The tumor virus landscape of AIDS-related lymphomas

Aaron Arvey¹, Akinyemi Ojesina²*, Chandra Sekhar Pedamallu², Gianna Ballon⁴, Joonil Jung², Fujiko Duke², Lorenzo Leoncini³, Giulia De Falco³, Eric Bressman⁴, Wayne Tam⁴, Amy Chadburn⁵, Matthew Meyerson², Ethel Cesarman⁴

1. Memorial Sloan Kettering Cancer Center, New York, NY
2. Broad Institute of MIT and Harvard, Cambridge, MA and Dana-Farber Cancer Institute, Boston, MA
3. University of Siena, Siena, Italy
4. Weill Cornell Medical College, New York, NY.
5. Northwestern University, Chicago, IL
* These authors contributed equally

Corresponding Author:
Ethel Cesarman
ecesarm@med.cornell.edu
Dept of Pathology and Laboratory Medicine
Weill Cornell Medical College
1300 York Ave
New York, NY 10065

Running title: Viral contribution to AIDS Lymphoma

Classification: Biological Sciences – Microbiology, genetics, systems biology

Keywords: Oncovirus, tumor virology, lymphoma, Epstein-Barr virus, HIV
**Key Points:**

We evaluated a cohort of ARL samples for known and novel tumor viruses, revealing the oncogenic herpesvirus EBV as the sole infectious agent.

Heterogeneous viral gene expression suggests that variable host immunosurveillance of EBV latency may influence lymphomagenesis.
ABSTRACT

Immunodeficiency dramatically increases susceptibility to cancer as a result of reduced immune surveillance and enhanced opportunities for virus-mediated oncogenesis. While AIDS-related lymphomas (ARLs) are frequently associated with known oncogenic viruses, many cases contain no known transforming virus. To discover novel transforming viruses, we profiled a set of ARL samples using whole transcriptome sequencing. We determined that Epstein-Barr virus (EBV) was the only virus detected in the tumor samples of this cohort, suggesting that if unidentified pathogens exist in this disease, they are present in fewer than 10% of cases. To evaluate the role of EBV in ARL pathogenesis, we analyzed viral gene expression and found highly heterogeneous patterns of viral transcription across samples. We also found significant heterogeneity of viral antigen expression across a large cohort, with many patient samples presenting with restricted type I viral latency, indicating that EBV latency proteins are under increased immunosurveillance in the post-CART era. Furthermore, EBV infection of lymphoma cells in HIV-positive individuals was associated with a distinct host gene expression program. These findings provide insight into the joint host-virus regulatory network of primary ARL tumor samples and expand our understanding of virus-associated oncogenesis. Our findings may also have therapeutic implications, as treatment may be personalized to target specific viral and virus-associated host processes that are only present in a subset of patients.
INTRODUCTION

Evidence that viral infection could lead to the development of cancer came early in the 20th century and many associations between viruses and hematologic malignancies have been identified in a variety of animal models. It has been estimated that close to 12% of all cancers in humans worldwide are caused by viral infection\(^1\). However, only a few viruses have been shown to genuinely cause cancer in humans, and only three have been shown to directly cause lymphomas: Epstein-Barr virus [EBV; formally designated human herpesvirus 4 (HHV-4)], Kaposi sarcoma-associated herpesvirus [KSHV; human herpesvirus 8 (HHV-8)], and human T cell lymphotropic virus 1 (HTLV-1).

Immunodeficient individuals have greatly increased cancer incidence, the pathogenesis of which is thought to be related to disrupted immune surveillance, chronic antigenic stimulation, genetic alterations, cytokine dysregulation and viral infection\(^2\)-\(^4\). The viral contribution to these malignancies has been extensively studied and two tumor viruses have been discovered after a targeted search based on the epidemiologic association of specific cancer types and AIDS: KSHV and Merkel cell polyomavirus\(^5,6\). These previous findings have raised the likelihood of novel tumor viruses being discovered in the context of immunosuppression. Improved combined antiretroviral therapies (CART) as HIV treatment options have resulted in decreased immunosuppression and concomitantly reduced incidence of the highly immunogenic virus-associated malignancies, such as Kaposi’s sarcoma\(^7\)-\(^10\). As a result, the proportion of non-Hodgkin’s lymphomas has increased as a fraction of the total number of AIDS-related cancers\(^11\).

Non-Hodgkin’s lymphomas constitute more than 50% of all AIDS defining cancers in developed countries and are the most common cause of cancer-related death in HIV-infected individuals\(^12\). AIDS-related lymphomas (ARLs) are phenotypically and histologically similar to lymphomas of the immunocompetent; however, ARL cases are more frequently associated with known virus-mediators of oncogenesis\(^13\)-\(^15\). For example, while EBV is present in ~30% of all ARLs, it is more common among DLBCL with immunoblastic morphology (over 80%) than those with centroblastic morphology (20-
40%), and it is present in the vast majority of DLBCLs with presentation in the central nervous system\textsuperscript{16}. In addition, approximately 5\% of ARLs are associated with KSHV infection. However, the remaining ~60\% of ARLs have no known viral contribution to lymphomagenesis. Given that non-Hodgkin’s lymphoma is increased ~15-fold in the context of AIDS\textsuperscript{17}, this implies that >55\% of ARLs have an etiology directly related to the immunosuppressed host and cannot be ascribed to any known virus.

