Epstein Barr virus genomes reveal population structure and type 1 association with endemic Burkitt lymphoma

Yasin Kaymaz¹, Cliff I. Oduor^{2,3}, Ozkan Aydemir³, Micah A. Luftig⁴, Juliana A. Otieno⁵, John Michael Ong'echa², Jeffrey A. Bailey^{3*}, Ann M. Moormann^{6*}

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.

67 **Importance**

- 68 • Improved viral enrichment methods conclusively demonstrate EBV type 1 as more
69 prevalent in eBL patients compared to geographically matched healthy controls, which
70 previously underrepresented the prevalence of EBV type 2.
- 71 • Genome-wide association analysis between cases and controls identifies 6 eBL-
72 associated nonsynonymous variants in EBNA1, EBNA2, BcLF1, and BARF1 genes.
- 73 • Analysis of population structure reveals that EBV type 2 exists as two genomic sub
74 groups, and was more commonly found in females than male eBL patients.

75 Introduction

76 EBV infects more than 90% of the world's population and typically persists as a chronic
77 asymptomatic infection (1). While most individuals endure a lifelong infection with minimal
78 effect, EBV is associated with ~1% of all human malignancies worldwide. EBV was first isolated
79 from an endemic Burkitt lymphoma (eBL) tumor which is the most prevalent pediatric cancer in
80 sub-Saharan Africa (2). Repeated *Plasmodium falciparum* infections during childhood appear to
81 drive this increased incidence (3). Malaria causes polyclonal B-cell expansion and increased
82 expression of activation-induced cytidine deaminase (AID) dependent DNA damage leading to
83 the hallmark translocation of the *MYC* gene under control of the constitutively active
84 immunoglobulin enhancer (4-6). How EBV potentiates eBL is incompletely understood,
85 however, the clonal presence of this virus in almost every eBL tumor suggests a necessary role.

86 EBV strains are categorized into two types based on the high degree of divergence in
87 the *EBNA2* and *EBNA3* genes (7-9). This long standing evolutionary division is also present in
88 orthologous primate viruses (10), yet remains unexplained. While EBV type 1 has been
89 extensively studied (11, 12), because it causes acute infectious mononucleosis and other
90 diseases in the developed world, type 2 virus studies have not kept pace since infected
91 individuals are less frequent and found primarily in sub-Saharan Africa. This view is changing as
92 several recent studies have reported a significant prevalence of type 2 circulating in western
93 countries, suggesting a greater role for type 1 and greater potential for interactions between the
94 EBV types worldwide (13, 14). For understanding endemic Burkitt, the African context provides
95 a direct opportunity to examine viral variation because type 1 and type 2 are found in both eBL
96 patients as well as healthy individuals (8, 15, 16). Viral variation has been shown to impact
97 differential transformation and growth, and capacity to block apoptosis or immune recognition
98 (7, 17, 18). However, studies focusing on only certain genomic regions/proteins potentially miss
99 disease associations of other loci (19, 20). Although new studies have been conducted (21, 22),

100 genome-wide examinations in case-control studies are few and often lack typing the virus. A
101 recent study which investigated whole EBV genomes for variant associations with
102 nasopharyngeal carcinoma among Chinese patients has discovered two variants associated
103 with increased disease risk (23). Similarly, another study investigated genome-wide variants of
104 HIV genomes in quest of finding the associations with drug resistance (24). However, to the
105 best of our knowledge, such viral genome-wide variant association for EBV and eBL remains to
106 be explored. To address this shortfall and provide a proof of concept to the field, we sequenced
107 a set EBV genomes in a disease/control setting.

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

118 **Materials and Methods**

119 **Ethical approval and sample collection**

120 For this study, we recruited children between 2009 and 2012 with suspected eBL,
121 between 2-14 years of age, undergoing initial diagnosis at Jaramogi Oginga Odinga Teaching
122 and Referral Hospital (JOOTRH; Kisumu), which is a regional referral hospital for pediatric
123 cancer in western Kenya (31). We also enrolled healthy age-matched children residing in the

124 same malaria endemic regions of Kenya as controls. We obtained written informed consent
125 from parents to enroll their child in this study. Ethical approval was obtained from the
126 Institutional Review Board at the University of Massachusetts Medical School and the Scientific
127 and Ethical Review Unit at the Kenya Medical Research Institute. From eBL patients, tumor
128 biopsies were collected using fine needle aspirates (FNA) and transferred into RNAlater at the
129 bedside, prior to induction of chemotherapy. From all children, peripheral blood samples were
130 collected and fractionated by centrifugation prior to freezing into plasma and cell pellets. All
131 samples were stored at -80°C prior to nucleic acid extraction.

132 **Cell cultures and controlled mixtures**

133 BL cultured cell lines, Namalwa, Daudi, Raji, and Jijoye were grown in a complete
134 growth medium, RPMI 1640 (Life Technologies), with 2mM L-glutamine adjusted to contain
135 1.5g/L sodium bicarbonate, 4.5g/L glucose, 10mM HEPES, 1.0mM sodium pyruvate, and 7.5%
136 fetal bovine serum. We used Jijoye and Daudi as representative genomes of type 1 and type 2
137 strains. For mixing experiments, we created relative ratios of Jijoye:Daudi of 10:90, 25:75,
138 75:25, 90:10 in addition to sequencing each strain individually.

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

140 We used Allprep DNA/RNA/Protein mini kit (Qiagen) for DNA isolations from FNAs and
141 QIAamp DNA Kit for blood and plasma. We developed an improved multi-step amplification and
142 enrichment process for the GC-rich EBV genome, particularly in samples with low viral copies.
143 We used EBV-specific whole genome amplification (sWGA) to provide sufficient material and
144 targeted enrichment with hybridization probes after the library preparation. For this, we designed
145 3'-protected oligos following the instructions from Leichthy et al. (32). For low viral load samples,
146 we added a multiplex long-range PCR amplification (mlrPCR) step, comprising two sets of non-
147 overlapping EBV-specific primers tiling across the genome (33). To increase viral DNA content
148 in low abundant specimens, we applied an initial amplification with long range PCR using a

