Epstein-Barr Virus Genomes Reveal Population Structure and Type 1 Association with Endemic Burkitt Lymphoma

Yasin Kaymaz
University of Massachusetts Medical School

Let us know how access to this document benefits you.
Follow this and additional works at: https://escholarship.umassmed.edu/umccts_pubs

Part of the Genetic Phenomena Commons, Genetics and Genomics Commons, Hemic and Lymphatic Diseases Commons, Neoplasms Commons, Pediatrics Commons, Population Biology Commons, Translational Medical Research Commons, Virology Commons, and the Viruses Commons

Repository Citation

This material is brought to you by eScholarship@UMassChan. It has been accepted for inclusion in UMass Center for Clinical and Translational Science Supported Publications by an authorized administrator of eScholarship@UMassChan. For more information, please contact Lisa.Palmer@umassmed.edu.
Epstein Barr virus genomes reveal population structure and type 1 association with endemic Burkitt lymphoma

Yasin Kaymaz1, Cliff I. Oduor2,3, Ozkan Aydemir2, Micah A. Luftig4, Juliana A. Otieno5, John Michael Ong’echa2, Jeffrey A. Bailey3*, Ann M. Moormann6* 

1 Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA USA 
2 Center for Global Health Research, Kenya Medical Research Institute, Kisumu, Kenya 
3 Department of Pathology and Laboratory Medicine, Warren Alpert Medical School, Brown University, Providence, RI, USA 
4 Department of Molecular Genetics and Microbiology and Center for Virology, Duke University School of Medicine, Durham, NC USA 
5 Jaramogi Oginga Odinga Teaching and Referral Hospital, Ministry of Health, Kisumu, Kenya 
6 Division of Infectious Diseases and Immunology, Department of Medicine, University of Massachusetts Medical School, Worcester, MA 

*shared last authorship 

†Corresponding authors: Ann Moormann, PhD, MPH 

Department of Medicine 
Division of Infectious Diseases and Immunology 
University of Massachusetts Medical School 
364 Plantation Street, LRB 313 
Worcester, MA 01605 USA 
Office: 508-856-8826 
Email: ann.moormann@umassmed.edu

Jeffrey A Bailey, MD, PhD 
Department of Pathology and Laboratory Medicine 
Warren Alpert Medical School 
Brown University 
Providence, Rhode Island 
Office: 401-446-4652 
Email: jeffrey_bailey@brown.edu

Key words: Epstein Barr virus, genome sequencing, genetic variation, endemic Burkitt Lymphoma, EBV type 1, EBV type 2.
Abstract

Endemic Burkitt lymphoma (eBL), the most prevalent pediatric cancer in sub-Saharan Africa, is distinguished by its inclusion of Epstein-Barr virus (EBV). In order to better understand the impact of EBV variation in eBL tumorigenesis, we improved viral DNA enrichment methods and generated a total of 98 new EBV genomes from both eBL cases (N=58) and healthy controls (N=40) residing in the same geographic region in Kenya. Using our unbiased methods, we found that EBV type 1 was significantly more prevalent in eBL patients (74.5%) compared to healthy children (47.5%) (OR=3.24, 95% CI=1.36 - 7.71, P=0.007), as opposed to similar proportions in both groups. Controlling for EBV type, we also performed a genome-wide association study identifying 6 nonsynonymous variants in the genes EBNA1, EBNA2, BcLF1, and BARF1 that were enriched in eBL patients. Additionally, viruses isolated from plasma of eBL patients were identical to their tumor counterpart consistent with circulating viral DNA originating from the tumor. We also detected three intertypic recombinants carrying type 1 EBNA2 and type 2 EBNA3 regions as well as one novel genome with a 20 kb deletion resulting in the loss of multiple lytic and virion genes. Comparing EBV types, viral genes displayed differential variation rates as type 1 appeared to be more divergent while type 2 demonstrated novel substructures. Overall, our findings highlight the complexities of EBV population structure and provide new insight into viral variation, potentially deepening our understanding of eBL oncogenesis.
Importance

- Improved viral enrichment methods conclusively demonstrate EBV type 1 as more prevalent in eBL patients compared to geographically matched healthy controls, which previously underrepresented the prevalence of EBV type 2.
- Genome-wide association analysis between cases and controls identifies 6 eBL-associated nonsynonymous variants in EBNA1, EBNA2, BcLF1, and BARF1 genes.
- Analysis of population structure reveals that EBV type 2 exists as two genomic subgroups, and was more commonly found in females than male eBL patients.
EBV infects more than 90% of the world's population and typically persists as a chronic asymptomatic infection (1). While most individuals endure a lifelong infection with minimal effect, EBV is associated with ~1% of all human malignancies worldwide. EBV was first isolated from an endemic Burkitt lymphoma (eBL) tumor which is the most prevalent pediatric cancer in sub-Saharan Africa (2). Repeated Plasmodium falciparum infections during childhood appear to drive this increased incidence (3). Malaria causes polyclonal B-cell expansion and increased expression of activation-induced cytidine deaminase (AID) dependent DNA damage leading to the hallmark translocation of the MYC gene under control of the constitutively active immunoglobulin enhancer (4-6). How EBV potentiates eBL is incompletely understood, however, the clonal presence of this virus in almost every eBL tumor suggests a necessary role.

EBV strains are categorized into two types based on the high degree of divergence in the EBNA2 and EBNA3 genes (7-9). This long standing evolutionary division is also present in orthologous primate viruses (10), yet remains unexplained. While EBV type 1 has been extensively studied (11, 12), because it causes acute infectious mononucleosis and other diseases in the developed world, type 2 virus studies have not kept pace since infected individuals are less frequent and found primarily in sub-Saharan Africa. This view is changing as several recent studies have reported a significant prevalence of type 2 circulating in western countries, suggesting a greater role for type 1 and greater potential for interactions between the EBV types worldwide (13, 14). For understanding endemic Burkitt, the African context provides a direct opportunity to examine viral variation because type 1 and type 2 are found in both eBL patients as well as healthy individuals (8, 15, 16). Viral variation has been shown to impact differential transformation and growth, and capacity to block apoptosis or immune recognition (7, 17, 18). However, studies focusing on only certain genomic regions/proteins potentially miss disease associations of other loci (19, 20). Although new studies have been conducted (21, 22),
genome-wide examinations in case-control studies are few and often lack typing the virus. A recent study which investigated whole EBV genomes for variant associations with nasopharyngeal carcinoma among Chinese patients has discovered two variants associated with increased disease risk (23). Similarly, another study investigated genome-wide variants of HIV genomes in quest of finding the associations with drug resistance (24). However, to the best of our knowledge, such viral genome-wide variant association for EBV and eBL remains to be explored. To address this shortfall and provide a proof of concept to the field, we sequenced a set EBV genomes in a disease/control setting.

Whole genome sequencing of EBV is now attainable from tumor, blood, or saliva using targeted viral DNA capture methods (25-30). However, studying EBV from the blood of healthy individuals remains challenging due to low viral abundance relative to human DNA (1-10 EBV copy/ng blood DNA). In addition, EBV’s GC-rich genome is inefficiently amplified using conventional library preparation methods. Here, we present improved methods for EBV genome enrichment that allow us to sequence viruses directly from eBL patients and healthy children. Leveraging these samples, we sought to define the viral population structure and characterize viral subtypes collected from children hailing from the same region of western Kenya. Additionally, we performed the first genome wide association study to identify viral variants that correlate with eBL pathogenesis.

