High risk Epstein-Barr virus variants characterized by distinct polymorphisms in the EBER locus are strongly associated with nasopharyngeal carcinoma

Short Title: High risk EBV variants of NPC

Authors: *Kwai Fung Hui* [1¶], *Tsz Fung Chan* [1¶], *Wanling Yang* [1,2], *Jiangshan Jane Shen* [1], *Ki Pui Lam* [1], *Hin Kwok* [3], *Pak C Sham* [2,3], *Sai Wah Tsao* [2,4], *Dora L Kwong* [2,5], *Maria Li Lung* [2,5] *and Alan Kwok Shing Chiang* [1, 2]

Author's Affiliations: [1]Department of Paediatrics and Adolescent Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Queen Mary Hospital, Pokfulam, Hong Kong SAR, China. [2]Center for Nasopharyngeal Carcinoma Research, The University of Hong Kong, Hong Kong SAR, China. [3]Centre for Genomic Sciences, The Hong Kong Jockey Club Building for Interdisciplinary Research, The University of Hong Kong, Hong Kong, China. [4]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [5]Department of Clinical Oncology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China.

[¶]These authors contributed equally to this work.

Corresponding author: Alan KS Chiang, Department of Paediatrics and Adolescent Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Queen Mary Hospital, Pokfulam, Hong Kong SAR, China. Tel: 852-22554091; Fax: 852-28551523; Email: chiangak@hku.hk

Keywords: Epstein-Barr virus, Nasopharyngeal carcinoma, Genome-wide analysis, Sequencing, EBV-encoded small RNA, NPC-EBERvar

Abbreviation: EBV – Epstein-Barr virus; NPC – Nasopharyngeal Carcinoma; GWAS – Genome-wide association study; SNP – single nucleotide polymorphisms; PCA – Principal component analysis; LMM – Linear mixed model; FDR – False discovery rate; GRS – Genetic risk score; NPC-EBERvar – high risk EBV variants with polymorphisms in the EBER locus; EBER-del – a four-base-deletion polymorphism downstream of EBER2

Article Category: Infectious Causes of Cancer

Novelty and Impact: This study reveals, for the first time, the presence of high risk EBV variants of NPC at the level of both loci and lineages. Further characterization of specific pathogenic mutations in the high risk EBV variants and survey of viral variants amongst different geographic populations will be important goals of future research. Analysis of EBV variants harboured in heathy individuals could be potentially developed as a novel screening method to assess the risk of NPC.

Abstract

Whether certain variants of Epstein-Barr virus (EBV) are linked to the pathogenesis of nasopharyngeal carcinoma (NPC), which shows a marked geographic restriction, remains an unresolved issue. We performed a case-control study comparing genomic sequences of EBV isolated from saliva samples of 142 population carriers with those from primary tumour biopsies derived from 62 patients with NPC of Hong Kong. Cluster analysis discovered five EBV subgroups 1A-C and 2A-B amongst the population carriers in contrast to the predominance of 1A and -B in the majority of NPC. Genome-wide association study (GWAS) identified a panel of NPC-associated single nucleotide polymorphisms (SNPs) and indels in the EBER locus. The most significant polymorphism, which can be found in 96.8% NPC cases and 40.1% population carriers of Hong Kong, is a four-base-deletion polymorphism downstream of EBER2 (EBER-del) from coordinates 7188-7191 ($p = 1.91\times10^{-7}$). In addition, the predicted secondary structure of EBER2 is altered with likely functional consequence in nearly all NPC cases. Using the SNPs and indels associated with NPC, genetic risk score is assigned for each EBV variant. EBV variants with high genetic risk score are found to be much more prevalent in Hong Kong Chinese than individuals of other geographic regions and in NPC than other EBV-associated cancers. We conclude that high risk EBV variants with polymorphisms in the EBER locus, designated as NPC-EBERvar, are strongly associated with NPC. Further investigation of the biological function and potential clinical application of these newly identified polymorphisms in NPC and other EBV-associated cancers is warranted.

Introduction

Epstein-Barr virus (EBV) is a human gammaherpesvirus that infects the majority of the world's population and is strongly associated with Burkitt's lymphoma and epithelial malignancies such as nasopharyngeal carcinoma (NPC) and a subset of gastric carcinoma.[1] EBV is a double-stranded DNA virus whose genome is of approximately 170 kb in size and contains >86 open reading frames. The virus genome contains four major internal repeats (IR1 to IR4) and terminal repeats (TR). Nine latent proteins, including EBV nuclear antigen 1 (EBNA1), EBNA2, EBNA3A, -3B, -3C, EBNA-LP and latent membrane protein 1 (LMP1) and LMP2A, -2B are encoded by genes situated in the unique regions of the genome.[1] Other open reading frames encode capsid proteins, transcription factors as well as lytic proteins of various functions.[1] The virus also encodes non-coding EBV RNAs, such as EBV-encoded small RNA 1 (EBER1) and 2 (EBER2), BART-derived microRNAs (miRNAs-BARTs) and BHRF1 microRNAs (miRNAs-BHRF1).[2]