149 strategy consisting of two multiplexed sets of primers which combined tiled the viral genome as
150 designed by Kwok et al. To this, we added EBV type 2 specific primers. Following the initial
151 multiplex long range (mlr) PCRs, we mixed two independent reactions and then performed
152 specific whole genome amplification (sWGA) using phi29 polymerase with EBV-specific oligos.
153 Overall DNA quality and quantities were assessed with NanoDrop and Picogreen and purified
154 with 2x XP-Ampure magnetic beads. We prepared two reaction solutions with separate primer
155 pools (2uL of 10uM each), using 2.5uL 10X long range PCR buffer Mg^{2+} (Qiagen), 1.25uL dNTP
156 (10mM each), 0.15uL long range PCR enzyme mix (Qiagen), 5uL 5X Q-solution (Qiagen) to
157 which we added 10ng of input DNA in 14 uL. The reaction conditions involved initial
158 denaturation at 95°C for 3 min followed by 20 cycles (95°C for 30 sec, gradient annealing 58°C
159 to 49°C for 15 sec each, extension at 72°C for 7 min) and a final extension at 68°C for 10 min.
160 We then mixed two independent reactions, denatured at 95°C for 3 min, and then added sWGA
161 reaction buffer that contains 7uL 10X phi29 reaction buffer (NEB), 3uL dNTP mix (final
162 concentrations: 30mM dGTP and dCTP; 10mM dATP and dTTP), 7uL EBV specific protected
163 oligo Mix (10uM each), 2uL phi29 polymerase (10U/uL, NEB), 0.7uL BSA (0.1ug/uL), and 0.3uL
164 H₂O. For samples with higher viral loads that did not require PCR amplification prior to sWGA,
165 we denatured DNA using the same conditions but replaced the reaction buffer with TE or TE-Q-
166 solution mix. We incubated the sWGA at 30°C for 16h followed by incubation at 65°C for 15 min
167 to stop the reaction. Instead of random hexamers for MDA (multiple strand displacement
168 amplification) reaction, we used EBV specific hexamers with 3'-end modification to protect
169 against phi29 exonuclease activity (see supplementary document for primer sequences). For
170 WGA with Genomiphi v2 kit, we followed the manufacturer's instructions modified by adding
171 extra 2x dGTP and dCTP. For hybrid capture, we followed MyBait protocol as manufacturer's
172 recommendations. Following incubation at 65°C for 72 hours, we purified hybridization products
173 with streptavidin beads and used Kapa HiFi to amplify the captured library. We quantified viral
174 content with bi-plex qPCR using primers for viral BALF5 and human beta-actin gene (34). For

175 validation of EBV subtypes, we used primers spanning EBNA3C gene producing 153bp and
176 246bp products for type 1 and type 2, respectively (see **Supplemental Table 1** for all primers).

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

183 **Sequence design for RNA baits**

184 Capture bait sequences were designed using in-house scripts to target both type 1 and
185 type 2. In addition to type 1 and type 2 references, we also designed against other available
186 complete genomes including Mutu I, Akata, GD1 and GD2 to ensure the capture of divergent
187 regions. Specifically, the design consisted of overlapping 120nt probes tiling every 30 bases
188 (4x overlapping tiling) across the genomic sequences with increased probes for regions with
189 elevated GC content (>65%). Additional probes were added based on the sequential analysis
190 of additional genomes, when current probes were greater than 5% divergent or there was a
191 gap in coverage for a specific region (**Supplemental Table 2**).

192 **Sequencing library preparation and hybrid capture enrichment**

193 Illumina sequencing library preparation steps consisted of DNA shearing, blunt-end
194 repair (Quick Blunting kit, NEB), 3'-adenylation (Klenow Fragment 3' to 5' exo-, NEB), and
195 ligation of indexed sequencing adaptors (Quick Ligation kit, NEB). We PCR amplified libraries to
196 a final concentration with 10 cycles using KAPA HiFi HotStart ReadyMix and quantified them
197 using bioanalyzer. We then pooled sample libraries balancing them according to their EBV
198 content and proceeded to target enrichment hybridization using custom EBV-specific
199 biotinylated RNA probes (MyBaits, Arbor Biosciences). We performed sequencing using

200 Illumina MiSeq, HiSeq 2000, and NextSeq 500 platforms with 1x75bp, 2x100bp, and 2x150bp,
201 respectively.

202 **Sequence preprocessing and de novo genome assembly**

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

214 **Diversity and variant association analysis**

215 We used Mafft (v7.215) (39) to generate multiple sequence alignment (msa) of
216 genomes, and masked the repetitive regions predefined in EBV reference genome,
217 NC_007605, in addition to repeat regions larger than 100nt detected by miropeat (40). For
218 analyses based on viral genes, we extracted the coding region sequences from msa of
219 assemblies according to the reference genome Genbank annotations. The substitutions in
220 coding regions were translated in protein sequences based on standard genetic code. The
221 genetic distance between the sequences were calculated using Kimura-2-parameter method
222 based on transition and transversion frequencies. We calculated dN/dS rates per gene based
223 on pairwise Nei-Gojobori algorithm using the python functions provided by

224 https://github.com/a1ultima/hpcleap_dnds/ after excluding frameshift insertions and ambiguous
225 bases. We constructed whole genome phylogenetic trees based on neighbor-joining method
226 and protein sequence trees based on maximum likelihood method with Jukes-Cantor
227 substitution model using MEGA (v6.0) (41). We determined variant sites of each isolate in
228 reference to the EBV reference genome, NC_007605, based on msa using snp-sites (v2.3.2)
229 (42). For principal coordinate analysis (PCoA), we used R package dartR (v1.0.5) (43). We
230 performed the variant association analysis using the 'v-assoc' function from PSEQ/PLINK (44).
231 To control for multiple testing, we calculated empirical p-values with one million permutations
232 (pseq proj v-assoc --phenotype eBL --fix-null --perm 1000000) with EBV type stratification which
233 permutes within types (--strata EBVtype).

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

239 Results

240 Study participant characteristics

241 The objective of this study was to examine EBV genetic variation in a region of western
242 Kenya with a high incidence of eBL (31) and determine if any variants are associated with eBL
243 pathogenesis. We leveraged specimens from eBL patients and healthy children residing in the
244 same geographic area (**Figure 1C**) (31). We sequenced the virus isolated from 58 eBL cases
245 and 40 healthy Kenyan children, as controls. Patients between 1 and 13 years of age were
246 predominantly male (74%), consistent with the sex ratio of eBL (**Table 2**) (31). Healthy controls
247 had similar levels of malaria exposure based on previous epidemiologic studies (45). Control
248 samples ranged in age from 1 to 6 years. This difference in age was necessarily due to the

249 finding that younger, healthy yet malaria-exposed children have higher average viral loads
250 compared to older children who have developed immune control over this chronic viral infection
251 (34). Therefore, it was infeasible to age-match controls with the eBL cases who tend to be older.
252 We make the assumption that children are infected with the same herpes virus throughout life
253 and therefore genomes from younger children reflect EBV genomes that would be found in older
254 children.

