VIRUS EVOLUTION

Virus Evolution, 2021, 7(1): veab010

doi: 10.1093/ve/veab010
Research Article

A comprehensive analysis of genetic diversity of EBV reveals potential high-risk subtypes associated with nasopharyngeal carcinoma in China

Wen-Qiong Xue,[1] Tong-Min Wang,[1] Jing-Wen Huang,[2] Jiang-Bo Zhang,[1] Yong-Qiao He,[1] Zi-Yi Wu,[1] Ying Liao,[1] Lei-Lei Yuan,[2] Jianbing Mu,[3] and Wei-Hua Jia[1,2,*,†]

[1]State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Sun Yat-Sen University Cancer Center, Guangzhou, Guangdong 510060, China, [2]School of Public Health, Sun Yat-Sen University, Guangzhou, Guangdong 510080, China and [3]Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, NIH, Rockville 20852, MD, USA

*Corresponding author: E-mail: jiawh@sysucc.org.cn

†http://orcid.org/0000-0002-0528-8715

Abstract

Epstein–Barr virus (EBV), a widespread oncovirus, is associated with multiple cancers including nasopharyngeal carcinoma (NPC), gastric cancer and diverse lymphoid malignancies. Recent studies reveal that specific EBV strains or subtypes are associated with NPC development in endemic regions. However, these NPC specific subtypes were only identified in a portion of infected individuals due possibly to the limited samples size studied or the complicated population structures of the virus. To identify additional high-risk EBV subtypes, we conducted a comprehensive genetic analysis of 22 critical viral proteins by using the largest dataset of 628 EBV genomes and 792 sequences of single target genes/proteins from GenBank. The phylogenetic, principal component and genetic structure analyses of these viral proteins were performed through worldwide populations. In addition to the general Asia-Western/Africa geographic segregation, population structure analysis showed a 'Chinese-unique' cluster (96.57% isolates from China) was highly enriched in the NPC patients, compared to the healthy individuals (89.6% vs. 44.5%, $P < 0.001$). The newly identified EBV subtypes, which contains four Chinese-specific NPC-associated amino acid substitutions (BALF2 V317M, BNRF1 G696R, V1222I and RPMS1 D51E), showed a robust positive association with the risk of NPC in China (Odds Ratio = 4.80, 20.00, 18.24 and 32.00 for 1, 2, 3 and 4 substitutions, respectively, P_{trend} <0.001). Interestingly, the coincidence of positively selected sites with NPC-associated substitutions suggests that adaptive nonsynonymous mutation on critical proteins, such as BNRF1, may interact with host immune system and contribute to the carcinogenesis of NPC. Our findings provide a comprehensive overview of EBV genetic structure for worldwide populations and offer novel clues to EBV carcinogenesis from the aspect of evolution.

Key words: Epstein–Barr virus; nasopharyngeal carcinoma; phylogeny; principal component analysis; population genetic structure; positive selection.

1. Introduction

Epstein–Barr virus (EBV) belongs to the gamma-herpesviruses family and is ubiquitous in adults across the world. To some extent, it can be considered as the most successful oncovirus. After primary infection usually occurred before teenagers, EBV stays largely dormant in its life-long persistent infection. It can be reactivated under certain condition (Xu et al. 2012; He et al. 2019), which may be related to a wide spectrum of malignant diseases, including epithelial tumors like nasopharyngeal carcinoma (NPC), gastric cancer and diverse lymphoid malignancies such as Burkitt's lymphoma (BL), Hodgkin's lymphoma (HL) and NK/T-cell tumors (Young and Rickinson 2004). These EBV-associated malignancies show a great disparity of geographical distribution. For example, NPC is extremely high prevalent in South China and Southeast Asia, while BL is highly endemic in Africa and EBV-associated NK/T-cell lymphoma is relatively prevalent in East Asia, especially in Japan and Korea (Shannon-Lowe and Rickinson 2019).

Several factors such as environmental agents, behavior factors and human genetic susceptibility have been associated with the substantial geographic disparity of NPC. The role of EBV genetic diversity on NPC development, which has long been speculated, however, was only depicted recently with the identification of several geographically specific high-risk EBV subtypes for NPC. For example, individuals infected with EBV subtype BALF2-CCT showed 11 times higher risk in NPC development than the carriers of low-risk BALF2-ATC in South China (Xu et al. 2019). In Hong Kong population, the high-risk EBV subtypes are identified by the genetic variation in EBV-encoded small RNA locus (Hui et al. 2019) as well as the genome-wide single nucleotide variant profiles (Lam et al. 2020). Other risk EBV subtypes appear to be associated with replication-competent since the identified mutations are located in genes that promote EBV replication, including *BZLF1*, *EBNA1* and *BRLF1* (Dheekollu et al. 2017; Bristol et al. 2018; Zhang et al. 2018). However, these NPC-specific EBV subtypes were only identified in a portion of infected individuals due possibly to the limited study samples or the concurrence of geographical variants that are not disease associated (Lin et al. 2019). For instance, a previous identified risk variant for NPC (e.g.G155391A) in China was found to be scarce in NPC from Indonesia (Correia et al. 2018), indicating the complexity of virus genetic populations in the different areas (Wegner et al. 2019). Thus, it is important to clarify the genetic population structure with sufficient EBV strains from both the controls and cases before exploring NPC associated variants/subtypes.

To this end, we retrieved all available unique EBV genomes (N = 628) from GenBank including 231 cancer cases and 397 controls. Since the high-risk EBV subtypes identified to date seem all related with the virus replication, we, therefore, selected 22 genes that are involved in the EBV replication from the EBV genome(Kenney and Mertz 2014; Al Moustafa et al. 2018) for the genetic diversity analysis. These genes include (1) *EBNAs*, *LMP1*, *LMP2* and *BARF1*, which inhibit viral replication and function in carcinogenesis of EBV-associated cancers (Chakravorty et al. 2019), such as activating NF-κB signaling (*LMP1*), promoting Lyn/Syk kinases (*LMP2A*) and suppression of apoptosis (*BARF1*) (Tsang et al. 2019); (2) *BZLF1* and *BRLF1*, two viral immediate-early proteins which are associated with EBV reactivation (Morales-Sanchez and Fuentes-Panana 2018; Wu et al. 2018); and (3) genes that are vital to persistent EBV infection and carcinogenesis through recently pan-cancer analysis (Chakravorty et al. 2019), including *A73*, *BALF2*, *BALF3*, *BALF5*, *BARF0*, *BZLF1*, *BRLF1*, *EBNA1*, *EBNA3B*, *EBNA3C* and *RPMS1*. These viral genes were expressed in nearly all cancer type, indicating their functional roles in cancer pathophysiology (Hu et al. 2016; Borozan et al. 2018; Chakravorty et al. 2019; Peng et al. 2019). In this study, we conducted phylogenetic and principal component analyses for all selected EBV genes and identified high-risk EBV subtypes that are associated with NPC in China. Positive selection for candidate proteins provide further support for these NPC associated high-risk subtypes.

