

RESEARCH ARTICLE

Infectious Causes of Cancer

Incorporation of Epstein–Barr viral variation implicates significance of Latent Membrane Protein 1 in survival prediction and prognostic subgrouping in Burkitt lymphoma

Isaac E. Kim Jr.^{1,2}  | Cliff Oduor¹ | Julian Stamp¹ | Micah A. Luftig^{3,4} | Ann M. Moormann⁵ | Lorin Crawford^{1,6}  | Jeffrey A. Bailey^{1,2,7} 

¹Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, USA

²The Warren Alpert Medical School, Brown University, Providence, Rhode Island, USA

³Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, North Carolina, USA

⁴Center for Virology, Duke University School of Medicine, Durham, North Carolina, USA

⁵Division of Infectious Diseases and Immunology, Department of Medicine, University of Massachusetts Chan Medical School, Worcester, Massachusetts, USA

⁶Microsoft Research, Cambridge, Massachusetts, USA

⁷Department of Pathology and Laboratory Medicine, Brown University, Providence, Rhode Island, USA

Correspondence

Jeffrey A. Bailey, Department of Pathology and Laboratory Medicine, Brown University, Box G-E5, Providence, RI 02912, USA.
Email: jeffrey_bailey@brown.edu

Lorin Crawford, Microsoft Research, 1 Memorial Drive, Cambridge, MA 02142, USA.
Email: lorin_crawford@brown.edu; lcrawford@microsoft.com

Funding information

Blavatnik Family Foundation; National Institutes of Health, Grant/Award Numbers: R01 CA189806, R01 CA234348

Abstract

Although Epstein–Barr virus (EBV) plays a role in Burkitt lymphoma (BL) tumorigenesis, it is unclear if EBV genetic variation impacts clinical outcomes. From 130 publicly available whole-genome tumor sequences of EBV-positive BL patients, we used least absolute shrinkage and selection operator (LASSO) regression and Bayesian variable selection models within a Cox proportional hazards framework to select the top EBV variants, putative driver genes, and clinical features associated with patient survival time. These features were incorporated into survival prediction and prognostic subgrouping models. Our model yielded 22 EBV variants, including seven in latent membrane protein 1 (LMP1), as most associated with patient survival time. Using the top EBV variants, driver genes, and clinical features, we defined three prognostic subgroups that demonstrated differential survival rates, laying the foundation for incorporating EBV variants such as those in LMP1 as predictive biomarker candidates in future studies.

KEYWORDS

Burkitt lymphoma, Epstein–Barr virus, LMP1

What's New?

Although Epstein–Barr virus (EBV) plays a role in Burkitt lymphoma tumorigenesis, it is unclear if EBV genetic variation impacts its clinical outcomes. Here, the authors incorporated EBV variation along with clinical features and driver gene mutations into survival models for Burkitt lymphoma patients with EBV-positive tumors. The findings offer key insights into the role of specific EBV variants in patient survival time and their interactions with driver genes and clinical features in Burkitt lymphoma. The study also highlights the potential use of EBV variants, especially those in latent membrane protein 1, as predictive biomarker candidates.

1 | INTRODUCTION

Burkitt lymphoma (BL) comprises 30%–50% of all pediatric cancers in equatorial Africa.¹ Molecularly characterized by a

translocation of the immunoglobulin enhancer upstream of the MYC gene, BL has traditionally been categorized into three epidemiologic subtypes (endemic BL, sporadic BL, and HIV-BL).² Sporadic BL is most often diagnosed in adolescents, adults, and elderly patients residing in non-tropical countries and has a survival rate greater than 90%, unlike endemic Burkitt lymphoma

Lorin Crawford and Jeffrey A. Bailey share senior authorship.

(eBL), which is primarily diagnosed in children younger than 9 years of age living in equatorial Africa and has a lower survival rate of 50%, likely largely due to access to care.³ It has been hypothesized that repeated infections with the malarial parasite, *Plasmodium falciparum*, combined with early-age infection with Epstein–Barr virus (EBV), induces polyclonal B-cell expansion and activation-induced cytidine deaminase-dependent DNA damage, increasing the probability of the MYC translocation.^{4,5} More importantly, the studies have increasingly supported the idea that the presence or absence of EBV may represent a more clinically/molecularly relevant means of subtyping as opposed to traditional epidemiologic divisions.^{6–8}

Known systematically as human herpesvirus 4, EBV has a double-stranded DNA genome measuring 172 kb.⁹ During initial infection of the epithelial cells, which primarily occurs via saliva,¹⁰ EBV has two primary stages: lytic and latent. Of these, the latent stage is most critical for cell immortalization and tumorigenesis and has limited states of protein expression involving the latent membrane proteins (LMP1, LMP2A, LMP2B) and the EBV nuclear proteins (EBNA1, EBNA2, EBNA-3A, EBNA-3B, EBNA-3C, EBNA-LP), of which EBNA2 and EBNA3 show deep divergence, classifying the virus into type 1 and type 2.¹¹

Historically, studies have conformed to the dogma that EBV in BL is confined to the latency I state in which only the EBNA1 gene as well as BART and EBER miRNAs are expressed.¹² Recent studies are finding, however, that heterogeneity across and within BL tumors may enable alternative latency states such as latency III where other genes including LMP1 are expressed.^{13,14} Willard and colleagues even reported spontaneous lytic BL cell lines that expressed LMP1.¹³

EBV type and variation are extensive and known to affect tumorigenesis.¹⁵ For example, whole-genome sequencing (WGS) revealed from 101 BL tumors that viral presence drove mutations through the activation-induced cytidine deaminase mechanisms and defined a set of 72 driver genes, which were differentially expressed among the three epidemiologic subtypes of BL.⁶ Kaymaz et al.⁷ demonstrated that about a third of eBL tumors contain EBV type 2, whereas two-thirds are infected with EBV type 1 along with intertypic recombinants containing type 1 and type 2 regions. Recent work has examined EBV in tumorigenesis and patient survival time, but so far it has focused on the major division of types 1 and 2. Thus, there is significant variation beyond type-specific variation that is currently unexplored. For example, Thomas et al.¹⁶ established genetic subtypes within BL based on the driver genes and compared EBV status and survival among their genetic subtypes but did not examine the relationship between driver genes and specific EBV variants. A recent study also found that some EBV-positive Sporadic BLs may be molecularly similar to EBV-positive eBLs.¹⁷ Thus, within EBV-positive BL, we sought to elucidate the role of specific EBV variants in affecting patient survival time, determine the association between specific EBV variants and driver genes in regard to patient survival time, and define new prognostic subtypes incorporating both EBV variants and driver genes.

2 | MATERIALS AND METHODS

2.1 | Study population

Based on the known EBV statuses of the publicly available BL tumor WGS samples from Thomas et al. and Panea et al.,^{6,16} we downloaded only the EBV-positive samples as processed binary alignment map files from the BL Genome Sequencing Project (BLGSP)^{8,16} via the Genomic Data Commons Data Portal (Project ID: CGCI-BLGSP, dbGaP Study Accession: phs000235) and our own publicly available samples from Panea et al.⁶ via the European Genome-Phenome Archive (Accession Number: EGAD00001005781) on August 31, 2023 (Figure S1). The BLGSP samples were obtained from patients in Uganda, the United States, France, Canada, and Brazil, whereas our samples were obtained from patients in Kenya (Table 1). Due to the growing shift away from traditional epidemiological subtypes, instead of categorizing patients as endemic or sporadic, patients were classified based on the country of origin's income classification in the United Nations' World Economic Situation and Prospects 2024.¹⁸ These categories included high-income countries, including Canada, France, and the United States, and low-middle-income countries, including those classified as low-income (Uganda), lower-middle-income (Kenya), and upper-middle-income (Brazil).¹⁹

2.2 | Sequencing read preprocessing and variant-calling

To conduct a uniform process for variant-calling, all reads were preprocessed with the Genome Analysis Toolkit (GATK) 4.2.2.0 and

TABLE 1 Patient characteristics.