The incidence of NPC has a remarkable geographical distribution being 100-fold more frequent in Southeast Asia, North Africa and Alaska than the rest of the world[3], which prompted studies to investigate whether distinct variants of EBV might contribute to this disease. EBV genomes can be broadly classified into two distinct types, type 1 and type 2. The EBV types can be distinguished by the polymorphisms in EBNA2 and EBNA3A-C. Intertypic EBV containing type 1 EBNA2 and type 2 EBNA3A-C had also been reported.[4, 5] EBV variants had been investigated in NPC tumours using various polymorphic genotype markers in the EBER1 and -2, LMP1, BHRF1, BZLF1 and EBNA1 loci in samples from different populations.[6-10] Particularly, the mutations in the transforming LMP1 gene were thought to be important in contributing to the pathogenesis of NPC.[11] However, studying genetic variations in a small number of candidate EBV genes is not sufficient to accurately assess the association between EBV genomic variations and NPC. A recent genomic study of EBV isolated from samples derived from different geographic areas and EBV-related diseases revealed worldwide EBV genetic diversity and suggested that NPC-derived EBV from endemic regions may be distinct from EBV variants derived from other regions.[5] Because genomic data are only available from a small number of NPC-derived EBV and those from healthy donor-derived EBV of the same population are largely absent, one cannot distinguish whether genetic differences between EBV derived from NPC and non-NPC samples are due to geographical variations or linked to pathogenesis. A genome-wide analysis of EBV variants isolated from NPC biopsies and those derived from the carriers of the same population is necessary to address this question.

Here we report the first case-control study comparing genomic sequences of EBV isolated from saliva samples of 142 healthy carriers with those from primary tumour biopsies derived from 62 patients with NPC of Hong Kong. Our data reveal the presence of high risk EBV variants in most of the NPC patients (96.8%) and a subset of population carriers in Hong Kong (40.1%). These high-risk variants carry non-synonymous mutations in EBV lytic proteins and mutations which may affect the secondary structure of EBER2.

Materials and Methods

Participant recruitment

We had recruited 894 subjects upon obtaining written consent to donate saliva samples for sequencing of EBV genomes harboured in population carriers of Hong Kong Chinese. All Hong Kong residents greater than 18 years old with no history of cancers or autoimmune diseases, but who may have medical conditions unrelated to EBV such as hypertension, are eligible to participate in the study. We obtained 62 NPC biopsies from the NPC Tissue bank of the Centre for Nasopharyngeal Carcinoma Research (CNPCR) and an established NPC tumour bank of our laboratory. The collection of the NPC biopsies and saliva was approved by the Institutional Review Board (IRB) of The University of Hong Kong (HKU)/Hospital Authority Hong Kong West Cluster (IRB ref. no. UW 08-156) for the purpose of EBV genome sequencing. High secretors of EBV, whose saliva samples contain >/= 1 x 10^5 copies of virus/ml, were selected for target capture of EBV genomes and sequencing. 180 saliva samples could meet the viral load criteria. The characteristics of the population carriers and NPC patients are shown in Supplementary Fig 1 and Supplementary Table 1.

DNA sample preparation

DNA of tumour and saliva samples was extracted using AllPrep DNA/RNA Micro Kit and Qiagen Blood and Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer’s protocol. NanoDrop spectrophotometer (Thermo Scientific) and Qubit dsDNA High Sensitivity (HS) Assay Kit (Life Technologies) were used to determine the concentration of the DNA samples.

Assessment of viral loads by quantitative PCR

The viral loads of tumour and saliva samples were quantified by detecting the EBV BamH1W repeats in EBV genomes by quantitative PCR using ABI PRISM 7900 sequence detector (AppliedBiosystems, Life Technologies, CA).

Library preparation, target capture and sequencing

100 ng DNA of each sample was used for library preparation by NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (New England Biolabs) followed by target capture of EBV genomes by Integrated DNA Technologies (IDT) xGen® Lockdown® probes and IDT Hybridization Kits. The IDT xGen® Lockdown® probes were 120 bp DNA oligos designed across the whole genome of type 1 EBV and selected regions of type 2 EBV. These oligos covered the genome at end-to-end (1x) coverage. The HiSeq 2500 Sequencer (100-bp paired-end) platform at the Centre for Genomic Sciences (CGS) of HKU was used to sequence the EBV genomes.

