nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0436-5

Genome sequencing analysis identifies Epstein–Barr virus subtypes associated with high risk of nasopharyngeal carcinoma

Miao Xu[1,2,15], Youyuan Yao[1,3,15], Hui Chen[4,15], Shanshan Zhang[1,15], Su-Mei Cao[1], Zhe Zhang[5], Bing Luo[6], Zhiwei Liu[7], Zilin Li[2], Tong Xiang[1], Guiping He[1], Qi-Sheng Feng[1], Li-Zhen Chen[1], Xiang Guo[1,8], Wei-Hua Jia[1], Ming-Yuan Chen[1], Xiao Zhang[1], Shang-Hang Xie[1], Roujun Peng[1], Ellen T. Chang[9,10], Vincent Pedergnana[4], Lin Feng[1], Jin-Xin Bei[1], Rui-Hua Xu[1], Mu-Sheng Zeng[1], Weimin Ye[7], Hans-Olov Adami[7,11], Xihong Lin[2], Weiwei Zhai[4,12,13]*, Yi-Xin Zeng[1]* and Jianjun Liu[4,14]*

Epstein–Barr virus (EBV) infection is ubiquitous worldwide and is associated with multiple cancers, including nasopharyngeal carcinoma (NPC). The importance of EBV viral genomic variation in NPC development and its striking epidemic in southern China has been poorly explored. Through large-scale genome sequencing of 270 EBV isolates and two-stage association study of EBV isolates from China, we identify two non-synonymous EBV variants within *BALF2* that are strongly associated with the risk of NPC (odds ratio (OR) = 8.69, $P = 9.69 \times 10^{-25}$ for SNP 162476_C; OR = 6.14, $P = 2.40 \times 10^{-32}$ for SNP 163364_T). The cumulative effects of these variants contribute to 83% of the overall risk of NPC in southern China. Phylogenetic analysis of the risk variants reveals a unique origin in Asia, followed by clonal expansion in NPC-endemic regions. Our results provide novel insights into the NPC endemic in southern China and also enable the identification of high-risk individuals for NPC prevention.

EBV was discovered in 1964 (refs. [1,2]) and is the first human virus to be associated with cancers, including NPC, a subset of gastric carcinoma and several kinds of lymphomas[3]. Although EBV infection is ubiquitous in human populations worldwide, its most closely associated malignancy, NPC, has a unique geographical distribution. Rare in most of the world, NPC is a very common cancer in southern China, where the incidence rate can reach 20 to 40 cases per 100,000 individuals per year[4]. Multiple human susceptibility loci, including *HLA*, *CDKN2A* and *CDKN2B*, *TNFRSF19*, *MECOM* and *TERT* loci, have been discovered for NPC, but the contributions of these loci to overall risk are limited[5–8]. Moreover, the risk variants at these loci are widely distributed in the Chinese population and therefore cannot explain the unique endemism of NPC to southern China. Thus, the cause of NPC, commonly known as the Cantonese cancer, remains unknown.

Since the first EBV genome sequence, B95-8, was published in 1984 (ref. [9]), more than 100 EBV genomes have been sequenced in spontaneous lymphoblastoid cell lines and patients with EBV-associated diseases. These studies revealed important genomic variations among EBV isolates from different geographical origins[10–15]. Although the importance of EBV genome variation in the risk of EBV-associated diseases has been explored[15–18], these studies suffered from the confounding effect of geographical distribution and insufficient sample sizes. As a result, robust epidemiological and genetic evidence that links specific EBV strains to the pathogenesis of NPC is lacking.

In the current study, we performed large-scale whole-genome sequencing (WGS) of 215 EBV isolates from patients diagnosed with EBV-associated cancers (including NPC, gastric carcinoma and lymphomas) and 54 isolates from healthy controls recruited from both NPC-endemic and non-endemic regions of China. Through a comprehensive and systematic association analysis of EBV genomic variation and subsequent replication analysis in an independent sample, we identified two non-synonymous variants in the *BALF2* gene associated with high risk for NPC. These two variants explain 83% of the overall risk in NPC-endemic southern China. In addition, phylogenetic analysis of EBV isolates from the current study and worldwide strains suggest a unique Asian origin followed by a clonal expansion of the two NPC-high-risk variants in southern China. Thus, we have discovered the high-risk EBV subtypes that contribute significantly to the overall risk of NPC, as well as its unique epidemic in southern China.

[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, China. [2]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [3]Department of Comprehensive Medical Oncology, Key Laboratory of Head & Neck Cancer Translational Research of Zhejiang Province, Zhejiang Cancer Hospital, Hangzhou, China. [4]Human Genetics, Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [5]Department of Otolaryngology/Head and Neck Surgery, First Affiliated Hospital of Guangxi Medical University, Nanning, China. [6]Department of Medical Microbiology, Qingdao University Medical College, Qingdao, China. [7]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [8]Department of Nasopharyngeal Carcinoma, Sun Yat-sen University Cancer Center, Guangzhou, China. [9]Center for Health Sciences, Exponent, Menlo Park, CA, USA. [10]Stanford Cancer Institute, Stanford, CA, USA. [11]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [12]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. [13]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China. [14]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [15]These authors contributed equally: Miao Xu, Youyuan Yao, Hui Chen, Shanshan Zhang. *e-mail: weiweizhai@ioz.ac.cn; zengyx@sysucc.org.cn; liuj3@gis.a-star.edu.sg

Results

EBV whole-genome sequencing. Using a capture-based protocol, we obtained EBV genome sequences from 215 samples of tumor, saliva and plasma from patients with EBV-associated cancer (NPC, gastric carcinoma and lymphomas) and 54 saliva samples from healthy donors, as well as one genome from NPC cell line C666-1 (for an overview of the study, see Supplementary Fig. 1, Supplementary Tables 1–4 and Methods). Of the 270 EBV isolates, 221 were obtained from the NPC-endemic region of southern China (Guangdong and Guangxi provinces), and 49 were from NPC-non-endemic regions of China. The average sequencing depth of all of the isolates was 1,282×, and on average, 95% of the EBV genome was covered with at least 10× coverage (Supplementary Fig. 2). Using B95-8 as the reference, we identified a total of 8,469 variants (8,015 SNPs, 454 INDELs) across the EBV genome (for variant statistics, see Supplementary Table 5 and Supplementary Fig. 2). The number of variants identified in each sample ranged from 1,006 to 2,104, with *EBNA-2*, *EBNA-3A*, *EBNA-3B* and *EBNA-3C*, and *LMP-2A* and *LMP-2B* being the most polymorphic genes (Supplementary Fig. 3), consistent with other reports[14–16]. To explore the accuracy in sequencing and variant calling, we compared the re-sequenced C666-1 EBV genome against the published record and found a high concordance rate of 97.9% (ref. [19]) (Supplementary Table 6). In addition, when subsets of variants discovered by EBV WGS were re-genotyped by Sanger sequencing and MassArray iPLEX assay, 97.55% and 99.99% of tested variants were confirmed, respectively (Supplementary Tables 7,8). Both results indicate that our sequencing and variant calling procedures were highly accurate.

To understand intra-host polymorphism within an individual, two EBV fragments were amplified and sequenced in paired saliva and tumor samples from 25 patients with NPC. The variant difference between the paired saliva and tumor samples (median 1.1%, first to third quartile: 0–3.4%) was substantially lower than the between-host difference (median 13.5%, first to third quartile: 3.7–16.9%) (Supplementary Fig. 4). In addition, we sequenced the EBV whole genomes from the same patient with NPC in paired tumor and saliva samples and observed a 99.27% concordance between the variants in EBV tumor and saliva isolates (Supplementary Table 9). Taken together, these observations suggest that paired saliva and tumor samples from the same subject had the same EBV genome or strain. Therefore, we combined the genome sequence information from tumor and saliva samples from NPC cases in subsequent analyses.