	No. patients (%)
Sex	
Male	87 (66.92)
Female	43 (33.08)
Median age (range)	8.33 (0.94–90)
EBV type	
Type 1	103 (79.23)
Type 2	27 (20.77)
Country income level	
High	27 (20.77)
Low-middle	103 (79.23)
Country	
Uganda	70 (53.85)
Kenya	25 (19.23)
United States	16 (12.31)
Canada	10 (7.69)
Brazil	8 (6.15)
France	1 (0.77)

Abbreviation: EBV, Epstein–Barr virus.

SAMtools 1.16.1 using the human (GRCh38), EBV1 (NC_007605), and EBV2 (NC_009334) reference genomes per GATK best practices (details found in Figure S2).^{20–22} Since the EBV strains from Thomas et al.'s samples were not typed, we uniformly typed all samples, including those from Panea et al., with BEDtools and SAMtools based on the number of reads mapping to the *EBNA2* and *EBNA3* genes of the reference genomes for each type (Figure S3).^{22,23}

To determine the major EBV variants, the uniformly preprocessed reads across all samples were jointly variant-called using NC_007605 with three rounds of GATK4 HaplotypeCaller per GATK4's best practices for germline short variant discovery akin to Kaymaz et al., albeit with haploid settings and revised functions based on the latest GATK updates (Figure S4).^{7,24} SNPs and indels were also filtered based on standard metrics (details found in Figure S4). EBV variants were then annotated with SnpEff²⁵ and filtered with SnpSift to remove synonymous EBV variants (Figure S5).²⁶ Variant sites were further filtered (details found in Figure S5). Missing EBV variants for a given sample were imputed with the mode of the variant in question across all samples.

All publicly available samples had human somatic mutation data for recurrently mutated genes, hereon referred to as driver genes, previously defined and detected from their respective publications.^{6,16} Using this, somatic tumor mutation data in common driver genes for the 130 WGS samples were extracted and preprocessed with in-house scripts. Then, for each sample, each driver gene was classified as mutated (1) if the sample had one or more missense, frameshift, nonsense, or splice site mutation or unmutated (0) otherwise. Driver genes were also filtered and imputed with the mode (details found in Figure S5).

To determine the geographic distribution of the LMP1 amino acid variants, all publicly available LMP1 protein sequences and complete EBV genomes were pulled from the National Center for Biotechnology Information's Nucleotides database on January 18, 2025.²⁷ The LMP1 protein sequences were aligned using MAFFT v7.526's multiple sequence alignment algorithm²⁸ and called for amino acid variants based on the LMP1 sequence from the EBV1 (NC_007605) reference genome. Consistent with our previous income classification categories, geography was split into two groups: Africa/South America (A/SA) and Europe/North America (E/NA) with sequences from Asia and Oceania being filtered out. Statistical significance was determined using Fisher's exact test and the chi-squared test on Stata/SE 18.0.²⁹

2.3 | Variable selection and survival prediction models

Along with clinical features including patient sex, age, EBV type, country, and country classification by income (high-income or low-middle-income), these filtered EBV variants and previously defined driver genes were then used as features in least absolute shrinkage and selection operator (LASSO) regularized and Bayesian variable selection models, both of which were implemented within a Cox proportional hazards framework with the glmnet³⁰ and

BVSNLP³¹ packages, respectively, in R³² (Figure S5). These two variable selection approaches are commonly used to find the most important features in high-dimensional data where the number of samples n is significantly smaller than the number of features p .

Separate variable selection models were implemented for each of the following categories, hereinafter referred to as feature sets: (1) clinical features, (2) driver genes, (3) clinical features and driver genes, (4) EBV variants, (5) EBV variants and driver genes, (6) EBV variants and clinical features (EBV/clinical features model), and (7) EBV variants, driver genes, and clinical features (complete model). LASSO Cox and Bayesian Cox then determined which features were associated with patient survival time. For each feature set, features with both non-zero coefficients in LASSO Cox and posterior inclusion probabilities above zero in Bayesian Cox were considered the top features used for prediction and prognostic subgrouping. These top features were then used as the inputs to a Cox proportional hazards survival prediction model. Using the coxph package³³ in R, each feature set's prediction model was run over 1000 replicates with each iteration using a random 90–10 training–testing split of all samples. The variable selection and survival prediction models were implemented on both the overall cohort including all patients as well as the African cohort consisting of patients only from African countries (Kenya and Uganda). For each feature, the reported hazard ratio was the median value across all 1000 replicates. Predictive performances were measured with Harrell's concordance index, where a value of 1 corresponded to perfect prediction of which of any given two patients would die first across all possible pairs of patients, while a value of 0.5 was an uninformed prediction.³⁴

2.4 | Prognostic subgrouping survival model

Using the top features in the complete feature set, the optimal number of prognostic subgroups was determined with the silhouette score. The silhouette score measures the mean distance between points within a subgroup compared to the mean distance between points across neighboring subgroups. Final prognostic subgrouping was conducted with non-negative matrix factorization using the NMF package in R,³⁵ and Kaplan–Meier curves were performed to compare each prognostic subgroup's overall survival.

3 | RESULTS

3.1 | Identification of publicly available EBV-positive BL samples

Of 331 publicly available BL tumor WGS samples from Thomas et al. ($n = 230$) and Panea et al. ($n = 101$),^{6,16} we extracted only 105 and 25 EBV-positive samples, respectively, resulting in a final sample size of 130 EBV-positive BL samples (Figure S1). Of these 130 EBV-positive BL patients, 103 (79.2%) were from low-middle-income countries, whereas 27 (20.8%) were from high-income countries

(Table 1). Countries of origin included Uganda ($n = 70$, 53.8%), Kenya ($n = 25$, 19.2%), the United States ($n = 16$, 12.31%), Canada ($n = 10$, 7.7%), Brazil ($n = 8$, 6.15%), and France ($n = 1$, 0.77%). EBV type 1 was the dominant type, representing 79.2% ($n = 103$) of all genotyped samples, with type 2 being found in 14.8% and 22.3% of patients in high-income and low-middle-income countries, respectively. The median survival time for patients from low-middle-income countries was 1.9 years, whereas those from high-income countries did not have median survival times due to a greater than 50% survival rate.

3.2 | EBV coverage and variants

Across all 130 samples, the median depth and breadth of read coverage of the non-repetitive EBV genome were 797.80% and 99.97%, respectively. For the repetitive regions, the median depth and breadth were 633.11% and 99.97%. From an initial 10,347 EBV variants and 16,033 mutated genes, we extracted 724 EBV variants and 73 potential driver genes for further analysis after filtration (Section 2).