**Materials and Methods**

**Ethical approval and sample collection**

For this study, we recruited children between 2009 and 2012 with suspected eBL, between 2-14 years of age, undergoing initial diagnosis at Jaramogi Oginga Odinga Teaching and Referral Hospital (JOOTRH; Kisumu), which is a regional referral hospital for pediatric cancer in western Kenya (31). We also enrolled healthy age-matched children residing in the
same malaria endemic regions of Kenya as controls. We obtained written informed consent  
from parents to enroll their child in this study. Ethical approval was obtained from the  
Institutional Review Board at the University of Massachusetts Medical School and the Scientific  
and Ethical Review Unit at the Kenya Medical Research Institute. From eBL patients, tumor  
biopsies were collected using fine needle aspirates (FNA) and transferred into RNAlater at the  
bedside, prior to induction of chemotherapy. From all children, peripheral blood samples were  
collected and fractionated by centrifugation prior to freezing into plasma and cell pellets. All  
samples were stored at -80°C prior to nucleic acid extraction.

**Cell cultures and controlled mixtures**

BL cultured cell lines, Namalwa, Daudi, Raji, and Jijoye were grown in a complete  
growth medium, RPMI 1640 (Life Technologies), with 2mM L-glutamine adjusted to contain  
1.5g/L sodium bicarbonate, 4.5g/L glucose, 10mM HEPES, 1.0mM sodium pyruvate, and 7.5%  
fetal bovine serum. We used Jijoye and Daudi as representative genomes of type 1 and type 2  
strains. For mixing experiments, we created relative ratios of Jijoye:Daudi of 10:90, 25:75,  
75:25, 90:10 in addition to sequencing each strain individually.

**Improved enrichment of GC-rich EBV in low abundance samples**

We used Allprep DNA/RNA/Protein mini kit (Qiagen) for DNA isolations from FNAs and  
QIAamp DNA Kit for blood and plasma. We developed an improved multi-step amplification and  
enrichment process for the GC-rich EBV genome, particularly in samples with low viral copies.  
We used EBV-specific whole genome amplification (sWGA) to provide sufficient material and  
targeted enrichment with hybridization probes after the library preparation. For this, we designed  
3’-protected oligos following the instructions from Leichty et al. (32). For low viral load samples,  
we added a multiplex long-range PCR amplification (mlrPCR) step, comprising two sets of non-  
overlapping EBV-specific primers tiling across the genome (33). To increase viral DNA content  
in low abundant specimens, we applied an initial amplification with long range PCR using a
strategy consisting of two multiplexed sets of primers which combined tiled the viral genome as designed by Kwok et al. To this, we added EBV type 2 specific primers. Following the initial multiplex long range (mlr) PCRs, we mixed two independent reactions and then performed specific whole genome amplification (sWGA) using phi29 polymerase with EBV-specific oligos. Overall DNA quality and quantities were assessed with NanoDrop and Picogreen and purified with 2x XP-Ampure magnetic beads. We prepared two reaction solutions with separate primer pools (2uL of 10uM each), using 2.5uL 10X long range PCR buffer Mg^{2+} (Qiagen), 1.25uL dNTP (10mM each), 0.15uL long range PCR enzyme mix (Qiagen), 5uL 5X Q-solution (Qiagen) to which we added 10ng of input DNA in 14 uL. The reaction conditions involved initial denaturation at 95°C for 3 min followed by 20 cycles (95°C for 30 sec, gradient annealing 58°C to 49°C for 15 sec each, extension at 72°C for 1 min) and a final extension at 68°C for 10 min. We then mixed two independent reactions, denatured at 95°C for 3 min, and then added sWGA reaction buffer that contains 7uL 10X phi29 reaction buffer (NEB), 3uL dNTP mix (final concentrations: 30mM dGTP and dCTP; 10mM dATP and dTTP), 7uL EBV specific protected oligo Mix (10uM each), 2uL phi29 polymerase (10U/uL, NEB), 0.7uL BSA (0.1ug/uL), and 0.3uL H2O. For samples with higher viral loads that did not require PCR amplification prior to sWGA, we denatured DNA using the same conditions but replaced the reaction buffer with TE or TE-Q-solution mix. We incubated the sWGA at 30°C for 16h followed by incubation at 65°C for 15 min to stop the reaction. Instead of random hexamers for MDA (multiple strand displacement amplification) reaction, we used EBV specific hexamers with 3’-end modification to protect against phi29 exonuclease activity (see supplementary document for primer sequences). For WGA with Genomiphi v2 kit, we followed the manufacturer’s instructions modified by adding extra 2x dGTP and dCTP. For hybrid capture, we followed MyBait protocol as manufacturer’s recommendations. Following incubation at 65°C for 72 hours, we purified hybridization products with streptavidin beads and used Kapa HiFi to amplify the captured library. We quantified viral content with bi-plex qPCR using primers for viral BALF5 and human beta-actin gene (34).
validation of EBV subtypes, we used primers spanning EBNA3C gene producing 153bp and
246bp products for type 1 and type 2, respectively (see Supplemental Table 1 for all primers).

We improved the amplification yield by adding extra 2x dGTP/dCTP to the amplification
buffers, especially for low EBV inputs (10 EBV copy/µL) (Table 1). We also tested the effect of
Q-solution (Qiagen) on sWGA yield and found that EBV yields were almost doubled (Figure 1A
and 1B). In addition, we found that prolonged sWGA incubation time (16 hours) improved
amplification yield compared with relatively shorter time (8 hours). Combining the above
methods allowed for adequate input for hybrid capture even from low viral load healthy controls.

Sequence design for RNA baits

Capture bait sequences were designed using in-house scripts to target both type 1 and
type 2. In addition to type 1 and type 2 references, we also designed against other available
complete genomes including Mutu I, Akata, GD1 and GD2 to ensure the capture of divergent
regions. Specifically, the design consisted of overlapping 120nt probes tiling every 30 bases
(4x overlapping tiling) across the genomic sequences with increased probes for regions with
elevated GC content (>65%). Additional probes were added based on the sequential analysis
of additional genomes, when current probes were greater than 5% divergent or there was a
gap in coverage for a specific region (Supplemental Table 2).

Sequencing library preparation and hybrid capture enrichment

Illumina sequencing library preparation steps consisted of DNA shearing, blunt-end
repair (Quick Blunting kit, NEB), 3’-adenylation (Klenow Fragment 3’ to 5’ exo-, NEB), and
ligation of indexed sequencing adaptors (Quick Ligation kit, NEB). We PCR amplified libraries to
a final concentration with 10 cycles using KAPA HiFi HotStart ReadyMix and quantified them
using bioanalyzer. We then pooled sample libraries balancing them according to their EBV
content and proceeded to target enrichment hybridization using custom EBV-specific
biotinylated RNA probes (MyBaits, Arbor Biosciences). We performed sequencing using
Illumina MiSeq, HiSeq 2000, and NextSeq 500 platforms with 1x75bp, 2x100bp, and 2x150bp, respectively.

Sequence preprocessing and de novo genome assembly

We checked the sequence quality using FastQC (v0.10.1) after trimming residual adapter and low quality bases (<20) using cutadapt (v1.7.1) (35) and prinseq (v0.20.4) (36), respectively. After removing reads that mapped to the human genome (hg38), we de novo assembled the remaining reads into contigs with VelvetOptimiser (v2.2.5) (37) using a kmer search ranging from 21 to 149 to maximize N50. We then ordered and oriented the contigs guided by the reference genomes (NC_007605.1 for type 1 and NC_009334.1 for type 2) using ABACAS, extended with read support using IMAGE (38), and merged the overlapping contigs to form larger scaffolds (using in-house scripts). By aligning reads back to scaffolds, we assessed contig quality requiring support from ≥5 unique reads. We created a final genome by demarcating repetitive and missing regions due to low coverage with sequential ambiguous “N” nucleotides. We excluded minor variants (<5% of reads) in final assemblies.