***BALF2* gene region showing strongest association.** To investigate the impact of EBV genomic variations on NPC risk, we performed a two-stage genome-wide association study (GWAS). In the discovery phase, we included the EBV genomes from 156 NPC cases and 47 controls from the 270 EBV WGS isolates. These isolates included in the discovery phase are exclusively from Guangdong and Guangxi provinces in the NPC-endemic region of southern China. A principal component analysis (PCA) of the human genome variation of all of the cases and controls with the reference population samples from the 1000 Genomes project[20] confirmed their ancestral origin and the genetic match between cases and controls (Supplementary Fig. 5). We also performed PCA of EBV genomes using all of the 270 strains from the current study together with 97 publicly available genomes. The distribution of the EBV strains along the first principal component was continuous, ranging from Africa and Europe to Asia (Fig. 1a). Within Asia, the second principal component showed a partial separation of the isolates from NPC-endemic and NPC-non-endemic regions of Asia (Fig. 1a,d).

To control for the potential impact of the population structures of both the human and EBV genomes, the GWAS was performed using a generalized linear mixed model, with age, sex, the first four human principal components and previously reported NPC human GWAS SNPs (rs2860580 and rs2894207 at the *HLA* locus; Supplementary Table 10, see Methods) as fixed effects and the genetic relatedness matrix of EBV genomes as random effects[21]. The discovery analysis revealed multiple association signals along the EBV genome. The strongest association was in the *BALF2* region (NC_007605.1:162507C>T, $P=9.17\times10^{-5}$) without any indication of inflation due to genetic structure (genomic control inflation factor $\lambda_{GC}=1.03$; Fig. 2a, Supplementary Table 11 and Supplementary Fig. 6). We also investigated evidence of association of the recently reported EBER2 variants[18] in our discovery data set. NC_007605.1:7048A>C, which was a leading variant for the reported associations at the EBER2 region, showed significant association in our genome-wide analysis ($P=1.25\times10^{-7}$) but the significance was largely abolished by controlling for population structure ($P=1.52\times10^{-2}$; Supplementary Fig. 6).

In addition, we performed a multi-SNP GWAS using Bayesian variable-selection regression by piMASS[22], which provided consistent and strong evidence for the association in the *BALF2* region (posterior probability = 0.86; Fig. 2b). When we evaluated the statistical significance of association using a permutation test (see Methods), only the associations within the *BALF2* region reached genome-wide significance (suggestive genome-wide significance, $P<4.07\times10^{-4}$). Consistent with the extensive linkage disequilibrium (LD) in the EBV genome (Supplementary Fig. 7), conditioning on the genetic effects of the SNPs in the *BALF2* region greatly reduced the extensive associations across the entire EBV genome (Supplementary Fig. 8).

Fine-mapping and validation of *BALF2* variants. We performed a Bayesian fine-mapping analysis to prioritize potentially causal SNPs in the *BALF2* gene region using PAINTOR and found that only the three non-synonymous coding variants (NC_007605.1:162215C>A, 162476T>C and 163364C>T) were significantly associated (Supplementary Fig. 9 and Supplementary Table 12). We genotyped these variants in an independent sample of 483 NPC cases and 605 age- and sex-matched healthy controls (validation phase; Supplementary Table 13). To reduce the potential impact of population stratification, all the cases and controls were recruited from the single NPC-endemic region, Zhaoqing county, in the Guangdong province of China. All three *BALF2* SNPs were significantly associated with NPC risk in the independent sample ($P<0.017$, 0.05 out of 3), consistent with the discovery phase results (Table 1). The meta-analysis of the combined discovery and validation samples confirmed the associations with the three SNPs of *BALF2* with genome-wide significance according to both permutation analysis (162215_C, OR = 7.60, $P=1.42\times10^{-18}$; 162476_C, OR = 8.69, $P=9.69\times10^{-25}$; and 163364_T, OR = 6.14, $P=2.40\times10^{-32}$; Table 1). All three SNPs showed significant LD (Supplementary Fig. 10) but conditional analysis revealed that the associations with SNPs 162215C>A and 162476T>C were correlated, whereas SNP 163364C>T showed an independent association that also reached genome-wide significance (Table 1).

We further explored the association between the haplotypes (strains) composed of SNPs 162215C>A, 162476T>C and 163364C>T, and the risk of NPC. When the haplotype composed of the three low-risk variants (A-T-C) were used as a reference, we found no association for the haplotype carrying the high-risk variant for SNP 162215_C (haplotype C-T-C: OR = 1.12; $P=0.78$), although the number of haplotypes for comparison was limited (Table 2 and Supplementary Table 14). Both the haplotypes carrying the high-risk variants of either all three SNPs or only SNPs 162215_C and 162476_C showed a strong risk effect (haplotype C-C-T, OR = 11.71, $P=2.39\times10^{-24}$; haplotype C-C-C, OR = 3.50, $P=1.22\times10^{-5}$; Table 2 and Supplementary Table 14), but haplotype C-C-T showed a significantly stronger effect than haplotype C-C-C ($P=2.07\times10^{-10}$), clearly indicating the additional risk effect of SNP