255 **Sequencing and assembly quality**

256 EBV is a large GC-rich double stranded DNA virus with a 172 kb genome of which ~20%
257 is a repetitive sequence. For the majority of eBL patients, we prepared sequencing libraries
258 directly from tumor DNA followed by hybrid capture enrichment. For low copy viral samples,
259 such as eBL plasma and healthy control blood, we designed and implemented additional viral
260 whole genome amplification and enrichment prior to library preparation and sequencing (**Figure**
261 **1C**). We generated a study set of 114 genomes including replicates from cell lines and primary
262 clinical samples, representing 98 cases and controls. In addition, we sequenced 20 technical
263 replicates for quality control purposes such as estimation of re-sequencing error or sWGA bias,
264 and sensitivity of detection of mixed infections. The baseline re-sequencing error rate was
265 limited to $\sim 1.1 \times 10^{-5}$ bases when our assemblies are compared with high-quality known strain
266 genomes (46) (**Table 3**). The mean error rate was $\sim 2.1 \times 10^{-5}$ bases for sWGA with GenomiPhi,
267 while it is $\sim 1.1 \times 10^{-4}$ bases when we used more sensitive mlrPCR-sWGA (Methods). We
268 obtained an average of ~5 million reads, resulting in an average 9,688 depth of coverage across
269 assemblies (**Supplemental Table 3**). De novo sequence assembly created large scaffolds
270 covering non-repetitive regions, except three isolates with low coverage, yielded a median of
271 137,887bp genomes (ranging 47,534bp - 146,920bp). We determined the types of each isolate
272 by calculating the nucleotide distance to both reference types in addition to read mapping rates
273 against type-specific regions. Despite our ability to experimentally detect mixed types at levels

274 as low as 10% (**Figure 2A**), we found no evidence of mixed infections in our cases and controls.
275 Also, to ensure that our sample inclusion was unbiased when selecting healthy individuals with
276 high enough viremias to sequence, we quantified the baseline viral loads with bi-plex qPCR
277 using primers for viral BALF5 and human beta-actin gene (see Supplemental methods). We
278 compared the viral loads and found no significant difference between type 1 and 2 ($P=0.529$,
279 **Figure 2B**).

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

281 Plasma EBV load has been studied to show its potential as a biomarker or as a
282 prognostic marker in various lymphomas including BL (47, 48). We included plasma specimens
283 along with the tumor biopsies from eBL patients in the whole genome sequencing set to
284 compare and contrast the two counterparts at the sequence level. Following the separate
285 sequencing and genome assembly of 6 pairs of plasma and tumor associated viruses from 6
286 patients, we confirmed that viral DNA in the plasma was representative of the virus in the tumor
287 cells (**Figure 2C**). Accounting for the sequencing errors, the pairs appeared to be identical
288 (**Supplemental Figure 1**). Out of these pairs, we further confirmed the subtype of three EBV
289 isolates (eBL-Tumor-0035, -0037, and -0038) from the plasma and tumor biopsies using type-
290 specific PCRs (see Supplemental method) in addition to 5 other samples (eBL-Tumor-0003, -
291 0019, -0022, -0029, and -0030). Overall, these findings demonstrate that viral DNA isolated
292 from eBL patient plasma represents the tumor virus and reflects its genome sequence to the
293 circulating system. This further assures the potential of plasma DNA for prognostic tools in
294 disease monitoring.

295 **Structural variation and intertypic recombinants**

296 First, we looked for large deletions within our viral genomes, but did not detect any of the
297 previously described deletions in EBNAs, even though we were able to detect, as positive
298 controls, EBNA3C deletion in Raji and the EBNA2 deletion in Daudi cell lines. However, in one

299 sample we did detect a novel 20kb deletion, spanning from 100 kb to 120 kb in the genome
300 (**Figure 2D**), which appears as the lack of sequencing read coverage while the rest of the
301 genome, even high GC regions, show high sequencing depth (> 6,000x in average). This
302 deleted region normally encodes multiple lytic phase genes, e.g. *BBRF1/2*, *BBLF1/3*,
303 *BGLF1/2/3/4/5*, and *BDLF2/3/4*. Interestingly, none of the latent genes were affected by this
304 deletion.

305 Next, we interrogated our isolates by comparing the pairwise similarities of each genome
306 against EBV type 1 and type 2 references. By traversing through the genome with a window, we
307 were able to delineate regions that were more similar to one type over the other (**Figure 2E**). As
308 expected, Jijoye, a type 2 strain, displayed less similarity against type 1 reference around its
309 *EBNA2* and *EBNA3* genes, the most divergent region between types, while Namalwa as a type
310 1 strain shows the same pattern of dissimilarity against type 2 reference around the same
311 regions. Interestingly, we found three patient-derived genomes, eBL-Tumor-0012, eBL-Tumor-
312 0033, and eBL-Plasma-0049, with mixed similarity trends. Similar to a previously detected
313 recombinant strain (LN827563.2_sLCL-1.18) (46), all of the intertypic isolates carried type 1
314 *EBNA2* and type 2 *EBNA3* genes. Although these new intertypic hybrids were all isolated from
315 eBL patients only as opposed to healthy controls, this finding does not reach to a statistical
316 significance ($P=0.268$, Chi-square).

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

319 Our samples present a unique opportunity to study population structure of EBV types
320 and their co-evolution within a geographically defined region. As expected, the major bifurcation
321 within the phylogenetic tree based on the entire genome occurs between type 1 and type 2
322 viruses (**Figure 3A**). Viruses from eBL patients as well as healthy controls appeared to be
323 intermixed almost randomly within the type 1 branch. Interestingly, within type 2 genomes 8

324 eBL-associated isolates formed a unique sub-cluster. The hybrid genomes are clustered with
325 type 2s, which is consistent with type 2 EBNA3s representing a greater amount of sequence
326 than the type 1 EBNA2 region.

327 We further explored viral population structure with principal coordinate analysis (PCoA)
328 of variation across the genome. While the first three components cumulatively explain 57.2% of
329 the total variance, the first component, which solely accounted for 43.9% of the variance,
330 separates genomes based on type 1 and type 2 (**Figure 3B, upper plot**). Similar to the
331 phylogenetic tree, intertypic genomes are positioned more closely to type 2s. Interestingly, the
332 second and predominantly third components separate type 2 viruses into two distinct clusters,
333 group A and B (**Figure 3B, lower plot**). These clusters were reflected, although not as
334 distinctly, in the structure of the phylogenetic tree in the Figure 3A as well. The PCoA loading
335 values, which accounts for 37.1% of the variance between the type 2 groups, are predominantly
336 driven by correlated variation spanning 70kb upstream of EBNA3C (**Figure 3C and D**).
337 Together these findings suggest that there are two EBV type 2 strains circulating within this
338 population. We also examined viral variation from the perspective of LMP1. Interestingly, the
339 vast majority of viruses were grouped into Alaskan and Mediterranean strains (**Supplemental**
340 **Figure 2**). The majority of genomes that carry Alaskan LMP1 are type 2 genomes, while
341 Mediterranean strain LMP1s are mostly from type 1 genomes. For all available LMP1 type 2
342 sequences, group A and group B correlated with Mediterranean and Alaskan, respectively.

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

344 We further explored the pattern and nature of genomic variation across the genome
345 comparing and contrasting EBV type 1 and type 2. Examining the pairwise divergence of coding
346 genes for all viral genomes, we found that the divergence was the highest in the type-specific
347 *EBNA* genes (*EBNA2* and *EBNA3s*), in particular, with *EBNA2* showing the greatest divergence
348 ($d=0.1313 \pm 2.3 \times 10^{-3}$) (**Figure 4A, upper panel**). Investigating each type separately, the

349 diversity within types was low for *EBNA2* and *EBNA3Cs*, consistent with type 1 and 2 being
350 separated by many fixed differences (**Figure 4A, middle panel**). In both types, intra-type
351 divergence was greatest for *EBNA1* and *LMP1*. Most remarkable was the fact that type 2
352 generally showed lower levels of divergence across the genome ($0.0047 \pm 3.7 \times 10^{-3}$ and 0.0025
353 $\pm 2.7 \times 10^{-3}$ for type 1 and type 2, respectively). We observed the same trend even with the
354 balanced sample sizes through random down-sampling (**Supplemental Figure 3A and B**).
355 Overall, these measures suggest that EBV gene evolutionary rates differ by types.

