 that are normally overexpressed in peripheral blood mononuclear cells (PBMCs) and lymph nodes.

We also examined the correlation of given mutations to clinical and molecular features. Supporting the lack of expression difference between tumor presentation sites, we observed no suggestive associations with mutations. A single gene GNAI2 has a potential association with in-hospital survival ($P = 0.021$) where one out of 18 survivors had a mutation in GNAI2 compared to three of five who died during initial hospitalization (Supplementary Table S4).

**Rates of Gene Mutations vary greatly based on EBV genome type**

EBV and other In terms of EBV type, PRRC2C, SMARCA4, PLCG2, and TFAP4 differed significantly ($P<0.05$) (Supplementary Table S4). Interestingly, for these genes, type 2 EBV tumors always had the higher proportion of mutations compared to eBL tumors containing EBV type 1 (Table 2). All of the mutations that TFAP4 carried were either deleterious (Met1?, Gln15*, Arg127*, Pro185Leu) or accumulated in the DNA-binding domain (Arg50Trp, Arg58Trp, Arg60Cys) (Supplementary Table S4). These suggest possible loss of function for the protein AP4 encoded by this gene mutated mostly in type 2 carrying eBLs (50%). SMARCA4 mutation rates in groups of BLs with type 1, type 2, and negative were roughly 5%, 38%, and 39%, respectively. Mutations in this gene were located in its important domains SNF2, helicase, and HSA domains (Supplementary Fig. S7). In general, the average number of mutated genes per tumor (including all 21 genes) was 2.9 in BL tumors infected with type 1 while 4.9 in BLs with type 2 EBV ($P < 0.01$, t-test, 2-tailed) (Fig. 5B). This was significant even excluding type 1 tumors that were sporadic and also the overall type 2 mutation rate was on par with that of EBV negative tumors. The only genes that appear to have significantly lower mutation rates in type 2 tumors compared to EBV negative tumors was ID3 and TCF3, which were on par with type 1. Overall, functions of genes with distinct mutation frequencies in these groups, in addition to the significantly different general mutation rates, supports type 1 EBV’s reputation regarding better transformation ability compared to type 2, which had almost equivalent levels of mutated genes per tumor as EBV negative BLs (4.9 and 4.4, respectively).
DISCUSSION

In this study of Kenyan children diagnosed with eBL, we examined the expression and mutational spectrum vis-a-vis clinical and molecular features as compared to publicly available data for sBL. We observed relative homogeneity of expression within tumors collected from our patient population suggesting no overt subtypes within eBL. We found minimal differences between tumors presenting in the jaw or abdomen but observed differences in expression that correlated with survival, viral presence and type of EBV.

We also detected previously undescribed somatically mutated genes and showed that the BL mutational spectrum appears to most greatly differ based on the type of EBV infecting the tumor rather than geographic origin of the patient. Tumors harboring EBV type 1 display a significantly different host mutational profile compared to BL tumors with EBV type 2 and without EBV.

Our first major finding is that eBL is a homogenous tumor with highly correlated expression profiles, regardless of tumor location within the body. This means there is no need to create diagnostic subdivisions based on tumor presentation site or refine staging based on gene expression profiles. This contradicts previous expression analyses that suggested greater heterogeneity within eBL compared to sBL or id-BL (47). Our unsupervised hierarchal clustering did not reveal major clades based on virus, viral type, in-hospital survival or tumor presentation site. Our analysis of tumor presentation site suggested minimal differences that could be explained by associated cellular microenvironment (e.g differences in endothelium presence, NOS3). On the other hand, gene expression comparison of tumors based on survival status of the patients revealed several candidate genes and gene sets providing potential prognostic biomarkers which are currently lacking for eBL. Survival rate difference could be attributed to delayed time to diagnosis of abdominal cases compared to more apparent facial tumors (48). Further studies are required to validate the clinical utility of such markers.