Diversity and variant association analysis

We used Mafft (v7.215) (39) to generate multiple sequence alignment (msa) of genomes, and masked the repetitive regions predefined in EBV reference genome, NC_007605, in addition to repeat regions larger than 100nt detected by miropeat (40). For analyses based on viral genes, we extracted the coding region sequences from msa of assemblies according to the reference genome Genbank annotations. The substitutions in coding regions were translated in protein sequences based on standard genetic code. The genetic distance between the sequences were calculated using Kimura-2-parameter method based on transition and transversion frequencies. We calculated dN/dS rates per gene based on pairwise Nei-Gojobori algorithm using the python functions provided by
https://github.com/a1ultima/hpcleap_dnds/ after excluding frameshift insertions and ambiguous bases. We constructed whole genome phylogenetic trees based on neighbor-joining method and protein sequence trees based on maximum likelihood method with Jukes-Cantor substitution model using MEGA (v6.0) (41). We determined variant sites of each isolate in reference to the EBV reference genome, NC_007605, based on msa using snp-sites (v2.3.2) (42). For principal coordinate analysis (PCoA), we used R package dartR (v1.0.5) (43). We performed the variant association analysis using the ‘v-assoc’ function from PSEQ/PLINK (44).

To control for multiple testing, we calculated empirical p-values with one million permutations (pseq proj v-assoc --phenotype eBL --fix-null --perm 1000000) with EBV type stratification which permutes within types (--strata EBVtype).

Deposited genomes can be accessed from European Nucleotide Archive (ENA) database with accession PRJEB38735 (study accession: ERP122181) and raw reads can be downloaded from Sequence Read Archive (SRA) database with Bio-project accession number PRJNA552587 (study accession: SRP212943).

Results

Study participant characteristics

The objective of this study was to examine EBV genetic variation in a region of western Kenya with a high incidence of eBL (31) and determine if any variants are associated with eBL pathogenesis. We leveraged specimens from eBL patients and healthy children residing in the same geographic area (Figure 1C) (31). We sequenced the virus isolated from 58 eBL cases and 40 healthy Kenyan children, as controls. Patients between 1 and 13 years of age were predominantly male (74%), consistent with the sex ratio of eBL (Table 2) (31). Healthy controls had similar levels of malaria exposure based on previous epidemiologic studies (45). Control samples ranged in age from 1 to 6 years. This difference in age was necessarily due to the
finding that younger, healthy yet malaria-exposed children have higher average viral loads compared to older children who have developed immune control over this chronic viral infection (34). Therefore, it was infeasible to age-match controls with the eBL cases who tend to be older. We make the assumption that children are infected with the same herpes virus throughout life and therefore genomes from younger children reflect EBV genomes that would be found in older children.

Sequencing and assembly quality

EBV is a large GC-rich double stranded DNA virus with a 172 kb genome of which ~20% is a repetitive sequence. For the majority of eBL patients, we prepared sequencing libraries directly from tumor DNA followed by hybrid capture enrichment. For low copy viral samples, such as eBL plasma and healthy control blood, we designed and implemented additional viral whole genome amplification and enrichment prior to library preparation and sequencing (Figure 1C). We generated a study set of 114 genomes including replicates from cell lines and primary clinical samples, representing 98 cases and controls. In addition, we sequenced 20 technical replicates for quality control purposes such as estimation of re-sequencing error or sWGA bias, and sensitivity of detection of mixed infections. The baseline re-sequencing error rate was limited to ~1.1x10^-5 bases when our assemblies are compared with high-quality known strain genomes (46) (Table 3). The mean error rate was ~2.1x10^-5 bases for sWGA with GenomiPhi, while it is ~1.1x10^-5 bases when we used more sensitive mlrPCR-sWGA (Methods). We obtained an average of ~5 million reads, resulting in an average 9,688 depth of coverage across assemblies (Supplemental Table 3). De novo sequence assembly created large scaffolds covering non-repetitive regions, except three isolates with low coverage, yielded a median of 137,887bp genomes (ranging 47,534bp - 146,920bp). We determined the types of each isolate by calculating the nucleotide distance to both reference types in addition to read mapping rates against type-specific regions. Despite our ability to experimentally detect mixed types at levels...
as low as 10% (Figure 2A), we found no evidence of mixed infections in our cases and controls.

Also, to ensure that our sample inclusion was unbiased when selecting healthy individuals with high enough viremias to sequence, we quantified the baseline viral loads with bi-plex qPCR using primers for viral BALF5 and human beta-actin gene (see Supplemental methods). We compared the viral loads and found no significant difference between type 1 and 2 ($P=0.529$, Figure 2B).

**Equivalence of tumor and plasma viral DNA in eBL cases**

Plasma EBV load has been studied to show its potential as a biomarker or as a prognostic marker in various lymphomas including BL (47, 48). We included plasma specimens along with the tumor biopsies from eBL patients in the whole genome sequencing set to compare and contrast the two counterparts at the sequence level. Following the separate sequencing and genome assembly of 6 pairs of plasma and tumor associated viruses from 6 patients, we confirmed that viral DNA in the plasma was representative of the virus in the tumor cells (Figure 2C). Accounting for the sequencing errors, the pairs appeared to be identical (Supplemental Figure 1). Out of these pairs, we further confirmed the subtype of three EBV isolates (eBL-Tumor-0035, -0037, and -0038) from the plasma and tumor biopsies using type-specific PCRs (see Supplemental method) in addition to 5 other samples (eBL-Tumor-0003, -0019, -0022, -0029, and -0030). Overall, these findings demonstrate that viral DNA isolated from eBL patient plasma represents the tumor virus and reflects its genome sequence to the circulating system. This further assures the potential of plasma DNA for prognostic tools in disease monitoring.

**Structural variation and intertypic recombinants**

First, we looked for large deletions within our viral genomes, but did not detect any of the previously described deletions in EBNAs, even though we were able to detect, as positive controls, EBNA3C deletion in Raji and the EBNA2 deletion in Daudi cell lines. However, in one
sample we did detect a novel 20kb deletion, spanning from 100 kb to 120 kb in the genome (Figure 2D), which appears as the lack of sequencing read coverage while the rest of the genome, even high GC regions, show high sequencing depth (> 6,000x in average). This deleted region normally encodes multiple lytic phase genes, e.g. BBRF1/2, BBLF1/3, BGLF1/2/3/4/5, and BDLF2/3/4. Interestingly, none of the latent genes were affected by this deletion.

Next, we interrogated our isolates by comparing the pairwise similarities of each genome against EBV type 1 and type 2 references. By traversing through the genome with a window, we were able to delineate regions that were more similar to one type over the other (Figure 2E). As expected, Jijoye, a type 2 strain, displayed less similarity against type 1 reference around its EBNA2 and EBNA3 genes, the most divergent region between types, while Namalwa as a type 1 strain shows the same pattern of dissimilarity against type 2 reference around the same regions. Interestingly, we found three patient-derived genomes, eBL-Tumor-0012, eBL-Tumor-0033, and eBL-Plasma-0049, with mixed similarity trends. Similar to a previously detected recombinant strain (LN827563.2_sLCL-1.18) (46), all of the intertypic isolates carried type 1 EBNA2 and type 2 EBNA3 genes. Although these new intertypic hybrids were all isolated from eBL patients only as opposed to healthy controls, this finding does not reach to a statistical significance (P=0.268, Chi-square).

Genomic population structure is driven by type differences with distinct substructure in type 2 viruses.

Our samples present a unique opportunity to study population structure of EBV types and their co-evolution within a geographically defined region. As expected, the major bifurcation within the phylogenetic tree based on the entire genome occurs between type 1 and type 2 viruses (Figure 3A). Viruses from eBL patients as well as healthy controls appeared to be intermixed almost randomly within the type 1 branch. Interestingly, within type 2 genomes
eBL-associated isolates formed a unique sub-cluster. The hybrid genomes are clustered with type 2s, which is consistent with type 2 EBNA3s representing a greater amount of sequence than the type 1 EBNA2 region.