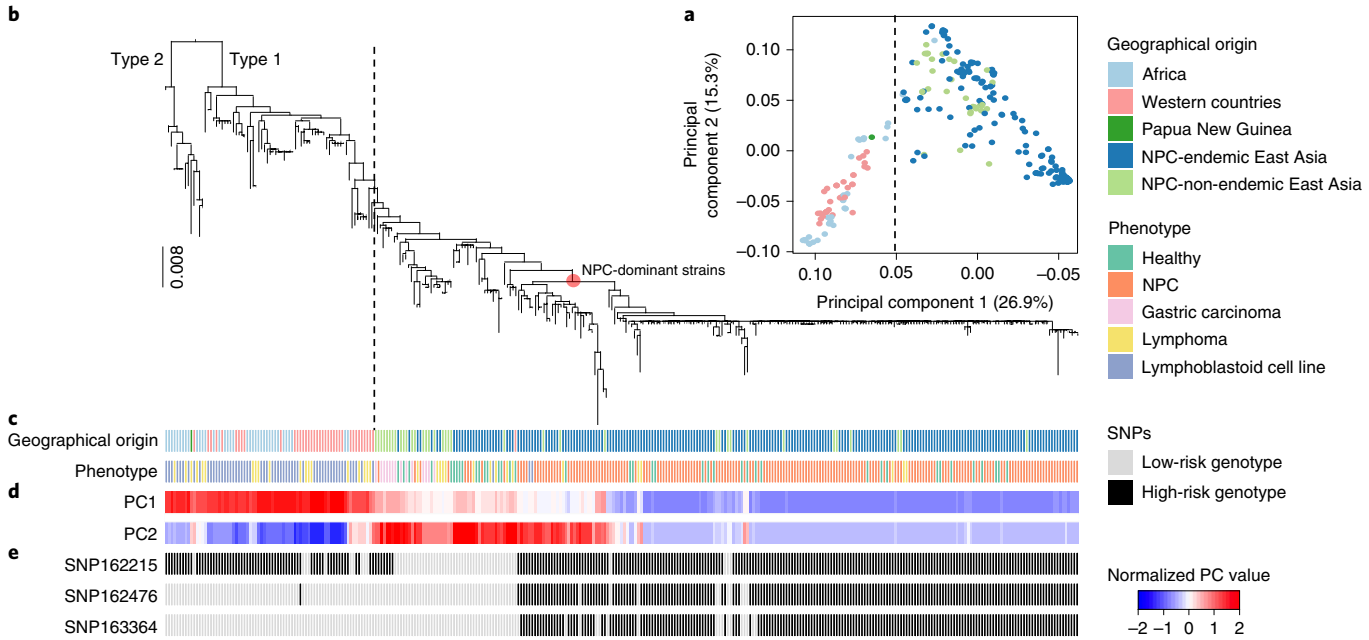

Fig. 1 | Principal component and phylogenetic analyses of EBV genomes. a, PCA of 270 EBV isolates sequenced in the current study and 97 published isolates. The first two principal-component scores (PC1 and PC2) are plotted. PC1 explains 26.9% of the total genomic variance, and discriminates between East Asian, and Western and African strains. PC2 explains 15.3% of the total variance. Western countries include the United Kingdom, United States and Australia. **b**, Phylogeny of 230 EBV single strains sequenced in the current study and 97 published strains. Macacine herpesvirus 4 genome sequence (NC_006146) was used as the outgroup to root the tree. Type 1 and type 2 EBV lineages are indicated. The red dot on the phylogeny indicates the lineage of the NPC-dominant EBV strains, where 22 of 37 strains from healthy controls in NPC-endemic regions in southern China were located. Dashed lines in **a** and **b** indicate the separation between East Asian, and Western and African strains. **c**, Geographical origins and phenotypes of samples from which EBV strains were sequenced. **d**, The normalized values of PC1 and PC2 scores are shown from blue to red. **e**, The genotypes of SNPs 162215C>A, 162476T>C and 163364C>T in each isolate.

163364_T. The haplotype analysis further confirmed that NPC risk is primarily associated with SNPs 162476_C and 163364_T, and that the association with SNP 162215_C needs to be evaluated further. We also performed pairwise interaction analysis that showed no evidence for an interaction between SNPs 162476T>C and 163364C>T ($P=0.93$). Finally, multiple regression analysis yielded independent risk effects (OR) of 3.31 for SNP 162476_C and 3.35 for SNP 163364_T (Supplementary Table 15), which were consistent with the risk effect of the haplotype carrying the two high-risk variants (haplotype C-C-T, OR = 11.71; Table 2).

Given the well-known function of *BALF2* as the single-stranded DNA binding protein, a core component of the viral DNA replication machinery[23–25], we also investigated oral EBV abundance and its association with different *BALF2* haplotypes in the 533 NPC cases and 651 controls. The viral DNA load varied widely across the samples, and viral DNA abundance in saliva was significantly lower in patients than in controls ($P=4.2\times10^{-13}$; Supplementary Fig. 11). In both cases and controls, we observed a consistent decrease in viral load among individuals infected by the high-risk subtypes (C-C-T or C-C-C), especially C-C-C ($P=0.056$), compared to the low-risk (A-T-C) haplotype (Supplementary Fig. 12), but the differences were marginally significant (Supplementary Table 16).

The evolution of the high-risk subtypes. In China, the frequency of the two high-risk haplotypes (C-C-T and C-C-C) was very high in the NPC-endemic region (93.27% in NPC cases and 63.04% in controls), but much lower in non-endemic areas (55% in NPC cases, 14.29% in controls; Supplementary Table 17). Interestingly, the two risk haplotypes were absent or extremely rare in non-Asian individuals from Africa and western countries including the United Kingdom, United States and Australia (Supplementary Table 17), which suggests an Asian origin of the EBV high-risk variants. To further explore the evolution of the EBV risk variants, we investigated the phylogenetic relationship among the EBV strains from the current study and from published sequences. By examining the frequency and distribution of heterozygous SNPs, we identified 230 EBV single-infection strains from the 270 WGS isolates (see Methods, Supplementary Fig. 13 and Supplementary Table 18). With the 230 EBV isolates from the current study and 97 publicly available genomes, we performed phylogenetic inference. The evolutionary relationship among all sequences was highly unbalanced, with a deep split between type 1 and type 2 EBV isolates (Fig. 1b). All type 2 EBV isolates were geographically restricted to Africa, as observed previously[14,15,26]. The type 1 EBV clade showed a continuous branching starting from Africa, Europe and Asia, and matching the overall distribution along the first prinicpal component in the PCA (Fig. 1b–d). As in previous studies[17,27], 97% of the 230 EBV single strains were found to be the China 1 subtype, and 2% were the China 2 subtype (defined by *LMP-1* classification; Supplementary Fig. 14). Within the Asian group, isolates from NPC-non-endemic areas clustered towards the basal position of the lineage, similarly to the pattern observed along the second principal component in the PCA map (Fig. 1b–d). The most striking pattern in the phylogenetic relationship was a rapid radiation of NPC-dominant strains in the endemic population from southern China. EBV genomes from patients with NPC appeared to have expanded recently from a common ancestor, and more than half (22 of 37) of healthy controls from this region were infected with NPC-dominant strains (Fig. 1b,c).

When mapping the three SNPs of *BALF2* (SNPs 162215C>A, 162476T>C and 163364C>T) onto the phylogenetic tree of the EBV genomes, we observed that allof the strains carrying the risk variants of SNPs 162476_C and 163364_T were within the Asian