Second, we found that the differences in tumor mutational spectrum were more striking when categorized by viral presence or absence rather than geographic origin. This was also supported by stronger differences in expression profiles in key pathways as well as involving likely downregulation of PTEN affecting the central pathway of PI3K-Akt signaling followed by mTORC1 activation. It has been previously shown that EBV can modulate mTOR pathway by LMP2A (49). Given the limited viral gene expression in latency I, EBV could be interacting with key host regulators especially through viral miRNAs (50) or by
regulating cellular miRNAs expressions (51). It has been recently shown that EBV microRNA Bart6-3p can inhibit PTEN translation (52). Even though this could be the mechanism by which the virus interferes with key cellular pathways, the consequences of viral miRNA interactions with PTEN and whether translational inhibition or mRNA degradation in EBV positive BL needs further clarification. Our data shows that the mRNA transcription of PTEN itself is not significantly differentially expressed between EBV positive and negative BLs. However, Ambrosio et al found that protein levels of PTEN are significantly lower in EBV positive BLs compared to negatives (53). This suggests a mechanism in which viral miRNAs interact with PTEN causing a translational inhibition. Combined, this suggests that the virus plays a key role in oncogenesis beyond the likely role in potentiating the translocation. However, further functional assays are needed in order to validate viral compensation for the less frequent ID3, TCF3 or CCND3 mutations found in EBV positive eBL tumors.

Third, we discovered new mutations occurring in genes previously not reported in other BL studies. While many of these additional genes are mutated at low frequencies (<10%), they support the key roles of previously identified pathways (e.g. BCL7A as another member of SWI/SNF). These genes also implicate DNA repair in terms of oncogenesis where we identify 3 previously undescribed genes (MSH6, RAD50, and PRKDC) involved in double strand repair and nonhomologous end joining. The analysis presented here of BL patients from Kisumu, Kenya was underway when a similar study was published involving 20 Ugandan patients (24). Therefore, we reevaluated our analysis to determine if we could validate their findings. We found similar mutation rates in ID3 and TCF3 genes and associated these with the lack of EBV positivity. However, we observed less frequent RHOA and no CCNF mutations in Kenyan eBL tumors. In contrast to the Ugandan study, we also failed to detect any significant trace of other herpes viruses such as KSHV or CMV other than EBV although this may be attributable to our FNAs which decrease the sampling of the connective tissue were these viruses were mainly present. Their study was also limited in its ability to examine EBV types for which we found significant differences in expression within our Kenyan tumors.

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

Finally, we observed that the viral expression pattern is consistent with viral latency I where EBV is essentially quiescent and maintained with EBNA1 expression. However, increased detection of lytic gene expression was suggestive of poor prognosis. It has been argued that this observed expression pattern may primarily be due to various levels of lytic reactivation (54). Consistent with the previous reports (55), our results demonstrated similar heterogeneous viral gene expression in BL suggesting that tumor cells could be targeted by antiviral immunotherapies. Dysfunctional T cell immunity has been reported for children diagnosed with eBL who were defective for EBNA1 specific IFN-gamma T cell responses (56). This defect putatively allows latency I tumors to escape from immune surveillance. Interestingly, we found that one third of the eBL tumors which carried type 2 EBV genome had significantly suppressed immunoproteasome complex gene transcriptions compared to eBLs with type 1 EBV. One explanation for this novel observation could be that type 2 EBV more readily infects immunocompromised individuals. Baarle et al., reported an increased prevalence of type 2 EBV among HIV patients (57). However, a larger cohort of HIV patients showed that T-cell impairment does not sensitize individuals for type 2 EBV infections (58). Thus, an alternate explanation is that type 2 EBV directly interacts with host regulatory components in order to interfere with immunoproteasome complex formation. Reduced transcriptional expression of these genes in EBV type 2 eBL tumors implies a mechanism in which viral components (coding or non-coding) suppress \textit{JAK/STAT1} mediated transcription. This could be an additional mechanism by which type 2 EBV is able to escape from immune-
surveillance by preventing T-cell responses (59). Since we did not observe any viral transcriptional pattern differences compared to type 1 genes, this phenomenon requires further investigation to confirm.