Genome assembly

We generated assembly that represents single dominant strain in every sample. Raw reads were trimmed using Trimmomatic (v0.36).[12] The first 2 bases and end bases with quality score below 28 were removed. Reads shorter than 70 bp were discarded. Reads were assembled into contigs using SPAdes (v3.11).[13] Kmer sizes of 69, 79 and 89 were used. Contigs longer than 500 bp and coverage >5 were output. We located the contigs using MiniMap2 (v2.11-r797).[14] We minimized the confounding effects of mixed infection by excluding samples with multiple contigs stacking outside repetitive sequences and OriLyt (Supplementary Fig 12). Contigs were mapped, oriented and sorted using ABACAS (v1.3.2)[15] against NC_007605.1. Unmapped contigs were discarded and then NG50 was calculated using 171,823 as reference genome size. Highly fragmented assemblies (NG50<8000) were excluded to provide higher confidence that each assembly represents a single EBV genome. Ten nucleotides in each end of contigs were trimmed and gaps were narrowed with reads using GapFiller (v1.10).[16] The assemblies were aligned to NC_007605.1 using MUMmer(v3.23)[17] and linearized according to the configuration of NC_007605.1. Then, mixed infection effects were further controlled by assessing the heterogeneity in each base. Reads were mapped back to the draft genomes using BWA-MEM[18], PCR duplicates were removed using PICARD (http://broadinstitute.github.io/picard/), and reads were piled up with BCFtools (v1.7).[19] Bases were considered ambiguous if supported by fewer than 50% of high-quality bases in BCFTools [i.e. (AD of the allele on the assembly) / (total AD for all alleles + 0.1) < 0.5]. Ambiguous bases were substituted by 'N' and treated as missing data in subsequent analyses. Sequencing depths were estimated by the number of mapping reads mapped to genomic regions outside repetitive regions and OriLyt using the same alignment method. A summary of NG50, viral copies and sequencing depth is available in Supplementary Fig 2 and Supplementary Table 1.

Variant calling and quality control

Assemblies were compared with NC_007605.1 using MiniMap2.[14] BCFtools[19] were used to call variants and mask repetitive regions and OriLyt to reduce inaccurate variant calls due to multiple copies of repeats and OriLyt. The EBV types were determined by the relative mismatch in EBNA2 and EBNA3 genes of the type 1 and type 2 reference genome (NC_007605.1 and NC_009334.1, respectively). Each sample is required to have fewer than 0.1 missing variants that have minor allele frequency (MAF) > 0.05 and missingness < 0.1. The pattern of linkage disequilibrium was computed and visualized in Haploview[20] (Supplementary Fig 13). After completing all quality control steps, the dataset of our study comprised of 142 controls and 62 NPC biopsies. The heterogeneity across the genome of controls was assessed by the fraction of reads supporting the allele at each position. Mean values of this fraction across is calculated (Supplementary Fig 14).

Analysis of geographic patterns between Hong Kong EBV and non-NPC endemic EBV

Published EBV genomes from non-NPC endemic regions were selected from Genbank. We excluded EBV with unknown origins as well as those collected from Chinese/Vietnamese ethnicity or areas. Strains B95-8, Jijoye and SNU719 were resequenced and assembled using the same method in this study. In total, 97 assemblies of non-endemic origins were obtained. In addition to our 204 assemblies, we carried out SNP calling and quality control procedures. Principal component analysis (PCA) was performed using PLINK v1.90.[21] The pairwise distance was computed using the genotypes in R. Hierarchical clustering and heatmap were computed with the default setting in R package pheatmap. We used VCFtools[22] to estimate the weighted F_{ST} values in 1000 nucleotide windows with 100 base-pair step size. The input genotypes were modified into homozygous diploid samples. In type 1 samples, EBV from non-endemic origins were compared with Hong Kong control to obtain genome-wide F_{ST}. Similarly, non-endemic EBV were compared with Hong Kong NPC. The regions with F_{ST} uniquely high in NPC were shown by subtracting the F_{ST} values for controls from NPC.

Cluster analysis

To infer the subpopulations of our dataset, an alignment was generated by concatenating the SNPs, where the SNPs in the repetitive regions are excluded to avoid artefacts due to assembly and alignment errors. The samples are then grouped into genetically similar clusters using Bayesian Analysis of Population Structure (BAPS) 6.0.[23] SNPs with MAF<0.05 (excluded in GWAS) were also included. The maximum number of populations was set to 7. Admixture analysis was performed using default parameters. Results were visualized using R package pophelper.[24] The Hamming distance was used in the distance matrix to generate a neighbour joining tree using R package ape.[25] F_{ST} statistics for each SNP and weighted F_{ST} statistics in sliding 1000 nucleotide windows were calculated between populations using VCFtools[22] as described above. The association between subgroups and NPC was assessed with chi-square test in R. The overrepresentation of NPC EBV in each subgroup was tested with hypergeometric test implemented in R.

Phylogenetic analysis

The alignment in the clustering, with missing genotypes coded as 'N', was used to construct the tree. The alignment was trimmed by automated algorithm in trimAl v1.4.[26] Maximum-likelihood trees were constructed under a general time reversal model with ascertainment bias correction (GTR+ASC) in IQ-TREE.[27] Fast tree search mode was chosen. The tree was visualised and annotated using ggtree.[28]

Principal component analysis (PCA) and association test for PCs and SNPs.

PCA and tests for PCs were performed in R package bugwas.[29] The same set of SNPs in association tests were used. Missing data was imputed using BEAGLE (v5.0).[30] The output from phylogenetic analysis was used for the tree input file. Bonferroni correction of p-value cut-off of 2.45×10^{-4} was used to assess the significance of the 204 principal components. SNPs were also assigned to their most correlated PCs by bugwas. To test the association between each SNP and NPC, we fit the each SNP in a logistic regression model adjusting for sex and age using PLINK v1.90[21], without including PCs as covariates in order to maximize the power for detecting variants correlated to NPC-associated lineages.