2. Materials and methods

2.1 Data sources and samples

A total of 781 sequences of entire or partial (>100 kb) EBV genome (released before Feb 2019) were downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank). The information including diseases/health condition, geographic origin, sample type was also retrieved. Sequences lacking the information of diseases/health condition and geographic origin were excluded. If multiple isolates were from identical cell lines or individuals, the more complete sequences were included. A total of 22 EBV proteins involving in viral replication and carcinogenesis were identified through literature mining (Supplementary Table S1). The complete coding sequences of these targeted genes were extracted. An additional set of EBV genomic sequences from China released between Feb and Oct in 2019 were acquired to confirm the cancer-related amino acid (AA) substitutions (Supplementary Table S2). The complete coding sequences of single target gene were obtained from GenBank (Supplementary Table S3). Duplicates in the combined dataset were excluded. In addition, a set of 244 sequences for *BRLF1* from our previous study were included, which contains 80 NPC cases, 80 controls from Guangdong province, and 84 healthy persons from Shanxi province, China (Zhang et al. 2018).

2.2 Phylogenetic and principal component analyses of EBV protein sequences

The coding sequences for each gene were aligned using MUSCLE following the instruction of MEGA-X and translated into AA sequences. The repeated regions in EBNA1 and EBNA3C were removed from the sequence alignments. The sequence from a chimaera of B95-8 and Raji EBV was defined as prototypic reference (Accession number NC007605). Maximum-likelihood phylogenetic tree was generated for each protein by IQ-TREE version1.6 with 1,000 bootstrap value using the best-fit model (Hoang et al. 2018; Zhou et al. 2018). Visualization and annotation of the trees were performed using treeio and ggtree packages in R (Wang et al. 2020; Yu et al. 2017). Principal component analysis (PCA) was conducted for each protein on AA substitutions (excluding repeated regions) using adegenet package (Jombart 2008).

2.3 Population structure analysis

Population structure analyses were conducted for different set of proteins. The analyses were performed using variable locus excluding incidental mutations (mutated AA = 1) by Structure software version 2.3 (Hubisz et al. 2009) with a burn-in period of 10,000 and 10,000 MCMC simulations at five iterations. Estimation for the most likely number of putative clusters follows the direction raised by Evanno et al. (2005) using

STRUCTURE HARVESTER web v0.6.94. The results were visualized using DISTRUCT (Rosenberg 2004).

2.4 Association analysis and identification of NPC-associated variants

The association analyses between single AA substitution and cancers or geographic origin were conducted using treeWAS (https://github.com/caitiecollins/treeWAS) in R (version 3.6.3) (Collins and Didelot 2018; San et al. 2019). Rare substitutions (minor frequency <0.05) were excluded before the analysis.

To further identify the variants associated with certain cancer such as NPC while not with geographical area, AA substitutions which were major in cancer-related cluster (>50%), while minor in other geographically distinct clusters (<20%) were extracted based on the proportions assigned to each cluster in population structure analyses.

In the present study, a novel mutation (BNRF1 V1222I, loc5399) was validated via nested PCR and Sanger sequencing using mouthwashes from 108 NPC cases and 179 controls in our EPI-NPC-2005 project (Xue et al. 2018). All samples from the cases were obtained before any therapy. The cases and controls were matched by age and gender.

2.5 Positive selection analysis

Positive selection analyses of viral proteins in East Asian strains (including the isolates from China, South Korea and Japan) were performed using the HyPhy software (Weaver et al. 2018) on the Datamonkey webserver (www.datamonkey.com). Four different methods were employed to detect the positively selected sites: Single-Likelihood Ancestor Counting, Fixed Effects Likelihood, Mixed Effects Model of Evolution, and Fast, Unconstrained Bayesian Approximation for inferring selection. Positively selected site was inferred and reported if it was predicted by any one of the methods.

3. Results

3.1 Characteristics of the target protein sequences

The final dataset of EBV genome contains 628 sequences that were released from January 2006 to 2019 (Supplementary Table S2). Nearly 40 per cent sequences were from China (N = 247), including 75 from NPC patients or cell lines (C666, M81), 16 from gastric cancer, 4 from lung cancer and 152 from healthy controls. A total of 164 sequences were from other East Asia countries, mainly Japan and South Korea. The isolates from lymphoma, gastric cancer and non-malignant disorders such as IM, PTLD and CAEBV were 45, 5 and 114, respectively. The sequences from Europe and America account to 54 and 43, respectively, most of which were from non-cancerous diseases (N = 41 for Europe; N = 30 for America). There were 57 sequences from Africa, including 34 from BL, 22 from non-cancerous individuals or cell lines and 1 from NPC. A total of 11 sequences were from Papua New Guinea in Oceania, with one from BL and ten from non-cancerous isolates (Table 1).

The complete coding regions of 22 target viral genes (*A73*, *BARF0*, *RPMS1*, *EBNA1*, *EBNA3B*, *EBNA3C*, *LMP1*, *LMP2A*, *BALF2*, *BALF3*, *BALF4*, *BALF5*, *BARF1*, *BHRF1*, *BNLF2A*, *BNLF2B*, *BNRF1*, *BILF1*, *BRLF1*, *BZLF1*, *LF1* and *LF2*) were extracted from the whole EBV genome and aligned. The additional sequences of individually downloaded genes or proteins were further incorporated into their corresponding alignments (Supplementary Table S3). The total number of sequences included in the phylogenetic and principal component analysis for each protein ranges from 574 to 873 (Supplementary Table S1).

3.2 Strong geographic segregation was corroborated in most viral proteins

The phylogenetic tree without branch lengths was built for each protein to detect potential virus population. For most of the proteins (18/22), we observe a strong geographic segregation between Asians and Western populations, which is consistent with a recent study using entire genome sequences of EBV (Wegner et al. 2019). The isolates from China tend to aggregate with those from Japan and South Korea, while sequences from western countries were clustered to those from Africa. The isolates from Southeast Asia were clustered with those from Papua New Guinea (Fig. 1, Supplementary S1–S3). Similar conclusion can be drawn from the PCA results (Fig. 2, Supplementary S1–S3). The isolates from Europe, America and Australia were assigned to the group of “Western world” in the ensuring analysis. This remarkable distinction between Asians and Western populations was also found in the genetic structure analysis (Fig. 3).

Although strong geographic segregation was detected among EBV strains, no cancer-specific clade was observed based on these trees. The EBV sequences from cancer patients or cell lines clustered together with those from non-cancerous isolates in the same geographic area (Fig. 1, Supplementary S1–S4). When the AA substitutions in analyzed proteins were compared between isolates from China and other populations, after correcting for population structure, fewer substitution were associated with geographical areas compared to the NPC-associated ones, especially in EBNA3B, EBNA3C and LMP1 (Supplementary Table S4), indicating the genetic diversity between these populations were mainly due to the clonal population structure.