The mutated genes we observed in BL tumors are also dysregulated or mutated in other cancer types with viral etiologies. DDX3X is activated by Hepatitis C Virus (HCV) leading to alteration of host cellular gene expressions (60). RHOA and CCND3 are dysregulated by Human T lymphophotropic virus type I (HTLV-I) (61). For the DNA tumor viruses, KSV tumors appear to be driven by viral programming in settings of immunocompromised with only a few described driver mutations including interleukin 1 receptor-associated kinase (IRAK1) in primary effusion lymphoma (62). In HPV-associated squamous cell cervical carcinoma the most common driver mutations are PIK3CA, EP300, TP53, FBXW7, and MAPK1 (63). PI3K mutations are common in epithelial derived EBV-positive nasopharyngeal carcinoma, with the most commonly somatically altered genes being TP53, CDKN2A/B, ARID1A, CCND1, SYNE1 and PI3KCA (64)(65). While these are comparable with BL the targeting of PI3K pathway, SWI/SNF and p53, the overall differences suggest that even between EBV malignancies the major factor in determining what genes are the lynch pins between normalcy and malignancy is the cell lineage and state. This concept is supported by the greater mutational commonality with other lymphomas and the fact that both virus positive and negative BL tumors have mutational commonality differing mainly in degree.

In summary, we have illustrated the key pathways implicated in BL oncogenesis integrating our analytical results with current literature (Fig. 5C). This expanded view of BL oncogenesis more clearly defines a role for EBV. Driving proliferation through PTEN/PI3K/AKT and CCND3 pathways may be a key step toward bypassing the lack of mutations in the TCF3/ID3/CCND3 axis that may include SWI/SNF interactions of SMARCA4 as well. MYC translocation provides the pivotal accelerant while gain of function mutations in CCND3 strengthens the pressure. Although the mutated genes functioning in B cell development and chromatin remodeling complexes might be contributing to this signaling, our results in terms of gene expression and pathway differences suggest a role for viral microRNAs which can inhibit PTEN function and cause activated BCR signaling via AKT. Other genes frequently mutated in BL play roles in distinct but relevant pathways such as DNA repair and focal adhesion. Overall, this combined model demonstrates
pathogenic mechanistic routes to tumorigenesis BL and introduces a defined role for EBV that warrants further interrogation.

ACKNOWLEDGEMENTS

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


28. Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res [Internet]. 2014;42. Available from: http://dx.doi.org/10.1093/nar/gku864


Table 1. Summarized clinical information for sequenced endemic BL tumors.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Total (N=28)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), median (range)</td>
<td>8.2 (2 - 14)</td>
</tr>
<tr>
<td>Gender, n (%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>20 (71%)</td>
</tr>
<tr>
<td>Female</td>
<td>8 (29%)</td>
</tr>
<tr>
<td>Tumor presentation site, n (%)</td>
<td></td>
</tr>
<tr>
<td>Abdomen</td>
<td>16 (57.1%)</td>
</tr>
<tr>
<td>Jaw</td>
<td>12 (42.9%)</td>
</tr>
<tr>
<td>In-Hospital Survival-Status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Died *</td>
<td>8 (28.6%)</td>
</tr>
<tr>
<td>Survived</td>
<td>16 (57.1%)</td>
</tr>
<tr>
<td>Died in remission ¥</td>
<td>2 (7.1%)</td>
</tr>
<tr>
<td>NA</td>
<td>2 (7.1%)</td>
</tr>
<tr>
<td>EBV infection status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>26 (92.8%)</td>
</tr>
<tr>
<td>Negative</td>
<td>2 (7.1%)</td>
</tr>
<tr>
<td>EBV Genome type, n (%)</td>
<td></td>
</tr>
<tr>
<td>Type I</td>
<td>18 (69.3%)</td>
</tr>
<tr>
<td>Type II</td>
<td>8 (30.7%)</td>
</tr>
</tbody>
</table>

* Only 5 of these patients started chemotherapy treatment, ¥ Cause of dead is either relapse or non-tumor related.
Table 2. Frequency of mutated genes in BL tumors classified based on EBV presence and genome type.