We further explored viral population structure with principal coordinate analysis (PCoA) of variation across the genome. While the first three components cumulatively explain 57.2% of the total variance, the first component, which solely accounted for 43.9% of the variance, separates genomes based on type 1 and type 2 (Figure 3B, upper plot). Similar to the phylogenetic tree, intertypic genomes are positioned more closely to type 2s. Interestingly, the second and predominantly third components separate type 2 viruses into two distinct clusters, group A and B (Figure 3B, lower plot). These clusters were reflected, although not as distinctly, in the structure of the phylogenetic tree in the Figure 3A as well. The PCoA loading values, which accounts for 37.1% of the variance between the type 2 groups, are predominantly driven by correlated variation spanning 70kb upstream of EBNA3C (Figure 3C and D).

Together these findings suggest that there are two EBV type 2 strains circulating within this population. We also examined viral variation from the perspective of LMP1. Interestingly, the vast majority of viruses were grouped into Alaskan and Mediterranean strains (Supplemental Figure 2). The majority of genomes that carry Alaskan LMP1 are type 2 genomes, while Mediterranean strain LMP1s are mostly from type 1 genomes. For all available LMP1 type 2 sequences, group A and group B correlated with Mediterranean and Alaskan, respectively.

**EBV type 2 has less diversity compared with type 1**

We further explored the pattern and nature of genomic variation across the genome comparing and contrasting EBV type 1 and type 2. Examining the pairwise divergence of coding genes for all viral genomes, we found that the divergence was the highest in the type-specific EBNA genes (*EBNA2* and *EBNA3s*), in particular, with *EBNA2* showing the greatest divergence (d=0.1313 ± 2.3x10^{-3}) (Figure 4A, upper panel). Investigating each type separately, the
diversity within types was low for EBNA2 and EBNA3Cs, consistent with type 1 and 2 being separated by many fixed differences (Figure 4A, middle panel). In both types, intra-type divergence was greatest for EBNA1 and LMP1. Most remarkable was the fact that type 2 generally showed lower levels of divergence across the genome (0.0047 ± 3.7x10^{-3} and 0.0025 ± 2.7x10^{-3} for type 1 and type 2, respectively). We observed the same trend even with the balanced sample sizes through random down-sampling (Supplemental Figure 3A and B).

Overall, these measures suggest that EBV gene evolutionary rates differ by types.

To explore signatures of evolutionary selection, we examined the dN/dS ratios within coding sequences (Figure 4A, lower panel). Overall most genes showed signals of purifying selection, as indicated by ω < 1.0, except LMP1, BARF0, and BKRF2 (only type 2). Interestingly, with dN/dS measures, EBNA2, BSLF1, BSLF2, and BLLF2 genes had relatively higher rates in type 2 compared to type 1 (P < 0.001). Having significantly different ω values for multiple genes (P < 0.001, t-test) can suggest the existence of differential evolutionary pressure on these two divergent types. This can be interpreted as an ongoing adaptation process of type 2 genomes (through certain genes) in the population while fixed functions of these genes for the type 1 genomes. Overall, the magnitude of average nonsynonymous and synonymous changes per gene, normalized by gene length, reflect the high-level diversity accumulated in certain genes (Supplemental Figure 4). Latency-associated genes generally have the highest non-synonymous variant rates, but they also have the highest synonymous rates consistent with longstanding divergence (Figure 4B). Other functional categories, including lytic genes, have relatively low levels of nonsynonymous mutations suggesting stronger purifying selection (see Supplemental Table 4 for functional categories).

Global context of Kenyan viruses

To more broadly contextualize our viral population from western Kenya, we examined the phylogeny of the Kenyan viruses along with other publicly available genomes from across
the world (Supplemental Table 5). Among all isolates, the most polymorphic genomic regions appeared to be around EBNA2 and EBNA3 genes (Supplemental Figure 5A). Phylogenetic tree shows that the major types, type 1 and type 2, are the main demarcation point regardless of the source or geographic location. The three intertypic genomes from our sample set neatly cluster with the previously isolated intertypic hybrid, sLCL-1.18 (Supplemental Figure 5B).

Type 1 genomes from our study were split into two groups, with one forming a sub-branch only with Kenyan type 1, including Mutu, Daudi, and several Kenyan LCLs. The second group interspersed with other African (Ghana, Nigeria, North Africa) and non-African isolates. In addition, a few of our genomes from healthy carriers clustered with a group of mainly Australian isolates, however; none of them clustered with the South Asian group. Our Kenyan EBV type 2s generally intermixed with other type 2 genomes.

Viral Genomic Variants and Associations with eBL

Previous studies, which examined viral sequence variants with relative frequencies, often lack properly controlled disease association analysis in genome-wide context (14, 49). After excluding the intertypic hybrids, we compared type frequencies of EBV genomes isolated from eBL patients and healthy controls. We observed a significant difference in frequencies with 74.5% of eBLs carrying type 1 while only 25.5% carried type 2 infections. In contrast, 47.5% vs. 52.5% of type 1 and type 2, respectively were found in healthy controls. EBV type 1 was associated with eBL (OR=3.24, 95% CI=1.36 - 7.71, P = 0.007, Fisher’s exact) (Figure 4C), independent of age and gender (all P>0.05, Supplemental Figure 6). The type 1 prevalence was still ~70% (and 30% type 2) among eBL children who are within the equivalent age range of their healthy counter-parts (1-6 years old). The breakdown of the subtype frequencies based on gender revealed that the female eBL patients most frequently carried type 2 (type 1 N=14 vs type 2 N=2) while their healthy counterparts showed the opposite trend (type 1 N=7 vs type 2 N=26) (Table 4). On the other hand, male individuals carried both subtypes with roughly...
equivalent frequencies regardless of their disease status (N=27 vs N=26 and N=12 vs N=16, type 1 and type 2 among eBL and healthy control groups, respectively). We then expanded the association analysis to all 6191 synonymous and non-synonymous single nucleotide variations across the entire genome (Figure 4D and Supplemental Table 6). We conducted an initial association test for each nonsynonymous variant and detected 133 significant associations (Supplemental Table 7 and Methods). The vast majority of these variants were located within the type1-type2 region given the highly correlated nature of this region (Supplemental Figure 7). We then stratified by type to detect variants independent of viral type. This yielded 6 variants solely associated with the disease (Table 5). Variant 37668T>C represents a serine residue change to a proline at the C-terminus of EBNA2 (S485P) which is carried by 24/54 eBL cases; while this variant was present in only 2/36 healthy controls. Two variants in EBNA1 at 95773A>T and 95778T>G (N38Y and H39Q, respectively) were both observed in 3/57 eBL isolates while their corresponding frequencies were 11/36 and 12/37 among healthy controls. These two variants fall into one of the two chromosome binding domains of the EBNA1 protein which plays a bridging role for tethering the viral episome to the host chromosome (50, 51). Other two significant variants we detected are within the BcLF1 gene which encodes for viral major capsid protein. This protein is the most essential component of the self-assembly structures for the viral capsid (52). Elevated substitution rates in viruses of healthy controls as opposed to eBL associated viruses comply with their role in capsid formation and pathogenesis. The BARF1 variant with higher frequencies in eBL associated viruses that replaces the valine residue with an alanine might provide a fitness advantage with its role as a soluble form of CSF-1 receptor that neutralizes effects of human CSF-1. With this residual change in the protein, BARF1 increases its sequence identity to human CD80 as both share the same amino acid, alanine, at the position 29 (53). Nucleotide variants in non-coding and promoter regions can affect regulation of viral gene expression and activity within host cells. BZLF1 is a regulator gene of lytic reactivation and
classified based on its promoter as prototype Zp-P (B95-8) and Zp-V3 (M81 strain). Zp-V3 variant of the promoter has recently been found to enhance lytic activity and over-represented in EBV-positive BLs (54). Therefore, we sought to find out whether our dataset can validate the association. We determined variants at seven positions in the upstream promoter region of BZLF1 (Supplemental Table 8). Interestingly, all of the Kenyan viruses carried C at positions both -525 and -274 (as in Zp-P) regardless of promoter type. We also found that -532 and -524 are variable in our isolates while these two are not variant in both promoter types. Our results show that only 12.5% (5/40) type 1 promoter sequences fully resembled Zp-V3 in eBL group as opposed to 22% (2/9) healthy genomes, while all of the type 2 genomes, without exception, carried Zp-V3 type promoter regardless of disease status.