Genome-wide association study (GWAS) using linear mixed model.

The association between NPC and the variants with MAF > 0.05 and missingness < 0.1 was tested by a GWAS. To control for population structure, the GWAS was performed using a linear mixed model implemented in GEMMA[31] provided by bugwas.[29] A relatedness matrix among genomes was generated with the variants. Then the variants were tested under linear mixed model adjusting for age and sex as fixed effects and relatedness matrix as random effects. The p-values from likelihood ratio tests were reported. Bonferroni correction was applied to set the genome-wide significance cut-off at 1.71×10^{-5}. False discovery rate (FDR) was controlled at 0.05. To rule out the potential bias of data analysis due to the variation of viral loads among the samples, we performed a GWAS using the same model to test the associations between the variations and viral copy numbers in log scale in the saliva samples (Supplementary Fig 15).

Assessment of heterogeneity at the deletion sites

To assess the heterogeneity of positions differentiating the control and high-risk variants, the fraction of sequencing reads with the 4-base-deletion downstream of EBER2 (EBER-del) out of all sequencing reads covering coordinates 7188-7191 was calculated. We also revisited the samples that were excluded due to difficulty in assembly due to mixed infection.

Haplotype association

The haplotypes formed by 26 significant SNPs and indels were identified in GWAS after Bonferroni correction for p-values and control for FDR at EBER locus. The presence or absence (coded as 1 or 0) of each haplotype was retested under the linear mixed model (LMM), including age and sex as fixed effects and the same relatedness matrix generated in GWAS as random effects. The p-values were also derived from likelihood ratio test.

PCR and Sanger Sequencing

A region of the EBER haplotype (coordinates 6776-7397) was verified by PCR amplification followed by subsequent dideoxy-DNA sequencing of 42 additional saliva samples derived from population carriers. The forward primer and reverse primer used for the PCR amplification are 5’- GTCTTCGGTCAAGTACCA -3’ and 5’-GTAGAGTGGGAGTGCTATC-3’, respectively. 100 ng of genomic DNA was added to a 50 µl reaction mixture, containing 1 µl each of 10 µM forward and reverse primers, 1 µl of 10 mM deoxynucleotide triphosphate (dNTPs), 0.25 µl HotstarTaq enzyme and 5 µl buffer provided by the kit (Qiagen). Fifty cycles of denaturation (94ºC for 30s), annealing (56ºC for 30s) and extension (72ºC for 30s) were carried out in automated thermal cycler. The PCR products were subjected to dideoxy-DNA sequencing at the CGS of HKU.

RNA secondary structure prediction

The secondary structure of EBER2 was analyzed using RNAstructure version 6.0.1 with default parameters.[32] The structures with lowest free energy are reported.

Genetic risk score calculation

For each sample, genetic risk score was calculated from a set of specified SNPs. It was calculated as the sum of beta weighted by the presence/absence (coded as 1 or 0) of risk alleles, divided by the number of non-missing genotypes the sample has. The beta resulting from GEMMA output in Wald test in GWAS was used. Other Chinese NPC sources include cell lines C666-1, M81, NPC43 (resequenced in this study), biopsies GD2 and D3201, and NPC saliva GD1. Chinese lung cancers included LC1-LC4. Chinese gastric cancers included EBVaGC1 to -9 and GC-variant-1 to -3 (Supplementary Table 11).

Data availability

Raw sequencing reads can be accessed on NCBI Sequencing Read Archive with accession number SRP152584. Assemblies were deposited on GenBank MH590370-MH590579.

Results

The genomes of EBV isolated in Hong Kong Chinese are genetically distinct from those derived from other geographic regions

We sequenced and assembled EBV genomes isolated in saliva of 142 population carriers and tumour biopsies of 62 NPC patients in Hong Kong, in which HKNPC1-9[33] were re-sequenced (Supplementary Table 1 and Supplementary Fig 1 and 2). We compared our data with 97 published EBV genomes from non-NPC endemic regions. The total 301 genomes captured 11,040 single nucleotide polymorphisms (SNPs). We performed principal component analysis (PCA) to visualize their genetic pattern (Fig 1a). The first two PCs account for 39.2% and 17.1% respectively of the total variance. The first PC separates the type 1 and type 2 EBV, which are defined by the polymorphisms in EBNA2 and EBNA3A-C. Four intertypic EBV genomes are found in our data, in which three of them carry type 1 EBNA2 and type 2 EBNA3A-C, consistent with the most commonly reported intertypic genotypes.[5] The remaining intertypic strain, found in an NPC biopsy, contains a rare combination of type 1 EBNA2 and EBNA3A in addition to type 2 EBNA3B-C.[34] The second PC displays geographic difference between EBV variants derived from Hong Kong and non-endemic regions which can be confirmed by clustering EBV genomes based on their genetic distance (Fig 1b, Supplementary Fig 3).