The permissive environment of the moderately immunosuppressed host and lack of known oncoviruses in a majority of cases motivated us to reexamine the viral and cellular transcriptomes of ARLs, in search of novel transforming viruses.

**METHODS**

*Patient samples*

Seventy-two formalin-fixed paraffin-embedded (FFPE) ARL samples were collected from New York Presbyterian Hospital–Weill Cornell, the University of Siena, and the AIDS Malignancy Consortium. Cases were included if > 80\% tumor cells were present and the diagnosis of B-cell lymphoma was confirmed. All samples were obtained with the approval of the institutional review boards at both institutions. Research was conducted in accordance with the Declaration of Helsinki. Sixteen cases with frozen tissue were collected from New York Presbyterian Hospital–Weill Cornell and confirmed to contain >80\% tumor cells. The 16 frozen and 9 of the FFPE samples were subjected to RNA extraction and transcriptome sequencing. Independent validation cases shown in Table 2 and Figure 2 were collected from the Hematopathology Laboratory of Weill Cornell Medical College Pathology/The New York Presbyterian Hospital (40 cases), from the AIDS Malignancy Consortium (12 cases), The University of Siena (20 cases), and the AIDS Cancer Specimen Resource (140 cases). Tissue diagnosis was made with the use of criteria from the World Health Organization\textsuperscript{18}. In cases of DLBCL, germinal center B-cell (GCB) versus non-GCB subtype was determined by the Hans algorithm\textsuperscript{19,20}. Our study of 25 ARL samples is well powered to contain at least one sample with a virus if that virus occurs in >10\% of total ARL cases (statistical power > 0.9).
Histology

Epstein-Barr Virus Probe ISH Kit (Leica Microsystems, Wetzlar, Germany; Vision BioSystems Novocastra, Newcastle-upon-Tyne, UK) was used for in situ hybridization for Epstein-Barr virus (EBV) RNA (EBER). Cases were considered positive when > 20% neoplastic cells were immune reactive, except for BCL-2 where cases were considered positive when > 50% of tumor cells had moderate to strong positivity.

Immunohistochemical studies were performed using monoclonal antibodies to CD-10, BCL-2, BCL-6, MUM-1(3H2E8; Santa Cruz Biotech, Santa Cruz, CA) and Ki-67 (MIB-1; DakoCytomation, Carpenteria, CA). Nuclear Ki-67 expression was semi-quantitative, and assessed as percentage of positive tumor cells. For EBV latency assessment, immunohistochemistry was done with antibodies to LMP1 (clones CS1-4, Abcam and clone OT21C, kind gift from Jaap Middeldorp from the VU University Medical Center in The Netherlands), EBNA2 (clone PE2, Dako), and LMP2A (15F9, Abcam).

Sequencing and analysis

Sequenced cDNA libraries were constructed from 25 cases using random hexamer primers, ligated to Illumina adaptors, and subjected to high throughput Illumina sequencing to generate 17-150 million 76-100bp paired-end sequence reads per sample. Reads were aligned to the EBV (HHV4, GenBank accession NC_007605) and human (UCSC hg19) genomes using Bowtie2 with the default “sensitive” parameters for alignment21. This approach allows multiple read alignment mismatches, which increases coverage of diverse EBV genotypes. We also aligned using BWA22 (both default and increased sensitivity parameters) and obtained similar results (Table S1). Reads that mapped to the viral genome with low alignment score or to the human genome were considered of ambiguous origin and discarded, as previously described23.

To detect other known and novel viruses, the PathSeq pipeline was employed, as previously described24. Briefly, reads were aligned against all known viral genomes, as downloaded from NCBI, using Bowtie2. To mitigate mono-clonal reads inflating counts, we compress all reads that align to a single position in the genome to a single read. This
provides more reliable evidence of presence/absence of a given virus or gene transcript, but may decrease the dynamic range of estimated gene expression for highly expressed transcripts. In search of novel pathogens, high quality unalignable reads were formed into contigs using Trinity\textsuperscript{25}.

Viral RNA gene expression values were estimated using cufflinks as reads per kilobase per million reads aligned (RPKM) across all known transcriptional units, as previously described\textsuperscript{23}. Cufflinks estimates overlapping transcript abundance by the read coverage of unique portions of each transcript. Track plots show read overlap counts normalized to total reads aligning to the viral genome. Normalized heatmaps show RPKM divided by $-\log(f)$, where $f$ is the fraction of reads uniquely aligning to the viral genome.