3.3 Phylogenetic analyses and PCA revealed diverse patterns among different proteins

Although most of the proteins showed similar geographic segregation, inconsistent patterns were observed among different viral proteins. The most distinct difference lies in the relationship between isolates from China and other districts. According to the phylogenetic analysis, twenty-two proteins were assigned into four sets as follows:

i. *The clustered Chinese set*: this set exhibited a cluster composed of the isolates from China, while largely excluding those from other Asians. Take protein BNRF1 as an example, 52.70 per cent of the isolates from China can be delineated as a separate cluster apart from most other Asians (Figs. 1a, 2a). Other proteins assigned to this set included: BALF2, BALF4, BALF5, BHRF1, LMP2 and RPMS1 (Supplementary Fig. S1).

ii. *The clustered Asian set:* six proteins included into this set are BALF3, BNLF2A, BNLF2B, EBNA1, LF1 and LMP1. The typical cluster (BALF3) showed that the isolates from China were mingled with the strains from other Asian isolates, but the segregation of Asia-Western/Africa remains conspicuous (Figs. 1b, 2b; Supplementary Fig. S2).

iii. *The partitioned Western set*: this was represented by the protein BARF0 (Figs. 1c, 2c). The isolates from Western world and Africa were partitioned into different clusters and some were close to those from Asians. This set also included four other proteins: BRLF1, BZLF1, EBNA3B, and EBNA3C (Supplementary Fig. S3).

Table 1. The distribution of disease types and geographical origins of whole EBV genome included in the phylogenic analysis.

Diseases	China	Southeast Asia	East Asia	Oceania	Europe	America	Australia	Africa
NPC	75	20						1
GC	16	1	5		4	1		
LC	4							
Lymphoma		3	45	1	9	12		34
BL			15	1	1	12		34
HL					8			
NK/TL		3	26					
Lymphoma			2					
Lymphoepithelioma			2					
Non-cancer	152	7	114	10	41	30	21	22
IM			15			11	5	
PTLD			14		4	2	16	
CAEBV			84			12		
SOT					17			
Health/Other benign diseases	152	7	1		20	5		22

BL, Burkitt's lymphoma; CAEBV, chronical active Epstein-Barr virus infection; IM, infectious mononucleosis; LC, lung cancer; NK/TL, NK/T lymphoma; NPC, nasopharyngeal carcinoma; GC, gastric carcinoma, HL, Hodgkin's lymphoma, PTLD, post-transplant lymphoproliferative disease; SOT, solid organ transplant recipient;.

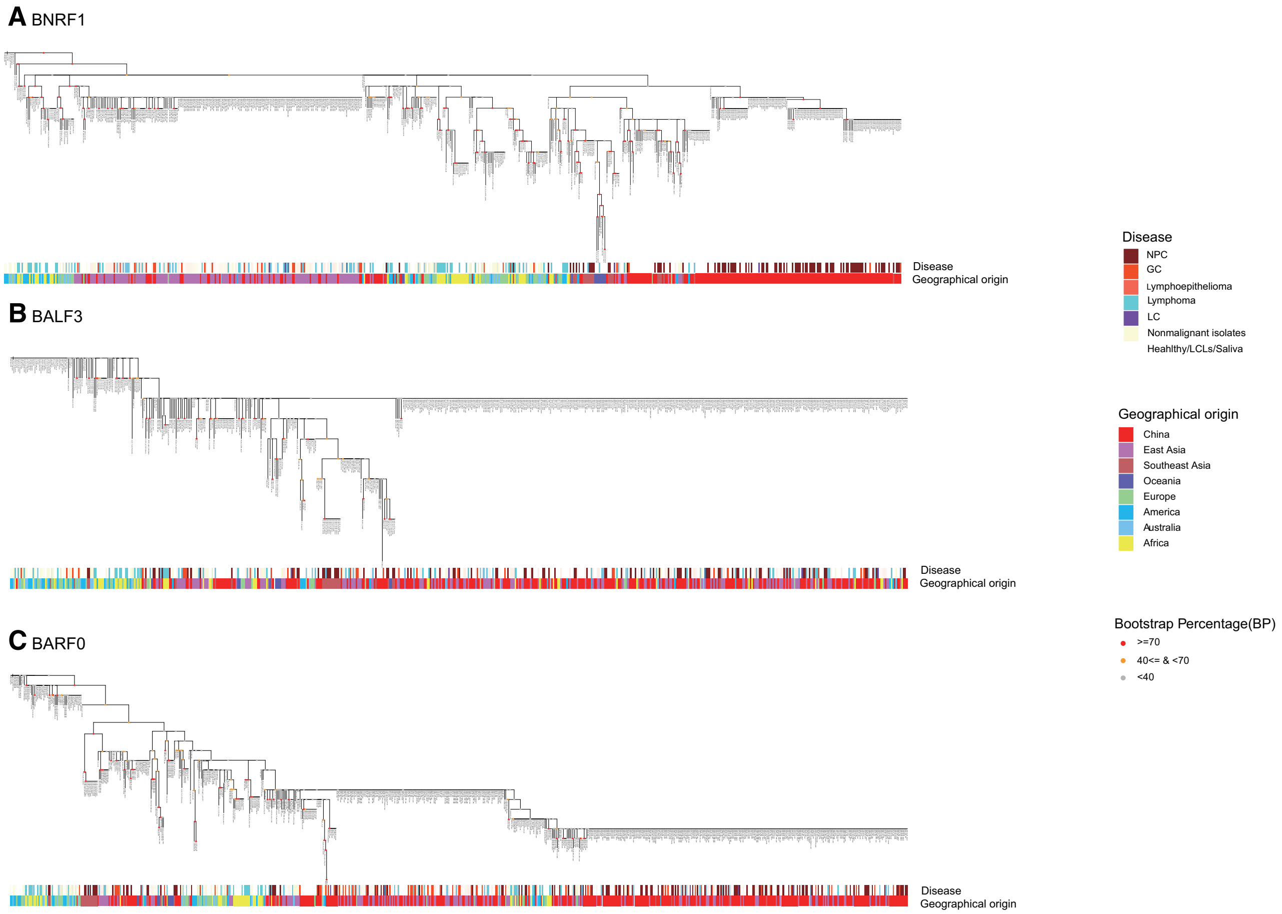

Figure 1. Phylogenetic trees of the exemplary EBV proteins from different cluster sets (A, BNRF1 for the clustered Chinese set; B, BALF3 for the clustered Asian set; C, BARF0 for the partitioned Western set) linked to geographic origin and disease type. Maximum-likelihood (ML) phylogenetic tree was built for each protein using AA sequences excluding repeat regions. The clades supported by a bootstrap value of <40 per cent, 40–70 per cent and >70 per cent were marked with grey, orange and red circles. GC, gastric carcinoma; LC, lung cancer, LCLs, lymphoblastoid cell lines.