<table>
<thead>
<tr>
<th>Gene</th>
<th>EBV Type 1 (N=22)</th>
<th>EBV Type 2 (N=8)</th>
<th>EBV Negative (N=26)</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYC</td>
<td>54.5% (12)</td>
<td>75.0% (6)</td>
<td>73.1% (19)</td>
<td>v-myc myelocytomatosis viral oncogene homolog</td>
<td>A TF that drives cell cycle progression and transformation. Translocation key initiating step of BL. Hypermutated secondary to juxtaposition to IgH. Mutated in numerous cancers.</td>
</tr>
<tr>
<td>ID3</td>
<td>36.4% (8)</td>
<td>25.0% (2)</td>
<td>69.2% (18)</td>
<td>Inhibitor of DNA binding 3</td>
<td>A HLH protein lacking DNA-binding domain and functions as a negative regulator of TCF3. Mutations are inactivating which decrease TCF3 interaction.</td>
</tr>
<tr>
<td>DDX3X</td>
<td>31.8% (7)</td>
<td>62.5% (5)</td>
<td>42.3% (11)</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked</td>
<td>ATP-dependent RNA helicase. DDX3X is mutated in T-cell ALL, CLL, and medulloblastoma. Decreased expression in viral hepatic cellular carcinoma.</td>
</tr>
<tr>
<td>CCND3</td>
<td>9.1% (2)</td>
<td>25.0% (2)</td>
<td>38.5% (10)</td>
<td>Cyclin D3</td>
<td>A regulator of progression through G1 phase during cell cycle. Loss of C terminal domain leads to constitutive activation.</td>
</tr>
<tr>
<td>SMARCA4*</td>
<td>4.5% (1)</td>
<td>37.5% (3)</td>
<td>38.5% (10)</td>
<td>BRG1, SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4</td>
<td>Chromatin remodeler required for transcriptional activation. Functions in B cell maturation and maintenance of IgH and TCF3 open chromatin. Loss of function mutations. Also mutated in ovarian cancer.</td>
</tr>
<tr>
<td>TCF3</td>
<td>9.1% (2)</td>
<td>12.5% (1)</td>
<td>38.5% (10)</td>
<td>Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)</td>
<td>TF that plays a critical role in lymphocyte development. Mutations lead to gain of function.</td>
</tr>
<tr>
<td>TP53</td>
<td>18.2% (4)</td>
<td>12.5% (1)</td>
<td>23.1% (6)</td>
<td>Tumor protein p53</td>
<td>Tumor suppressor that is key driver of apoptosis at cell cycle checks. Mutations are loss of function. Ubiquitously mutated in numerous cancers.</td>
</tr>
<tr>
<td>Gene</td>
<td>Frequency 1</td>
<td>Frequency 2</td>
<td>Frequency 3</td>
<td>Frequency 4</td>
<td>Function Description</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>GNA13</td>
<td>18.2%</td>
<td>12.5%</td>
<td>15.4%</td>
<td>(4)</td>
<td>Guanine nucleotide binding protein, alpha 13 Functions as modulators of various transmembrane signaling systems for cell migration/homing. Loss of function mutations.</td>
</tr>
<tr>
<td>GNAI2</td>
<td>9.1%</td>
<td>37.5%</td>
<td>15.4%</td>
<td>(4)</td>
<td>Guanine nucleotide binding protein, alpha inhibiting activity polypeptide 2 Involved in hormonal regulation of adenylate cyclase upstream of PI3K. Likely loss of function mutations. Not implicated in other cancers.</td>
</tr>
<tr>
<td>TFAP4 *‡</td>
<td>4.5%</td>
<td>50.0%</td>
<td>15.4%</td>
<td>(4)</td>