Human RNA gene expression was estimated using Cufflinks\textsuperscript{23}. Human genes without significant expression (i.e., those genes with RPKM < 1) were removed from subsequent analyses, resulting in ~8000 genes for analyses. Principal component analysis clustering identified histological subtype (DLBCL and BL) and sample preparation method (fresh frozen or archival FFPE) as covariates (Figure S5A,B). We thus used cuffdiff to analyze differential gene expression by subtype and by sample preparation method. Consistency of gene expression signatures was determined through overlap of significant differentially expressed genes (e.g., Figure 4C), scatter plot correlation (e.g., Figure S5C), and most differentially expressed rank overlap (e.g., Figure S5D). Overlap statistical enrichment was estimated using the hypergeometric distribution and implemented as a Fisher’s exact test. Human genome-wide gene expression was summarized by PCA on genes with variance greater than 0.01.

**RESULTS**

*AIDS-related lymphoma cohort description, histology, and sample selection*
To understand the viral landscape of ARL, and determine if previously unidentified oncogenic pathogens are present in this disease, we established a tissue sample cohort (n=88 distinct patients) that could be assayed for viral RNA. We selected 25 cases with tissue blocks available for accurate histological subclassification by WHO criteria, which included 19 diffuse large B cell lymphomas (DLBCL), 3 Burkitt’s lymphomas (BL), 2 follicular lymphomas, and one plasmablastic lymphoma (PBL) (Table 1). Six EBV positive cases, as determined by EBER in situ hybridization, were sequenced and analyzed blindly as a way to assess our ability to detect this virus. Due to known double viral infection in ARL (e.g., EBV and KSHV in primary effusion lymphomas), we also analyzed the EBV+ cases for novel pathogens. We excluded KSHV-positive cases from our cohort.

**PathSeq identifies Epstein-Barr virus as the only virus present in ARL samples**

To identify the viral populations in ARL samples, we performed massively parallel whole transcriptome sequencing of total RNA, enabling detection of non-poly-adenylated and other RNA species in addition to coding transcripts. This data was then analyzed using the PathSeq pipeline, which subtracts all reads that could potentially align to the human genome, aligns remaining reads to all known viruses, and assembles unaligned reads into contigs

To assess specificity and sensitivity of our assay, we performed a double-blind control where we analyzed RNA-seq data of samples that were positive and negative for EBV. We determined a simple read cutoff statistic, which was able to identify EBV in all samples that tested positive by EBER RNA in situ hybridization while maintaining perfect specificity (Figure 1B). The samples that tested negative for EBER, but had reads that aligned to EBER and other viral genes (Figure S1A), were found to contain tumor infiltrating EBV-infected lymphocytes (Figure 1C). This confirmed that PathSeq coupled with total RNA sequencing is a sensitive and specific assay for presence of viral RNA.

We next determined if other known viruses were present in our cohort. We identified a subset of samples that contained reads aligning to genic regions of known human
adenovirus C and polyomaviruses (Table S2). Based on previous experience and reports, we tested the sequence libraries for contamination\textsuperscript{26,27}. Subsequent PCR analysis of cDNA sequencing libraries and DNA re-extracted from the original tumor samples revealed that neither adenovirus nor polyomavirus were present in our tumor samples. Namely, the viral-aligning reads were assembled into >100nt contigs that could be efficiently amplified by PCR. Gel electrophoresis revealed that the sequencing libraries were positive for viral sequence, but the original and re-extracted cDNA were negative (Figure S1B-E). Conversely, PCR detected multiple EBV transcripts in the re-extracted cDNA (data not shown). Thus, the sequences represent low-level contamination in the libraries and the samples were in fact devoid of known adeno- and polyomaviruses.

To detect novel viruses that may be contributing to AIDS-related lymphomas, we assembled all unalignable reads into contigs. While many contigs were formed, all contigs aligned to known human, viral, or bacterial genomes (Table S3). To increase statistical power and sensitivity for detection, reads across all tumor samples were aggregated; however, this again resulted in contigs with negligible potential for encoding novel viral agents.

\textit{Epstein-Barr virus gene expression in primary ARL samples is heterogeneous}

To characterize viral persistence and potential maintenance of lymphomagenesis, we characterized the viral RNA and protein antigen landscape through transcript expression quantification, \textit{in situ} hybridization, and immunohistochemistry. EBV-mapping reads aligned to known transcribed regions of the viral genome with the most prominently expressed transcripts being the RNA Pol III transcribed EBER1 and EBER2 transcripts along with type I latent associated EBNA1 transcripts (Figure 2A). Additional type II and III viral latency gene transcripts were expressed, including immunogenic LMP1 and LMP2 (Figure 2B, 2C) and an unexpected proportion of samples contained lytic transcripts, most notably being transcribed at the BHLF1 locus (Figure 2D,S2A). Consistent with previous results\textsuperscript{28}, our findings suggested a diversity of latency-associated gene expression
programs with a subset of cells exhibiting expression of genes corresponding to the lytic program.