3.3 | EBV/clinical features model on overall cohort found a high number of LMP1 variants associated with patient survival time

Using an initial feature set of EBV variants and clinical features known as the EBV/clinical features model on the overall cohort, the variable selection models yielded 22 EBV variants, including seven in LMP1 alone (Met129Ile/M129I, His101Gln/H101Q, Ile152Leu/I152L, Ile63-Met/I63M, Gly212Ala/G212A, Gly212Ser/G212S, and Gly331Glu/G331Q) as most associated with patient survival time (Table 2). While only three of the 22 EBV variants, including Thr208Ala/T208A (BTRF1), Met12Leu/M12L (BDLF3), and Thr710Asn/T710N (BSLF1), were individually associated with patient survival time in the multivariable Cox proportional hazards analysis (Table 2), all 22 were used in the survival prediction model. There were no statistically significant differences in the geographic distribution of the seven LMP1 variants across our samples in Kenya and Uganda (Table S3).

3.4 | Some key LMP1 variants were differentially distributed across geographic regions

The 1678 LMP1 sequences were extracted from National Center for Biotechnology Information's Nucleotides database, of which 278 and 388 were from A/SA and E/NA, respectively, for a total of 666 final sequences. Of the seven LMP1 variants, three were differentially distributed across geographic regions. Compared to E/NA, A/SA had a higher proportion of LMP1 samples with the M129I and I152L LMP1 variants and a lower proportion of the G331Q variant (M129I: E/NA 73.17% vs. A/SA 90.16%; I152L: E/NA 13.41% vs. A/SA 40.44%;

G331Q: E/NA 12.76% vs. A/SA 0.72%; Table S4). The remaining four LMP1 variants (H101Q, I63M, G212A, and G212S) were not differentially distributed across the geographic regions.

3.5 | EBV/clinical features model on African cohort still indicated a high number of significant variants in LMP1 but not EBV type

When limited to patients in Africa ($n = 95$) for the African cohort, six of the 22 EBV variants from the overall cohort remained significantly associated with patient survival time (P308Q, I63M, H101Q, A339T, I152L, and G212A) with four of those coming from LMP1 alone (I63M, H101Q, I152L, and G212A) (Table S1). In this variable selection model, six of the 19 variants significantly associated with patient survival time were located in LMP1 (Leu25Ile/L25I, I63M, H101Q, I152L, G212A, Gln322His/Q322H). EBV type was still not significantly associated with patient survival time, and while there were three significant variants in EBNA2 in the initial analysis, removing all type-specific variants did not alter the lack of association between EBV type and patient survival time.

3.6 | Country income level was the strongest individual feature among all features associated with patient survival time but EBV type was not associated

In the EBV/clinical features model on the overall cohort, country income level was the single most significant feature of all features associated with patient survival time, with a hazard ratio of 12.10 (Table 2). Country-level income was also the only clinical feature associated with patient survival time, as EBV type, patient sex, age, and country were not significantly associated. When limited to African patients only, there was no statistically significant survival difference between patients in Kenya ($n = 25$) and patients in Uganda ($n = 70$, $p = .44$). Removing type-specific variants such as those within the EBNA2 and EBNA3 genes did not change the lack of association.

3.7 | Complete model on overall cohort found eight tumor driver genes and nine EBV variants, including four in LMP1 associated with patient survival time

In the complete model, which consisted of all clinical features, tumor driver genes, and EBV variants, the variable selection models identified eight driver genes, including Contactin-Associated Protein Family Member 3B (CNTNAP3B), ID3, HIST1H2BK, MAP3K9, IKG, ETS1, P2RY8, and TP53, in addition to nine EBV variants with four in LMP1 (H101Q, I152L, G212A, and G331Q) as most associated with patient survival time (Table 2), indicating that 13 of the 22 variants from the EBV variant-only feature set were no longer significant after incorporating driver genes. A complete feature set analysis on the African

TABLE 2 Multivariable cox proportional hazards analysis of Epstein–Barr virus (EBV) variants associated with overall survival.

EBV variant (gene)	Allele frequency/N (%)	aHR (95% CI)	p-Value
EBV feature set			
Thr208Ala/T208A (BTRF1)	123/130 (94.62)	0.004 (0.000–0.061)	8.095*10 ^{−5}
Met12Leu/M12L (BDLF3)	128/129 (99.22)	0.017 (0.001–0.259)	.004
Thr710Asn/T710N (BSLF1)	16/130 (12.31)	4.461 (1.133–17.574)	.033
Met129Ile/M129I (LMP1)	121/130 (93.08)	35.888 (0.953–1318.905)	.053
His101Gln/H101Q (LMP1)	11/130 (8.46%)	4.128 (0.945–17.792)	.060
Ile152Leu/I152L (LMP1)	55/129 (42.64)	4.111 (0.738–23.153)	.106
Ile63Met/I63M (LMP1)	54/130 (41.54)	0.143 (0.015–1.463)	.104
Gly212Ala/G212A (LMP1)	11/129 (8.53)	0.380 (0.061–2.349)	.298
Pro308Gln/P308Q (BVRF2)	8/129 (6.20)	3.120 (0.313–30.738)	.335
Gly118Arg/G118R (BLLF1)	15/130 (11.54)	1.765 (0.557–5.559)	.337
Val564_Ala568del/V564_A568del (EBNA-2)	16/127 (12.60)	0.454 (0.088–2.402)	.350
Thr124Pro/T124P (BZLF1)	39/129 (30.23)	0.621 (0.200–1.925)	.407
Pro622Ser/P622S (BBLF4)	37/130 (28.46)	1.348 (0.540–3.384)	.520
Ala339Thr/A339T (BBLF4)	122/129 (94.57)	0.627 (0.086–4.583)	.643
Arg648fs/R648fs (BHLF1)	19/128 (14.84)	1.236 (0.418–3.635)	.689
Glu36Gln/E36Q (BNRF1)	100/127 (78.74)	0.792 (0.214–2.959)	.706
Gly212Ser/G212S (LMP1)	106/128 (82.81)	1.218 (0.268–5.504)	.756
Ala768Val/A768V (BLLF1)	11/130 (8.46)	1.130 (0.106–12.977)	.874
Gly331Glu/G331Q (LMP1)	7/129 (5.43)	2.007*10 ^{−10} (0.000–N/A)	.997
Val479Ile/V479I (BRLF1)	7/130 (5.38)	7.003*10 ^{−8} (0.000–N/A)	.998
Val611Ile/V611I (BBLF4)	37/130 (28.46)	N/A	N/A
Pro486Ser/P486S (BRLF1)	7/130 (5.38)	N/A	N/A
Country income level			
Low–Middle	103 (79.23)	12.099 (1.318–104.407)	.027
Complete feature set			
Met12Leu/M12L (BDLF3)	128/129 (99.22)	0.002 (0.000–0.036)	2.625*10 ^{−9}
His101Gln/H101Q (LMP1)	11/130 (8.46%)	10.942 (3.273–36.860)	9.929*10 ^{−5}
Thr208Ala/T208A (BTRF1)	123/130 (94.62)	0.004 (0.000–0.074)	2.418*10 ^{−4}
Thr710Asn/T710N (BSLF1)	16/130 (12.31)	3.887 (1.614–9.395)	.002
Pro308Gln/P308Q (BVRF2)	8/129 (6.20)	4.173 (1.230–13.348)	.016
Ile152Leu/I152L (LMP1)	55/129 (42.64)	2.434 (1.181–5.009)	.016
Gly212Ala/G212A (LMP1)	11/129 (8.53)	0.260 (0.033–2.182)	.222
Val564_Ala568del/V564_A568del (EBNA-2)	16/127 (12.60)	0.653 (0.136–3.234)	.602
Gly331Glu/G331Q (LMP1)	7/129 (5.43)	1.515*10 ^{−9} (0–N/A)	.997
Driver gene			
CNTNAP3B	16/130 (12.31)	4.025 (1.172–14.025)	.027
ID3	55/130 (42.31)	0.424 (0.189–0.946)	.036
HIST1H2BK	14/130 (10.77)	2.329 (0.851–6.398)	.100
MAP3K9	12/130 (9.23)	0.194 (0.017–2.372)	.204
IGK	33/130 (25.38)	0.559 (0.222–1.405)	.217
ETS1	17/130 (13.08)	1.559 (0.608–4.035)	.352
P2RY8	10/130 (7.69)	0.928 (0.097–3.387)	.520
TP53	25/130 (19.23)	0.926 (0.330–2.610)	.797

TABLE 2 (Continued)

EBV variant (gene)	Allele frequency/N (%)	aHR (95% CI)	p-Value
Country income level			
Low-middle	103 (79.23)	37.166 (3.548–373.040)	.002

Abbreviation: LMP1, latent membrane protein 1.

cohort shared three of the nine EBV variants (P308Q, H101Q, and I152L) and three of the eight driver genes (HIST1H2BK, ETS1, and ID3) in common with the overall cohort (Table S2).