356 To explore signatures of evolutionary selection, we examined the dN/dS ratios within
357 coding sequences (**Figure 4A, lower panel**). Overall most genes showed signals of purifying
358 selection, as indicated by $\omega < 1.0$, except *LMP1*, *BARF0*, and *BKRF2* (only type 2).
359 Interestingly, with dN/dS measures, *EBNA2*, *BSLF1*, *BSLF2*, and *BLLF2* genes had relatively
360 higher rates in type 2 compared to type 1 ($P < 0.001$). Having significantly different ω values for
361 multiple genes ($P < 0.001$, t-test) can suggest the existence of differential evolutionary pressure
362 on these two divergent types. This can be interpreted as an ongoing adaptation process of type
363 2 genomes (through certain genes) in the population while fixed functions of these genes for the
364 type 1 genomes. Overall, the magnitude of average nonsynonymous and synonymous changes
365 per gene, normalized by gene length, reflect the high-level diversity accumulated in certain
366 genes (**Supplemental Figure 4**). Latency-associated genes generally have the highest non-
367 synonymous variant rates, but they also have the highest synonymous rates consistent with
368 longstanding divergence (**Figure 4B**). Other functional categories, including lytic genes, have
369 relatively low levels of nonsynonymous mutations suggesting stronger purifying selection (see
370 **Supplemental Table 4** for functional categories).

371 **Global context of Kenyan viruses**

372 To more broadly contextualize our viral population from western Kenya, we examined
373 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

399 equivalent frequencies regardless of their disease status (N=27 vs N=26 and N=12 vs N=16,
400 type 1 and type 2 among eBL and healthy control groups, respectively). We then expanded the
401 association analysis to all 6191 synonymous and non-synonymous single nucleotide variations
402 across the entire genome (**Figure 4D and Supplemental Table 6**). We conducted an initial
403 association test for each nonsynonymous variant and detected 133 significant associations
404 (**Supplemental Table 7 and Methods**). The vast majority of these variants were located within
405 the type1-type2 region given the highly correlated nature of this region (**Supplemental Figure**
406 **7**). We then stratified by type to detect variants independent of viral type. This yielded 6 variants
407 solely associated with the disease (**Table 5**). Variant 37668T>C represents a serine residue
408 change to a proline at the C-terminus of *EBNA2* (S485P) which is carried by 24/54 eBL cases;
409 while this variant was present in only 2/36 healthy controls. Two variants in *EBNA1* at
410 95773A>T and 95778T>G (N38Y and H39Q, respectively) were both observed in 3/57 eBL
411 isolates while their corresponding frequencies were 11/36 and 12/37 among healthy controls.
412 These two variants fall into one of the two chromosome binding domains of the EBNA1 protein
413 which plays a bridging role for tethering the viral episome to the host chromosome (50, 51).
414 Other two significant variants we detected are within the BcLF1 gene which encodes for viral
415 major capsid protein. This protein is the most essential component of the self-assembly
416 structures for the viral capsid (52). Elevated substitution rates in viruses of healthy controls as
417 opposed to eBL associated viruses comply with their role in capsid formation and pathogenesis.
418 The BARF1 variant with higher frequencies in eBL associated viruses that replaces the valine
419 residue with an alanine might provide a fitness advantage with its role as a soluble form of CSF-
420 1 receptor that neutralizes effects of human CSF-1. With this residual change in the protein,
421 BARF1 increases its sequence identity to human CD80 as both share the same amino acid,
422 alanine, at the position 29 (53).

423 Nucleotide variants in non-coding and promoter regions can affect regulation of viral
424 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

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

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

467 Importantly, we were able to explore EBV population genetics and compare and contrast
468 type 1 and type 2 because of their co-prevalence in Africa. As well described, the major
469 differentiation in terms of genetic variability was the variation correlated with type 1 and type 2
470 viruses. These viral types showed distinct population characteristics with type 1 harboring
471 greater diversity especially in functionally important latent genes. Combined with the observed
472 nucleotide diversity, latency genes appear to have long standing divergence that has
473 accumulated significant synonymous changes (as opposed to recent sweeps on
474 nonsynonymous changes that would erase synonymous variants). Global phylogenetic analysis

475 emphasizes this diversity by providing two main subgroups for type 1 genomes in our
476 sequencing set. One group represents core local Kenyan viruses while the second group is a
477 mixture of viruses from across the globe, with the exception of South Asian viruses that group
478 apart. While previously sequenced type 2 viruses intermingle with western Kenya isolates, the
479 majority of these originated from East Africa with only a few from West Africa. Interestingly,
480 intermingling is also true for type 2 as we observed two distinct groups. This is more apparent in
481 PCA where type 2 virus forms 2 clusters. Examination via PCA, the loading values are
482 determined by a broad stretch of the genome from the end of *EBNA3C* to *LMP1*, where
483 Mediterranean and Alaskan designations correlate. It remains to be determined whether this
484 substructure might be due to the introduction of previously geographically isolated viruses or
485 distinct evolutionary trajectories within the population. Further study is needed with broader
486 samplings to understand its significance but our findings suggest that there may be significant
487 epistasis potentially including *LMP1*.

488 By sequencing the virus directly from healthy controls, we were able to address the
489 question of relative tumorigenicity between EBV type 1 and 2. We evaluated the long-standing
490 presumption that type 1 virus is more strongly associated with eBL, in contrast to type 2. Our
491 work was able to more definitely answer this question as we were not reliant on LCLs from
492 healthy controls where type 1 bias in transformation might explain the lack of previous
493 associations. We earlier demonstrated, by mutational profiling of EBV positive and negative eBL
494 tumors, that the virus, especially type 1, might mitigate the necessity of certain driver mutations in
495 the host genome (16). In addition, our genome-wide results controlling for viral type substantiates
496 investigations of non-type associated variation that could also impart oncogenic risk, as we
497 found suggestive trends for several nonsynonymous variants as well. Supporting the putative
498 existence of EBV sub-strains that have increased oncogenic potential, we observed sub-cliques
499 of solely eBL or control isolates within the type 2 genomes. Although these sub-groups were
500 formed with only 7 or 8 members, the significance of this observation will be deciphered with

501 more extensive cohorts. On the other hand, only a small subset of type 1 viruses from eBL
502 patients carried *BZLF1* promoter variant, which leads to a gain of function (54), while all type 2
503 viruses carried this variant suggesting this promoter might be beneficial for type 2 but makes it
504 unlikely to be a driver of oncogenesis. It is essential to remind that the suggestive associations
505 we uncovered are in need of further validation with independent cohorts and treated cautiously.