LMP1 is one of the most important oncogenic proteins contributing to the pathogenesis of NPC.[35] Several variations in LMP1, particularly the 30-bp deletion between coordinates 167,808 to 167,837 (del-LMP1), were thought to be associated with NPC.[36] However, this del-LMP1 variant is highly prevalent in EBV derived from both population carriers and NPC tumours of Hong Kong, consistent with previous findings that the del-LMP1 represents a geographic variation rather than a disease-associated variation[37] (supplementary Fig 4). In addition, a number of variations which differentiates the EBV variants in Hong Kong (both control and NPC) and non-endemic regions is identified (Fig 1c, supplementary Table 2). Analysis of EBV genomic variations without comparison with population control would risk identifying the geographic variations as disease-associated variations. Interestingly, we observe genomic variations which may be specific to NPC-derived EBV in comparison with control EBV in the regions near EBERs and other genes such as LF3, BALF4 and BALF5 (Fig 1c, Supplementary Table 2).

Identification of five EBV subgroups 1A-C and 2A-B amongst the population carriers in contrast to the predominance of 1A and -B in the majority of NPC

We next employed two approaches, cluster analysis and PCA, to assess lineage-level association between EBV variants and NPC. First, we clustered all EBV strains derived from NPC cases and controls into subgroups of genetically similar genomes based on the whole viral genome data using Bayesian Analysis of Population Structure (BAPS) [23]. The repetitive regions and the OriLyt were excluded to avoid artefacts due to sequencing errors. We identified five clusters with sizes of 45, 73, 66, 4 and 16, which are designated as subgroups 1A, 1B, 1C, 2A and 2B, respectively (Fig 2a, Supplementary Table 3). Except for one genome classified in subgroup 1A which contains an intertypic EBV strain, all other genomes in subgroups 1A-C are type 1 EBV and those in subgroups 2A and -B are type 2 EBV. Whilst the five subgroups account for 16.9%, 24.6%, 44.4%, 2.8% and 11.3% of control saliva-derived EBV, respectively, almost all NPC-derived EBV variants belong to subgroups 1A (33.9%) and 1B (61.3%) with the remaining in subgroup 1C (4.8%). The strong association between subgroups and NPC ($p = 1.84 \times 10^{-10}$) and the overrepresentation of NPC in subgroups 1A and 1B ($p = 2.50\times 10^{-3}$ and $p = 1.42 \times 10^{-07}$, respectively) suggests that two major lineages of EBV are harboured in NPC. The neighbour joining tree of the subgroups shows that subgroups 1A and -B are more closely related than 1C (Supplementary Fig 5). To identify the genomic regions that characterize subgroups 1A and -B, we scanned with a 1000-nucleotide sliding window across the EBV genome and identified regions with high fixation index (F_{ST}), which reflects the population differentiation due to genetic structure (Supplementary Fig 5, Supplementary Table 3, 4). The windows near the EBER region (coordinates 6328-7355) best separate subgroups 1A and -B from 1C (F_{ST}=0.931).

Second, we examined the population structure using PCA, which projects the genotype data into different PCs. We observed consistent results between cluster analysis and PCA, where subgroups are well stratified in the first two PCs (Fig 2b). NPC-derived EBVs, which are characterized by subgroups 1A and -B, are found to cluster together in the phylogenetic tree corresponding to a region at a negative value of PC1 (Fig 2c). We used the R package bugwas software to assess the association between the PCs and NPC as well as assigned every tested SNP to its most correlated PCs[29] (Supplementary Fig 6). At the cut-off p-value of 2.45×10^{-4}, association between PC1 and NPC is just below the level of statistical significance ($p = 5 \times 10^{-4}$). We tested the association between SNPs and NPC with a logistic regression model adjusting for age and sex (Fig 2d). We found that most significant SNPs were assigned to PC1, supporting the significance of PC1. These SNPs constitute 83 SNPs involving EBER and a number of genes, amongst which 12 will lead to non-synonymous changes in amino acids encoding DNA binding protein (BALF2), tegument protein (BNRF1, BTRF1), viral nuclease (BALF3), capsid antigen (BCLF1), scaffold protein (BVRF2 and BDRF1) and glycoprotein (BALF4 and BDLF3) (Supplementary Table 5).

Genome-wide association study (GWAS) identifies the strongest NPC-associated polymorphisms in the EBER region