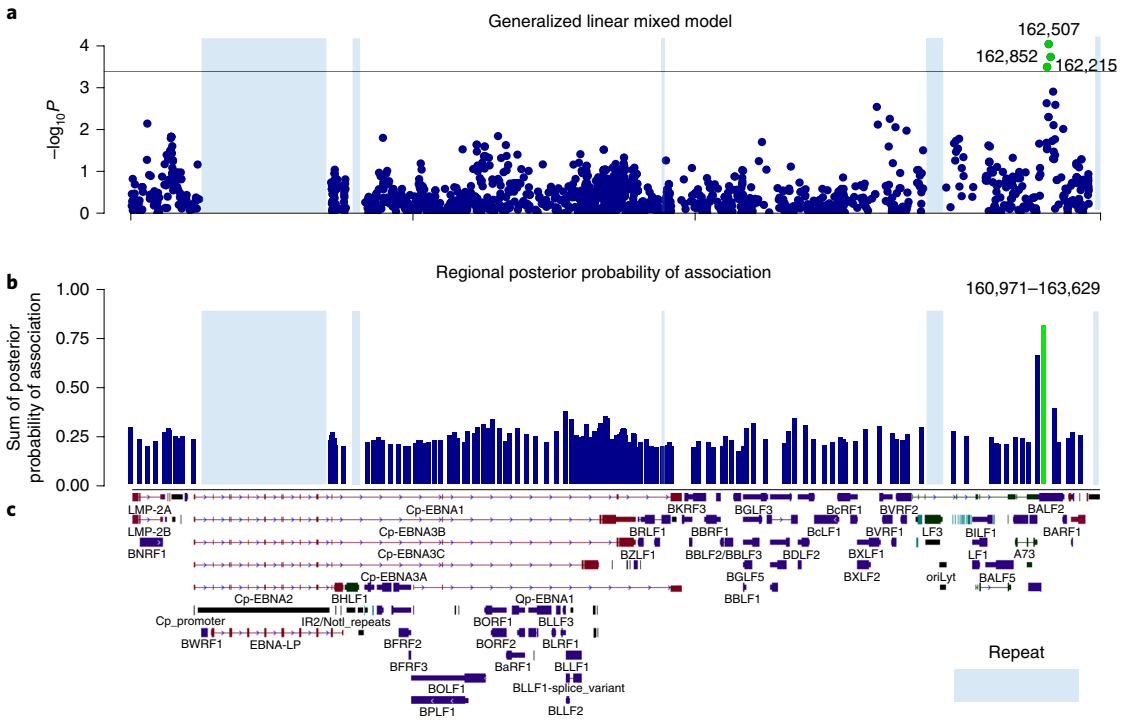

Fig. 2 | Genome-wide association analysis of EBV variants in 156 NPC cases and 47 controls. a, Manhattan plot of genome-wide *P* values from the association analysis using a generalizedlinear mixed model. The $-\log_{10}$-transformed *P* values (*y* axis) of 1,545 variants in 156 NPC cases and 47 controls are presented according to their positions in the EBV genome. The minimum *P* value (SNP 162507C>T) is 9.17×10^{-5}. The red line is the suggestive genome-wide significance *P*-value threshold of 4.07×10^{-4}. The three SNPs 162507C>T, 162852G>T and 162215C>A reaching genome-wide significance are shown in green. **b**, The regional plot of the posterior probabilities of association. The EBV genome was partitioned into overlapping 20-variant bins with 10-variant overlaps between adjacent bins. The sum of the posterior probabilities for variants was assigned to each region. The one region from position 160,971 to 163,629 with strong evidence (>0.85) for association with NPC risk is shown in green. **c**, Schematic of EBV genes. Repetitive regions in EBV genomes are masked by light blue.

Table 1 | The association of three non-synonymous SNPs in the *BALF2* gene and the risk for NPC

SNP	High-risk genotype	Discovery			Validation			Combined			OR	95% CI	*P* value conditional on SNPs		Annotation
		156 cases	47 controls	*P* value	483 cases	605 controls	*P* value	639 cases	652 controls	*P* value			163364	162476	
162215C>A	C	96.15%	65.96%	3.22×10^{-04}	95.03%	74.71%	9.92×10^{-16}	95.31%	74.08%	1.42×10^{-18}	7.60	4.97–11.62	7.78×10^{-05}	1.94×10^{-01}	*BALF2*, V700L
162476T>C	C	93.59%	61.70%	5.09×10^{-03}	94.00%	65.12%	1.94×10^{-23}	93.90%	64.88%	9.69×10^{-25}	8.69	5.79–13.03	1.10×10^{-06}	NA	*BALF2*, I613V
163364C>T	T	88.46%	48.94%	7.95×10^{-03}	83.85%	45.45%	6.92×10^{-32}	84.98%	45.71%	2.40×10^{-32}	6.14	4.59–8.22	NA	4.84×10^{-11}	*BALF2*, V317M

The association of three EBV SNPs with NPC risk was tested in discovery and validation samples, and with a meta-analysis of the discovery and validation samples combined. Frequencies of high-risk genotypes in discovery, validation and combined analyses are indicated. ORs conferred by high-risk genotypes and the 95% CIs were estimated from the meta-analysis of the combined discovery and validation phases. Conditional regression analyses were performed in combined samples, and *P* values of SNP associations in conditional analyses are listed. 'NA' represents the conditional SNPs.

subclade, whereas the carriers of the risk variant of SNP 162215_C had a much broader distribution (Fig. 1b,e). Within the Asian subclades, the carriers of SNPs 162476_C and 163364_T were enriched in the strains from patients with NPC (NPC-dominant strains). These results provided strong evidence for the Asian origin of SNPs 162476_C and 163364_T, and were consistent with their high-risk effect on NPC. The distribution of these genotypes also suggested that SNP 162215_C was less likely to be a risk variant for NPC than 162476T>C and 163364C>T, and its association effect may be due to its LD with SNP 162476_C (LD $R^2=0.67$).

Discussion

Owing to the ubiquity of EBV infection, the determinants of the distinctive geographical distribution of NPC have long puzzled the scientific community. Using large-scale sequencing and functional analyses, we discovered two EBV coding SNPs, 162476_C and 163364_T, that are the strongest known risk factors for NPC. The more than sixfold increase in NPC risk conferred by these two high-risk EBV variants is far greater than the effects of any other known risk factors for this disease, including human genetic variants (Table 1 and Supplementary Table 10). In particular, with a population frequency of 45% and an OR of 11.71 (95% CI, 7.44–19.26%), the EBV haplotype C-T of the two SNPs is the dominant NPC risk factor, contributing 71% (95% CI, 64–77%) of the overall risk of NPC in the endemic population of southern China. The second risk haplotype, C-C, also contributed approximately 10% of the risk, such that the two high-risk EBV haplotypes jointly accounted for 83% (95% CI, 76–90%) of NPC risk in this population (Supplementary Table 19). In non-endemic regions of China, the frequency of these high-risk haplotypes is much lower (approximately 10%), but they

Table 2 | EBV haplotypes composed of SNPs 162215C>A, 162476T>C and 163364C>T, and the risk for NPC

EBV subtype (162215-162476-163364)	639 cases		652 controls		Odds ratio[a]	95% CI	*P* value
	no.	%	no.	%			
L-L-L (A-T-C)	25	3.91%	171	26.23%	Reference	Reference	
H-H-H (C-C-T)	539	84.35%	293	44.94%	11.71	7.44–19.26	2.39×10^{-24}
H-H-L (C-C-C)	57	8.92%	118	18.10%	3.50	2.02–6.24	1.22×10^{-05}
H-L-L (C-T-C)	13	2.03%	65	9.97%	1.12	0.47–2.50	7.83×10^{-01}
Other subtypes	5	0.78%	5	0.77%	4.26	0.80–19.63	6.71×10^{-02}

[a]ORs of individual EBV subtypes and 95% CIs were estimated with a logistic model by categorizing each subtype as a single variable and adjusting for age, sex, the status of single- or multiple-infection, and human GWAS SNPs (rs2860580 and rs2894207) in the combined discovery and validation data sets. Subjects with EBV subtype A-T-C, a common low-risk subtype, were used as the reference category. H represents the high-risk genotype; L represents the low-risk genotype.

still contribute about 50% of the NPC risk driven by the strong risk effect. The frequency of the two high-risk EBV subtypes was not associated with the risk of developing other EBV-related cancers in our study, which suggests that their oncogenic effects might be specific to NPC. However, this observation would benefit from further work, as our study was only powered to explore NPC.

Mapping these two causal variants onto the phylogenetic tree of EBV genomes revealed a distinct subclade of EBV subtypes carrying the two high-risk variants in Asia. The carriers were found only in Asia, thereby indicating an Asian origin for these two risk variants. Most interestingly, the phylogenetic analysis suggests a clonal expansion of these unique high-risk EBV subtypes in southern China. This expansion is consistent with the current distribution of these subtypes in China, with a very high frequency in the NPC-endemic region (93.27% in NPC cases and 63.04% in controls), but much lower in the non-endemic areas (55% in NPC cases and 9.68% in other non-NPC samples; Supplementary Table 17). At this point, we do not know what kind of selective phenotypes have driven the clonal expansion. More studies are needed to understand this evolutionary process. Taken together, the strong risk effect, the confined geographical distribution, the clonal expansion and the extremely enriched frequency of these high-risk variants in the NPC-endemic region strongly suggest that these two EBV risk variants are the driving factors of the unique epidemic of NPC in southern China.