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

511 **Acknowledgements**

512 This work was supported by the US National Institutes of Health, National Cancer
513 Institute R01 CA134051, R01 CA189806 (A.M.M., J.A.B, C.I.O, Y.K.) and The Thrasher
514 Research Fund 02833-7 (A.M.M.), UMCCTS Pilot Project Program U1 LTR000161-04 (Y.K.,
515 J.A.B., and A.M.M.), Turkish Ministry of National Education Graduate Study Abroad Program
516 (Y.K.). We would like to thank the Kenyan children and their families who participated in this
517 study. Patrick Marsh for helping with EBV genotyping assays, Mercedeh Movassagh for sharing
518 genotyping primers. This publication was approved by the Director of KEMRI. The current
519 affiliation for Yasin Kaymaz is FAS Informatics and Scientific Applications, Harvard University,
520 Cambridge, MA

521 **Authorship Contributions**

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

526 **Disclosure of Conflicts of Interest**

527 The authors declare no competing financial interests.

528 References

- 529 1. Young LS, Rickinson AB. 2004. Epstein-Barr virus: 40 years on. *Nat Rev Cancer* 4:757-
530 768.
- 531 2. Crawford DH. 2001. Biology and disease associations of Epstein-Barr virus. *Philos*
532 *Trans R Soc Lond B Biol Sci* 356:461-473.
- 533 3. Moormann AM, Bailey JA. 2016. Malaria—how this parasitic infection aids and abets
534 EBV-associated Burkitt lymphomagenesis. *Curr Opin Virol*.
- 535 4. Torgbor C, Awuah P, Deitsch K, Kalantari P, Duca KA, Thorley-Lawson DA. 2014. A
536 multifactorial role for *P. falciparum* malaria in endemic Burkitt's lymphoma pathogenesis.
537 *PLoS Pathog* 10:e1004170.
- 538 5. Simone O, Bejarano MT, Pierce SK, Antonaci S, Wahlgren M, Troye-Blomberg M,
539 Donati D. 2011. TLRs innate immunoreceptors and *Plasmodium falciparum* erythrocyte
540 membrane protein 1 (PfEMP1) CIDR1 α -driven human polyclonal B-cell activation. *Acta*
541 *Trop* 119:144-150.
- 542 6. Robbiani DF, Deroubaix S, Feldhahn N, Oliveira TY, Callen E, Wang Q, Jankovic M,
543 Silva IT, Rommel PC, Bosque D, Eisenreich T, Nussenzweig A, Nussenzweig MC. 2015.
544 *Plasmodium* Infection Promotes Genomic Instability and AID-Dependent B Cell
545 Lymphoma. *Cell* 162:727-737.
- 546 7. Cohen JI, Wang F, Mannick J, Kieff E. 1989. Epstein-Barr virus nuclear protein 2 is a
547 key determinant of lymphocyte transformation. *Proc Natl Acad Sci U S A* 86:9558-9562.
- 548 8. Rowe M, Young LS, Cadwallader K, Petti L, Kieff E, Rickinson AB. 1989. Distinction
549 between Epstein-Barr virus type A (EBNA 2A) and type B (EBNA 2B) isolates extends to
550 the EBNA 3 family of nuclear proteins. *J Virol* 63:1031-1039.
- 551 9. Dambaugh T, Hennessy K, Chamnankit L, Kieff E. 1984. U2 region of Epstein-Barr virus
552 DNA may encode Epstein-Barr nuclear antigen 2. *Proc Natl Acad Sci U S A* 81:7632-
553 7636.
- 554 10. Cho YG, Gordadze AV, Ling PD, Wang F. 1999. Evolution of two types of rhesus
555 lymphocryptovirus similar to type 1 and type 2 Epstein-Barr virus. *J Virol* 73:9206-9212.
- 556 11. Zimmer U, Adldinger HK, Lenoir GM, Vuillaume M, Knebel-Doeberitz MV, Laux G,
557 Desgranges C, Wittmann P, Freese UK, Schneider U. 1986. Geographical prevalence of
558 two types of Epstein-Barr virus. *Virology* 154:56-66.
- 559 12. Apolloni A, Sculley TB. 1994. Detection of A-Type and B-Type Epstein-Bart Virus in
560 Throat Washings and Lymphocytes. *Virology* 202:978-981.
- 561 13. Sixbey JW, Shirley P, Chesney PJ, Buntin DM, Resnick L. 1989. Detection of a second
562 widespread strain of Epstein-Barr virus. *Lancet* 2:761-765.
- 563 14. Correia S, Palser A, Elgueta Karstegl C, Middeldorp JM, Ramayanti O, Cohen JI,
564 Hildesheim A, Fellner MD, Wiels J, White RE, Kellam P, Farrell PJ. 2017. Natural
565 variation of Epstein-Barr virus genes, proteins and pri-miRNA (revised). *J Virol*
566 doi:10.1128/JVI.00375-17.
- 567 15. Young LS, Yao QY, Rooney CM, Sculley TB, Moss DJ, Rupani H, Laux G, Bornkamm
568 GW, Rickinson AB. 1987. New type B isolates of Epstein-Barr virus from Burkitt's
569 lymphoma and from normal individuals in endemic areas. *J Gen Virol* 68 (Pt 11):2853-
570 2862.
- 571 16. Kaymaz Y, Oduor CI, Yu H, Otieno JA, Ong'echa JM, Moormann AM, Bailey JA. 2017.
572 Comprehensive Transcriptome and Mutational Profiling of Endemic Burkitt Lymphoma
573 Reveals EBV Type-Specific Differences. *Mol Cancer Res* 15:563-576.
- 574 17. Lucchesi W, Brady G, Dittrich-Breiholz O, Kracht M, Russ R, Farrell PJ. 2008.
575 Differential gene regulation by Epstein-Barr virus type 1 and type 2 EBNA2. *J Virol*
576 82:7456-7466.