It was unclear if the viral protein antigen landscape would reflect the heterogeneity of RNA gene expression, and how intratumoral heterogeneity would contribute to this phenomenon. To characterize viral antigens and intratumoral heterogeneity, we stained for the immunogenic LMP1, LMP2A, and EBNA2 proteins using immunohistochemistry (IHC) on ARL tissue microarrays. Many EBV+ cases were positive for at least one of these proteins (Table 2, Figure 2E, S2B); however, there was significant diversity in the intensity of protein staining and the proportion of tumor cells that stained positive (Figure S2C). Thus, at both the RNA and protein level, EBV latent gene expression is heterogeneous in EBV-positive ARL, and does not necessarily conform to classical latency.

Given the heterogeneity of the viral transcriptional and protein landscape, we sought to confirm our findings across a larger cohort of ARL cases. We characterized viral latency type by LMP1 and EBNA2 immunohistochemistry in a validation cohort of 212 ARL cases (Table 2), 94 (44%) of which were positive for EBV as determined by EBER ISH. While we expected a large proportion of DLBCLs to have a viral type II or III expression pattern, as suggested by previous studies,29,30 most cases were restricted to viral latency I programs, especially the majority of tumors with germinal center (GC) B cell origin. Consistent with previous studies, most cases of BL had a latency I pattern.

It was unclear if the EBV transcriptional regulation patterns were unique to the immunosuppressed host environment of ARL. Thus, we performed a meta-analysis of the publicly available viral transcriptome data in HIV-negative sporadic Burkitt’s lymphomas (sBL) 31, which are thought to express type I latency to limit immunogenicity. We found that EBV+ sBL samples presented with robust expression of EBNA1 and other viral genes compatible with type I latency, such as RPMS1 (Figure 3A, S3A). Consistent with previous reports,28,32,33 we also found significant RNA expression of LMP2 in all EBV-positive sBL samples and a significant subset also expressed LMP1, BHLF1, BHRF1, BMLF1, and BMRF2 (Figure 3B, S3B).
We confirmed the active translation of the viral RNA diversity in a subset of tumors of immunocompetent individuals. Immunohistochemistry staining demonstrated presence of LMP1 after stringent positive and negative controls (Figure 3C). We next compared AIDS-related BL to HIV-negative sBL by IHC staining for LMP2A as a test for latency II or III viral gene expression. We assembled a cohort of 84 EBV \(^+\) ARL and 59 EBV \(^+\) sBL cases and found that 3/59 EBV \(^+\) sBL samples and 9/84 ARL cases showed robust LMP2A staining (Table 3). Taken together, these findings demonstrate a broad set of transcriptional latency programs \textit{in vivo} in both the immunocompetent as well as the immunocompromised host.

\textit{EBV-positive ARL has a distinct host gene expression program}

Since viral gene products influence host gene expression patterns that can drive oncogenesis and inhibit apoptosis, we characterized the host expression program associated with the presence of EBV. Total RNA-seq reads were mapped to the human genome and expression values were estimated using cufflinks as reads per kilobase per million reads aligned (RPKM)\(^34\). We confirmed that expression of host genes previously identified as enhanced by EBV were upregulated in EBV-positive cases. These genes included the EBNA3A response gene HSP70B \((HSPA7)\)^35 and CD30 \((TNFRSF8)\), which acts as a biomarker for EBV-titer in immunosuppressed lymphomas and infectious mononucleosis\(^36\) (Figure 4A, S4A-C). Genome-wide, hundreds of host genes were quantitatively associated with EBV in at least one lymphoma histological subtype (DLBCL or BL) (Figure 4B, S5A-B; Table S4; Methods). These findings were robust to leave-one-out validation and host genes were associated with EBV in independent sets of samples (Figure S5C). Intriguingly, differentially regulated genes were significantly similar across the histological subtypes, supporting a consistent role for EBV across multiple ARL malignancies (Figures 4C, S5D).
We next wanted to understand how the immunodeficient environment influenced the EBV-host regulatory network. We compared the EBV-associated host gene expression changes in ARL versus lymphomas of immunocompetent individuals by analyzing EBV-positive HIV-negative sBL cases	extsuperscript{31}. We first confirmed the existence of a robust EBV-signature in sBL samples (Figure S6A). Host-genes differentially expressed in the presence of EBV in sBL had highly statistically significant overlap with the EBV-associated gene expression program in BL of HIV	extsuperscript{+} subjects, but not DLBCL (Figure S6B,C). To determine if the EBV-associated host gene expression program was enriched for lymphomagenesis pathways, we performed a gene set enrichment analysis on the molecular signatures database MSigDB3.0	extsuperscript{37}. EBV-positive samples had statistically significant differential regulation of genes with altered expression in the context of plasma cell differentiation, maintenance of hematopoietic stem cells, and primary effusion lymphomas (Figure 4D, S7). There also existed a significant correlation between EBV and host gene expression in lymphoblastoid cell lines (LCLs)	extsuperscript{23}. When taken together, these four datasets capture the influence of immunosurveillance and malignant cell-type morphology on the EBV-associated host gene expression signature. Accordingly, we found that sBL and AR BL cluster as do the LCL and AR DLBCL pathway signatures. Viral response pathways (e.g., HBV response genes) were more prevalent in sBL than other samples; whereas EBV-associated enhancement of the plasma cell differentiation program was present only in DLBCL and LCLs.