<td>Transcription factor AP-4 (activating enhancer binding protein 4) A TF which can also activate viral genes by binding to certain motifs. Loss of function mutations.</td>
</tr>
<tr>
<td>ARID1A</td>
<td>18.2%</td>
<td>0.0%</td>
<td>11.5%</td>
<td>(3)</td>
<td>AT rich interactive domain 1A SWI/SNF complex protein member. Loss of function mutations. Mutated in gastric, NPC, ovarian and endometrial cancer.</td>
</tr>
<tr>
<td>FBXO11</td>
<td>9.1%</td>
<td>25.0%</td>
<td>11.5%</td>
<td>(3)</td>
<td>F-box protein 11 Substrate recognition component of a SCF (SKP1-CUL1-F-box protein) E3 ubiquitin-protein ligase complex. Major target is BCL-6. Loss of function mutations. Mutated in and Hodgkin and DLBCL.</td>
</tr>
<tr>
<td>MSH6</td>
<td>0.0%</td>
<td>0.0%</td>
<td>19.2%</td>
<td>(5)</td>
<td>MutS homolog 6 Functions in DNA mismatch repair system. Loss of mismatch recognition may lead to loss of cell cycle checkpoint. Likely loss of function mutations. Germline mutations increase risk of multiple cancers.</td>
</tr>
<tr>
<td>PRRC2C *‡</td>
<td>4.5%</td>
<td>37.5%</td>
<td>3.8%</td>
<td>(1)</td>
<td>Proline-rich coiled-coil 2C Limited info about the function. Over-expressed in PBMCs.</td>
</tr>
<tr>
<td>BCL7A</td>
<td>13.6%</td>
<td>12.5%</td>
<td>0.0%</td>
<td>(0)</td>
<td>B-cell CLL/lymphoma 7A SWI/SNF protein complex member. Mutational effects unclear. Mutated in other non-Hodgkin lymphomas.</td>
</tr>
<tr>
<td>FOXO1</td>
<td>9.1%</td>
<td>0.0%</td>
<td>7.7%</td>
<td>(2)</td>
<td>Forkhead box O1 Key TF regulated by the PI3K/AKT pathway. Loss of function mutations may have a role in cell growth or escape from apoptosis. Mutated in DLBCL.</td>
</tr>
<tr>
<td>Gene</td>
<td>Type 1</td>
<td>Type 2</td>
<td>Type 3</td>
<td>Function</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>PLCG2 *</td>
<td>0.0%</td>
<td>25.0%</td>
<td>7.7%</td>
<td>Phospholipase C, gamma 2 A crucial enzyme in BCR signaling upstream of PI3K/AKT pathway. Mutated frequently in melanoma.</td>
<td></td>
</tr>
<tr>
<td>PRKDC</td>
<td>9.1%</td>
<td>12.5%</td>
<td>3.8%</td>
<td>Protein kinase, DNA-activated, catalytic polypeptide Functions in DSBR and V(D)J recombination and repair of double strand breaks. Mutation effects unclear. Mutated in DLBCL.</td>
<td></td>
</tr>
<tr>
<td>RAD50</td>
<td>4.5%</td>
<td>12.5%</td>
<td>7.7%</td>
<td>Double strand break repair protein A component of MRN complex which functions in DSBR and through recombination or non-homologous end joining. Likely loss of function. Mutations observed in breast and ovarian cancer.</td>
<td></td>
</tr>
<tr>
<td>RHOA</td>
<td>13.6%</td>
<td>12.5%</td>
<td>0.0%</td>
<td>Ras homolog family member A Regulation of signal transduction between membrane receptors and focal adhesion molecules. Likely loss of function with potential for tumor spread metastasis.</td>
<td></td>
</tr>
<tr>
<td>RPRD2</td>
<td>9.1%</td>
<td>0.0%</td>
<td>7.7%</td>
<td>Regulation of nuclear pre-mRNA domain containing 2 Involves in gene expression and transcriptional initiation pathways. Mutation effects unclear. Mutations not observed in other cancers.</td>
<td></td>
</tr>
</tbody>
</table>