Our findings provide novel biological insights into EBV-mediated NPC tumorigenesis. The two risk variants 162476_C and 163364_T encode amino acid alterations in *BALF2*, the EBV single-stranded DNA binding protein, which is an abundantly expressed early lytic protein and a core component of viral DNA replication machinery[23–25]. Studies have shown that antibodies against EBV early lytic antigens, including *BALF2*, were highly enriched in the antibody signature for NPC risk prediction[28,29], and *BALF2* is also a frequent target of the EBV-induced cytotoxic T cell response[30]. Because of the essential role of *BALF2* in EBV lytic DNA replication, these amino acid changes may influence the productive lytic cycle of EBV by alternating the function of *BALF2*. This is consistent with our observation and that of others[31] that the oral EBV abundance is lower in the NPC cases than the controls. In addition, we also observed a trend for the oral EBV DNA load to decrease with the EBV subtype carrying the high-risk *BALF2* haplotype, although this association is only marginally significant with a huge variation in saliva viral load among individuals. As demonstrated previously[32], this large variation of viral load in saliva is mainly due to the fact that EBV in buccal epithelium sporadically undergoes a periodic lytic cycle with a large variation in viral load at different time points for the same individual. Given the moderate impact of the *BALF2* haplotypes on the overall variation of viral load, a much larger number of samples will help to confirm the statistical difference of viral load among the carriers of EBV with different *BALF2* haplotypes.

Taken together, our results and those of others suggest that the regulation of the EBV lytic cycle has an important role in the development of NPC. More molecular and functional investigations are needed to test this hypothesis and to understand how the high-risk EBV subtypes and variants promote NPC tumorigenesis.

The discovery of these high-risk EBV variants also has important implications for public health efforts to reduce the burden of NPC, particularly in the endemic region of southern China. Testing for these high-risk EBV variants enables the identification of high-risk individuals for targeted implementation of routine clinical monitoring to detect NPC early. Primary prevention by developing vaccines against high-NPC-risk EBV strains is expected to lead to great attenuation of the Cantonese cancer in China.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0436-5.

Received: 12 August 2018; Accepted: 7 May 2019;
Published online: 17 June 2019

References

1. Epstein, M. A., Achong, B. G. & Barr, Y. M. Virus particles in cultured lymphoblasts from Burkitt's lymphoma. *Lancet* **1**, 702–703 (1964).
2. Epstein, A. Why and how Epstein-Barr virus was discovered 50 years ago. *Curr. Top. Microbiol Immunol.* **390**, 3–15 (2015).
3. Kieff, E. D. & Rickinson, A. B. in *Fields' Virology* 5th edn, Vol. 2 (eds Knipe, D. M. & Howley, P. M.) Ch. 68A, 2603–2654 (Lippincott Williams & Wilkins, Wolters Kluwer, 2007).
4. Zhang, L. F. et al. Incidence trend of nasopharyngeal carcinoma from 1987 to 2011 in Sihui county, Guangdong province, south China: an age-period-cohort analysis. *Chin. J. Cancer* **34**, 350–357 (2015).
5. Bei, J. X. et al. A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. *Nat. Genet* **42**, 599–603 (2010).
6. Bei, J. X. et al. A GWAS Meta-analysis and replication study identifies a novel locus within *CLPTM1L/TERT* associated with nasopharyngeal carcinoma in individuals of Chinese ancestry. *Cancer Epidemiol. Biomark. Prev.* **25**, 188–192 (2016).
7. Cui, Q. et al. An extended genome-wide association study identifies novel susceptibility loci for nasopharyngeal carcinoma. *Hum. Mol. Genet.* **25**, 3626–3634 (2016).
8. Tang, M. et al. The principal genetic determinants for nasopharyngeal carcinoma in China involve the HLA class I antigen recognition groove. *PLoS Genet.* **8**, e1003103 (2012).
9. Baer, R. et al. DNA sequence and expression of the B95-8 Epstein–Barr virus genome. *Nature* **310**, 207–211 (1984).
10. Zeng, M. S. et al. Genomic sequence analysis of Epstein-Barr virus strain GD1 from a nasopharyngeal carcinoma patient. *J. Virol.* **79**, 15323–15330 (2005).
11. Dolan, A., Addison, C., Gatherer, D., Davison, A. J. & McGeoch, D. J. The genome of Epstein–Barr virus type 2 strain AG876. *Virology* **350**, 164–170 (2006).

12. Liu, P. et al. Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. *J. Virol.* **85**, 11291–11299 (2011).
13. Lin, Z. et al. Whole-genome sequencing of the Akata and Mutu Epstein-Barr virus strains. *J. Virol.* **87**, 1172–1182 (2013).
14. Palser, A. L. et al. Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. *J. Virol.* **89**, 5222–5237 (2015).
15. Correia, S. et al. Natural Variation of Epstein-Barr Virus Genes, Proteins, and Primary MicroRNA. *J. Virol.* **91**, e00375-17 (2017).
16. Kwok, H. et al. Genomic diversity of Epstein-Barr virus genomes isolated from primary nasopharyngeal carcinoma biopsy samples. *J. Virol.* **88**, 10662–10672 (2014).
17. Edwards, R. H., Seillier-Moiseiwitsch, F. & Raab-Traub, N. Signature amino acid changes in latent membrane protein 1 distinguish Epstein–Barr virus strains. *Virology* **261**, 79–95 (1999).
18. Hui, K. F. et al. High risk Epstein–Barr virus variants characterized by distinct polymorphisms in the *EBER* locus are strongly associated with nasopharyngeal carcinoma. *Int. J. Cancer* **144**, 3031–3042 (2018).
19. Tso, K. K. et al. Complete genomic sequence of Epstein–Barr virus in nasopharyngeal carcinoma cell line C666-1. *Infect. Agent Cancer* **8**, 29 (2013).
20. Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
21. Chen, H. et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.* **98**, 653–666 (2016).
22. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780–1815 (2011).
23. Decaussin, G., Leclerc, V. & Ooka, T. The lytic cycle of Epstein–Barr virus in the nonproducer Raji line can be rescued by the expression of a 135-kilodalton protein encoded by the *BALF2* open reading frame. *J. Virol.* **69**, 7309–7314 (1995).
24. Zeng, Y., Middeldorp, J., Madjar, J. J. & Ooka, T. A major DNA binding protein encoded by *BALF2* open reading frame of Epstein–Barr virus (EBV) forms a complex with other EBV DNA-binding proteins: DNAase, EA-D, and DNA polymerase. *Virology* **239**, 285–295 (1997).
25. Mumtsidu, E. et al. Structural features of the single-stranded DNA-binding protein of Epstein–Barr virus. *J. Struct. Biol.* **161**, 172–187 (2008).
26. Rowe, M. et al. 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.* **63**, 1031–1039 (1989).
27. Li, D. J. et al. The dominance of China 1 in the spectrum of Epstein–Barr virus strains from Cantonese patients with nasopharyngeal carcinoma. *J. Med. Virol.* **81**, 1253–1260 (2009).
28. Coghill, A. E. et al. Identification of a Novel, EBV-based antibody risk stratification signature for early detection of nasopharyngeal carcinoma in Taiwan. *Clin. Cancer Res.* **24**, 1305–1314 (2018).
29. Paramita, D. K. et al. Native early antigen of Epstein–Barr virus, a promising antigen for diagnosis of nasopharyngeal carcinoma. *J. Med. Virol.* **79**, 1710–1721 (2007).
30. Steven, N. M. et al. Immediate early and early lytic cycle proteins are frequent targets of the Epstein–Barr virus-induced cytotoxic T cell response. *J. Exp. Med.* **185**, 1605–1617 (1997).
31. Xue, W. Q. et al. Decreased oral Epstein-Barr virus DNA loads in patients with nasopharyngeal carcinoma in Southern China: a case-control and a family-based study. *Cancer Med.* **7**, 3453–3464 (2018).
32. Hadinoto, V., Shapiro, M., Sun, C. C. & Thorley-Lawson, D. A. The dynamics of EBV shedding implicate a central role for epithelial cells in amplifying viral output. *PLoS Pathog.* **5**, e1000496 (2009).