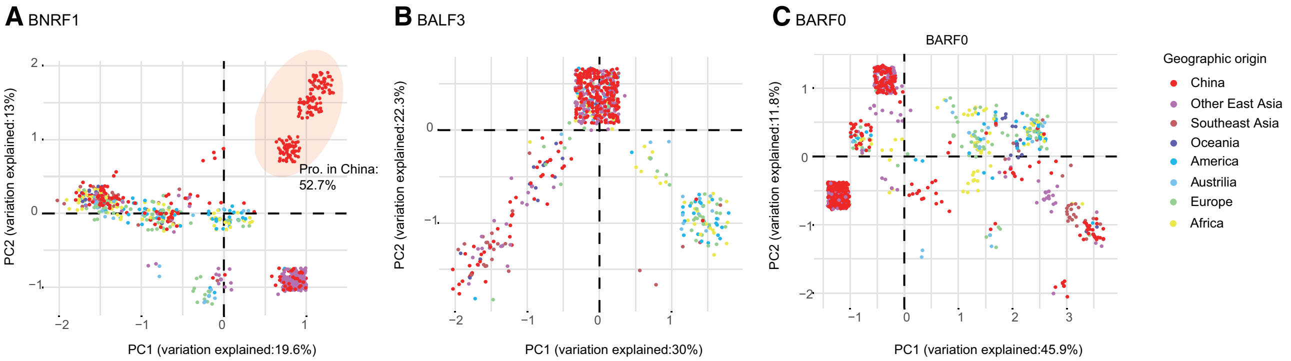

Figure 2. Principal component analysis of the exemplary EBV proteins from different cluster sets (a, BNRF1 for the clustered Chinese set; b, BALF3 for the clustered Asian set; c, BARF0 for the partitioned Western set) using AA substitutions data. The geographic origin of the dots is represented by different colors. Jittering is used to avoid overlapping dots. Red ellipses in plots of BNRF1 indicate the Chinese-unique clusters, and the proportion of the Chinese isolates assigned to this cluster was marked.

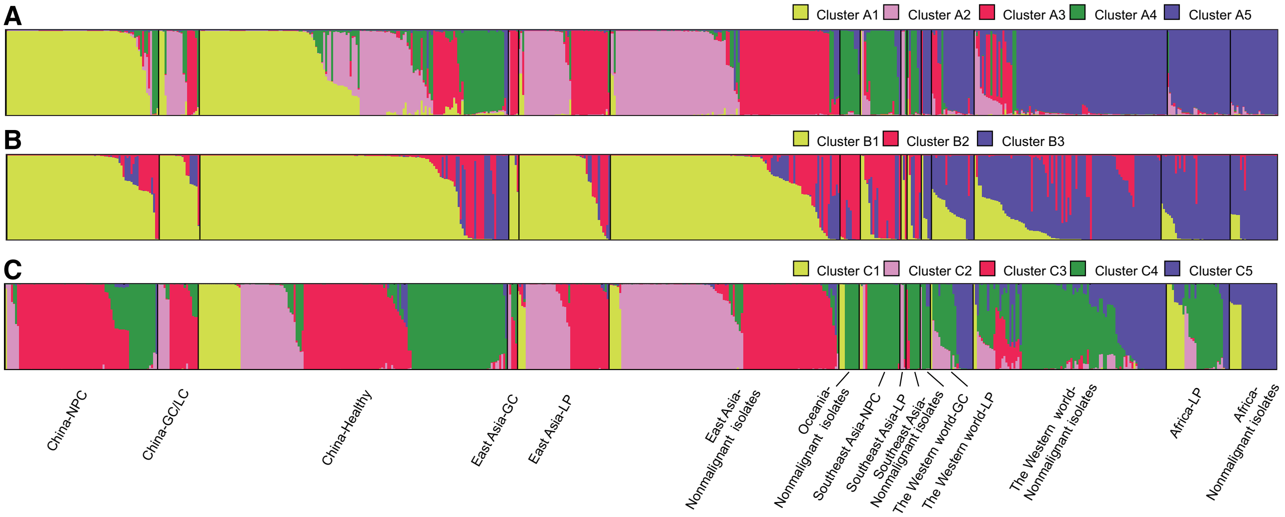

Figure 3. Population structure analysis on AA substitutions in different sets of EBV proteins (a, the clustered Chinese set; b, the clustered Asian set, c, the partitioned Western set). Western populations include individuals from America, Europe and Australia. GC, gastric carcinoma; LC, lung cancer; LP, lymphomas.

iv. *The ambiguous set*: it contains four proteins, namely A73, BARF1, BILF1 and LF2. The Asia-Western/Africa segregation were ambiguous in these trees (Supplementary Fig. S4), possible due to the relative conservativeness of the sequences among these genes.

3.4 Genetic structure analyses identified a Chinese-specific cluster enriched in the NPC patients

To identify the possible cancer associated with EBV, genetic structure analyses were further performed using the gene sets assigned above. Agreed with phylogenetic results, a 'Chinese-specific' cluster emerged from the first gene set analysis, in which the majority of the Chinese isolates (56.02%) were assigned. Particularly, in cluster A1, 96.57 per cent of isolates were from China (Fig. 3a). For the second gene set (clustered Asian set), 55.50 per cent isolates in cluster B1 were from China (Fig. 3b). In the plot of the third gene set (partitioned Western set), no Chinese-specific cluster was observed (Fig. 3c). The Western populations and Africans were assigned similarly by all three gene sets analysis, with corresponding percentages of 92.38 per cent, 78.22 per cent and 92.30 per cent for cluster A5, B3 and C5, respectively (Fig. 3).

Interestingly, the Chinese-specific cluster A1 was significantly more prevalent in NPC patients than in controls in China (89.6% vs. 44.5%, $P < 0.001$, Fig. 3a, Supplementary Table S5). However, the other relatively 'district-specific' clusters (the cluster A5, B3 and C5 in the Western populations and Africans) did not exhibit significant divergence between lymphomas and corresponding controls.

3.5 NPC risk EBV subtype defined by four substitutions conferred a much higher risk for NPC

To check whether specific EBV subtypes were related with NPC, we performed association study in Chinese-specific cluster A1 using treeWAS, which corrects the confounding effects, such as clonal population structure and recombination, and therefore enhances the statistical power of the association analysis (Collins and Didelot 2018; San et al. 2019). Four AA substitutions (BALF2 V317M, BNRF1 G696R, V1222I and RPMS1 D51E) were significantly associated with NPC (Supplementary Fig. S5), with Odd ratios (ORs) range from 5.38 to 21.10 (all $P < 0.001$, Table 2). These results were confirmed by newly released EBV genomes from 162 NPC patients and 38 controls in China, with the ORs range from 2.94 to 8.36 (all $P < 0.01$, Table 2). Particularly, the

BNRF1 V1222I mutation, which yields the highest OR of 21.10, was further validated using our independent samples from EPI-NPC-2005 project. The adjusted effect size was 5.55 [95% confidence interval (CI): 3.22–9.55, $P < 0.001$, Supplementary Table S6). When these substitutions were used as a combined indicator, a significant positive effect was observed in the original dataset ($P_{\text{trend}} < 0.001$). The ORs were 4.80, 20.00, 18.24 and 32.00 for the EBV isolates containing 1, 2, 3 and 4 AA substitutions, compared to the wild-type. These effects remained significant in additional Chinese sequence set, in which the ORs were 21.67 (95% CI: 2.51–187.16) and 7.99 (95% CI: 3.11–20.50) for EBV isolates with three and four AA substitutions, respectively (Table 2).