**DISCUSSION**

Using the sensitive technology of total RNA-seq and the PathSeq analysis pipeline, we have evaluated the transcribed viral contribution to AIDS-related lymphoma. While no novel viruses were detected, we discovered that EBV exploits the immuno-permissive environment of ARL to express viral antigens that are associated with an altered host gene expression program. Meta-analysis of sporadic Burkitt’s lymphomas in immunocompetent individuals revealed that EBV transcriptional regulation is broadly more promiscuous than previously appreciated; however, the immunocompetent immune system exerts a significant selection on viral antigen expression.
While extensive literature has speculated on the contribution of viruses to increased cancer risk in the immunosuppressed\textsuperscript{2,4,38}, our observation that no additional pathogens were found in a cohort of 25 ARLs, suggests that if any new pathogens exist in this disease, they are relatively rare or tightly latent, without significant RNA expression (Methods). Since we examined steady state RNA sequences in established tumors, we cannot rule out the possibility that EBV or other viruses are transient enablers of cellular transformation, from which frank malignancy can expand and be subsequently maintained independent of initial viral contribution (i.e., through so called “hit-and-run” mechanisms)\textsuperscript{39}. The potential for this mechanism is illustrated by EBV rescue of crippled germinal center centrocytes, allowing cells with transforming translocations to escape apoptosis\textsuperscript{40-42}, which can be followed by loss of viral episome after \textit{in vitro} culture of EBV\textsuperscript{+} tumors\textsuperscript{43-45}. AIDS-associated malignancies may also be affected, in some cases, by increased activation of lymphocytes, resulting in greater frequency of DNA-damaged uninfected germinal center B cells\textsuperscript{46,47}. This indirect mechanism could be considered analogous to cancers associated with Hepatitis C or \textit{Helicobacter pylori} infections, which may trigger transformation through chronic inflammation, but pathogen sequences or proteins are not required for maintenance of tumorigenesis. Beyond hit-and-run pathogens, lack of immunosurveillance required for removal of sporadic pre-tumor dysplastic tissue may contribute to cancer susceptibility; however, attributing the increase of AIDS-defining cancers to decreased tumor surveillance begs the question of why there exists such an extensive risk for specific malignancies while non-AIDS defining cancers are increased only moderately in the HIV-positive or post-transplant immunosuppressed individual\textsuperscript{17}.

Our findings point towards a dramatic increase of immunoselective pressure on viral antigen in the post-CART era, as evidenced by our cohort being predominantly type I latent EBV\textsuperscript{+} DLBCLs whereas this histological subclass pre-CART was associated with a type III latency immunophenotype\textsuperscript{30}. Furthermore, distinct viral oncogenic mechanisms may be employed in specific ARL histological subclasses. For instance, most EBV\textsuperscript{+} non-GC DLBCLs (63\%) maintained type II or III latency, in contrast to only 24\% of GCB
DLBCLs, suggesting that more active EBV gene expression programs may be required to maintain the proliferative program of this tumor subclass.

Lastly, our results reveal a diversity of viral and host gene expression programs that may be therapeutic targets in only a subset of cases. Better integration of molecular diagnostics that include viral gene expression, as well as immunohistochemistry, has the potential to identify patients that may benefit from in-development anti-latent EBV protein therapies and interrupt the host-virus joint regulatory network.
Acknowledgements: This project was funded by the Starr Cancer Consortium and NIH grant 1RC2CA148317 to MM and EC. AIO was supported by the Rebecca Ridley Kry Fellowship of the Damon Runyon Cancer Research Foundation. The AIDS Malignancy Clinical Trials Consortium (AMC, NIH grant U01CA121947) and the AIDS Cancer Specimen Resource (ACSR, NIH grant 1UM1CA181255) contributed cases of ARL.

Author contributions: AA, AIO, MM, and EC designed research. AA, AO and CSP performed statistical and computational analyses. GB and EB analyzed primary samples and sequencing libraries for contamination and EBV. JJ provided computational assistance. FD prepared samples for RNA-seq. LL, GDF, and AC contributed ARL samples. AA and EC write the manuscript with input from AIO and MM.

Conflict of interest disclosure: The authors have no conflicts of interest to declare.