Population structure can be a potential confounding factor in the analysis of NPC-associated variations.[38] For example, ethnically homogeneous groups are often recruited in human GWAS to minimize the systematic difference among samples and hence prevent false positive findings. However, as we have observed in the clustering and PCA results, the EBV in Hong Kong population is stratified into subpopulations rather than a homogenous group. Therefore, testing the associations between variants and NPC without considering the population structure is prone to false discovery. In order to address this potential problem and identify the most robust variations that are associated with NPC, we carried out a genome-wide association study (GWAS) using a linear mixed model (LMM) implemented in GEMMA.[31] We tested the significance of 2919 common SNPs/indels (minor allele frequency > 5%) under a LMM that includes age and sex as fixed co-variates and genetic similarities among samples as a random effect and reduces the genomic inflation factor, lambda, to 0.55 (Supplementary Fig 7). The chip heritability, which reflects the proportion of variance in the phenotype explained by the tested SNPs, was 18.25% with a standard error of 7.52%, supporting the contribution of EBV genetics. At the Bonferroni corrected cut-off p value of 1.71×10^{-5}, we found the most significant variants associated with NPC in the EBER region (Fig 3a, Table 1 and Supplementary Table 6). The 22 most significant SNPs/indels map to a region overlapping with EBER1, EBER2 and OriP (coordinates 6484-7327). The top association signal is a four-base-deletion downstream of EBER2 (EBER-del) from coordinates 7188 to 7191 ($p = 1.91\times10^{-7}$). The second most significant associations are SNPs located between EBER1 and EBER2 at coordinates 6866, 6884 and 6886 ($p = 2.60\times10^{-7}$). Since Bonferroni correction could be too stringent for EBV genomes which contain highly linked polymorphisms, we also adjusted p-values by controlling false discovery rate (FDR) at 0.05. This identified 3 additional SNPs near the EBER region (coordinates 5850, 6584 and 8568), 1 SNP (coordinate 5399) causing p.Val1222Ile in BNRF1, 2 six-base-insertions (after coordinates 59515 and 59518) that lead to in-frame insertions of two amino acids in BOLF1, and 1 SNP (coordinate 137316) causing p.His560Pro in BVRF2.

Whilst polymorphisms in EBER[39], EBNA1 (V-Val)[40], BZLF1 promoter (Zp-V3)[41], RPMS1[42] and LMP1[11, 43] had been proposed to be associated with NPC and other EBV-associated malignancies, these studies were based on data derived from pre-selected candidate genes and are prone to detect spurious association due to failure to adjust for the effects of population stratification.[38] To illustrate this point, we analyzed the association of these polymorphisms with NPC by chi-square test in our case-control dataset without considering the population structure. We found that there was no significant association between NPC and neither the A91006C in Zp-V3 ($p=2.89\times10^{-05}$) nor C97121T in V-Val ($p=2.91\times10^{-04}$), but there was significant association between NPC and either the G155391A in RPMS1 ($p=1.84\times10^{-17}$) or C167859T in LMP1 ($p=1.39\times10^{-18}$) (refer to Supplementary Fig 8, Supplementary Table 7 and 8). However, none of them reached genome-wide significant level of $p=1.71\times10^{-5}$ in the LMM where population stratification is corrected. In contrast, the variations in the EBER region remain highly significant.

NPC-associated haplotype in EBER region causes structural changes in EBER2.

A total of 26 SNPs/indels in the EBER region constitute 13 haplotypes, with one dominant haplotype in NPC and two dominant haplotypes in control saliva (refer to Fig 3b and Supplementary Table 9). Only 4 SNPs located in the intergenic region between EBER1 and -2 are conserved in former (refer to Fig 3b). The distributions of these major haplotypes were significantly different between NPC and controls. The dominant NPC haplotype was found in 93.6% (58/62) of NPC-derived EBV variants and 38.7% (55/142) of control saliva-derived EBV. Meanwhile, the two dominant control haplotypes, which are different from the NPC haplotype, accounted for only 1.6% (1/62) of NPC-derived EBV and 56.3% (80/142) of control saliva-derived EBV. Under the LMM, the association between the dominant NPC haplotype and disease is found to reach statistical significance ($p = 4.85\times10^{-6}$). To validate whether such distribution was biased by sequencing samples with high viral copy number, we analyzed 42 additional saliva samples from population carriers with relatively lower viral loads (5 x 10^3 to 3 x 10^5 viral copies per µg DNA) by Sanger sequencing of a portion of the haplotypes (coordinates 6776-7397). Out of these 42 samples, the NPC haplotype was found in 16 cases (38%) whilst the control haplotype was found in 26 cases (62%) (Supplementary Table 10). The ratio of NPC haplotype to control haplotype is similar to that observed in the original 142 EBV genomes derived from control saliva samples which contain higher viral loads (1 x 10^5 to 9 x 10^7 viral copies per µg DNA). These results confirm that the differences observed between NPC and controls are not related to the variation of viral loads between the samples.

Amongst the 26 variations in the EBER region, we hypothesized that the variations in EBER2 may have functional significance. EBER2 is a non-protein coding RNA thought to have direct interaction with both viral and cellular proteins or nucleic acids through its multiple stem loops.[44] The EBER2 variants dominant in NPC contain variations in RNA coordinates 44, 46, 57, 61, 93, 167, 168 and 170. We observed a shortened tail of EBER2 and structural changes in the first stem loop of EBER2 upon prediction by RNAStructure[32] (Fig 3c). The first stem loop of EBER2 of wild type EBV consists of an internal loop, a bulge loop and a hairpin loop whereas the internal loop changes to two bulge loops and the original bulge loop and hairpin loop are altered in both structure and sequence in that of EBER2 of NPC-derived EBV. The first stem loop of EBER2 may be involved in the recruitment of cellular transcription factor such as PAX5 to the terminal repeat region of EBV thus can potentially modulate the transcription of LMP1.[44]