3.6 The NPC-associated AA substitutions in BNRF1, BALF2 and RPMS1 are under positive selection

The entire coding sequences of proteins BNRF1, BALF2 and RPMS1 were tested for positive selection and the results are shown in Table 3. It is intriguing to find that many NPC-associated mutations, including BALF2 V317M and BNRF1 V1222I, were under positive selection. In BNRF1, there were fifteen positively selected sites and seven were found to be associated with the NPC. In BALF2, three positively selected sites (317, 700 and 1093) also associated with NPC (3/8). While one site (50) in RPMS1 was associated with NPC (1/2) (Table 3, Supplementary S4).

It is worth noting that not all detected positive sites are the determinant variants for NPC in China. For example, non-synonymous mutations at BALF2 codon 1093 and BNRF1 codon 797 showed positive association with NPC in China, but their proportions in non-endemic areas of NPC (East Asia, the Western world and Africa) achieved to 36–96 per cent, suggesting these could not be the determinant mutations associated with Chinese NPC. The substitutions at other seven sites showed inverse association with NPC in China. And most of these substitutions had relatively higher frequencies in non-endemic areas, which was consistent with the former result.

4. Discussion

In this study, we identified potential NPC-risk EBV subtypes in China through the extensive genetic variation analysis. With the rapid development of sequencing technologies and accumulating viral sequences, inferring the EBV inter- and intra-populations genetic structure across the world are only possible in recent years. Here, we performed, by far, the largest EBV population structure analysis, and detected its geographic segregation, as well as the disparity of the mutations among different proteins. Importantly, a set of proteins generated a 'Chinese-specific' genetic cluster that was enriched in NPC. From this cluster, four NPC-associated AA substitutions located on three proteins (BALF2, BNRF1 and RPMS1) were identified. A combined indicator, which contains the NPC risk subtype, yields the highest risk estimation (OR) of 32.00. In addition, nearly half of the positively selected sites on these proteins were overlapped with the NPC-associated substitutions, implying that adaptive mutations of EBV proteins in Chinese population may drive the development of NPC.

A geographic segregation was observed for most of the viral protein in this study. The isolates from East Asians clustered distantly to those from the Western populations and Africans. Similar patterns have been reported by the studies using 127–270 entire genome sequences of EBV (Palser et al. 2015; Chiara et al. 2016; Borozan et al. 2018; Bridges et al. 2019; Wegner et al. 2019; Xu et al. 2019). Such highly structured virus population may relate with the host immune selection. As suggested in recent study, the genetically homogenous host populations (like Chinese and other Asians, and among Western populations) may force the virus to evolve along similar trajectories (Wegner et al. 2019). Long-time hybridization with immigrants, recombination between individuals and genetic drift also contribute to its distinct yet connected population structure (Chiara et al. 2016). Interestingly, our results highlight the subtle divergence of genetic structure among different viral proteins. About one third of the proteins like BNRF1 and BALF2 generate a 'Chinese-unique' cluster, in which a large proportion of isolates were clustered separately apart from other Asians. One recent study also showed the different tree topologies among nine EBV genes using 188 sequences (Zanella et al. 2019), but the geographic discrepancy has not been mentioned. The theory of co-evolution of proteins under host immune selection may give clues to this divergence. Different proteins of EBV may induce varied levels and different types of host adaptive immune response (Taylor et al. 2015). Researches on non-cancerous carriers of EBV found that lytic proteins, mainly BZLF1 and BRLF1, and latent proteins would normally induce frequent and intensified $CD4^{+}$ T-cell responses, while the viral proteins such as BZLF1, BRLF1, BALF2, BALF4, BALF5 and EBNA3s induce a relatively large $CD8^{+}$ T-cell response. Similarly, a small-size but stable $CD8^{+}$ T-cell response could be induced by BNRF1 (Taylor et al. 2015). The nonsynonymous mutations in these immunogenic proteins showed strong linkage disequilibrium with each other, indicating co-evolution of these proteins (Wegner et al. 2019). Thus, we speculate that the distinction of immune selection pressure represented by MHC molecules from various populations may result in the population structure disparity among viral proteins, as indicated in HIV studies (Moore et al. 2002; Bhattacharya et al. 2007).

A Chinese-unique cluster A1 is highly enriched in the NPC patients but also exists in the controls (89.6% vs. 44.5%), which is consistent with previous studies in identification of NPC-associated variants (Hui et al. 2019; Xu et al. 2019). These potential risk subtype of EBV may partly contribute to the carcinogenesis. Here, we identified four NPC-associated AA substitutions located on BALF2, BNRF1 and RPMS1 from the cluster A1. These substitutions remain significantly associated with NPC in Chinese after correction of confounding factors, such as the population structure and recombination, indicating the highly potential of the causal association. Although two variates (BALF2 V317M and RPMS1 D51E) had been reported (Feng et al. 2015; Xu et al. 2019), the newly identified BNRF1 variant (V1222I) yields the highest OR. Importantly, the combination of these four variants generated the highest risk estimation of 32.00 (95% CI: 9.18–111.49), indicating that EBV subtype containing these risk mutations may contribute to carcinogenic virulence in Chinese population. It is noteworthy that the genetic variation of BALF2 V317M and BNRF1 V1222I showed the strong evidence of positive selection, which possibly caused by the viral evolution to avoid host immune surveillance. In other words, the interaction between host immune system and virus shapes potential NPC risk subtypes and both contributed to the carcinogenesis. Indeed, the human leucocyte antigen genes have shown strong association with NPC (Bei et al. 2012). Thus, future combined research of host genetic susceptibility and viral subtype with careful consideration of possible founder effects may provide sufficient evidence of carcinogenicity as well as more precise prediction for NPC.

Table 2. The frequency of the four Chinese-unique EBV amino acid substitutions in worldwide populations and their association with NPC risk in China.