**Discussion**

In this study, we investigated genomic diversity of EBV by sampling viruses from children in western Kenya where eBL incidence is high (45). Our improved methods allowed us to sequence asymptomatically infected healthy controls with relatively low peripheral blood viral loads, and thereby examine the virus in the population at large (34). We performed the first association study comparing viral genomes from eBL patients and geographically matched controls, without the need for viral propagation in LCLs; thus showing that type 1 EBV, as well as potentially several non-type specific variants, are associated with eBL. Furthermore, as the first study that characterized significant numbers of EBV type 2, we were able to compare and contrast both types and explore the viral population, thus discovering novel differences including population substructure and in female to male frequencies in EBV type 2. An extended cohort is required to further validate our results.

Our sequencing data demonstrated that EBV from plasma is representative of the tumor virus in eBL patients. This is consistent with the premise that peripheral EBV DNA originates
from apoptotic tumor cells given that cell-free EBV DNA in eBL patients are mostly unprotected against DNase (55), as opposed to being encapsidated during lytic reactivation, and that plasma EBV levels are associated with tumor burden and stage (48). These findings support the use of plasma viremia as a surrogate biomarker for tumor burden and the development of plasma-based prognostic tests with predictive models that could be used during clinical trials (48). The lack of mixed infections observed in our healthy controls could be due to the limit of detection in blood compared to viruses isolated from saliva (14). Further studies are needed to extrapolate and understand the coevolution and dynamics of both EBV types.

In addition, we detected three intertypic recombinant EBV genomes solely found within our eBL patients; similar to those previously described in other cancers (56). It is unclear whether the intertypic genomes represent a common event with subsequent mutation and recombination or multiple independent events. If the latter is true, it supports more frequent mixed-type infections given that both parents have to be present in the same cell (57-59). It is interesting that all four intertypics observed to date carry the same type EBNA2/EBNA3 combinations with the type 2 genes being so closely related (Supplemental Figure 8). Thus, if multiple events have generated these viruses, it suggests that certain strains may have a greater proclivity to recombine. Further studies will be needed to better define the intertypic population, their origins and their association with disease.

Importantly, we were able to explore EBV population genetics and compare and contrast type 1 and type 2 because of their co-prevalence in Africa. As well described, the major differentiation in terms of genetic variability was the variation correlated with type 1 and type 2 viruses. These viral types showed distinct population characteristics with type 1 harboring greater diversity especially in functionally important latent genes. Combined with the observed nucleotide diversity, latency genes appear to have long standing divergence that has accumulated significant synonymous changes (as opposed to recent sweeps on nonsynonymous changes that would erase synonymous variants). Global phylogenetic analysis
emphasizes this diversity by providing two main subgroups for type 1 genomes in our sequencing set. One group represents core local Kenyan viruses while the second group is a mixture of viruses from across the globe, with the exception of South Asian viruses that group apart. While previously sequenced type 2 viruses intermingle with western Kenya isolates, the majority of these originated from East Africa with only a few from West Africa. Interestingly, intermingling is also true for type 2 as we observed two distinct groups. This is more apparent in PCA where type 2 virus forms 2 clusters. Examination via PCA, the loading values are determined by a broad stretch of the genome from the end of EBNA3C to LMP1, where Mediterranean and Alaskan designations correlate. It remains to be determined whether this substructure might be due to the introduction of previously geographically isolated viruses or distinct evolutionary trajectories within the population. Further study is needed with broader samplings to understand its significance but our findings suggest that there may be significant epistasis potentially including LMP1.

By sequencing the virus directly from healthy controls, we were able to address the question of relative tumorigenicity between EBV type 1 and 2. We evaluated the long-standing presumption that type 1 virus is more strongly associated with eBL, in contrast to type 2. Our work was able to more definitely answer this question as we were not reliant on LCLs from healthy controls where type 1 bias in transformation might explain the lack of previous associations. We earlier demonstrated, by mutational profiling of EBV positive and negative eBL tumors, that the virus, especially type 1, might mitigate the necessity of certain driver mutations in the host genome (16). In addition, our genome-wide results controlling for viral type substantiates investigations of non-type associated variation that could also impart oncogenic risk, as we found suggestive trends for several nonsynonymous variants as well. Supporting the putative existence of EBV sub-strains that have increased oncogenic potential, we observed sub-cliques of solely eBL or control isolates within the type 2 genomes. Although these sub-groups were formed with only 7 or 8 members, the significance of this observation will be deciphered with
more extensive cohorts. On the other hand, only a small subset of type 1 viruses from eBL patients carried *BZLF1* promoter variant, which leads to a gain of function (54), while all type 2 viruses carried this variant suggesting this promoter might be beneficial for type 2 but makes it unlikely to be a driver of oncogenesis. It is essential to remind that the suggestive associations we uncovered are in need of further validation with independent cohorts and treated cautiously.

Overall, our population-based study provides the groundwork to unravel the complexities of EBV genome structure and insight into viral variation that influences oncogenesis. Genomic and mutational analysis of BL tumors identified key differences based on viral content suggesting new avenues for the development of prognostic molecular biomarkers and the potential for antiviral therapeutic interventions.
Acknowledgements

This work was supported by the US National Institutes of Health, National Cancer Institute R01 CA134051, R01 CA189806 (A.M.M., J.A.B, C.I.O, Y.K.) and The Thrasher Research Fund 02833-7 (A.M.M.), UMCCTS Pilot Project Program U1 LTR00161-04 (Y.K., J.A.B., and A.M.M.), Turkish Ministry of National Education Graduate Study Abroad Program (Y.K.). We would like to thank the Kenyan children and their families who participated in this study. Patrick Marsh for helping with EBV genotyping assays, Mercedeh Movassagh for sharing genotyping primers. This publication was approved by the Director of KEMRI. The current affiliation for Yasin Kaymaz is FAS Informatics and Scientific Applications, Harvard University, Cambridge, MA

Authorship Contributions

Contribution: Y.K., C.I.O., and O.A. designed and performed experiments; Y.K. and C.I.O analyzed and interpreted results; Y.K. made the figures; Y.K., J.A.B. and A.M.M. designed the research and wrote the paper, C.I.O, J.A.O., J.M.O., and A.M.M. organized clinical sample acquisition.