Classification of risk variants of EBV for NPC

Finally, we constructed a genetic risk score (GRS) based on the statistical effect sizes derived from GWAS to compare EBV variants isolated from different types of samples and geographic regions. Using the 29 SNPs/indels outlined in Table 1, we found that the GRS of EBV from non-endemic regions is generally lower than that of NPC-derived EBV, with the exception of the variant known as VGO from Brazil (Fig 4a, Supplementary Table 11). On the other hand, 5 out of 6 EBV genomes isolated from cell lines (M81, C6661 and NPC43), biopsies (GD2, D3201) and saliva (GD1, lowest GRS) derived from NPC patients have very high GRS. In contrast, EBVs derived from all reported Chinese EBV-positive gastric cancers have zero GRS and only one of 4 EBV-positive lung cancers has high GRS. A similar trend is observed when we include the SNPs assigned to PC1 (refer to Supplementary Table 5) that are found to be significant (Fig 4b, Supplementary Table 10). The difference in GRS between EBV derived from NPC and non-NPC samples indicates the possibility of using EBV genomic variations in classifying risk variants of EBV for NPC.

Discussions

Previous case-control studies based on the analyses of candidate EBV genes have reported association between various genetic polymorphisms and NPC.[11, 39, 41, 42] However, these studies are prone to detect spurious association due to the failure to adjust for the confounding effects of population stratification.[38] To minimize the confounding effects, it is essential to analyze individual SNP of EBV genome in the context of genome-wide sequence data in order to accurately identify significant variations associated with NPC. Whilst a number of SNPs, including G155391A in RPMS1 and C167859T in LMP1, are found to be associated with NPC by chi-square test in our case-control dataset (refer to Supplementary Fig 8 and Supplementary Table 7), neither reach genome-wide significant level of 1.71×10^{-5} in the linear mixed model (LMM) where population stratification is corrected.

The results of our GWAS suggested that the variations in the EBER region, designated as NPC-EBERvar, are the most robust ones in the context of the genetic background of EBV variants harboured in Hong Kong Chinese. Apparent association between some of these variations in the EBER region and NPC by a candidate gene approach in individuals of South China has been reported.[39] Functionally, several studies have indicated the link between EBERs and NPC by their abilities to promote cell growth and survival[45, 46] and modulate the expression of LMP1 and LMP2 genes.[44] Based on previous functional experiments and the predicted alteration in the secondary structure of EBER2 (refer to Fig 3c), we postulate that EBER2 may play an important role in the development of NPC. In addition to the SNPs in the EBER region, a panel of SNPs that may be associated with the development of NPC is also identified (refer to Fig 2d and Supplementary Table 5). Twelve of these SNPs will lead to non-synonymous changes in amino acids of EBV lytic proteins which include DNA binding protein, tegument protein, viral nuclease, capsid antigen, scaffold protein and glycoprotein.

The major limitation of the current study is that we are not able to assemble multiple EBV genomes in a sample with mixed infections. In order to ensure the accuracy of the assemblies, 12 samples with mixed infections are excluded in our study. The comparison of the 142 dominant EBV genomes derived from control saliva samples and 62 NPC biopsies reveals the presence of high risk EBV variants (NPC-EBERvar) in 96.8% of NPC samples and 40.1% of control saliva samples (refer to Table 1). However, one may suspect that low level infection of NPC-EBERvar would exist in the control samples and affect our results. We, therefore, further assessed the proportion of high-risk and control variants in each sample (including both 142 assembled genomes and the 12 excluded samples with mixed infection) using the percentage of raw reads containing the four-base-deletion polymorphism (EBER-del) as an indicator (Supplementary Fig 9). EBER-del is highly homogeneous in 147 out of the 154 control saliva samples. There are only seven samples, including HKHD64 and six excluded samples, harboring a mixture of viral populations of high risk and control variants. When including the excluded samples, there are 63/154 controls (40.9%) with detectable reads with EBER-del, comparable to the analysis of the 142 samples with dominant EBV genomes. Similarly, EBER-del is highly homogeneous in all the 62 EBV genomes derived from NPC biopsies, consistent with the known monoclonal nature of EBV infection in tumours. In addition, we have compared the EBV genomes isolated from NPC biopsies with their matched saliva samples of 24 NPC patients and found that the EBV genomes harboured in the matched biopsies and saliva samples are almost identical in most of the patients, suggesting that the dominant EBV strains harboured in the saliva might contribute to the carcinogenesis of NPC in the patients (Supplementary Fig 10 & 11).

Our results may not be able to reveal all possible NPC risk loci on the EBV genome. First, we may lose statistical power for detecting other risk loci when correcting for population structure, in exchange for lower false discovery rate. However, these missed genetic variants are likely to be assigned to PC1 as shown in Fig 2d and Supplementary Table 5. Meta-analysis of studies with larger sample size as well as similar studies in other NPC-endemic regions will increase the chance to discover these variations. Second, the variations in the repetitive regions are not examined due to the limitation of short-read sequencing. Further studies utilizing long-range sequencing platforms will help to interrogate the association between variations in repetitive regions and NPC. Third, interaction between host genetics[47-50], environmental exposures and viral variants has not been explored. A joint study of EBV and host genomics may provide new insights into such interactions. Last, sequencing of EBV genomes is feasible in about 20% of the saliva samples due to the requirement of high viral copy number for efficient capture of EBV DNA. Further optimization of the target capture protocol will enable sequencing of EBV in a higher percentage of samples.