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

Figure 1

Expression heatmap for all known EBV genes for 26 eBL, 4 sBL and 3 long-term BL cultures (Daudi, Raji, Namalwa) that were found to be EBV positive. This correlation based clustering heatmap using log2 transformed FPKM values demonstrates a predominant expression pattern resembling latency I for most of the BLs while two eBLs (eBL_23 and eBL_25) and cell lines have elevated expression in other genes. eBL_02 and eBL_20 show intermediate level lytic genes such as BMRF1, BALF2, and BSLF2/BMLF1 in addition to the two eBLs that cluster with cell lines.

Figure 2

Sample to sample clustering of BL tumors based on expression profiles of top 10,000 genes with highest correlation of variation (CV) values (calculated using regularized log transformed expression data). While 2 sBLs (sBL_u1 and sBL_u2) separate out from 28 eBLs, eBL tumor expression profiles demonstrate greater correlation within eBLs ($r > 0.95$, Pearson correlation; Dark red is 1.0) compared to sBLs, which might be due to differences in biopsy and preservation methods or biology. Overall gene expression correlations between eBLs does not reveal significant clustering consistent with no major underlying molecular subtypes nor clustering correlating with tumor presentation site, treatment outcome, or EBV type.

Figure 3

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

**Figure 4**

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

**Figure 5**

Mutational landscape of BL tumors. **A)** This panel demonstrates mutated gene distribution in each tumor sample (columns) are tumors and rows are top frequently mutated genes (>10% of
samples mutated at least once). Tumor samples were grouped based on their EBV content and second color bar shows the subtype of the tumor. Red squares represent mutated gene while blue is for no mutation detected. Barplot on the right measures the frequency of mutated tumor samples and compares regarding the subtype of BLs (Percent frequency). Similarly, barplot on the left compares the mutated tumor frequencies for each gene stratified by EBV status (* $P < 0.05$, Fisher’s Exact). **B** Average number of mutated genes per BL tumor by EBV type. Error bars represent standard error (*** $P < 0.01$, t-test). **C** Schematic overview of the proposed key pathways and frequently and mutated genes in eBL pathogenesis. Genes in the boxes are found to be frequently mutated (gain of function and likely loss of function, represented by green and light-red boxes, respectively). Likely key interactions with EBV components are shown are in red connections. Possible interactions are shown in grey.
Figure 2

EBV Type

Tumor Site

Treatment Outcome

Abdomen
Jaw
NA
NotClear
Not_Received_Died
Received_Died
Received_Survived

EBV Type

Negative
Type_2
Type_1

Correlation

1
0.98
0.96
0.94
0.92
Figure 3

A

B
Figure 4

A

**PTEN_DN.V2_DN**

Running Enrichment Score (RES)

- Peak at 2959
- Zero crossing at 9437

```
EBV_Negative
```

```
EBV_Positive
```

ES = 0.438  NES = 1.74 Nom. p-val= 0  FWER= 0.122  FDR= 0.0959

B

**HALLMARK_MTORC1_SIGNALING**

Running Enrichment Score (RES)

- Peak at 12090
- Zero crossing at 9437

```
EBV_Negative
```

```
EBV_Positive
```

ES = -0.439  NES = 1.54 Nom. p-val= 0  FWER= 0.461  FDR= 0.151
Figure 5

A

Percent Mutated Tumors

EBV Status
Positive
Negative

Subtype
EndemicBL
SporadicBL

Percent Mutated Tumors

80 60 40 20 0

* *

B

Number of Mutated Gene per Tumor

EBV Type 1 EBV Type 2 Negative

***

C

DNA Repair
RAD50 PRKDC MSH6

Focal Adhesion
GNA13 RHOA

Escape repair & checkpoints
BCL6 FBOX11

Invasion / Metastasis

Evading Apoptosis

Growth / Proliferation

MYC Translocation

BCR / B cell Dev.

GNAI2 PLCG2

Chromatin Remodelling

SMARCA4 BCL7A AIRD1A
Molecular Cancer Research

Comprehensive Transcriptome and Mutational Profiling of Endemic Burkitt Lymphoma Reveals EBV Type-specific Differences


Mol Cancer Res  Published OnlineFirst January 30, 2017.

Updated version
Access the most recent version of this article at:
doi:10.1158/1541-7786.MCR-16-0305

Supplementary Material
Access the most recent supplemental material at:
http://mcr.aacrjournals.org/content/suppl/2017/06/21/1541-7786.MCR-16-0305.DC1
http://mcr.aacrjournals.org/content/suppl/2017/07/12/1541-7786.MCR-16-0305.DC2

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link http://mcr.aacrjournals.org/content/early/2017/01/28/1541-7786.MCR-16-0305. Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.