Acknowledgements

We thank all of the participants for their generous support of the current study. We would also thank R. Sun, C. Wang, H. Chen, J. Shen and C. Jie for helpful discussions on viral biology and genetic statistical, evolutionary and phylogenetic analyses, W.-S. Liu and X. Zuo for providing code support, Z. Lin (Tulane University) for kindly sharing EBV genome annotation files and J.-Y. Shao from Sun Yat-sen University Cancer Center for providing the MassArray iPlex platform. This work was supported by the National Natural Science Foundation of China (81430059 to Y.-X.Z. and 81872228 to M.X.), the National Key R&D Program of China (2016YF0902000 to Y.-X.Z., and 2018YFC1406902 and 2018YFC0910400 to W.Z.), the National Cancer Institute at the US National Institutes of Health (NIH) (R01CA115873-01 to H.-O.A. and Y.-X.Z., and R35-CA197449, P01-CA134294, U01-HG009088 and U19-CA203654 to X.L.) and the Agency of Science, Technology and Research (A*STAR), Singapore (to J.L.).

Author contributions

Y.-X.Z., J.L. and W.Z. were the principal investigators who conceived the study. Y.-X.Z., J.L., W.Z. and M.X. designed and oversaw the study. J.L. and X.L. supervised the viral genome-wide association studies. W.W. supervised phylogenetic analysis. M.X. contributed to sample preparation, sequencing, genotyping, variant calling and genetic statistical analyses. Y.Y. contributed to sequencing, genotyping and variant calling. H.C. contributed to phylogenetic analyses. S.Z. contributed to genotyping and genetic statistical analyses. Z.Li contributed to genetic statistical analyses. Z.Z. contributed to collection of samples from the First Affiliated Hospital of Guangxi Medical College. B.L. contributed to collection of samples from the Affiliated Hospital of the Qingdao University. X.G., M.-Y.C., R.P. and R.-H.X. contributed to collection of samples from Sun Yat-sen University Cancer Center. H.-O.A., W.Y. and Y.-X.Z. supervised the design and implementation of the population-based case–control study in Zhaoqing. W.Y., E.T.C., S.-M.C., S.-H.X. and Z.Liu participated in the case–control study. The manuscript was drafted by M.X., J.L., W.Z. and Y.-X.Z., and revised by V.P. and E.T.C. All authors critically reviewed the article and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0436-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to W.Z., Y.-X.Z. or J.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Study participants and samples. Participants in the current study were enrolled through two rounds of recruitment. The first was a hospital-based study, which enrolled patients with EBV-related cancers, including NPC, Burkitt lymphoma, Hodgkin lymphoma, natural killer (NK) or T cell lymphoma, and gastric carcinoma, as well as healthy controls from the Sun Yat-sen University Cancer Center in Guangdong province, the First Affiliated Hospital of Guangxi Medical College in Guangxi province, and the Affiliated Hospital of the Qingdao University in Shandong province, China. The geographical origin of the participants covers the NPC-endemic area of southern China (Guangdong and Guangxi provinces), where NPC has the highest incidence (20–40 cases per 100,000 individuals per year), and non-endemic regions in China, where NPC is rare. After measuring the EBV DNA level, we selected 170 samples of tumor, saliva and plasma with a real-time qPCR threshold cycle (Ct) value < 30 from the first round of recruitment for EBV WGS.

The second round of recruitment was a population-based NPC case–control study that enrolled patients with NPC and healthy control subjects from Zhaoqing county, Guangdong province, China (an NPC-endemic region). Cases and controls were matched by age and sex. Saliva samples were collected from all of the subjects. After measuring saliva EBV DNA load in the second study, we selected 99 saliva DNA samples with a Ct value < 30 from 53 cases and 46 controls for EBV WGS (details can be found in the Supplementary Note). Written informed consent was obtained from each participant before any study-related procedures were undertaken, and both studies were approved by the institutional ethics committee of Sun Yat-sen University Cancer Center.

Detailed sample information, including the geographical origins of the 270 isolates used for WGS, is summarized in Supplementary Tables 1–4. For the discovery phase of the EBV GWAS with NPC, we included 156 cases and 47 controls exclusively from the NPC-endemic region from the 270 EBV WGS isolates. For the validation phase, 990 NPC cases and 1,105 healthy controls from the endemic population-based case–control study were used by genotyping GWAS candidate SNPs (for details, see Supplementary Note).

Sample processing. Saliva samples were collected into vials containing lysis buffer (50 mM Tris, pH 8.0, 50 mM EDTA, 50 mM sucrose, 100 mM NaCl, 1% SDS). Tumor specimens were obtained from biopsy samples collected during surgical treatment and confirmed by histopathological examination. All saliva, tumor and plasma specimens were stored at −80 °C. DNA was extracted from the saliva using the Chemagic STAR workstation (Hamilton Robotics), and from the tumor biopsy, plasma and NPC cell line C666-1 using a DNeasy Blood and Tissue Kit (Qiagen).

EBV genome quantification, whole-genome sequencing and variant calling. Using real-time PCR targeting of a DNA fragment at the *BALF5* gene (5′ and 3′ primers, GGTCACAATCTCCACGCTGA and CAACGAGGCTGACCTGATCC), we measured the EBV DNA concentration in each DNA sample with a quantitative PCR (qPCR) standard curve. Samples with EBV DNA concentrations of higher than 2,500 copies per microliter were selected for viral WGS (for detailed information see the Supplementary Note).

The EBV genomes were captured using the MyGenostics GenCap Target Enrichment Protocol (GenCap Enrichment, MyGenostics). After capture enrichment, DNA libraries were prepared and sequenced using the Illumina HiSeq 2000 platform according to standard protocols (Illumina). After raw sequence processing and quality control, paired-end reads were aligned to the EBV B95-8 reference genome (NC_007605.1) using the Burrows–Wheeler Aligner (BWA, version 0.7.5a)[33,34]. The average sequencing depth was 1,282 (range, 32 to 6,629). High genome coverage (average, 98.02%; range, 94.44% to 99.91%) was achieved (Supplementary Fig. 2).

Following the GATK best practice workflows (version 3.2-2), an initial set of 8,469 variants was first called after base and variant recalibration[35]. To avoid inaccurate calling, we further filtered out variants that had low coverage (depth <10×) or were in repetitive elements or within 5 bp of an indel; 7,962 variants were retained for subsequent EBV phylogenetic, principal component and association analyses. The functional annotation of the EBV variants was performed using the SNPEff package according to the reference genome (NC_007605.1, NCBI annotation, November 2013)[36]. A complete description of the sequencing and variant calling is presented in the Supplementary Note. No outlier was detected among the EBV isolates sequenced based on sequencing and variant statistics in the current study (Supplementary Fig. 2).

To evaluate the accuracy of our sequencing and variant calling, subsets of EBV variants were validated using either the Sanger sequencing or MassAarray iPLEX assay (Agena Bioscience). Two independent technologies can provide orthogonal evaluations of the sequencing accuracy. We amplified 299 PCR fragments from 53 randomly selected EBV isolates and re-sequenced them using Sanger sequencing. The SNPs called by WGS and by Sanger sequencing were 97.55% concordant (Supplementary Table 7). Similarly, the variants called by WGS and by the MassArray iPLEX assay were 99.99% concordant when genotyping 37 variants in 239 samples (Supplementary Table 8). In addition, when comparing the re-sequenced C666-1 EBV genome against the publicly available sequence[19], the concordance was 97.93% (Supplementary Table 6).