- 577 18. Kaye KM, Izumi KM, Kieff E. 1993. Epstein-Barr virus latent membrane protein 1 is
578 essential for B-lymphocyte growth transformation. *Proc Natl Acad Sci U S A* 90:9150-
579 9154.
- 580 19. Wohlford EM, Asito AS, Chelimo K, Sumba PO, Baresel PC, Oot RA, Moormann AM,
581 Rochford R. 2013. Identification of a novel variant of LMP-1 of EBV in patients with
582 endemic Burkitt lymphoma in western Kenya. *Infect Agent Cancer* 8:34.
- 583 20. Chang CM, Yu KJ, Mbulaiteye SM, Hildesheim A, Bhatia K. 2009. The extent of genetic
584 diversity of Epstein-Barr virus and its geographic and disease patterns: a need for
585 reappraisal. *Virus Res* 143:209-221.
- 586 21. Chiara M, Manzari C, Lionetti C, Mechelli R, Anastasiadou E, Chiara Buscarinu M,
587 Ristori G, Salvetti M, Picardi E, D'Erchia AM, Pesole G, Horner DS. 2016. Geographic
588 Population Structure in Epstein-Barr Virus Revealed by Comparative Genomics.
589 *Genome Biol Evol* 8:3284-3291.
- 590 22. Zhou L, Chen J-N, Qiu X-M, Pan Y-H, Zhang Z-G, Shao C-K. 2017. Comparative
591 analysis of 22 Epstein-Barr virus genomes from diseased and healthy individuals. *J Gen*
592 *Virol* 98:96-107.
- 593 23. Xu M, Yao Y, Chen H, Zhang S, Cao S-M, Zhang Z, Luo B, Liu Z, Li Z, Xiang T, He G,
594 Feng Q-S, Chen L-Z, Guo X, Jia W-H, Chen M-Y, Zhang X, Xie S-H, Peng R, Chang ET,
595 Pedergrana V, Feng L, Bei J-X, Xu R-H, Zeng M-S, Ye W, Adami H-O, Lin X, Zhai W,
596 Zeng Y-X, Liu J. 2019. Genome sequencing analysis identifies Epstein-Barr virus
597 subtypes associated with high risk of nasopharyngeal carcinoma. *Nat Genet* 51:1131-
598 1136.
- 599 24. Power RA, Davaniah S, Derache A, Wilkinson E, Tanser F, Gupta RK, Pillay D, de
600 Oliveira T. 2016. Genome-wide association study of HIV whole genome sequences
601 validated using drug resistance. *bioRxiv* doi:10.1101/076216.
- 602 25. Depledge DP, Palser AL, Watson SJ, Lai IY-C, Gray ER, Grant P, Kanda RK, Leproust
603 E, Kellam P, Breuer J. 2011. Specific capture and whole-genome sequencing of viruses
604 from clinical samples. *PLoS One* 6:e27805.
- 605 26. Kwok H, Wu CW, Palser AL, Kellam P, Sham PC, Kwong DLW, Chiang AKS. 2014.
606 Genomic diversity of Epstein-Barr virus genomes isolated from primary nasopharyngeal
607 carcinoma biopsy samples. *J Virol* 88:10662-10672.
- 608 27. Liu Y, Yang W, Pan Y, Ji J, Lu Z, Ke Y. 2016. Genome-wide analysis of Epstein-Barr
609 virus (EBV) isolated from EBV-associated gastric carcinoma (EBVaGC). *Oncotarget*
610 7:4903-4914.
- 611 28. Wang S, Xiong H, Yan S, Wu N, Lu Z. 2016. Identification and Characterization of
612 Epstein-Barr Virus Genomes in Lung Carcinoma Biopsy Samples by Next-Generation
613 Sequencing Technology. *Sci Rep* 6:26156.
- 614 29. Lei H, Li T, Li B, Tsai S, Biggar RJ, Nkrumah F, Neequaye J, Gutierrez M, Epelman S,
615 Mbulaiteye SM, Bhatia K, Lo S-C. 2015. Epstein-Barr virus from Burkitt Lymphoma
616 biopsies from Africa and South America share novel LMP-1 promoter and gene
617 variations. *Sci Rep* 5:16706.
- 618 30. Parras-Moltó M, López-Bueno A. 2018. Methods for Enrichment and Sequencing of Oral
619 Viral Assemblages: Saliva, Oral Mucosa, and Dental Plaque Viromes. *Methods Mol Biol*
620 1838:143-161.
- 621 31. Buckle G, Maranda L, Skiles J, Ong'echa JM, Foley J, Epstein M, Vik TA, Schroeder A,
622 Lemberger J, Rosmarin A, Remick SC, Bailey JA, Vulule J, Otieno JA, Moormann AM.
623 2016. Factors influencing survival among Kenyan children diagnosed with endemic
624 Burkitt lymphoma between 2003 and 2011: A historical cohort study. *Int J Cancer*
625 139:1231-1240.
- 626 32. Leichty AR, Brisson D. 2014. Selective whole genome amplification for resequencing
627 target microbial species from complex natural samples. *Genetics* 198:473-481.

- 628 33. Kwok H, Tong AHY, Lin CH, Lok S, Farrell PJ, Kwong DLW, Chiang AKS. 2012.
629 Genomic sequencing and comparative analysis of Epstein-Barr virus genome isolated
630 from primary nasopharyngeal carcinoma biopsy. *PLoS One* 7:e36939.
- 631 34. Moormann AM, Chelimo K, Sumba OP, Lutzke ML, Ploutz-Snyder R, Newton D, Kazura
632 J, Rochford R. 2005. Exposure to holoendemic malaria results in elevated Epstein-Barr
633 virus loads in children. *J Infect Dis* 191:1233-1238.
- 634 35. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing
635 reads. *EMBnetjournal* 17:10-12.
- 636 36. Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic
637 datasets. *Bioinformatics* 27:863-864.
- 638 37. Consortium VB, Others. 2012. Velvetoptimiser. Available: bioinformatics.net.au/software/velvetoptimiser.shtml Accessed 22.
- 639
640 38. Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, Otto TD. 2012. A post-assembly
641 genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat*
642 *Protoc* 7:1260-1284.
- 643 39. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
644 improvements in performance and usability. *Mol Biol Evol* 30:772-780.
- 645 40. Parsons JD. 1995. Miropeats: graphical DNA sequence comparisons. *Comput Appl*
646 *Biosci* 11:615-619.
- 647 41. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics
648 Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* 33:1870-1874.
- 649 42. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016.
650 SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb*
651 *Genom* 2:e000056.
- 652 43. Gruber B, Unmack PJ, Berry OF, Georges A. 2018. dartr: An r package to facilitate
653 analysis of SNP data generated from reduced representation genome sequencing. *Mol*
654 *Ecol Resour* 18:691-699.
- 655 44. Anonymous. PLINK/SEQ genetics library.
- 656 45. Rainey JJ, Mwanda WO, Wairimu P, Moormann AM, Wilson ML, Rochford R. 2007.
657 Spatial distribution of Burkitt's lymphoma in Kenya and association with malaria risk.
658 *Trop Med Int Health* 12:936-943.
- 659 46. Palser AL, Grayson NE, White RE, Corton C, Correia S, Ba Abdullah MM, Watson SJ,
660 Cotten M, Arrand JR, Murray PG, Allday MJ, Rickinson AB, Young LS, Farrell PJ,
661 Kellam P. 2015. Genome diversity of Epstein-Barr virus from multiple tumor types and
662 normal infection. *J Virol* 89:5222-5237.
- 663 47. Muncunill J, Baptista M-J, Hernandez-Rodríguez Á, Dalmau J, Garcia O, Tapia G,
664 Moreno M, Sancho J-M, Martínez-Picado J, Feliu E, Mate J-L, Ribera J-M, Navarro J-T.
665 2019. Plasma Epstein-Barr Virus Load as an Early Biomarker and Prognostic Factor of
666 Human Immunodeficiency Virus-related Lymphomas. *Clin Infect Dis* 68:834-843.
- 667 48. Westmoreland KD, Montgomery ND, Stanley CC, El-Mallawany NK, Wasswa P, van der
668 Gronde T, Mtete I, Butia M, Itimu S, Chasela M, Mtunda M, Kampani C, Liomba NG,
669 Tomoka T, Dhungel BM, Sanders MK, Krysiak R, Kazembe P, Dittmer DP, Fedoriw Y,
670 Gopal S. 2017. Plasma Epstein-Barr virus DNA for pediatric Burkitt lymphoma diagnosis,
671 prognosis and response assessment in Malawi. *Int J Cancer* doi:10.1002/ijc.30682.
- 672 49. Correia S, Bridges R, Wegner F, Venturini C, Palser A, Middeldorp JM, Cohen JI,
673 Lorenzetti MA, Bassano I, White RE, Kellam P, Breuer J, Farrell PJ. 2018. Sequence
674 Variation of Epstein-Barr Virus: Viral Types, Geography, Codon Usage, and Diseases. *J*
675 *Virol* 92.
- 676 50. Sears J, Ujihara M, Wong S, Ott C, Middeldorp J, Aiyar A. 2004. The amino terminus of
677 Epstein-Barr virus (EBV) nuclear antigen 1 contains at hooks that facilitate the