Disclosure of Conflicts of Interest

The authors declare no competing financial interests.
References


44. Anonymous. PLINK/SEQ genetics library.


50. Sears J, Uijhara M, Wong S, Ott C, Middeldorp J, Aiyar A. 2004. The amino terminus of Epstein-Barr virus (EBV) nuclear antigen 1 contains at hooks that facilitate the


### Table 1. Optimization of mirPCR-sWGA and whole genome amplification reactions.

<table>
<thead>
<tr>
<th>Input EBV copies</th>
<th>Denaturation Buffer</th>
<th>Average EBV Input (genome copy/μL)</th>
<th>Average Human Input (B-actin/μL)</th>
<th>dNTP composition (G/C/T/A ratio)</th>
<th>Whole Genome Amplification Incubation</th>
<th>EBV DNA Output (copy/ng)</th>
<th>Human DNA Output (B-actin/ng)</th>
<th>EBV Enrichment (EBVpost / EBVpre)</th>
<th>Human Enrichment (Humanpost / Humanpre)</th>
<th>EBV Enrichment / Human Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000</td>
<td>TE</td>
<td>62,448</td>
<td>3,456</td>
<td>3:0:0:0</td>
<td>16 hours at 30°C</td>
<td>4.5±3</td>
<td>7</td>
<td>1.06</td>
<td>0.03</td>
<td>30.33</td>
</tr>
<tr>
<td>1,000</td>
<td>TE</td>
<td>10,038</td>
<td>13,911</td>
<td>3:0:0:0</td>
<td>16 hours at 30°C</td>
<td>555</td>
<td>59</td>
<td>2.01</td>
<td>0.08</td>
<td>12.63</td>
</tr>
<tr>
<td>100</td>
<td>TE</td>
<td>467</td>
<td>10,983</td>
<td>3:0:0:0</td>
<td>16 hours at 30°C</td>
<td>55</td>
<td>70</td>
<td>2.25</td>
<td>0.12</td>
<td>19.00</td>
</tr>
<tr>
<td>10</td>
<td>TE</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>16 hours at 30°C</td>
<td>4</td>
<td>68</td>
<td>8.26</td>
<td>0.12</td>
<td>68.83</td>
</tr>
<tr>
<td>1</td>
<td>TE</td>
<td>6</td>
<td>19,483</td>
<td>3:0:0:0</td>
<td>16 hours at 30°C</td>
<td>0</td>
<td>55</td>
<td>0.21</td>
<td>0.06</td>
<td>3.50</td>
</tr>
<tr>
<td>0</td>
<td>TE</td>
<td>11,551</td>
<td>30/30/10/10</td>
<td>3:0:0:0</td>
<td>30/30/10/10</td>
<td>76</td>
<td>76</td>
<td>1.00</td>
<td>0.03</td>
<td>23.00</td>
</tr>
<tr>
<td>10,000</td>
<td>TE</td>
<td>62,448</td>
<td>3,456</td>
<td>3:0:0:0</td>
<td>16 hours at 30°C</td>
<td>4.847</td>
<td>2</td>
<td>1.12</td>
<td>0.01</td>
<td>112.00</td>
</tr>
<tr>
<td>1,000</td>
<td>TE</td>
<td>10,038</td>
<td>13,911</td>
<td>3:0:0:0</td>
<td>16 hours at 30°C</td>
<td>544</td>
<td>75</td>
<td>1.99</td>
<td>0.11</td>
<td>9.91</td>
</tr>
<tr>
<td>100</td>
<td>TE</td>
<td>467</td>
<td>10,983</td>
<td>3:0:0:0</td>
<td>16 hours at 30°C</td>
<td>72</td>
<td>65</td>
<td>0.12</td>
<td>0.12</td>
<td>25.00</td>
</tr>
<tr>
<td>10</td>
<td>TE</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>16 hours at 30°C</td>
<td>3</td>
<td>85</td>
<td>5.97</td>
<td>0.11</td>
<td>54.27</td>
</tr>
<tr>
<td>1</td>
<td>TE</td>
<td>6</td>
<td>19,483</td>
<td>3:0:0:0</td>
<td>16 hours at 30°C</td>
<td>0</td>
<td>81</td>
<td>0.22</td>
<td>0.09</td>
<td>2.44</td>
</tr>
<tr>
<td>0</td>
<td>TE</td>
<td>11,551</td>
<td>30/30/5/5</td>
<td>3:0:0:0</td>
<td>30/30/5/5</td>
<td>71</td>
<td>71</td>
<td>0.21</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>10,000</td>
<td>TE</td>
<td>62,448</td>
<td>3,456</td>
<td>3:0:0:0</td>
<td>15/15/5/5</td>
<td>3,483</td>
<td>16</td>
<td>0.81</td>
<td>0.07</td>
<td>11.57</td>
</tr>
<tr>
<td>1,000</td>
<td>TE</td>
<td>10,038</td>
<td>13,911</td>
<td>3:0:0:0</td>
<td>15/15/5/5</td>
<td>3,488</td>
<td>84</td>
<td>0.97</td>
<td>0.12</td>
<td>8.08</td>
</tr>
<tr>
<td>100</td>
<td>TE</td>
<td>467</td>
<td>10,983</td>
<td>3:0:0:0</td>
<td>15/15/5/5</td>
<td>73</td>
<td>109</td>
<td>3.04</td>
<td>0.19</td>
<td>16.00</td>
</tr>
<tr>
<td>10</td>
<td>TE</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>15/15/5/5</td>
<td>0</td>
<td>80</td>
<td>2.71</td>
<td>0.14</td>
<td>19.36</td>
</tr>
<tr>
<td>1</td>
<td>TE</td>
<td>6</td>
<td>19,483</td>
<td>3:0:0:0</td>
<td>15/15/5/5</td>
<td>0</td>
<td>103</td>
<td>0.25</td>
<td>0.11</td>
<td>2.27</td>
</tr>
<tr>
<td>0</td>
<td>TE</td>
<td>11,551</td>
<td>15/15/5/5</td>
<td>3:0:0:0</td>
<td>15/15/5/5</td>
<td>118</td>
<td>118</td>
<td>0.36</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>10,000</td>
<td>TE</td>
<td>62,448</td>
<td>3,456</td>
<td>3:0:0:0</td>
<td>15/15/2/2</td>
<td>3,324</td>
<td>12</td>
<td>0.77</td>
<td>0.05</td>
<td>15.40</td>
</tr>
<tr>
<td>1,000</td>
<td>TE</td>
<td>10,038</td>
<td>13,911</td>
<td>3:0:0:0</td>
<td>15/15/2/2</td>
<td>533</td>
<td>77</td>
<td>1.06</td>
<td>0.11</td>
<td>9.64</td>
</tr>
<tr>
<td>100</td>
<td>TE</td>
<td>467</td>
<td>10,983</td>
<td>3:0:0:0</td>
<td>15/15/2/2</td>
<td>78</td>
<td>105</td>
<td>3.24</td>
<td>0.19</td>
<td>17.05</td>
</tr>
<tr>
<td>10</td>
<td>TE</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>15/15/2/2</td>
<td>4</td>
<td>105</td>
<td>7.8</td>
<td>0.18</td>
<td>43.33</td>
</tr>
<tr>
<td>1</td>
<td>TE</td>
<td>6</td>
<td>19,483</td>
<td>3:0:0:0</td>
<td>15/15/2/2</td>
<td>0</td>
<td>110</td>
<td>5.9</td>
<td>0.12</td>
<td>0.00</td>
</tr>
<tr>
<td>0</td>
<td>TE</td>
<td>11,551</td>
<td>15/15/2/2</td>
<td>3:0:0:0</td>
<td>15/15/2/2</td>
<td>108</td>
<td>108</td>
<td>0.32</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td>TE</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>8 hours at 30°C</td>
<td>2</td>
<td>40</td>
<td>6.91</td>
<td>0.14</td>
<td>40.07</td>
</tr>
<tr>
<td>10</td>
<td>TE</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>8 hours at 30°C</td>
<td>2</td>
<td>40</td>
<td>8.14</td>
<td>0.14</td>
<td>62.43</td>
</tr>
<tr>
<td>10</td>
<td>TE</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>8 hours at 35°C</td>
<td>2</td>
<td>40</td>
<td>8.14</td>
<td>0.14</td>
<td>62.43</td>
</tr>
<tr>
<td>10</td>
<td>TE</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>8 hours at 35°C</td>
<td>1</td>
<td>22</td>
<td>4.05</td>
<td>0.08</td>
<td>50.63</td>
</tr>
<tr>
<td>10</td>
<td>TE</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>8 hours at 35°C</td>
<td>2</td>
<td>22</td>
<td>6.14</td>
<td>0.08</td>
<td>76.75</td>
</tr>
<tr>
<td>10</td>
<td>TE+Q sol</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>8 hours at 30°C</td>
<td>2</td>
<td>41</td>
<td>9.64</td>
<td>0.14</td>
<td>67.43</td>
</tr>
<tr>
<td>10</td>
<td>TE+Q sol</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>8 hours at 30°C</td>
<td>3</td>
<td>60</td>
<td>13.94</td>
<td>0.21</td>
<td>66.38</td>
</tr>
<tr>
<td>10</td>
<td>TE+Q sol</td>
<td>10</td>
<td>11,326</td>
<td>3:0:0:0</td>
<td>8 hours at 35°C</td>
<td>3</td>
<td>28</td>
<td>10.72</td>
<td>0.1</td>
<td>107.20</td>
</tr>
</tbody>
</table>

* total EBV genomes 20 μL of input for sWGA
Table 2. Characteristics of children included in EBV sequencing analysis.