To understand viral genomes from multiple sample types from the same patient, two EBV fragments (position 80,089 to 80,875 and position 81,092 to 81,829) containing 89 SNPs were resequenced using the Sanger method from paired saliva and tumor samples from the same set of patients. Across 25 NPC patients with paired tumor and saliva samples, the pairwise difference (defined as the genotype discordance rate at the 89 SNPs) between the tumor samples of the 25 patients (inter-host difference) as well as between the paired tumor and saliva samples of the same patient (intra-host difference) were calculated and compared (Supplementary Fig. 4). The median inter-patient difference was 13.5% (first to third quartile, 3.7–16.9%), and the median intra-host difference was only 1.1% (first to third quartile, 0–3.4%). The high concordance between variants from saliva and from tumors suggests that EBV sequences from paired saliva and tumor samples from the same patient are highly similar.

Genotyping analysis of EBV and human genetic variants by MassArray iPLEX. To genotype the EBV variants in the 990 cases and 1,105 controls from Zhaoqing, the customized primers and the protocol recommended by the Agena Bioscience MassArray iPLEX platform were used. A fixed position in the human albumin gene was used as a positive control. Because the genotyping success rate strongly correlates with the EBV DNA abundance (Supplementary Fig. 15), approximately half of the validation samples (483 of the cases and 605 of the controls) could be successfully genotyped for all three GWAS candidate markers (that is, SNPs 162215C>A, 162476T>C and 163364C>T). The slightly lower success rate in the cases is consistent with the fact that the EBV DNA abundance was lower in the saliva from patients than from controls. For detailed information, see Supplementary Note.

Seven previously reported human SNPs in *HLA* (rs2860580, rs2894207 and rs28421666), *CDKN2A* and *CDKN2B* (rs1412829), *TNFRSF19* (rs9510787), *TERT* (rs31489) and *MECOM* (rs6774494) we re genotyped using customized primers and following the protocol recommended by the Agena Bioscience MassArray iPLEX platform in the 990 cases and 1,105 controls from Zhaoqing. A fixed position in the human albumin gene was used as a positive control. The genotyping completion rate for all seven human SNPs was >95%. Associations with NPC were assessed with logistic regression under an additive model adjusted for sex and age.

Determining single versus multiple EBV infections. The EBV genome usually undergoes clonal expansion in NPC tumors and other malignancies[37–39]. During clonal expansion, the EBV genome is stable, the intra-host mutation rate is often low, and heterozygous variants, as a result of quasi-species evolution within a host, are not frequent[12,19,40]. On the contrary, EBV isolates from specimens with multiple infections will have a higher number of heterozygous variants. We plotted the percentage of heterozygous variants across all of the 270 samples from the WGS analysis and observed that heterozygosity (defined as a percentage of heterozygous variants) across all of the samples showed two different distributions, with low and high numbers of heterozygous variants. By fitting two curves to the lower and higher quantiles of the empirical distribution, we defined the reflection point (that is, the intersection of the two distributions) as the cutoff value (Supplementary Fig. 13). Samples with the proportion of heterozygous variants lower than the cutoff value were identified as single-infection samples, whereas samples above this threshold were identified as multi-infection samples. For the validation cohort, samples with the homozygous calls at all three EBV SNPs were regarded as a single EBV subtype defined by *BALF2* haplotypes. For samples with infection by multiple EBV subtypes, haplotypes of the three SNPs were inferred by Beagle 4.1 (ref. [41]). For details, see Supplementary Note.

Phylogenetic and principal component analyses of EBV genome sequences. The phylogenetic analysis and PCA were performed using EBV isolates sequenced by the current study, and publicly accessible EBV genomes. For the phylogenetic analysis, we first created the fasta sequence for each resequenced isolate using the variant data extracted from variant calling. The 230 EBV single-infection whole genomes were subsequently combined with the 97 public genomes and multiple sequence alignment was carried out using the multiple alignment program MAFFT (version 7)[42]. After masking the regions of repetitive sequences and poor coverage in resequencing, the maximum likelihood of the phylogenetic relationship was inferred using Randomized Axelerated Maximum Likelihood (RAxML version 8), assuming a general time reversible (GTR) model[43]. The inferred phylogeny was subsequently rooted using the Evolutionary Placement Algorithm (EPA) algorithm[44] from RAxML using a Macacine herpesvirus 4 genome sequence (NC_006146) as the outgroup.

In the PCA, genomic variation from the 97 public genomes was generated by global pairwise sequence alignment of published genome sequences against the B95-8 reference genome (NC_007605.1) using EMBOSS Stretcher[45]. The variant set was then combined with the variation data extracted from WGS. A combined set of 12,182 SNPs from the 270 newly sequenced isolates and 97 published ones were then used for PCA. During PCA, SNPs were first filtered by allele frequency (minor genotype frequency of >0.05) and LD (pruning with a pairwise correlation R^2 value > 0.6 within a 1,000-bp sliding window). In total, 495 SNPs were included in the PCA using the R package SNPRelate (version 1.10.2)[46].

Principal component analysis of cases and controls. To assess the human population structure of the 156 cases and 47 healthy controls used for the EBV GWAS discovery phase, the human DNA of these samples wasgenotyped using OmniZhongHua-8 Chip (Illumina). Sample filtering was carried out using the following criteria: (i) call rate of >95%; (ii) SNP filtering by minor allele frequency of >5%; (iii) Hardy–Weinberg equilibrium ($P > 1 \times 10^{-6}$); and (iv) LD-based SNP pruning ($R^2 < 0.1$ and not within the five high-LD regions[5]). PCA was then performed using PLINK (version 1.9) based on the discovery samples alone or by combining them with reference samples from the 1000 Genomes project[20].

Association analysis. Genetic associations of EBV variants were analyzed by testing either single or multiple variants. Single-variant association analysis used a generalized linear mixed model with EBV genetic relatedness matrix as random effects[21]. Sex and age were included as fixed effects, as well as four human principal components and previously reported human NPC GWAS loci (rs2860580 and rs2894207) at the *HLA* locus to correct for any potential impact of human population structures and genetics on the association results. Both single- and multiple-infection samples were included in the association analysis with the status of single- or multiple-infection being a covariate to correct for any potential confounding effect of multiple infections. The genome-wide discovery analysis was performed by testing 1,545 EBV variants (with missing rate of <10%, minor genotype frequency of >0.05 and heterozygosity of <0.1) in 156 cases and 47 healthy controls. The validation analysis was performed by testing three EBV non-synonymous coding SNPs 162215C>A, 162476T>C and 163364C>T in *BALF2* in an additional 483 cases and 605 population controls matched to the cases by age and sex from the case–control study in Zhaoqing county. The logistic regression model was used for validation, adjusting for age, sex, the human SNPs (rs2860580 and rs2894207 in the *HLA* locus) and the status of single- or multiple-infection of EBV. The meta-analysis of the discovery and validation phases was performed with the *z*-score pooling method. Considering the extensive LD across the EBV genome, to obtain a suggestive genome-wide significance of association, we used permutations of a logistic model adjusting for age, sex, status of single- or multiple-infection, and the human and EBV population structures. The genome-wide significance (4.07×10^{-4}) was determined with a 5% quantile of the empirical distribution of minimum *P* values from 10,000 permutations as the data-driven threshold to control for family-wise error rate under multiple correlated tests.