3.8 | Interactions between driver genes, specific EBV variants, and country income levels required for the best survival predictive performance

To better understand variable selection across features, we explored additional models, examining a total of seven variable selection models. For each, we incorporated the corresponding features to predict survival and compared their performance. Each model had the following number of features: clinical features = 1, driver genes = 33, clinical features/driver genes = 34, EBV variants = 22, EBV variants/driver genes = 16, EBV variants/clinical features = 23, and EBV variants/driver genes/clinical features = 18. Since EBV type, patient sex, age, and country were not associated with survival from the variable selection models, these clinical features were not applied in the survival prediction models. The survival model with the lowest predictive performance used country income level as its only feature and had a testing concordance index (c-index) of 0.56 (Table 3). The next best performing variable selection and survival prediction model only used driver genes in its feature set and had a testing c-index of 0.57 ($p = .11$ compared to country income level-only model). When both driver genes and country income level were used for the feature set, the model had an improved testing c-index of 0.61 ($p < .01$ compared to driver gene-only model). Compared to the model with driver genes and country income level, the model trained with EBV variants only had an even better testing c-index of 0.72 ($p < .01$ compared to driver gene + country income level model). The model using both EBV variants and driver genes performed similarly with a testing c-index of 0.72 ($p = .39$). In comparison, the model with EBV variants and country income level had an improved testing c-index of 0.73 ($p = .04$). Finally, a model trained on the complete feature set of EBV variants, driver genes, and country income level had the best testing c-index at 0.76 ($p < .01$), which was significantly higher than all previous models, indicating that interactions between EBV variants, driver genes, and clinical features were required for the best survival predictive performance.

When limited to African patients only, the EBV/clinical features model had a testing c-index of 0.67. The complete model on the African cohort observed a performance boost to 0.71 for the testing c-index.

3.9 | Three prognostic subgroups were defined with differential survival

Using the complete feature set of EBV variants, driver genes, and clinical features, the silhouette score indicated that three prognostic subgroups yielded the highest cluster stability. These subgroups were defined by different meta-features, which were composed of various driver genes and EBV variants (Figure S6). Subgroup 1 was defined by a meta-feature composed of driver genes TP53, CNTNAP3B, MAP3K9, P2RY8, and IGHK and the EBV variants G331Q (LMP1) and H101Q (LMP1) (Figure 1). Subgroup 2 was defined by a meta-feature consisting of the driver gene ID3 and a different LMP1 variant I152L (LMP1) and EBV variant T710N (BSLF1). Finally, Subgroup 3 was defined by a different meta-feature that included the driver genes ETS1 and HIST1H2BK as well as a different LMP1 variant G212A and EBV variants Val564_Ala568del/V564_A568del (EBNA-2), P308Q (BVRF2), M12L (BDLF3), and T208A (BTRF1). In a Kaplan–Meier survival analysis, Subgroup 1 observed the highest overall survival, followed by Subgroup 2, and then finally Subgroup 3 ($p = .03$, Figure 2).

These prognostic subgroups included geographical differences ($p < .01$). For example, Subgroup 1 was predominantly composed of North American and European patients (78.26%, $n = 18$), whereas Subgroups 2 and 3 had higher proportions of African and South American patients (Subgroup 2: 97.22%, $n = 70$ and Subgroup 3: 80.00%, $n = 28$). Notably, geography was not absolute among these subgroups, meaning that there were some African and South American patients in Subgroup 1 (21.74%, $n = 5$) and North American and European patients (Subgroup 2: 2.78%, $n = 2$ and Subgroup 3: 20.00%, $n = 7$) in Subgroups 2 and 3 (Table S5).

4 | DISCUSSION

BL is an aggressive B-cell lymphoma that is fatal if left untreated. While the prognosis of BL is known to depend on factors such as the stage of the disease, other molecular and genetic interactions, including viral variation, have not been comprehensively explored. In this study, we incorporated EBV variation along with clinical features and driver gene mutations into survival models for patients with EBV-positive tumors. EBV variants, particularly those in LMP1, were found to be associated with survival. Not surprisingly, we found that study participants' country income level (a likely surrogate for clinical access and care) was the strongest individual feature associated with BL patient survival time. An analysis limited to patients in Africa shared many of the same variants as that of the complete cohort and still

Feature set	Training set C-index	Testing set C-index	p-Value ^a
Complete (EBV + tumor + clinical features)	0.8504 ± 0.0003	0.7572 ± 0.0038	<.0001
EBV + clinical features	0.8356 ± 0.0003	0.7324 ± 0.0040	.0358
EBV + tumor	0.7968 ± 0.0004	0.7226 ± 0.0038	.3912
EBV	0.8184 ± 0.0005	0.7210 ± 0.0041	<.0001
Tumor + clinical features	0.8144 ± 0.0005	0.6101 ± 0.0047	<.0001
Tumor	0.7881 ± 0.0005	0.5739 ± 0.0046	.1098
Clinical features	0.6085 ± 0.0005	0.5601 ± 0.0124	N/A

Abbreviation: EBV, Epstein-Barr virus.

^aT-test comparing the testing set C-index of the current feature set relative to that of the next-best performing feature set.

TABLE 3 Survival prediction model performances.