<table>
<thead>
<tr>
<th></th>
<th>eBL Patients (N=58)</th>
<th>Healthy Controls (N=40)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age at collection, N (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;6 (yrs)</td>
<td>16 (27.6)</td>
<td>39 (97.5)</td>
</tr>
<tr>
<td>7 - 13 (yrs)</td>
<td>42 (72.4)</td>
<td>1 (2.5)</td>
</tr>
<tr>
<td><strong>Sex, N (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female/Male</td>
<td>15/43 (25.9/74.1)</td>
<td>20/20 (50.0/50.0)</td>
</tr>
<tr>
<td><strong>Obtained Specimen, N (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tumor biopsy</td>
<td>41 (41.8)</td>
<td>-</td>
</tr>
<tr>
<td>Blood</td>
<td>-</td>
<td>40 (100.0)</td>
</tr>
<tr>
<td>Plasma</td>
<td>14 (14.2)</td>
<td>-</td>
</tr>
<tr>
<td>New cultured eBL</td>
<td>3 (3.0)</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 3. Estimated sequencing error rates based on replicates and controls.

<table>
<thead>
<tr>
<th>Prior Amplification</th>
<th>Control Assembly</th>
<th>Experimental Assembly</th>
<th>Number of Substitution Errors</th>
<th>Number of Correct Bases</th>
<th>Error Rate (per base)</th>
<th>Mean Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>No amplification</td>
<td>Jijoye</td>
<td>Jiyoye Assembly*</td>
<td>1</td>
<td>134,118</td>
<td>7.46E-06</td>
<td>1.13E-05</td>
</tr>
<tr>
<td>No amplification</td>
<td>Daudi</td>
<td>Daudi Assembly*</td>
<td>0</td>
<td>132,780</td>
<td>0.00E+00</td>
<td></td>
</tr>
<tr>
<td>No amplification</td>
<td>Raji_Rep1</td>
<td>Raji Assembly*</td>
<td>0</td>
<td>130,560</td>
<td>0.00E+00</td>
<td></td>
</tr>
<tr>
<td>No amplification</td>
<td>Raji_Rep2</td>
<td>Raji Assembly*</td>
<td>5</td>
<td>132,736</td>
<td>3.77E-05</td>
<td></td>
</tr>
<tr>
<td>GenomiPhi-WGA</td>
<td>Raji_wga</td>
<td>Raji Assembly*</td>
<td>2</td>
<td>132,836</td>
<td>1.51E-05</td>
<td></td>
</tr>
<tr>
<td>GenomiPhi-WGA</td>
<td>eBtumor-01_EBV_type1_wga</td>
<td>eBtumor-01_EBV_type1#</td>
<td>4</td>
<td>140,069</td>
<td>7.14E-06</td>
<td>2.13E-05</td>
</tr>
<tr>
<td>GenomiPhi-WGA</td>
<td>eBtumor-02_EBV_type2_wga</td>
<td>eBtumor-02_EBV_type2#</td>
<td>1</td>
<td>142,154</td>
<td>3.53E-05</td>
<td></td>
</tr>
<tr>
<td>GenomiPhi-WGA</td>
<td>eBtumor-03_EBV_type2_wga</td>
<td>eBtumor-03_EBV_type2#</td>
<td>1</td>
<td>142,154</td>
<td>3.53E-05</td>
<td></td>
</tr>
<tr>
<td>GenomiPhi-WGA</td>
<td>eBtumor-04_EBV_type1_wga</td>
<td>eBtumor-04_EBV_type1#</td>
<td>5</td>
<td>141,643</td>
<td>3.53E-05</td>
<td></td>
</tr>
<tr>
<td>mlrPCR-sWGA</td>
<td>Raji_Rep1_mlrPCR-sWGA</td>
<td>Raji Assembly*</td>
<td>6</td>
<td>98,507</td>
<td>6.09E-05</td>
<td>1.09E-04</td>
</tr>
<tr>
<td>mlrPCR-sWGA</td>
<td>Raji_Rep2_mlrPCR-sWGA</td>
<td>Raji Assembly*</td>
<td>11</td>
<td>130,538</td>
<td>8.43E-05</td>
<td></td>
</tr>
<tr>
<td>mlrPCR-sWGA</td>
<td>Jiyoye_Rep1_mlrPCR-sWGA</td>
<td>Jiyoye Assembly*</td>
<td>13</td>
<td>132,045</td>
<td>9.84E-05</td>
<td></td>
</tr>
<tr>
<td>mlrPCR-sWGA</td>
<td>Daudi_mlrPCR-sWGA</td>
<td>Daudi Assembly*</td>
<td>25</td>
<td>130,160</td>
<td>1.92E-04</td>
<td></td>
</tr>
</tbody>
</table>

*Reference assemblies from Palser et al. 2015 (Jijoye: LN827800, Daudi: LN827545, Raji: KF717093); #The isolates were sequenced without any amplification. Preprocess denotes whether sample DNA was amplified prior to sequencing library preparation. The number of substitutions were determined by pairwise whole genome alignments of control and reference assemblies. Error rates were referred to the average mismatches to reference assemblies after normalizing to total covered genomic regions. GenomiPhi-WGA: whole genome amplification using EBV specific protected hexamers, mlrPCR-sWGA: pre-amplification with PCR primer pools followed by sWGA.
Table 4. Characteristics of children included in our study and viral genome subtypes.

<table>
<thead>
<tr>
<th>Group</th>
<th>Gender</th>
<th>Type</th>
<th>Mean age (years)</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>eBL</td>
<td>Female</td>
<td>1</td>
<td>7.57</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>1</td>
<td>6.81</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>7.08</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1/2</td>
<td>8.33</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Healthy Control</td>
<td>Female</td>
<td>1</td>
<td>3.36</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>3.58</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>1</td>
<td>3.3</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>3.74</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