Correspondence: Ethel Cesarman, Dept of Pathology and Laboratory Medicine, Weill Cornell Medical College, 1300 York Ave New York, NY 10065. Email: ecesarm@med.cornell.edu.
Table 1. Virus- and host- aligning RNA-seq reads in AIDS-related lymphoma patient cohort. Columns shown are internal ID, histological subclass diagnosis, sample storage, EBV EBER ISH status, total RNA-seq reads, and number of stringently unique reads that align to human and EBV.

<table>
<thead>
<tr>
<th>ID</th>
<th>Diagnosis</th>
<th>Storage</th>
<th>EBV status</th>
<th>Total Reads</th>
<th>Unique Human reads</th>
<th>Unique EBV reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>LY002</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>40,642,580</td>
<td>4,179,804</td>
<td>4</td>
</tr>
<tr>
<td>LY003</td>
<td>BL</td>
<td>Frozen</td>
<td>EBV+</td>
<td>58,461,328</td>
<td>6,082,519</td>
<td>1,094</td>
</tr>
<tr>
<td>LY004</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>45,676,046</td>
<td>14,737,943</td>
<td>0</td>
</tr>
<tr>
<td>LY005</td>
<td>Follicular lymphoma</td>
<td>Frozen</td>
<td>EBV-</td>
<td>66,222,036</td>
<td>19,694,654</td>
<td>0</td>
</tr>
<tr>
<td>LY006</td>
<td>Follicular lymphoma</td>
<td>Frozen</td>
<td>EBV-</td>
<td>51,384,920</td>
<td>6,113,426</td>
<td>0</td>
</tr>
<tr>
<td>LY007</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>45,299,480</td>
<td>5,195,300</td>
<td>0</td>
</tr>
<tr>
<td>LY008</td>
<td>DLBCL</td>
<td>FFPE</td>
<td>EBV-</td>
<td>70,393,682</td>
<td>21,924,559</td>
<td>8</td>
</tr>
<tr>
<td>LY009</td>
<td>DLBCL</td>
<td>FFPE</td>
<td>EBV-</td>
<td>74,302,054</td>
<td>24,897,813</td>
<td>971</td>
</tr>
<tr>
<td>LY010</td>
<td>DLBCL</td>
<td>FFPE</td>
<td>EBV-</td>
<td>80,747,336</td>
<td>22,747,406</td>
<td>0</td>
</tr>
<tr>
<td>LY011</td>
<td>DLBCL</td>
<td>FFPE</td>
<td>EBV-</td>
<td>79,507,540</td>
<td>28,720,368</td>
<td>0</td>
</tr>
<tr>
<td>LY012</td>
<td>DLBCL</td>
<td>FFPE</td>
<td>EBV+</td>
<td>65,902,794</td>
<td>12,861,869</td>
<td>114,938</td>
</tr>
<tr>
<td>LY013</td>
<td>DLBCL</td>
<td>FFPE</td>
<td>EBV+</td>
<td>107,114,152</td>
<td>6,542,704</td>
<td>0</td>
</tr>
<tr>
<td>LY014</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>165,832,378</td>
<td>9,712,379</td>
<td>0</td>
</tr>
<tr>
<td>LY015</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>139,967,864</td>
<td>7,754,150</td>
<td>2</td>
</tr>
<tr>
<td>LY016</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>66,708,934</td>
<td>8,132,534</td>
<td>0</td>
</tr>
<tr>
<td>LY017</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>62,481,036</td>
<td>8,862,120</td>
<td>1,286</td>
</tr>
<tr>
<td>LY018</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV+</td>
<td>64,447,628</td>
<td>4,411,390</td>
<td>20</td>
</tr>
<tr>
<td>LY019</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>62,411,645</td>
<td>9,436,092</td>
<td>512</td>
</tr>
<tr>
<td>LY020</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>48,797,764</td>
<td>9,209,151</td>
<td>50</td>
</tr>
<tr>
<td>LY021</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>81,856,928</td>
<td>9,976,478</td>
<td>78</td>
</tr>
<tr>
<td>LY022</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>61,246,428</td>
<td>12,269,534</td>
<td>81</td>
</tr>
<tr>
<td>LY023</td>
<td>DLBCL</td>
<td>Frozen</td>
<td>EBV-</td>
<td>52,597,350</td>
<td>19,920,974</td>
<td>36</td>
</tr>
<tr>
<td>LY024</td>
<td>BL</td>
<td>FFPE</td>
<td>EBV-</td>
<td>52,770,210</td>
<td>30,181,777</td>
<td>23,584</td>
</tr>
<tr>
<td>LY025</td>
<td>PBL</td>
<td>FFPE</td>
<td>EBV+</td>
<td>16,461,068</td>
<td>10,901,688</td>
<td>66,897</td>
</tr>
<tr>
<td>LY026</td>
<td>DLBCL</td>
<td>FFPE</td>
<td>EBV+</td>
<td>63,740,374</td>
<td>16,484,786</td>
<td>7,318</td>
</tr>
</tbody>
</table>

Table 2. Pathological subclassification and EBV assessment by EBER ISH and immunohistochemistry for LMP1 and EBNA2 in AIDS-related lymphoma. These 212 cases were classified as latency I in EBER⁺ cases when no LMP1 or EBNA2 was expressed, as latency II when these were positive for LMP1 but negative for EBNA2, and as latency III where there was expression of both EBNA2 and LMP1.