	Chinese EBV genomes from the original worldwide set				The additional set of EBV genomes from China				NPC isolates in Southeast Asia N (%)	Non-malignant isolates in other populations			
	NPC case N (%)	Control N (%)	OR (95% CI)	P-value	NPC case N (%)	Control N (%)	OR (95% CI)	P-value		Southeast Asia N (%)	East Asia N (%)	Western N (%)	Africa N (%)
BALF2 V317M													
V	10 (13.33)	90 (59.21)	1		19 (11.73)	20 (52.63)	1		18 (90.00)	7 (100.00)	100 (88.50)	79 (85.87)	22 (100.00)
M	63 (84.00)	62 (40.79)	9.15 (4.36–19.20)	<0.001	143 (88.27)	18 (47.37)	8.36 (3.77–18.55)	<0.001	2 (10.00)	0	0	2 (2.17)	0
Gap/equivocal	2 (2.67)	0	–		0	0	–		0	0	13 (11.50)	11 (11.96)	0
BNRF1 G696R													
G	29 (38.67)	116 (76.32)	1		38 (23.46)	18 (47.37)	1		19 (95.00)	7 (100.00)	112 (99.12)	89 (96.74)	22 (100.00)
R	43 (47.33)	32 (21.05)	5.38 (2.91–9.92)	<0.001	124 (76.54)	20 (52.63)	2.94 (1.41–6.11)	0.004	1 (5.00)	0	0	0	0
Gap/equivocal	3 (4.00)	4 (2.63)	–		0	0			0	0	1 (0.88)	3 (3.26)	0
BNRF1 V1222I													
V	5 (6.67)	90 (59.21)	1		21 (12.96)	18 (47.37)	1		19 (95.00)	7 (100.00)	112 (99.12)	89 (96.74)	22 (100.00)
I	68 (90.67)	58 (38.16)	21.10 (8.03–55.46)	<0.001	141 (87.04)	20 (52.63)	6.04 (2.76–13.24)	<0.001	1 (5.00)	0	0	0	0
Gap/equivocal	2 (2.67)	4 (2.63)	–		0	0	–		0	0	1 (0.88)	3 (3.26)	0
RPMS1 D51E													
D	7 (9.33)	77 (50.66)	1		19 (11.73)	16 (42.11)	1		19 (95.00)	7 (100.00)	106 (93.81)	86 (93.48)	22 (100.00)
N	67 (89.33)	75 (49.34)	9.83 (4.24–22.78)	<0.001	143 (88.27)	22 (57.89)	5.47 (2.45–12.21)	<0.001	1 (5.00)	0	0	1 (1.09)	0
Gap/equivocal	1 (1.33)	0	–		0	0	–		0	0	7 (6.19)	5 (5.43)	0
Number of changed AA													
0	3 (4.00)	72 (47.37)	1		12 (7.41)	13 (34.21)	1		18 (90.00)	7 (100)	94 (100)	79 (96.34)	22 (100.00)
1	3 (4.00)	15 (9.87)	4.80 (0.88–26.12)	0.07	5 (3.09)	3 (7.89)	1.81 (0.35–9.24)	0.478	1 (5.00)	0	0	3 (3.66)	0
2	5 (6.67)	6 (3.95)	20.00 (3.82–104.77)	<0.001	7 (4.32)	5 (13.16)	1.52 (0.38–6.09)	0.557	0	0	0	0	0
3	19 (25.33)	25 (16.45)	18.24 (4.97–66.92)	<0.001	20 (12.35)	1 (2.63)	21.67 (2.51–187.16)	<0.001	0	0	0	0	0
4	40 (53.33)	30 (19.74)	32.00 (9.18–111.49)	<0.001	118 (72.84)	16 (42.11)	7.99 (3.11–20.50)	<0.001	1 (5.00)	0	0	0	0
P for trend				<0.001				<0.001					

OR and P-value are calculated using logistic regression model. NPC, nasopharyngeal carcinoma.

Table 3. Positive selection analysis of BALF2, BNRF1 and RPMS1 protein using four methods and the frequency of these positively selected mutation in worldwide populations.

Protein	Positive selection analysis									Chinese EBV genomes from the original worldwide set			NPC isolates in Southeast Asia	Non-malignant isolates in other populations			
	Method	SLAC		FEL		MEME		FUBAR		NPC cases	Controls	P-value[*]		Southeast Asia	East Asia	Western	Africa
	Codon	dN-dS	P-value[#]	dN-dS	P-value[#]	omega+	P-value[#]	dN-dS	Post. Pr.[#]	N (%)	N (%)		N (%)	N (%)	N (%)	N (%)	N (%)
BALF2	181	13.762	0.445	4.015	0.137	0.398	0.67	10.513	**0.906**	0	0	NA	0	0	6 (5.31)	0	0
	278	6.66	0.746	1.283	0.714	>100	**0.04**	−1.036	0.585	0	1 (0.01)	1.000	0	0	0	0	0
	317	25.6	0.263	6.015	0.256	>100	0.67	23.914	**0.944**	63 (84.00)	62 (40.79)	**<0.001**	2 (10.00)	0	0	2(2.17)	0
	700	22.902	0.221	5.514	0.208	0.626	0.67	21.212	**0.950**	5 (6.67)	58 (38.16)	**<0.001**	1 (5.00)	1 (14.29)	42 (37.17)	16 (17.39)	0
	1065	34.348	0.135	5.751	0.231	>100	0.67	26.478	**0.955**	71 (94.67)	148 (97.37)	0.962	15 (75.00)	5 (71.43)	97 (85.84)	28 (30.43)	16 (72.73)
	1067	20.647	0.296	4.792	**0.096**	>100	0.67	19.233	**0.978**	2 (2.67)	3 (1.97)	0.715	2 (10.00)	0	5 (0.04)	0	0
	1093	62.437	**0.046**	10.137	0.264	0.428	0.67	41.879	**0.975**	66 (88.00)	65 (42.76)	**<0.001**	11 (55.00)	3 (42.86)	6 (5.31)	6 (6.52)	8 (36.36)
	1100	13.721	0.447	3.21	0.475	>100	**0.03**	7.955	0.839	0	0	NA	0	0	1 (0.88)	0	0
BNRF1	91	6.593	0.438	2.66	0.153	>100	**0.04**	1.374	0.667	0	0	NA	0	0	1 (0.01)	0	0
	405	4.829	0.642	0	1	>100	**0.03**	0.676	0.627	0	0	NA	0	0	0	0	0
	456	21.616	0.134	6.982	0.104	>100	0.67	16.861	**0.981**	5 (6.67)	40 (26.32)	**<0.001**	16 (80.00)	5 (71.43)	8 (7.08)	9 (9.78)	13 (59.09)
	485	34.672	**0.039**	14.487	**0.043**	>100	0.67	30.566	**0.993**	2 (2.66)	7 (4.61)	0.481	11 (55.00)	5 (71.43)	1 (0.88)	1 (1.09)	0
	486	12.925	0.302	3.653	0.217	>100	0.67	6.876	**0.925**	0	3 (1.97)	0.221	9 (45.00)	5 (71.43)	2 (1.77)	1 (1.09)	0
	497	27.703	0.111	8.759	0.153	0.693	0.67	19.549	**0.980**	4 (5.33)	39(25.66)	**<0.001**	6 (30.00)	0	7 (6.19)	3 (3.26)	0
	549	16.653	0.232	5.685	**0.073**	0.561	0.67	13.031	**0.986**	0	30 (19.74)	**<0.001**	0	0	104 (92.04)	12 (13.04)	0
	573	15.112	0.342	6.036	0.277	>100	0.67	12.195	**0.957**	1 (1.33)	4 (2.63)	0.531	0	0	0	0	0
	736	19.066	0.262	6.243	0.107	0.279	0.67	14.063	**0.986**	5 (6.67)	60 (39.47)	**<0.001**	19 (95.0)	6 (85.71)	12 (10.62)	68 (73.91)	16 (72.73)
	756	12.882	0.305	5.008	0.235	>100	0.67	9.682	**0.942**	1 (1.33)	18 (11.84)	**0.007**	1 (5.00)	0	0	1 (1.09)	0
	797	22.479	0.211	6.541	0.261	>100	0.67	14.337	**0.966**	72 (96.00)	131 (86.18)	**0.010**	18 (90.00)	5 (71.43)	108 (95.58)	45 (48.91)	16 (72.73)
	835	12.913	0.303	4.667	0.249	>100	0.67	9.514	**0.941**	1 (1.33)	3 (1.97)	0.730	0	0	1 (0.88)	1 (1.09)	0
	944	14.722	0.912	4.979	0.589	0.526	0.67	11.895	**0.903**	2 (2.67)	9 (5.92)	0.283	6 (30.00)	1 (14.29)	1 (0.88)	3 (3.26)	0
	1222	12.9	0.307	4.981	**0.088**	>100	0.11	10.104	**0.965**	68 (90.67)	58 (38.16)	**<0.001**	1 (5.00)	0	0	0	0
	1289	30.349	**0.059**	12.237	**0.062**	NA	1	25.745	**0.990**	1 (1.33)	9 (5.92)	0.113	12 (60.00)	4 (57.14)	3 (2.65)	4 (4.35)	0
RPMS1	50	20.449	0.615	59.125	**0.096**	>100	0.67	22.879	**0.910**	1 (1.33)	25 (16.45)	**0.001**	14 (70.00)	5 (71.43)	54 (47.79)	7 (7.61)	12 (54.55)
	81	36.249	0.416	80.111	**0.084**	>100	0.67	25.191	**0.914**	0	0	NA	0	0	1 (0.01)	0	0