The genome-wide multi-variant-based association analysis was performed by testing 1,477 bi-allelic EBV variants using Bayesian variable selection regression implemented in piMASS (version 0.90)[22]. Age, sex, four human prinicipal components, two EBV principal components and the human SNPs (rs2860580 and rs2894207) were included as covariates. The analysis was performed by partitioning the EBV genome into the regions of a 20-SNP sliding window with 10 overlapping SNPs. The sum of the posterior probabilities of the SNPs being associated within a window was calculated as the 'region statistic' indicating the strength of the evidence for genetic associations in that region.

To further prioritize potentially causal SNPs in the top hit *BALF2* gene region for validation, we applied further fine-mapping analysis using Bayesian multiple-variable selection by PAINTOR3.1 (ref. [47]). Functional annotation of SNPs was used as the prior probability to compute the probability of being causal for each variant in the region. We assumed a single causal variant in *BALF2* genes and calculated a 95% credible set that contains the minimum set of variants that jointly have at least a 95% probability of including the causal variant.

We also evaluated the association of seven previously reported human GWAS SNPs with NPC in our combined samples of 639 cases and 652 controls. Of the seven SNPs, two within the *HLA* locus, rs2860580 and rs2894207, showed significant associations with consistent ORs after correction for multiple testing (Supplementary Table 10). For the remaining SNPs in the *HLA* (rs28421666), *CDKN2A* and *CDKN2B* (rs1412829), *TERT* (rs31489), *TNFRSF19* (rs9510787) and *MECOM* (rs6774494) loci, the ORs in our samples were consistent with the values reported previously, although the results were not statistically significant after correction for multiple testing (Supplementary Table 10). Therefore, in the association analyses of EBV variants, we included the two significant human GWAS SNPs (rs2860580 and rs2894207) as covariates. The results with the two human SNPs are very similar to the results without them as covariates (Supplementary Table 20). These findings clearly indicate that the reported human GWAS loci do not affect our association evidences for the EBV risk variants. A Life Sciences Reporting Summary for this paper is available.

Estimation of the population attributable fraction of risk. The proportion of NPC risk explained by the effect of the two high-risk haplotypes (*X*) of SNPs 162476T>C and 163364C>T (C-T and C-C) was estimated in the validation sample. The attributable fraction of risk and 95% CI were estimated in a logistic regression model adjusting for confounders (*Z*), age and sex, with the R package AF (version 0.1.4) (ref. [48]). Because NPC is not a common disease (prevalence of fewer than 40 cases per 100,000 individuals per year), the risk ratio can be approximated by OR. Thus, the population attributable fraction (AF) of NPC risk (probability of $Y=1$) can be approximated by $AF \approx 1 - E_Z\{OR^{-X}(Z)|Y=1\}$

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The EBV sequencing data are deposited in the US National Center for Biotechnology Information (NCBI) database under BioProject ID PRJNA522388. EBV sequences are released in NCBI database under GenBank IDs MK540241–MK540470.

References

33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
34. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
35. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
36. Cingolani, P. 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* **6**, 80–92 (2012).
37. Raab-Traub, N. & Flynn, K. The structure of the termini of the Epstein–Barr virus as a marker of clonal cellular proliferation. *Cell* **47**, 883–889 (1986).
38. Pathmanathan, R., Prasad, U., Sadler, R., Flynn, K. & Raab-Traub, N. Clonal proliferations of cells infected with Epstein–Barr virus in preinvasive lesions related to nasopharyngeal carcinoma. *N. Engl. J. Med.* **333**, 693–698 (1995).
39. Neri, A. et al. Epstein–Barr virus infection precedes clonal expansion in Burkitt's and acquired immunodeficiency syndrome-associated lymphoma. *Blood* **77**, 1092–1095 (1991).
40. Weiss, E. R. et al. Early Epstein-Barr virus genomic diversity and convergence toward the B95.8 Genome in primary infection. *J. Virol.* **92**, e01466-17 (2018).
41. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
42. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
43. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
44. Berger, S. A., Krompass, D. & Stamatakis, A. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.* **60**, 291–302 (2011).
45. Li, W. et al. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* **43**, W580–W584 (2015).
46. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
47. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
48. Dahlqwist, E., Zetterqvist, J., Pawitan, Y. & Sjolander, A. Model-based estimation of the attributable fraction for cross-sectional, case-control and cohort studies using the R package AF. *Eur. J. Epidemiol.* **31**, 575–582 (2016).

nature research

Corresponding author(s): Jianjun Liu, Yi-Xin Zeng, Weiwei Zhai

Last updated by author(s): Apr 4, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	NA
Data analysis	All softwares and packages used in current study have been stated and cited in the Methods section.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The EBV sequencing data are deposited in NCBI database under BioProject ID PRJNA522388. EBV sequences are released in NCBI database under GenBank ID MK540241-MK540470.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was largely determined by sample and data availability. The genome-wide association discovery analysis was performed using all the tissue and saliva samples from independent individuals in NPC-endemic southern China collected by the hospital-based recruitment and saliva samples with sufficient EBV DNA abundance (realtime PCR Ct value < 30) collected by the population-based case-control study. With 156 cases and 47 controls, the genome-wide discovery analysis has at least 90% power for detecting strong risk variants with a MAF of 30% and odds ratio of 6 and 80% power for detecting variants with a MAF of 30% and odds ratio of 5 at the suggestive significant level of 0.0004 (determined by permutation). The validation analysis was performed with independent samples from the NPC cases and population controls. With 483 cases and 605 controls (more than tripling the size of the discovery analysis), the validation analysis has sufficient power to overcome any potential "winner's curse" effect and replicate the finding(s) from the genome-wide discovery analysis. Giving that the EBV risk variants identified in this study are common with large effect size (odds ratio > 6), both the discovery and validation analyses of the current study have sufficient power to discover such associations.
Data exclusions	In term of data exclusion, the EBV variants were filtered with standard quality control metrics in the filed that only the 1,545 EBV variants with missing rate < 10%, minor genotype frequency > 0.05 and heterozygosity < 0.1 were included in the discovery phase of genome-wide association analysis.
Replication	Technical replication of variant calling was performed by using Sanger sequencing and MassArray, evaluating the quality of variant callings through orthogonal technologies. Biological replication of the findings from the WGS based genome-wide association analysis was performed by the analysis of independent samples using different technology (MassArray).
Randomization	This is a case-control study. Randomization is not applicable.
Blinding	All sequencing and genotyping analyses were performed by researchers who were blinded to disease status. And, the investigators involved in the data analysis and result interpretation were not involved in sample recruitments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	NPC cases and healthy controls for association analyses were matched with regard to age, sex and ethnicity. Baseline characteristic are shown in the Supplementary Tables 1, 3, 4 and 13.
Recruitment	Participants of the current study were enrolled through two recruitments. The first one was a hospital cohort, enrolling patients diagnosed with EBV-related cancers (including NPC, Burkitt lymphoma, Hodgkin lymphoma, NK/T cell lymphoma and gastric carcinomas) and healthy controls from three hospitals: the Sun Yat-sen university Cancer Center in Guangdong Province, the First Affiliated Hospital of Guangxi Medical College in Guangxi Province, and the Affiliated Hospital of the Qingdao University in Shandong Province of China. The second recruitment was a population cohort, enrolling NPC cases and population control subjects from Zhaoqing County, Guangdong Province of China (NPC-endemic region). For details of recruitment, see Methods and Supplementary Note.
Ethics oversight	Sun Yat-sen University Cancer Center

Note that full information on the approval of the study protocol must also be provided in the manuscript.