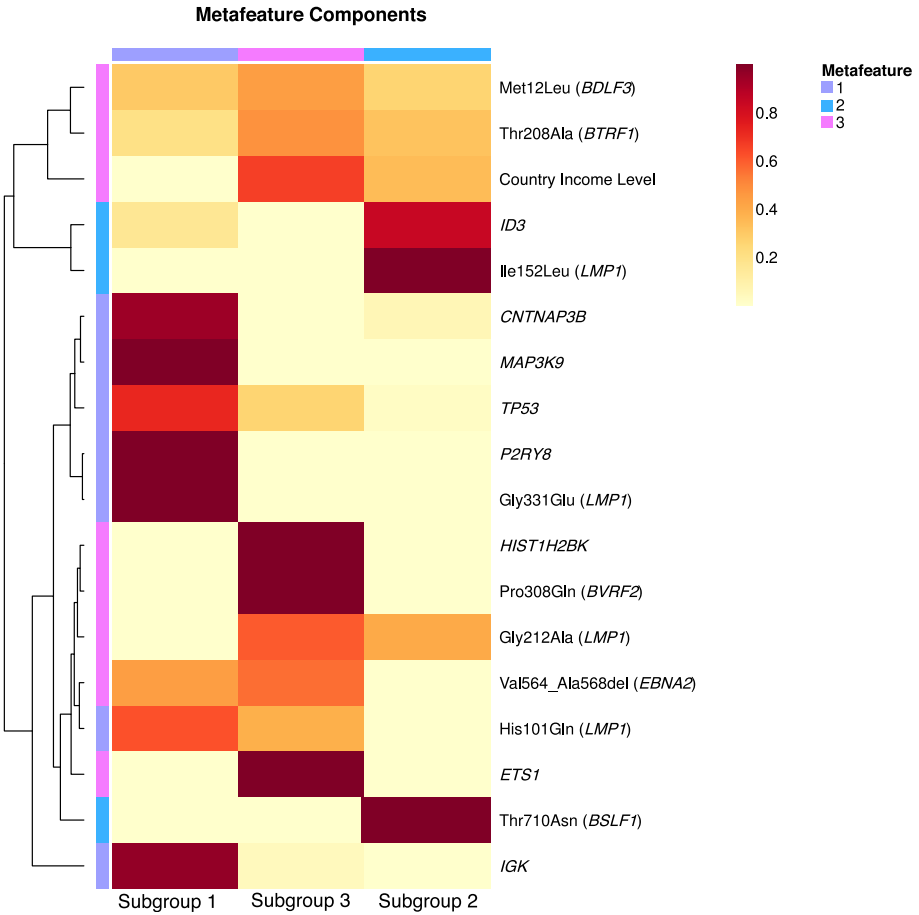


FIGURE 1 Epstein-Barr virus (EBV) variants/driver genes/country income level contributions to metafeatures. Each metafeature was defined by different dominant genes. Metafeature 1 consisted of CNTNAP3B, MAP3K9, TP53, P2RY8, IGK, G331Q (LMP1), and H101Q (LMP1). Metafeature 2 was defined by ETS1, HIST1H2BK, G212A (LMP1), V564_A568del (EBNA2), P308Q (BVRF2), M12L (BDLF3), and T208A (BTRF1). Finally, Metafeature 3 was composed of ID3, I152L (LMP1), and T710N (BSLF1). Note that all three metafeatures included a different LMP1 variant. LMP1, latent membrane protein 1. [Color figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

found a disproportionately higher number of significant variants in LMP1, but not EBV type, to be associated with patient survival time, indicating that the addition of patients outside of Africa did not drastically alter our models. We also defined potential interactions between specific EBV variants and driver genes in the context of patient survival time and determined three new prognostic subgroups that showed differing survival.

The EBV variants associated with patient survival were especially enriched for LMP1 variants. LMP1 is the main transforming oncoprotein in EBV and has a range of biological roles, including inhibiting

apoptosis and promoting cell proliferation through activation of the NF-κB pathway.³⁶ Amino acid 212 in LMP1 was especially intriguing, as G212S had a hazard ratio greater than 1, whereas G212A had a hazard ratio less than 1. A previous study demonstrated that the G212S mutation in EBV-positive post-transplant lymphoproliferative disorder (PTLD) B-cell lines was associated with increased extracellular signal-regulated kinase/mitogen-activated protein kinases activation and c-Fos expression, indicating this variant may be a gain-of-function mutation.³⁷ Furthermore, a clinical trial of 872 pediatric transplant recipients by Martinez et al.³⁸ reported that G212S

FIGURE 2 Overall survival of prognostic subgroups. The prognostic subgroups observed differential survival with patients from Subgroup 1 experiencing the best survival, followed by those in Subgroup 2, and finally the patients in Subgroup 3. [Color figure can be viewed at wileyonlinelibrary.com]

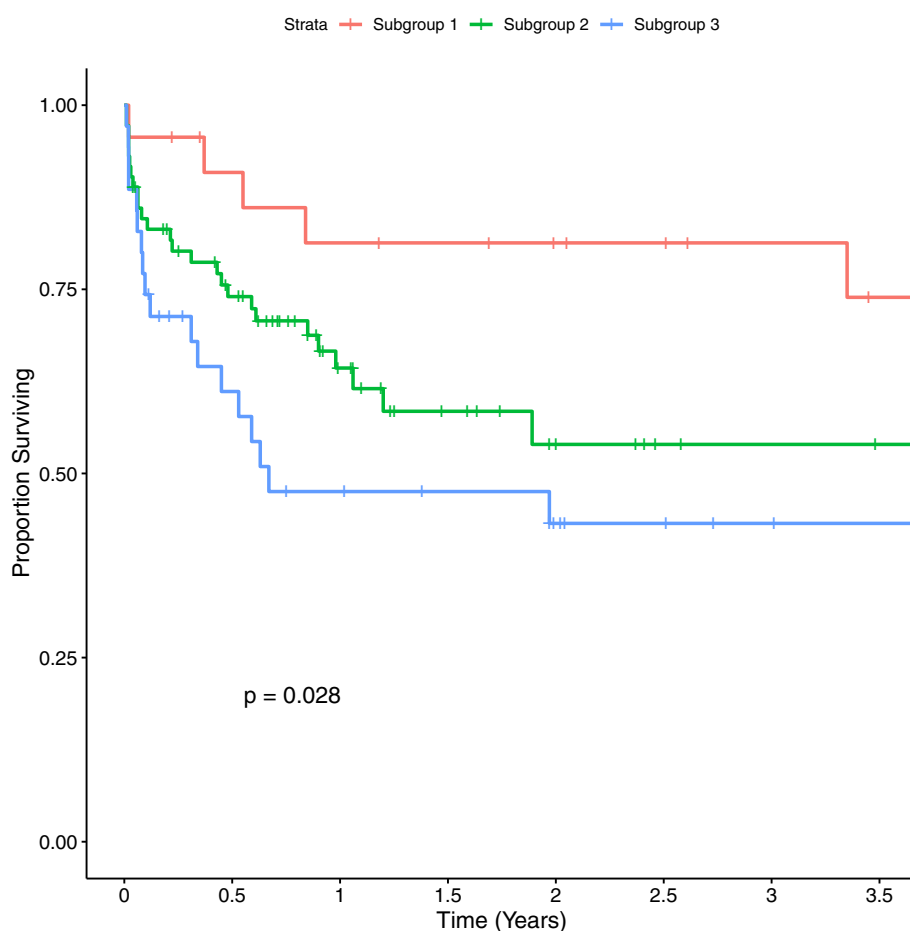
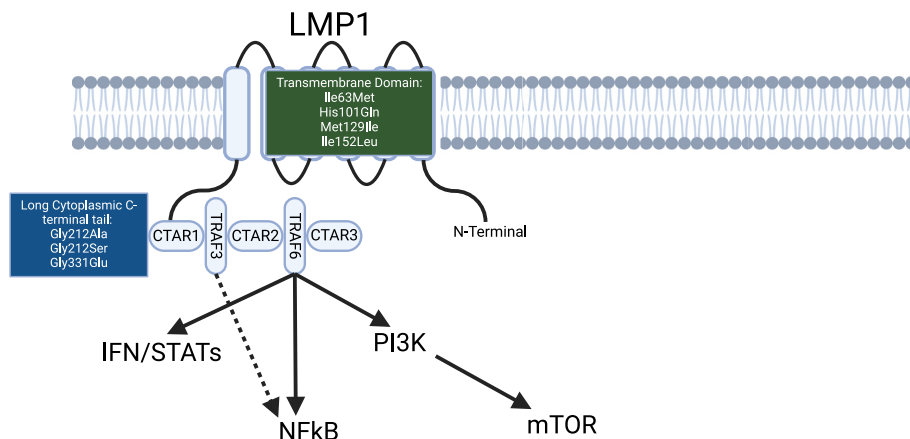