EBV subtypes are represented as 1, 2 and 1/2 for inter-typic.
Table 5. Single nucleotide variants associated with eBL.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Position</th>
<th>Ref</th>
<th>Alt</th>
<th>AA Change</th>
<th>eBLs Genotypes</th>
<th>eBLs Alt Count</th>
<th>Healthy Controls Genotypes</th>
<th>Healthy Controls Alt Count</th>
<th>P</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBNA2</td>
<td>37668</td>
<td>T</td>
<td>C</td>
<td>S485P</td>
<td>54</td>
<td>24</td>
<td>36</td>
<td>2</td>
<td>0.000328</td>
<td>0.1</td>
</tr>
<tr>
<td>EBNA1</td>
<td>95773</td>
<td>A</td>
<td>T</td>
<td>N38Y</td>
<td>57</td>
<td>3</td>
<td>36</td>
<td>11</td>
<td>0.001322</td>
<td>6.67213</td>
</tr>
<tr>
<td>EBNA1</td>
<td>95778</td>
<td>T</td>
<td>G</td>
<td>H39Q</td>
<td>57</td>
<td>3</td>
<td>37</td>
<td>12</td>
<td>0.000538</td>
<td>7.16129</td>
</tr>
<tr>
<td>BcLF1</td>
<td>124703</td>
<td>T</td>
<td>G</td>
<td>K159T</td>
<td>56</td>
<td>1</td>
<td>34</td>
<td>7</td>
<td>0.003178</td>
<td>12.7377</td>
</tr>
<tr>
<td>BcLF1</td>
<td>124709</td>
<td>G</td>
<td>A</td>
<td>A157V</td>
<td>56</td>
<td>1</td>
<td>34</td>
<td>7</td>
<td>0.003092</td>
<td>12.7377</td>
</tr>
<tr>
<td>BARF1</td>
<td>165131</td>
<td>T</td>
<td>C</td>
<td>V29A</td>
<td>57</td>
<td>36</td>
<td>36</td>
<td>10</td>
<td>0.004082</td>
<td>0.349462</td>
</tr>
</tbody>
</table>

Single nucleotide variant association test results with P < 0.01 after type stratification. Table summarizes the statistically significant single nucleotide variant associations and their effects in the coding regions. Reference is the genotype based on the consensus of all genomes in the sequencing set and variant position denotes the projection to type 1 reference genome (NC_007605). The association test has been performed for every variant position comparing the frequency of reference and alternative (minor allele) bases among eBL patient and healthy control children (Fisher’s exact test). Empirical p values were based on one million permutations. *Genomes with missing data (Ns, lack of coverage) were excluded. Ref: reference allele, Alt: alternative/variant allele, AA: amino acid, P: p-value, OR: odds ratio.
Figure Legends

Figure 1. Optimized EBV genome sequencing from tumors and primary clinical samples.
A) Optimization results of various dNTP concentrations in mlrPCR-sWGA reaction measured as EBV copy increase normalized by overall DNA increase. B) Incubation buffer (TE: Tris-EDTA, or Q solution), time (8 or 16 hours), and temperature optimization for better EBV copy increase. C) Overview of sample collection and methods for sequencing virus from Kenyan children diagnosed with eBL and healthy children as controls. Hybrid capture was universally performed along with additional amplification and enrichment steps to overcome low amounts of virus and input DNA. mlrPCR-sWGA; multiplexed long range PCR - specific whole genome amplification.

Figure 2. Sequencing and detection quality, plasma-tumor pairs, and atypical genome isolates.
A) Controls for putative mixed infections and sampling bias against EBV types. The sensitivity of EBV genome typing approach measured by accurate type assignments of in-lab mixtures with predefined ratios. Each mixture of Daudi (type 1) and Jijoye (type 2) with varying ratios was prepared in replicates. Following the genome assembly, type of the major strain was determined as judged by the distance to both reference viral genomes. B) Comparison of viral load levels of individuals carrying different EBV types (P=0.529, t-test). The non-significant difference suggests unbiased sampling among either type regardless of viral loads. We quantified viral loads with bi-plex qPCR using primers for viral BALF5 and human beta-actin gene. C) Comparison of virus from paired tumor (filled pink circles) and plasma samples (hollow pink circles) at diagnosis shows viral DNA circulating in the peripheral blood represents the virus in the tumor. The Neighbor-Joining tree is scaled (0.001 substitutions per site) and includes standard reference genomes for type 1 (NC007605, blue diamond) and type 2 (NC009334, red diamond). D) The depth of coverage showing an absence of reads from approximately 100 kb to 120 kb is indicative of a large deletion in the virus from an eBL tumor (top panel). In the middle and lower panels, although we did not detect any in our tumor or...
control viruses, we detected the deletions previously described in tumor lines including EBNA3C deletion in Raji and EBNA2 deletion in Daudi strains. E) Three intertypic viruses were detected by scanning across the genomes for percent identity in 1kb windows to both type 1 and type 2 references (NC_007605, NC_009334, respectively). Top two graphs (grey) represent controls, Jijoye and Namalwa, followed by 3 intertypic viruses from this study and one publicly available intertypic virus (LN827563.2_sLCL-1.18 in grey).

Figure 3. Diversity and phylogenetic analysis of EBV genomes in Kenyan population.

A) Phylogenetic tree of the western Kenya EBV genomes demonstrating the major type 1 and type 2 demarcation (blue and red branches, respectively). Pairwise distance calculations were based on Jukes-Cantor nucleotide substitution model, and the tree was constructed with the simple Neighbor-Joining method. Genomes are colored based on sample type: healthy children blood (green squares), eBL tumors (full pink circles), plasma of eBL children (hollow pink circles), and new and previous cell lines (pink and yellow triangles, respectively). Low coverage genomes are excluded. B) Principal coordinates analysis plots of nucleotide variations among whole genome sequences with first and second axes (upper plot, colored by sample type), and second and third axes (lower plot, colored by EBV subtype and shapes represent case and control), which separates type 2 genomes into groups A and B (dashed red ellipses). Color coding is the same as in A. C) First and second axises of PCoA using only type 2 genomes showing the separation of two groups. D) Absolute loading values of Axis 1 from PCoA with all variants are plotted throughout the genome. Values are averaged across 1 kb window. Dashed arrow marks the region for sequence variations that predominantly drives the separation in the PCoA.
Figure 4. Diversity in EBV coding genes, significant associations of EBV type 1 genomes and single nucleotide variants with eBL.

A) Genetic distance metrics of each EBV gene calculated based on the Kimura-2-parameter method averaged across all genomes (upper panel) or type 1 / type 2 separately (middle panel). Lower panel shows nonsynonymous to synonymous change (dN/dS) ratios of viral protein coding genes averaged across all pairwise comparisons within each group separately. Error bars represent standard error of mean. (Three intertypic genomes are excluded).

B) Average synonymous and non-synonymous variants in genes are summarized as functional categories of genes. Variant level represents the number of variants per gene normalized by gene length in kb.

C) The frequency of type 1 and type 2 genomes identified from eBL patients and healthy control children (excluding the three intertypic hybrid genomes) is significantly different ($P=0.007$, Fisher’s exact).

D) Manhattan plot for genome-wide associations of all single nucleotide variants tested for frequency differences between cases and controls controlling for type specific variants. The significance of each locus association is represented with an empirical p-value (negative log10 scale) that was calculated by 1 million permutations with random label swapping. Permutations were stratified for EBV genome type and adjusted for the missing genotypes due to lack of coverage. All significant variants associated with eBL cases are shown in red ($P < 0.01$). Nucleotide positions are according to type 1 reference genome.
Figure 1

A

B

C

DNA Isolation

Sample collection

Primary Tumor Biopsy

Plasma

Blood

EBV children (N=58)

Healthy children (N=40)

sWGA

mPCR-sWGA

No Amplification

Library Prep & Hybrid Capture

EBV specific probes

Sequencing

Human

EBV

Denaturation Buffer

TE

TE + Q sol

Incubation Conditions

EBV Copy Increase / Total DNA Increase

EBV Copy Increase / Total DNA Increase

30/30/10/10

30/20/5/5

15/15/5/5

15/15/2/2

dNTP ratios (mM, dGdC/dAA/dT)

Input:

EBV copy/ul.:

6

467

10

10,038

62,448
Figure 2

A

B

C

D

E

Viral Load (copy/mL)

Type 1
Type 2

"<20Kb Deletion"

EBNA1/C Deletion
EBNA2 Deletion

Source of EBV
- eBL Tumor
- eBL Plasma
- Type 1 Reference
- Type 2 Reference

Percent Similarity

Genomic Position (kb)

Genomic Position (kb)
Figure 3

A

B

C

D

Only Type2 Genomes

mean PC Axis Loadings

Genomic Position (Kb)