<table>
<thead>
<tr>
<th>AIDS lymphoma subtype</th>
<th>No.</th>
<th>EBV positive (%) EBER-ISH</th>
<th>Immunephenotype</th>
<th>Latency I LMP1⁺EBNAN⁻ (% of EBV⁺)</th>
<th>Latency II LMP1⁺EBNA2⁻ (% of EBV⁺)</th>
<th>Latency III LMP1⁺EBNA2⁺ (% of EBV⁺)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLBCL nonGC</td>
<td>48</td>
<td>27 (56%)</td>
<td></td>
<td>10 (37%)</td>
<td>8 (30%)</td>
<td>8 (30%)</td>
</tr>
<tr>
<td>DLBCL GCB</td>
<td>98</td>
<td>25 (25%)</td>
<td></td>
<td>19 (76%)</td>
<td>3 (12%)</td>
<td>3 (12%)</td>
</tr>
<tr>
<td>DLBCL null</td>
<td>13</td>
<td>4 (31%)</td>
<td></td>
<td>1 (25%)</td>
<td>3 (75%)</td>
<td></td>
</tr>
<tr>
<td>BL</td>
<td>19</td>
<td>10 (53%)</td>
<td></td>
<td>9 (90%)</td>
<td>1 (10%)</td>
<td></td>
</tr>
<tr>
<td>PBL</td>
<td>9</td>
<td>8 (89%)</td>
<td></td>
<td>8 (100%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PEL (solid variant)</td>
<td>14</td>
<td>12 (86%)</td>
<td></td>
<td>11 (92%)</td>
<td>1 (8%)</td>
<td></td>
</tr>
<tr>
<td>BCL-U</td>
<td>4</td>
<td>3 (75%)</td>
<td></td>
<td>3 (100%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polymorphic LPD</td>
<td>7</td>
<td>5 (71%)</td>
<td></td>
<td>1 (20%)</td>
<td>1 (20%)</td>
<td>3 (60%)</td>
</tr>
<tr>
<td>Total</td>
<td>212</td>
<td>94 (44%)</td>
<td></td>
<td>61 (65%)</td>
<td>15 (16%)</td>
<td>17 (18%)</td>
</tr>
</tbody>
</table>

Controls: 8 BL and 13 DLBCL (7 GC, 6 nonGC). All HIV-negative control DLBCLs were negative for EBV, and 5 cases of HIV-negative BL were EBV⁺ (all negative for LMP1 and EBNA2). Abbreviations: DLBCL: diffuse large B cell lymphoma, BL: Burkitt’s lymphoma, PBL: plasmablastic lymphoma, PEL: primary effusion lymphoma, BCL-U: B cell lymphoma, unclassifiable, with features of DLBCL and BL, LPD: lymphoproliferative disorder, GC: germinal center, DLBCL null: negative for CD10, BCL6 and MUM1.
Table 3. LMP2A assessment by IHC in AIDS-related and HIV-negative Burkitt’s lymphoma

<table>
<thead>
<tr>
<th></th>
<th>EBV positive by EBER-ISH</th>
<th>LMP2A IHC positive (% of EBV-positive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV-positive</td>
<td>84</td>
<td>9 (11%)</td>
</tr>
<tr>
<td>HIV-negative</td>
<td>59</td>
<td>3 (5%)</td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1. Transcriptome sequencing coupled with the PathSeq computational analysis pipeline is highly sensitive and specific for detecting presence of novel and known viruses in primary AIDS-related lymphoma samples. (A) The PathSeq pipeline profiles total RNA by random hexamer priming and high throughput sequencing. Reads that do not map to the human genome are aligned against known viral and bacterial genomes. The remaining unmappable reads are assembled into contigs to detect transcripts originating from novel viruses. (B) PathSeq is sensitive and specific for detecting the Epstein-Barr virus in tumor cells. Viral read count fraction (y-axis) is shown for all ARL samples (x-axis). Using a threshold of 0.01%, we were able to uniquely identify all EBV-positive cases, which were confirmed by in situ hybridization for the EBV encoded RNA 1 (EBER1) transcript. A complete listing of read counts, viral ISH detection, and case identifiers can be found in Table 1. (C) Samples with less than 0.01% of reads aligning to EBV were profiled by in situ hybridization for EBER1 to determine if any of these cases were mislabeled as EBV-negative by PathSeq. While scattered EBV+ cells infiltrated the tumor, tumor cells were in fact negative for EBV. This is exemplified by cases LY09, LY20, and LY24 with morphologically non-transformed tumor infiltrating lymphocytes highlighted. LY18 is shown to demonstrate EBER1 staining of EBV+ samples with EBV RNA in tumor-cells.