FIGURE 3 Model of LMP1 pathways and variants. Four of the seven significant LMP1 variants (Ile633Met, His101Gln, Met129Ile, and Ile152Leu) corresponded to the transmembrane domain, whereas the other three (Gly212Ala, Gly212Ser, and Gly331Glu) belonged to the domain of the long cytoplasmic C-terminal tail, which is involved in the NF κ B and PI3K/mTOR pathways. Adapted from sources 55–57. LMP1, latent membrane protein 1. [Color figure can be viewed at wileyonlinelibrary.com]



increased the risk of EBV-positive PTLD by nearly 12-fold. Furthermore, Martinez et al.³⁸ found that 0% of EBV-positive PTLD patients had G212A, but 96.9% carried G212S, suggesting that this non-synonymous substitution may significantly dampen the aggressiveness of the resulting tumor. Despite the traditional dogma that BL is classically latency I, our findings support those of more recent studies highlighting the expression and potential significance of LMP1 in BL¹³ and highlight the potential for using specific EBV variants such as G212S and G212A as predictive biomarkers in EBV-positive BL (Figure 3).

Overall, the pattern of LMP1 variation was consistent with our previous findings and those of previous studies where low-middle-income countries including Kenya, Uganda, and Brazil best fit with the Alaskan and Mediterranean classifications, and high-income countries such as Canada, the United States, and France fit with the North Carolina strain.^{39–44} The majority of our samples originated from Uganda and Kenya, which observed no differences in the distribution of the seven significant LMP1 variants. Three of these seven variants were also found in the classification proposed by Mainou et al.³⁹ Specifically, G212S was located in the Alaskan, China 1, and China 2 strains,

I63M was found in the Alaskan strain, and G331Q was detected in the North Carolina and China 2 strains. G331Q was the only variant preferentially distributed in high-income countries and was not significantly associated with survival in the EBV and complete feature analysis on the African-only cohort, suggesting a potential for geographical confounding with this variant. Collectively, these patterns of variation are consistent with our previous finding in Kenya that African LMP1 sequences are likely best classified as Alaskan and Mediterranean.⁴⁰

When we restricted our variable selection models to patients in Africa, we identified six of the same 22 EBV variants from the original cohort and still found a lack of association between EBV type and survival time, indicating that the incorporation of patients outside of Africa did not completely change our model and supporting the growing number of studies that propose traditional epidemiologic subtypes corresponding to patient geography may not be the best classification of BL.^{6,7} While a few LMP1 variants were disproportionately enriched in certain geographic regions across the publicly available sequences, that LMP1 variants were still outputted from the variable selection model on the Africa-only cohort and country income level was accounted for in the model on the overall cohort collectively indicate that these variants may be biologically and clinically significant irrespective of geography. Importantly, we continued to find a high number of variants associated with patient survival time in the LMP1 gene including G212A and I152L and a lack of association between EBV type and patient survival time, despite previous studies showing molecular differences between Types 1 and 2 in the context of tumorigenesis.^{6,40,45}

Of all features, country income level was the most significant contributor to patient survival time, with patients in low-middle-income countries experiencing the highest hazard ratios in all relevant feature sets. In this study, patients from low-middle-income countries, namely countries on the continents of Africa and South America, had a median survival time of 1.89 years compared to those from high-income countries, which observed a greater than 50% survival rate. This finding supports those of many previous studies, as BL patients in equatorial Africa have a survival rate nearly half of that of BL in non-tropical countries.³ Country income level is essentially a proxy for access to healthcare, and since BL is a highly aggressive tumor that requires intensive and consistent treatment, limited access to care, such as delays in diagnosis and insufficient clinical supportive care, can have potentially devastating consequences.⁴⁶ Altogether, our findings suggest that while genetic features of the human tumor cells and virus potentially underlying patient geography play a significant role in patient survival time, socioeconomic factors are by far the most important contributor.

Besides EBV variation, our variable selection and survival prediction models highlighted eight driver genes, two of which were individually associated with patient survival time as supported by previous studies.^{6,16,47} CNTNAP3B, which was associated with decreased patient survival time, plays an integral role in cell adhesion, and abnormalities in the gene have been implicated in other cancers such as B-cell acute lymphoblastic leukemia.⁴⁷ Furthermore, due to its association with immune checkpoint expression, CNTNAP3B has been used

in a prognostic model for bladder cancer.⁴⁸ ID3 was associated with increased patient survival time, and Thomas et al.¹⁶ similarly reported that patients with ID3 mutations observed increased survival compared to those who did not when controlling for TP53 mutations. Panea et al.⁶ found that ID3 was one of the most commonly silenced genes, and in vivo knockout of ID3 amplified the tumorigenic effects of MYC. Thus, while ID3 mutations may play a key role in the initial tumorigenesis, they could lead to less aggressive disease once the tumor has been established. Our results do differ from those of some previous studies. For instance, previous studies found that BL patients with TP53 mutations were more likely to relapse or experienced worse survival.⁴⁹ While our multivariable Cox proportional hazards analysis did not find TP53 mutations to have a statistically significant association with patient survival time, our cohort consisted of EBV-positive BL patients, most of whom originated from Africa and South America, while most previous studies studied both EBV-negative and EBV-positive patients mostly in North America or Europe. Thus, given that EBV infection alters the driver genes in BL tumorigenesis, TP53 may not be significantly associated with clinical outcomes in EBV-positive BL patients. Regardless, interactions between all three feature sets including driver genes, various specific EBV variants, and country income level were required for the best possible survival predictive performance.

Having defined interactions between specific EBV variants and driver genes, we found clinically relevant prognostic subgroups that demonstrated differences in survival and included a different LMP1 variant and various driver genes in each subgroup. Subgroup 1 consisted of the LMP1 variants Gly331Glu and His101Gln along with the driver genes TP53 and MAP3K9. Wang and colleagues determined that the TP53-encoded p53 protein activates LMP1, which in turn blocks many effects of p53 during viral transformation and enables EBV-infected cells to endure cellular damages.⁵⁰ Subgroup 2 included the LMP1 variant I152L and driver gene ID3. ID3 is a well-characterized tumor suppressor gene in BL that amplifies the effects of MYC, thereby causing tumorigenesis.⁶ Recent studies have reported increased methylation and silencing of ID3 in EBV-infected B cells via an LMP1-involved mechanism.⁵¹ Lastly, Subgroup 3 was composed of the LMP1 variant G212A, EBNA2 variant V564_A568del, and the driver gene ETS1. Kim et al.⁵² reported that LMP1 induced the expression of ETS1, thereby increasing the invasiveness of the tumor in nasopharyngeal carcinoma. Zhang et al.⁵³ found that LMP1 and EBNA2, which are critically involved in latent viral transcription, comprise the minimum set of EBV genes required for tumorigenesis. On a larger cohort of both EBV-negative and EBV-positive BL samples, Thomas et al. also defined three subgroups using tumor mutations only. Given that most of our samples came from Thomas et al.,¹⁶ their subgroups shared significant similarities yet had some key differences. For example, two of their three subgroups were defined by the same key driver genes: ID3 and TP53. However, unlike our subgroups, their subgroups did not incorporate viral variation and did not observe statistically significant differences in survival. With recent studies indicating that EBV status may be

more defining than epidemiologic subtype,⁶ our findings suggest the presence of subgroups within EBV-positive BL with each having a different prognosis. Altogether, our findings portend the significance of the virus and its variants in impacting patient survival time and prognostic subgrouping.