All the positively selected codon for the three proteins were present, apart from the four NPC-associated amino acid substitutions reported in Tables 2 and 3. The positively selected codon was identified by at least one method with P-value <0.1 or FUBAR posterior probability >0.9. SLAC, single-likelihood ancestor counting; FEL, fixed effects likelihood; MEME, mixed effects model of evolution; FUBAR, fast, unconstrained Bayesian approximation for inferring selection; NA, not applicable; NPC, nasopharyngeal carcinoma.

[*] P-values are estimated using chi-square tests. Bold text indicates statistically significant difference between NPC cases and controls with *P*-value less than 0.05.

[#] Bold values indicate positively selected codon inferred by different methods.

The protein BNRF1 exhibit largely coincidence between the NPC-associated AA substitutions (7/12) and its positively selected sites. There seems to be two tendencies, BNRF1 V1222I showing positive association with NPC only presented in China, while five others enrolling lower risks was relatively more frequent in non-endemic areas. BNRF1 is a major tegument protein which plays the roles in the inhibition of host intrinsic defenses against viral reactivation, virus entry and host chromosomal instability (Shumilov et al. 2017; Tsai et al. 2011). We speculate that some adaptive mutations on BNRF1 may have effect on the development of NPC in Southern China. Further experimental investigations are needed in order to validate the roles of these mutations in NPC.

There are some limitations in this study. First, the detailed ethnic information of the host was not available. Some sequences from immigrants and multiethnic areas such as Southeast Asians, may be misallocated, although the EBV sequences from these areas were relatively less (N = 30). The results of PCA showed distinct cluster for these isolates from other Asians, indicating different origins. Second, in this study, no lymphoma-associated cluster in any populations was found, which may be due to the mixture of pathological subtypes and small sample size. Lastly, the phylogenic tree construction was limited by the available sequences, and the trees without detailed branch length cannot reflect evolutionary order and distance. Therefore, phylogenetic analyses were used only as a genetic classification method and the conclusions of genetic diversity between populations were drawn.

In summary, we have proposed three distinct patterns of population structure for different EBV proteins on the basis of the general Asia-Western/Africa segregation. These potential NPC risk EBV subtypes were identified from a Chinese-unique cluster, indicating geographic- and disease-specific feature concurrently. The evidence of positive selection suggests that viral adaptive mutation on some critical proteins like BNRF1 may involve in NPC development. Our findings provide a comprehensive overview of EBV population structure worldwide, implicating that viral disease-specific mutations should be investigated jointly with the population-specific variations. More importantly, our results provide the novel insights into the EBV carcinogenesis in NPC from the aspect of viral adaptive evolution and acknowledge the significance of combined research on host immunity-related genes and EBV.

Acknowledgments

This work was supported by grants from the Sino-Sweden Joint Research Program (81861138006), the National Natural Science Foundation of China (81803319, 81973131, 81903395, 81502056); the National Key Research and Development Program of China (2016YFC1302700), the Science and Technology Planning Project of Guangzhou, China (201804020094); the Natural Science Foundation of Guangdong Province (2017A030312003). The funding bodies had no role in the study design, data collection, data analysis, data interpretation, or writing of this article.

Data availability

GenBank accession numbers of EBV genome and single gene/protein sequences in this study are listed in Supplementary Table S2 and S3. The R scripts for tree visualization, PCA, association analysis and the annotation information of the sequences are included in Supplemented file 3 and available at https://github.com/xuewq-2020/Genetic_Diversity_of_EBV.git. The raw data of the validation samples have been uploaded to the Research Data Deposit (RDD) public platform (www.researchdata.org.cn), with the approval RDD number of RDDA2021001894.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: All the authors declare that there are no conflict of interests.

References

Al, M., A. E., et al. (2018) 'Editorial: EBV-Associated Carcinomas: Presence, Role, and Prevention Strategies', *Front Oncol*, 8: 528.

Bei, J. X., Jia, W. H., and Zeng, Y. X. (2012) 'Familial and Large-Scale Case-Control Studies Identify Genes Associated with Nasopharyngeal Carcinoma', *Seminars in Cancer Biology*, 22: 96–106.

Bhattacharya, T. et al. (2007) 'Founder Effects in the Assessment of HIV Polymorphisms and HLA Allele Associations', *Science*, 315: 1583–6.

Borozan, I. et al. (2018) 'Analysis of Epstein-Barr Virus Genomes and Expression Profiles in Gastric Adenocarcinoma', *Journal of Virology*, 92: 17. e01239-

Bridges, R. et al. (2019) 'Essential Role of Inverted Repeat in Epstein-Barr Virus IR-1 in B Cell Transformation; Geographical Variation of the Viral Genome', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374: 20180299.

Bristol, J. A. et al. (2018) 'A Cancer-Associated Epstein-Barr Virus BZLF1 Promoter Variant Enhances Lytic Infection', *PLoS Pathogens*, 14: e1007179.

Chakravorty, S. et al. (2019) 'Integrated Pan-Cancer Map of EBV-Associated Neoplasms Reveals Functional Host-Virus Interactions', *Cancer Research*, 79: 6010–23.

Chiara, M. et al. (2016) 'Geographic Population Structure in Epstein-Barr Virus Revealed by Comparative Genomics', *Genome Biology and Evolution*, 8: 3284–91.

Collins, C., and Didelot, X. (2018) 'A Phylogenetic Method to Perform Genome-Wide Association Studies in Microbes That Accounts for Population Structure and Recombination', *PLoS Computational Biology*, 14: e1005958.