- 678 replication and partitioning of latent EBV genomes by tethering them to cellular
679 chromosomes. *J Virol* 78:11487-11505.
- 680 51. Kanda T, Horikoshi N, Murata T, Kawashima D, Sugimoto A, Narita Y, Kurumizaka H,
681 Tsurumi T. 2013. Interaction between basic residues of Epstein-Barr virus EBNA1
682 protein and cellular chromatin mediates viral plasmid maintenance. *J Biol Chem*
683 288:24189-24199.
- 684 52. Henson BW, Perkins EM, Cothran JE, Desai P. 2009. Self-assembly of Epstein-Barr
685 virus capsids. *J Virol* 83:3877-3890.
- 686 53. Tarbouriech N, Ruggiero F, de Turenne-Tessier M, Ooka T, Burmeister WP. 2006.
687 Structure of the Epstein-Barr virus oncogene BARF1. *J Mol Biol* 359:667-678.
- 688 54. Bristol JA, Djavadian R, Albright ER, Coleman CB, Ohashi M, Hayes M, Romero-
689 Masters JC, Barlow EA, Farrell PJ, Rochford R, Kalejta RF, Johannsen EC, Kenney SC.
690 2018. A cancer-associated Epstein-Barr virus BZLF1 promoter variant enhances lytic
691 infection. *PLoS Pathog* 14:e1007179.
- 692 55. Mulama DH, Bailey JA, Foley J, Chelimo K, Ouma C, Jura WG, Otieno J, Vulule J,
693 Moormann AM. 2014. Sickle cell trait is not associated with endemic Burkitt lymphoma:
694 An ethnicity and malaria endemicity-matched case--control study suggests factors
695 controlling EBV may serve as a predictive biomarker for this pediatric cancer.
696 *International Journal of Cancer* 134:645-653.
- 697 56. Cho S-G, Lee W-K. 2000. Analysis of Genetic Polymorphisms of Epstein-Barr Virus
698 Isolates from Cancer Patients and Healthy Carriers. *J Microbiol Biotechnol* 10:620-627.
- 699 57. Burrows JM, Khanna R, Sculley TB, Alpers MP, Moss DJ, Burrows SR. 1996.
700 Identification of a naturally occurring recombinant Epstein-Barr virus isolate from New
701 Guinea that encodes both type 1 and type 2 nuclear antigen sequences. *J Virol* 70:4829-
702 4833.
- 703 58. Yao QY, Tierney RJ, Croom-Carter D, Cooper GM, Ellis CJ, Rowe M, Rickinson AB.
704 1996. Isolation of intertypic recombinants of Epstein-Barr virus from T-cell-
705 immunocompromised individuals. *J Virol* 70:4895-4903.
- 706 59. Skare J, Farley J, Strominger JL, Fresen KO, Cho MS, zur Hausen H. 1985.
707 Transformation by Epstein-Barr virus requires DNA sequences in the region of BamHI
708 fragments Y and H. *J Virol* 55:286-297.
- 709

710

711 **Tables:**712 **Table 1. Optimization of mlrPCR-sWGA and whole genome amplification reactions.**

	EBV Input copies for sWGA*	Denaturation Buffer	Average EBV Input (genome copy/uL)	Average Human Input (B-actin/uL)	dNTP composition (G/C/T/A ratio)	Whole Genome Amplification Incubation	EBV DNA Output (copy/ng)	Human DNA Output (B-actin/ng)	EBV Enrichment (EBVpost / EBVpre) / (DNApost / DNApre)	Human Enrichment (Humanpost / Humanpre) / (DNApost / DNApre)	EBV Enrichment / Human Enrichment
Input EBV copies, dNTP Composition	10,000	TE	62,448	3,456	30/30/10/10	16 hours at 30C	4,573	7	1.06	0.03	35.33
	1,000	TE	10,038	13,911	30/30/10/10	16 hours at 30C	505	59	1.01	0.08	12.63
	100	TE	467	10,883	30/30/10/10	16 hours at 30C	55	70	2.28	0.12	19.00
	10	TE	10	11,326	30/30/10/10	16 hours at 30C	4	68	8.26	0.12	68.83
	1	TE	6	19,483	30/30/10/10	16 hours at 30C	0	55	0.21	0.06	3.50
	0	TE		11,551	30/30/10/10	16 hours at 30C		76		0.23	0.00
	10,000	TE	62,448	3,456	30/30/5/5	16 hours at 30C	4,847	2	1.12	0.01	112.00
	1,000	TE	10,038	13,911	30/30/5/5	16 hours at 30C	544	75	1.09	0.11	9.91
	100	TE	467	10,883	30/30/5/5	16 hours at 30C	72	65	3	0.12	25.00
	10	TE	10	11,326	30/30/5/5	16 hours at 30C	3	65	5.97	0.11	54.27
	1	TE	6	19,483	30/30/5/5	16 hours at 30C	0	81	0.22	0.09	2.44
	0	TE		11,551	30/30/5/5	16 hours at 30C		71		0.21	0.00
	10,000	TE	62,448	3,456	15/15/5/5	16 hours at 30C	3,483	16	0.81	0.07	11.57
	1,000	TE	10,038	13,911	15/15/5/5	16 hours at 30C	488	84	0.97	0.12	8.08
	100	TE	467	10,883	15/15/5/5	16 hours at 30C	73	109	3.04	0.19	16.00
	10	TE	10	11,326	15/15/5/5	16 hours at 30C	1	80	2.71	0.14	19.36
	1	TE	6	19,483	15/15/5/5	16 hours at 30C	0	103	0.25	0.11	2.27
	0	TE		11,551	15/15/5/5	16 hours at 30C		118		0.35	0.00
	10,000	TE	62,448	3,456	15/15/2/2	16 hours at 30C	3,324	12	0.77	0.05	15.40
	1,000	TE	10,038	13,911	15/15/2/2	16 hours at 30C	533	77	1.06	0.11	9.64
	100	TE	467	10,883	15/15/2/2	16 hours at 30C	78	105	3.24	0.19	17.05
	10	TE	10	11,326	15/15/2/2	16 hours at 30C	4	105	7.8	0.18	43.33
	1	TE	6	19,483	15/15/2/2	16 hours at 30C	0	110	0	0.12	0.00
	0	TE		11,551	15/15/2/2	16 hours at 30C		108		0.32	0.00
Denaturation Buffer, Incubation time and temperature	10	TE	10	11,326	30/30/5/5	8 hours at 30C	2	40	6.31	0.14	45.07
	10	TE	10	11,326	30/30/5/5	16 hours at 30C	2	40	8.74	0.14	62.43
	10	TE	10	11,326	30/30/5/5	8 hours at 35C	1	22	4.05	0.08	50.63
	10	TE	10	11,326	30/30/5/5	16 hours at 35C	2	22	6.14	0.08	76.75
	10	TE+Q sol	10	11,326	30/30/5/5	8 hours at 30C	2	41	9.44	0.14	67.43
	10	TE+Q sol	10	11,326	30/30/5/5	16 hours at 30C	3	60	13.94	0.21	66.38
	10	TE+Q sol	10	11,326	30/30/5/5	8 hours at 35C	3	28	10.72	0.1	107.20
	10	TE+Q sol	10	11,326	30/30/5/5	16 hours at 35C	4	49	15.13	0.17	89.00