Overall, our study identified specific EBV variants significantly associated with BL patient survival time. The main limitation of this study was its small sample size, as only 130 of the 331 publicly available BL samples were EBV-positive, of which 103 were African or South American. However, our variable selection models were designed to address this limitation by decreasing the number of possible features associated with patient survival. As a result, we found a high number of significant variants in the LMP1 gene and a lack of association with EBV type, findings which were confirmed in a sub-analysis of only patients from Africa. We also determined that country income level, which corresponds to access to healthcare, was the most important factor in patient survival time. We also described potential interactions between driver genes and specific EBV variants in BL in the context of patient survival time and proposed a model that accurately predicts survival based on clinical features, driver genes, and EBV variants as well as defined prognostic subgroups. Although previous studies have investigated the relationship between patient survival time and clinical features as well as driver genes,^{3,54} none have examined the role of EBV variation and its interactions with the aforementioned factors.

5 | CONCLUSIONS

Our study offers key insight into the role of specific EBV variants in BL patient survival time and their interactions with driver genes and clinical features, potentially introducing specific EBV variants as predictive biomarker candidates in future studies. We also define a survival prediction model and prognostic subgroups that may be used to assess prognosis for EBV-positive BL patients. Validation of these findings on larger patient cohorts as they become available is warranted.

AUTHOR CONTRIBUTIONS

Isaac E. Kim Jr.: Conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing – original draft; writing – review and editing. **Cliff Oduor:** Data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing – original draft; writing – review and editing. **Julian Stamp:** Formal analysis; methodology; writing – original draft; writing – review and editing. **Micah A. Luftig:** Formal analysis; funding acquisition; investigation; supervision; writing – original draft; writing – review and editing. **Ann M. Moormann:** Data curation; formal analysis; funding acquisition; investigation; methodology; supervision; writing – original draft; writing – review and editing. **Lorin Crawford:** Formal analysis; funding acquisition; investigation; methodology; software; supervision; validation; writing – original draft; writing – review and editing. **Jeffrey A. Bailey:** Conceptualization; data curation; formal analysis;

funding acquisition; investigation; methodology; supervision; writing – original draft; writing – review and editing.

FUNDING INFORMATION

This study was funded by two grants from the National Institutes of Health (R01 CA189806 and R01 CA234348) and the Brown University Blavatnik Family Graduate Fellowship in Biology and Medicine.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sources and handling of the publicly available datasets used in this study are described in the Section 2. All coding scripts used in the pipelines (preprocessing, EBV typing, variant-calling, survival prediction, and prognostic subgrouping) are available on GitHub at <https://github.com/bailey-lab/BL-Survival-Prediction-and-Prognostic-Subgrouping>.

ETHICS STATEMENT

This study used publicly available whole-genome sequencing data that was securely accessed and maintained during analysis in accordance with the Declaration of Helsinki.

ORCID

Isaac E. Kim Jr.  <https://orcid.org/0000-0002-9737-8100>

TWITTER

Lorin Crawford  [lorin_crawford](https://twitter.com/lorin_crawford)

Jeffrey A. Bailey  [JBaileyOmics](https://twitter.com/JBaileyOmics)

REFERENCES

- Magrath I. Epidemiology: clues to the pathogenesis of Burkitt lymphoma. *Br J Haematol*. 2012;156:744-756.
- Woroniecka R, Rymkiewicz G, Szafron LM, et al. Cryptic MYC insertions in Burkitt lymphoma: new data and a review of the literature. *PLoS One*. 2022;17:e0263980.
- Buckle G, Maranda L, Skiles J, et al. Factors influencing survival among Kenyan children diagnosed with endemic Burkitt lymphoma between 2003 and 2011: a historical cohort study. *Int J Cancer*. 2016;139(6):1231-1240. doi:10.1002/ijc.30170
- Torgbor C, Awuah P, Deitsch K, Kalantari P, Duca KA, Thorley-Lawson DA. A multifactorial role for *P. falciparum* malaria in endemic Burkitt's lymphoma pathogenesis. *PLoS Pathog*. 2014;10:e1004170.
- Robbiani DF, Deroubaix S, Feldhahn N, et al. Plasmodium infection promotes genomic instability and AID-dependent B cell lymphoma. *Cell*. 2015;162(4):727-737. doi:10.1016/j.cell.2015.07.019
- Panea RI, Love CL, Shingleton JR, et al. The whole-genome landscape of Burkitt lymphoma subtypes. *Blood*. 2019;134(19):1598-1607. doi:10.1182/blood.2019001880
- Kaymaz Y, Oduor CI, Yu H, et al. Comprehensive transcriptome and mutational profiling of endemic Burkitt lymphoma reveals EBV type-specific differences. *Mol Cancer Res*. 2017;15:563.
- Grande BM, Gerhard DS, Jiang A, et al. Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma. *Blood*. 2019;133(12):1313-1324. doi:10.1182/blood-2018-09-871418