Figure 2. Transcriptional regulation of Epstein-Barr virus gene expression demonstrates diversity of latency types. (A) Gene expression is shown as read counts per million reads aligned (y-axis) at canonical type I latency genes EBER1, EBER2, and EBNA1. Rows show individual samples and are colored alternating red/blue to emphasize sample-track pairing. (B) Immunogenic viral genes associated with type II and III latency are also expressed, such as LMP1. (C) EBV gene expression is highly diverse across ARL samples. The heatmap shows gene expression profiles of latency-associated genes (y-axis) across samples (x-axis). Gene expression is in units of RPKM normalized by fraction of reads aligning to EBV (Methods). (D) Genes associated with lytic reactivation, such as transcripts at the BamHI H locus, were expressed in primary ARL tumor samples. (E)
Immunohistochemical staining for viral proteins EBNA2 and LMP1 was used to classify viral gene expression latency in ARL cases. A BL sample with latency I (top) and a DLBCL sample with latency III (bottom) are representative examples. A complete analysis can be found in Table 2.

Figure 3. Viral antigen expression variation is comparable in immunodeficient and immunocompetent hosts. (A) Analysis of polyadenylated RNA-seq data confirms that viral expression is detectable in HIV-negative sporadic Burkitt’s lymphoma in immunocompetent hosts. Expression of latency I associated EBNA1 is robustly detected in primary samples (prefixed by SLN) and cell lines. (B) LMP1 and LMP2 RNA were detected in most primary HIV-negative sBL tumor samples. SLN2521 is excluded in the right panel due to low read counts. (C) Viral RNA is translated into immunogenic protein expression in a subset of cases (additional analysis in Table 3). IHC stains demonstrate that ARL (left) and sBL (right) cases contain highly immunogenic viral proteins. All samples are positive for the respective stains.

Figure 4. EBV-positive ARL cases have a distinct host gene expression signature. (A) HSPA7 is an example of a gene that is more highly expressed in EBV+ samples (above) compared to EBV- samples (below). (B) Genes consistently up- and down-regulated in the presence of EBV are shown in DLBCL cases. Axes show log2 (x+1) where x is the gene expression level in units of RPKM as quantified by cufflinks. (C) EBV-associated host gene expression changes are significantly similar in AIDS-related BL and DLBCL. Significantly differentially expressed upregulated (top) and downregulated (bottom) genes in EBV-positive samples are part of the EBV-signature in both BL and DLBCL (Hypergeometric test). (D) EBV-associated host gene expression changes cluster into pathways, including plasma cell differentiation, maintenance of hematopoietic stem cells, and genes dysregulated in primary effusion lymphoma. Heatmap shows pathways (y-axis) enriched for host expression signatures associated with EBV in a variety of cellular and host settings (x-axis), including HIV-negative sporadic Burkitt’s lymphoma (sBL), AIDS-related BL (AR BL), AIDS-related DLBCL (AR DLBCL), and lymphoblastoid cell lines.
(LCL) derived from whole peripheral blood infected with high titers of EBV. Enrichment significance is determined by one-sided KS tests.
References:

Figure 1

(A) Schematic representation of the workflow for identifying viral contributions in AIDS-related lymphoma. Total RNA-seq was performed, and reads were divided into known and unknown origins. Known reads were further classified into viral and human reads, while unknown reads were subjected to contig assembly. Novel infectious agents were identified in the unknown reads.

(B) Bar graph showing the fraction of reads aligning to Epstein-Barr Virus (EBV) in tumor cells compared to in situ samples. The bars represent different ARL samples, with LY018, LY020, LY009, and LY024 shown.

(C) Representative images of immunohistochemical staining for EBV in different samples: LY18, LY09, LY20, and LY24.
Figure 3

A

B

C

HIV-positive AIDS-related lymphomas

HIV-negative Burkitt’s lymphomas

LMP1

LMP2
Figure 4

A. Expression levels of different EBV-Positive and EBV-Negative samples.

B. Scatter plot showing the log2(x+1) human gene expression for EBV+ DLBCL samples.

C. Venn diagram showing ARL EBV-associated up-regulated and down-regulated transcripts.

D. Heatmap displaying the differential gene expression across different cell types.

Figure 4
The tumor virus landscape of AIDS-related lymphomas

Aaron Arvey, Akinyemi Ojesina, Chandra Sekhar Pedamallu, Gianna Ballon, Joonil Jung, Fujiko Duke, Lorenzo Leoncini, Giulia De Falco, Eric Bressman, Wayne Tam, Amy Chadburn, Matthew Meyerson and Ethel Cesarman