713 * total EBV genomes 20 uL of input for sWGA

714

715 **Table 2. Characteristics of children included in EBV sequencing analysis.**

		eBL Patients (N=58)	Healthy Controls (N=40)
Age at collection, N (%)	<6 (yrs)	16 (27.6)	39 (97.5)
	7 - 13 (yrs)	42 (72.4)	1 (2.5)
Sex, N (%)	Female/Male	15/43 (25.9/74.1)	20/20 (50.0/50.0)
Obtained Specimen, N (%)	Tumor biopsy	41 (41.8)	-
	Blood	-	40 (100.0)
	Plasma	14 (14.2)	-
	New cultured eBL	3 (3.0)	-

716

717

718 Table 3. Estimated sequencing error rates based on replicates and controls.

Prior Amplification	Control Experimental Assembly	Reference Assembly	Number of Substitution Errors	Number of Correct Bases	Error Rate (per base)	Mean Error Rate (per base)
No amplification	Jijoye	Jijoye Assembly*	1	134,118	7.46E-06	1.13E-05
No amplification	Daudi	Daudi Assembly*	0	132,780	0.00E+00	
No amplification	Raji_Rep1	Raji Assembly*	0	130,560	0.00E+00	
No amplification	Raji_Rep2	Raji Assembly*	5	132,736	3.77E-05	
GenomiPhi-WGA	Raji_wga	Raji Assembly*	2	132,836	1.51E-05	2.13E-05
GenomiPhi-WGA	eBLtumor-01_EBV_type1_wga	eBLtumor-01_EBV_type1#	1	140,069	7.14E-06	
GenomiPhi-WGA	eBLtumor-02_EBV_type2_wga	eBLtumor-02_EBV_type2#	4	142,932	2.80E-05	
GenomiPhi-WGA	eBLtumor-03_EBV_type2_wga	eBLtumor-03_EBV_type2#	1	142,154	7.03E-06	
GenomiPhi-WGA	eBLtumor-04_EBV_type1_wga	eBLtumor-04_EBV_type1#	5	141,643	3.53E-05	
GenomiPhi-WGA	eBLtumor-05_EBV_type1_wga	eBLtumor-05_EBV_type1#	5	141,643	3.53E-05	1.09E-04
mlrPCR-sWGA	Raji_Rep1_mlrPCR-sWGA	Raji Assembly*	6	98,507	6.09E-05	
mlrPCR-sWGA	Raji_Rep2_mlrPCR-sWGA	Raji Assembly*	11	130,538	8.43E-05	
mlrPCR-sWGA	Jijoye_Rep1_mlrPCR-sWGA	Jijoye Assembly*	13	132,045	9.84E-05	
mlrPCR-sWGA	Daudi_mlrPCR-sWGA	Daudi Assembly*	25	130,160	1.92E-04	

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

725
726

727 **Table 4. Characteristics of children included in our study and viral genome subtypes.**
728

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

729 EBV subtypes are represented as 1, 2 and 1/2 for inter-typic.
730
731

732 **Table 5. Single nucleotide variants associated with eBL.**

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

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

742

743 Figure Legends

744 **Figure 1. Optimized EBV genome sequencing from tumors and primary clinical samples.**

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

752

753 **Figure 2. Sequencing and detection quality, plasma-tumor pairs, and atypical genome isolates.**

754 **A)** Controls for putative mixed infections and sampling bias against EBV types. The sensitivity of EBV
755 genome typing approach measured by accurate type assignments of in-lab mixtures with predefined
756 ratios. Each mixture of Daudi (type 1) and Jijoye (type 2) with varying ratios was prepared in
757 replicates. Following the genome assembly, type of the major strain was determined as judged by the
758 distance to both reference viral genomes. **B)** Comparison of viral load levels of individuals carrying
759 different EBV types ($P=0.529$, t-test). The non-significant difference suggests unbiased sampling
760 among either type regardless of viral loads. We quantified viral loads with bi-plex qPCR using primers
761 for viral BALF5 and human beta-actin gene. **C)** Comparison of virus from paired tumor (filled pink
762 circles) and plasma samples (hollow pink circles) at diagnosis shows viral DNA circulating in the
763 peripheral blood represents the virus in the tumor. The Neighbor-Joining tree is scaled (0.001
764 substitutions per site) and includes standard reference genomes for type 1 (NC007605, blue
765 diamond) and type 2 (NC009334, red diamond). **D)** The depth of coverage showing an absence of
766 reads from approximately 100 kb to 120 kb is indicative of a large deletion in the virus from an eBL
767 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 axes 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.

792 **Figure 4. Diversity in EBV coding genes, significant associations of EBV type 1 genomes**
793 **and single nucleotide variants with eBL.**

794 **A)** Genetic distance metrics of each EBV gene calculated based on the Kimura-2-parameter
795 method averaged across all genomes (upper panel) or type 1 / type 2 separately (middle panel).
796 Lower panel shows nonsynonymous to synonymous change (dN/dS) ratios of viral protein
797 coding genes averaged across all pairwise comparisons within each group separately. Error
798 bars represent standard error of mean. (Three intertypic genomes are excluded). **B)** Average
799 synonymous and non-synonymous variants in genes are summarized as functional categories
800 of genes. Variant level represents the number of variants per gene normalized by gene length in
801 kb. **C)** The frequency of type 1 and type 2 genomes identified from eBL patients and healthy
802 control children (excluding the three intertypic hybrid genomes is significantly different ($P=0.007$,
803 Fisher's exact). **D)** Manhattan plot for genome-wide associations of all single nucleotide variants
804 tested for frequency differences between cases and controls controlling for type specific
805 variants. The significance of each locus association is represented with an empirical p-value
806 (negative log10 scale) that was calculated by 1 million permutations with random label
807 swapping. Permutations were stratified for EBV genome type and adjusted for the missing
808 genotypes due to lack of coverage. All significant variants associated with eBL cases are shown
809 in red ($P < 0.01$). Nucleotide positions are according to type 1 reference genome.

Figure 1

Figure 2

Figure 3

Figure 4