9. Baer R, Bankier AT, Biggin DPL, et al. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. *Nature*. 1984;310:207-211. <https://pubmed.ncbi.nlm.nih.gov/6087149/>
10. Thorley-Lawson DA, Allday MJ. The curious case of the tumour virus: 50 years of Burkitt's lymphoma. *Nat Rev Microbiol*. 2008;6:913-924. <https://pubmed.ncbi.nlm.nih.gov/19008891/>
11. Rowe M, Young LS, Cadwallader K, Petti L, Kieff E, Rickinson AB. Distinction between Epstein-Barr virus type A (EBNA 2A) and type B (EBNA 2B) isolates extends to the EBNA 3 family of nuclear proteins. *J Virol*. 1989;63:1031-1039.
12. Münz C. Immune escape by non-coding RNAs of the Epstein Barr virus. *Front Microbiol*. 2021;12:657387.
13. Willard KA, Barry AP, Oduor CI, et al. Viral and host factors drive a type 1 Epstein-Barr virus spontaneous lytic phenotype. *MBio*. 2023;14:e0220423. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10746244/610.1128/mbio.02204-23>
14. Kelly G, Bell A, Rickinson A. Epstein-Barr virus-associated Burkitt lymphomagenesis selects for downregulation of the nuclear antigen EBNA2. *Nat Med*. 2002;8:1098-1104.
15. Smatti MK, Al-Sadeq DW, Ali NH, Pintus G, Abou-Saleh H, Nasrallah GK. Epstein-Barr virus epidemiology, serology, and genetic variability of LMP-1 oncogene among healthy population: an update. *Front Oncol*. 2018;8:364512.
16. Thomas N, Dreval K, Gerhard DS, et al. Genetic subgroups inform on pathobiology in adult and pediatric Burkitt lymphoma. *Blood*. 2023; 141(8):904-916. doi:10.1182/blood.2022016534
17. Richter J, John K, Staiger AM, et al. Epstein-Barr virus status of sporadic Burkitt lymphoma is associated with patient age and mutational features. *Br J Haematol*. 2022;196:681-689. <https://pubmed.ncbi.nlm.nih.gov/34617271/>
18. Department of Economic and Social Affairs. *World Economic Situation and Prospects 2024*. Stylus Publishing, LLC; 2024:307.
19. Oleribe OO, Momoh J, Uzochukwu BSC, et al. Identifying key challenges facing healthcare systems in Africa and potential solutions. *Int J Gen Med*. 2019;12:395-403. doi:10.2147/ijgm.s223882
20. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491-498. <https://pubmed.ncbi.nlm.nih.gov/21478889/>
21. Van der Auwera GA, O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, 496; 2020.
22. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10:giab008.
23. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841-842.
24. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. 2018. <https://www.biorxiv.org/content/10.1101/201178v3.abstract>
25. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. 2012;6:80-92.
26. Cingolani P, Patel VM, Coon M, et al. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet*. 2012;3:35. doi:10.3389/fgene.2012.00035
27. Sayers EW, Bolton EE, Brister JR, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022;50:D20-D26.
28. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30:772-780.
29. Stata Statistical Software: Release 18. College Station, TX: Stata-Corp LLC.
30. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1-22.
31. Nikooienejad A, Wang W, Johnson VE. Bayesian variable selection for survival data using inverse moment priors. *Ann Appl Stat*. 2020; 14:809-828. doi:10.1214/20-aos1325
32. Data Science Chair, University of Würzburg, IPA Group, Berlin H-U zu, KDE Group, University of Kassel, L3S Research Center (Germany). R: A Language and Environment for Statistical Computing. 2022. Accessed December 15, 2023. <https://www.bibsonomy.org/bibtex/7469ffee3b07f9167cf47e7555041ee7>
33. Therneau TM. Survival Analysis [R package survival version 3.5-7]. 2023. Accessed December 15, 2023. <https://CRAN.R-project.org/package=survival>
34. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. 1982;247:2543-2546. <https://pubmed.ncbi.nlm.nih.gov/7069920/>
35. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*. 2010;11:367.
36. Wang D, Liebowitz D, Kieff E. An EBV membrane protein expressed in immortalized lymphocytes transforms established rodent cells. *Cell*. 1985;43:831-840. <https://pubmed.ncbi.nlm.nih.gov/3000618/>
37. Vaysberg M, Hatton O, Lambert SL, et al. Tumor-derived variants of Epstein-Barr virus latent membrane protein 1 induce sustained Erk activation and c-Fos. *J Biol Chem*. 2008;283:36573.
38. Martinez OM, Krams SM, Robien MA, et al. Mutations in latent membrane protein 1 of Epstein-Barr virus are associated with increased risk of posttransplant lymphoproliferative disorder in children. *Am J Transplant*. 2023;23:611-618. <https://pubmed.ncbi.nlm.nih.gov/36796762/>
39. Mainou BA, Raab-Traub N. LMP1 strain variants: biological and molecular properties. *J Virol*. 2006;80:6458-6468.
40. Kaymaz Y, Oduor CI, Aydemir O, et al. Epstein-Barr virus genomes reveal population structure and type 1 association with endemic Burkitt lymphoma. *J Virol*. 2020;94:e02007-19. doi:10.1128/JVI.02007-19
41. Banko AV, Lazarevic IB, Folc MM, et al. Characterization of the variability of Epstein-Barr virus genes in nasopharyngeal biopsies: potential predictors for carcinoma progression. *PLoS One*. 2016;11:e0153498.
42. Edwards RH, Seillier-Moisewitsch F, Raab-Traub N. Signature amino acid changes in latent membrane protein 1 distinguish Epstein-Barr virus strains. *Virology*. 1999;261:79-95.
43. Miller WE, Edwards RH, Walling DM, Raab-Traub N. Sequence variation in the Epstein-Barr virus latent membrane protein 1. *J Gen Virol*. 1994;75(Pt 10):2729-2740. doi:10.1099/0022-1317-75-10-2729
44. Sung NS, Edwards RH, Seillier-Moisewitsch F, Perkins AG, Zeng Y, Raab-Traub N. Epstein-Barr virus strain variation in nasopharyngeal carcinoma from the endemic and non-endemic regions of China. *Int J Cancer*. 1998;76:207-215.
45. Rickinson AB, Young LS, Rowe M. Influence of the Epstein-Barr virus nuclear antigen EBNA 2 on the growth phenotype of virus-transformed B cells. *J Virol*. 1987;61:1310-1317. <https://pubmed.ncbi.nlm.nih.gov/3033261/>
46. El-Mallawany NK, Mutai M, Mtete I, et al. Beyond endemic Burkitt lymphoma: navigating challenges of differentiating childhood lymphoma diagnoses amid limitations in pathology resources in Lilongwe, Malawi. *Glob Pediatr Health*. 2017;4:2333794X17715831.
47. Ryan SL, Peden JF, Kingsbury Z, et al. Whole genome sequencing provides comprehensive genetic testing in childhood B-cell acute lymphoblastic leukaemia. *Leukemia*. 2023;37:518-528.
48. Yang YJ, Xu XQ, Zhang YC, Hu PC, Yang WX. Establishment of a prognostic model related to tregs and natural killer cells infiltration in bladder cancer. *World J Clin Cases*. 2023;11:3444-3456. <https://pubmed.ncbi.nlm.nih.gov/37383920/>
49. Newman AM, Zaka M, Zhou P, et al. Genomic abnormalities of TP53 define distinct risk groups of paediatric B-cell non-Hodgkin lymphoma. *Leukemia*. 2021;36:781-789.

50. Wang Q, Lingel A, Geiser V, Kwapnoski Z, Zhang L. Tumor suppressor p53 stimulates the expression of Epstein-Barr virus latent membrane protein 1. *J Virol*. 2017;91:e00312-17. doi:[10.1128/jvi.00312-17](https://doi.org/10.1128/jvi.00312-17)
51. Hernandez-Vargas H, Gruffat H, Cros MP, et al. Viral driven epigenetic events alter the expression of cancer-related genes in Epstein-Barr-virus naturally infected Burkitt lymphoma cell lines. *Sci Rep*. 2017;7. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5517637/>:5852.
52. Kim KR, Yoshizaki T, Miyamori H, et al. Transformation of Madin-Darby canine kidney (MDCK) epithelial cells by Epstein-Barr virus latent membrane protein 1 (LMP1) induces expression of Ets1 and invasive growth. *Oncogene*. 2000;19:1764-1771. <https://pubmed.ncbi.nlm.nih.gov/10777210/>
53. Zhang J, Sommermann T, Li X, et al. LMP1 and EBNA2 constitute a minimal set of EBV genes for transformation of human B cells. *Front Immunol*. 2023;14:1331730.
54. Mukhtar F, Boffetta P, Risch HA, et al. Survival predictors of Burkitt's lymphoma in children, adults and elderly in the United States during 2000–2013. *Int J Cancer*. 2017;140:1494-1502.
55. Kieser A, Kaiser C, Hammerschmidt W. LMP1 signal transduction differs substantially from TNF receptor 1 signaling in the molecular functions of TRADD and TRAF2. *EMBO J*. 1999;18:2511-2521.
56. Münz C. *Epstein Barr Virus Volume 2: One Herpes Virus: Many Diseases*. Springer; 2015. https://books.google.com/books/about/Epstein_Barr_Virus_Volume_2.html?id=Mp-oCgAAQBAJ
57. Afrasiabi A, Parnell GP, Fewings N, et al. Evidence from genome wide association studies implicates reduced control of Epstein-Barr virus infection in multiple sclerosis susceptibility. *Genome Med*. 2019;11:1-13.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kim IE Jr., Oduor C, Stamp J, et al. Incorporation of Epstein–Barr viral variation implicates significance of Latent Membrane Protein 1 in survival prediction and prognostic subgrouping in Burkitt lymphoma. *Int J Cancer*. 2025;156(11):2188-2199. doi:[10.1002/ijc.35384](https://doi.org/10.1002/ijc.35384)