Correia, S. et al. (2018) 'Sequence Variation of Epstein-Barr Virus: Viral Types, Geography, Codon Usage, and Diseases', *Journal of Virology*, 92: 18. e01132-

Dheekollu, J. et al. (2017) 'Carcinoma-Risk Variant of EBNA1 Deregulates Epstein-Barr Virus Episomal Latency', *Oncotarget*, 8: 7248–64.

Evanno, G., Regnaut, S., and Goudet, J. (2005) 'Detecting the Number of Clusters of Individuals Using the Software STRUCTURE: A Simulation Study', *Molecular Ecology*, 14: 2611–20.

Feng, F. T. et al. (2015) 'A Single Nucleotide Polymorphism in the Epstein-Barr Virus Genome is Strongly Associated with a High Risk of Nasopharyngeal Carcinoma', *Chinese Journal of Cancer*, 34: 563–72.

He, Y. Q. et al. (2019) 'Association between Environmental Factors and Oral Epstein-Barr Virus DNA Loads: A Multicenter Cross-Sectional Study in China', *The Journal of Infectious Diseases*, 219: 400–9.

Hoang, D. T. et al. (2018) ‘UFBoot2: Improving the Ultrafast Bootstrap Approximation’, *Molecular Biology and Evolution*, 35: 518–22.

Hu, L. et al. (2016) ‘Profiling of EBV Gene Expression in Nasopharyngeal Carcinoma through Paired-End Transcriptome Sequencing’, *Frontiers of Medicine*, 10: 61–75. ’Comprehensive

Hubisz, M. J. et al. (2009) ‘Inferring Weak Population Structure with the Assistance of Sample Group Information’, *Molecular Ecology Resources*, 9: 1322–32.

Hui, K. F. et al. (2019) ‘High Risk Epstein-Barr Virus Variants Characterized by Distinct Polymorphisms in the EBER Locus Are Strongly Associated with Nasopharyngeal Carcinoma’, *International Journal of Cancer*, 144: 3031–42.

Jombart, T. (2008) ‘Adegenet: A R Package for the Multivariate Analysis of Genetic Markers’, *Bioinformatics*, 24: 1403–5.

Kenney, S. C., and Mertz, J. E. (2014) ‘Regulation of the Latent-Lytic Switch in Epstein-Barr Virus’, *Seminars in Cancer Biology*, 26: 60–8.

Lam, W. K. J. et al. (2020) ‘Sequencing Analysis of Plasma Epstein-Barr Virus DNA Reveals Nasopharyngeal Carcinoma-Associated Single Nucleotide Variant Profiles’, *Clinical Chemistry*, 66: 598–605.

Lin, N. et al. (2019) ‘Genome-Wide Analysis of Epstein-Barr Virus Isolated from Extranodal NK/T-Cell Lymphoma’, *The Oncologist*, 24: e905–e913.e13.

Moore, C. B. (2002) ‘Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level’, *Science*, 296: 1439–43.

Morales-Sanchez, A., and Fuentes-Panana, E. M. (2018) ‘The Immunomodulatory Capacity of an Epstein-Barr Virus Abortive Lytic Cycle: Potential Contribution to Viral Tumorigenesis’, *Cancers (Cancers,)*, 10: 98.

Palser, A. L. et al. (2015) ‘Genome Diversity of Epstein-Barr Virus from Multiple Tumor Types and Normal Infection’, *Journal of Virology*, 89: 5222–37.

Peng, R. J. et al. (2019) ‘Genomic and Transcriptomic Landscapes of Epstein-Barr Virus in Extranodal Natural Killer T-Cell Lymphoma’, *Leukemia*, 33: 1451–62.

Rosenberg, N. A. (2003) ‘DISTRUCT: A Program for the Graphical Display of Population Structure’, *Molecular Ecology Notes*, 4: 137–8.

San, J. E. et al. (2019) ‘Current Affairs of Microbial Genome-Wide Association Studies: Approaches’, *Frontiers in Microbiology*, 10: 3119.

Shannon-Lowe, C., and Rickinson, A. (2019) ‘The Global Landscape of EBV-Associated Tumors’, *Frontiers in Oncology*, 9: 713.

Shumilov, A. et al. (2017) ‘Epstein-Barr Virus Particles Induce Centrosome Amplification and Chromosomal Instability’, *Nature Communications*, 8: 14257.

Taylor, G. S. et al. (2015) ‘The Immunology of Epstein-Barr Virus-Induced Disease’, *Annual Review of Immunology*, 33: 787–821.

Tsai, K. et al. (2011) ‘EBV Tegument Protein BNRF1 Disrupts DAXX-ATRX to Activate Viral Early Gene Transcription’, *PLoS Pathogens*, 7: e1002376.

Tsang, C. M. et al. (2020) ‘Translational Genomics of Nasopharyngeal Cancer’, *Seminars in Cancer Biology*, 61: 84–100.

Wang, L. G. et al. (2020) ‘Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data’, *Molecular Biology and Evolution*, 37: 599–603.

Weaver, S. et al. (2018) ‘Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes’, *Molecular Biology and Evolution*, 35: 773–7.

Wegner, F. et al. (2019) ‘Co-Evolution of Sites under Immune Selection Shapes Epstein-Barr Virus Population Structure’, *Molecular Biology and Evolution*, 36: 2512–21.

Wu, C. C. et al. (2018) ‘Perspective: Contribution of Epstein-Barr Virus (EBV) Reactivation to the Carcinogenicity of Nasopharyngeal Cancer Cells’, *Cancers (Basel)*, 10: 120.

Xu, F. H. et al. (2012) ‘An Epidemiological and Molecular Study of the Relationship between Smoking, Risk of Nasopharyngeal Carcinoma, and Epstein-Barr Virus Activation’, *JNCI: Journal of the National Cancer Institute*, 104: 1396–410.

Xu, M. et al. (2019) ‘Genome Sequencing Analysis Identifies Epstein-Barr Virus Subtypes Associated with High Risk of Nasopharyngeal Carcinoma’, *Nature Genetics*, 51: 1131–6.

Xue, W. Q. et al. (2018) ‘Decreased Oral Epstein-Barr Virus DNA Loads in Patients with Nasopharyngeal Carcinoma in Southern China: A Case-Control and a Family-Based Study’, *Cancer Medicine*, 7: 3453–64.

Young, L. S., and Rickinson, A. B. (2004) ‘Epstein-Barr Virus: 40 Years On’, *Nature Reviews Cancer*, 4: 757–68.

Yu, G. C. et al. (2017) ‘GGTREE: An R Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data’, *Methods in Ecology and Evolution*, 8: 28–36.

Zanella, L. et al. (2019) ‘A Reliable Epstein-Barr Virus Classification Based on Phylogenomic and Population Analyses’, *Scientific Reports*, 9: 9829.

Zhang, J. B. et al. (2018) ‘Variations in BRLF1 Promoter Contribute to the Elevated Reactivation Level of Epstein-Barr Virus in Endemic Areas of Nasopharyngeal Carcinoma’, *EBioMedicine*, 37: 101–9. ’Natural

Zhou, X. et al. (2018) ‘Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets’, *Molecular Biology and Evolution*, 35: 486–503.