In conclusion, we have identified high risk EBV variants of NPC, designated as NPC-EBERvar, at the level of both loci and lineages through a genome-wide case-control study. The presence of NPC-EBERvar in the majority of NPC biopsies and a proportion of population carriers of Hong Kong support an important role of EBV genetic variations in the pathogenesis of NPC and explain the high incidence of NPC in endemic area such as Hong Kong. Further characterization of specific pathogenic mutations in the high risk EBV variants and survey of viral variants amongst different geographic populations will be important goals of future research.

Acknowledgements

This study is funded by Heath and Medical Research Fund (Project 02131706), NPC Area of Excellence (AoE/M 06/08 Center for Nasopharyngeal Carcinoma Research), and Epstein-Barr virus research (#20004525) grants of Alan Kwok Shing Chiang. We would like to thank Tissue Bank of NPC Area of Excellence (AoE/M 06/08 Centre of Nasopharyngeal Carcinoma Research) for providing additional saliva and tumour biopsy samples for the current study.

References

1. Knipe DM, Howley PM, Griffin DE, Lamb RA, Martin MA, Roizman B, Straus SE. Field's Virologyed.: Lippincott Williams & Wilkins 2007.
2. Swaminathan S. Noncoding RNAs produced by oncogenic human herpesviruses. J Cell Physiol 2008;216:321-6.
3. Chang CM, Yu KJ, Mbulaiteye SM, Hildesheim A, Bhatia K. The extent of genetic diversity of Epstein-Barr virus and its geographic and disease patterns: a need for reappraisal. Virus Res 2009;143:209-21.
4. Burrows JM, Khanna R, Sculley TB, Alpers MP, Moss DJ, Burrows SR. Identification of a naturally occurring recombinant Epstein-Barr virus isolate from New Guinea that encodes both type 1 and type 2 nuclear antigen sequences. J Virol 1996;70:4829-33.
5. Palser AL, Grayson NE, White RE, Corton C, Correia S, Ba abdullah MM, Watson SJ, Cotten M, Arrand JR, Murray PG, Allday MJ, Rickinson AB, et al. Genome Diversity of Epstein-Barr Virus from Multiple Tumor Types and Normal Infection. J Virol 2015;89:5222-37.
6. Grunewald V, Bonnet M, Boutin S, Yip T, Louzir H, Levrero M, Seigneurin JM, Raphael M, Touitou R, Martel-Renoir D, Cochet C, Durandy A, et al. Amino-acid change in the Epstein-Barr-virus ZEBRA protein in undifferentiated nasopharyngeal carcinomas from Europe and North Africa. International journal of cancer 1998;75:497-503.

7. Sacaze C, Henry S, Icart J, Mariame B. Tissue specific distribution of Epstein-Barr virus (EBV) BZLF1 gene variants in nasopharyngeal carcinoma (NPC) bearing patients. Virus Res 2001;81:133-42.

8. Dardari R, Khyatti M, Cordeiro P, Odda M, ElGueddari B, Hassar M, Menezes J. High frequency of latent membrane protein-1 30-bp deletion variant with specific single mutations in Epstein-Barr virus-associated nasopharyngeal carcinoma in Moroccan patients. International journal of cancer 2006;118:1977-83.

9. See HS, Yap YY, Yip WK, Seow HF. Epstein-Barr virus latent membrane protein-1 (LMP-1) 30-bp deletion and Xho I-loss is associated with type III nasopharyngeal carcinoma in Malaysia. World J Surg Oncol 2008;6:18.

10. Zhang X-S, Wang H-H, Hu L-F, Li A, Zhang R-H, Mai H-Q, Xia J-C, Chen L-Z, Zeng Y-X. V-val subtype of Epstein-Barr virus nuclear antigen 1 preferentially exists in biopsies of nasopharyngeal carcinoma. Cancer Letters 2004;211:11-18.

11. Cheung ST, Leung SF, Lo KW, Chiu KW, Tam JS, Fok TF, Johnson PJ, Lee JC, Huang DP. Specific latent membrane protein 1 gene sequences in type 1 and type 2 Epstein-Barr virus from nasopharyngeal carcinoma in Hong Kong. International journal of cancer 1998;76:399-406.

12. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114-20.

13. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 2012;19:455-77.

14. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018.

15. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 2009;25:1968-69.

16. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biology 2012;13.

17. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biology 2004;5:R12.

18. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. . arXiv preprint arXiv:1303.3997 2013.

19. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;27:2987-93.

20. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2004;21:263-65.

21. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 2015;4.

22. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. The variant call format and VCFtools. Bioinformatics 2011;27:2156-58.

23. Corander J, Marttinen P. Bayesian identification of admixture events using multilocus molecular markers. Molecular Ecology 2006;15:2833-43.

24. Francis RM. pophelper: an R package and web app to analyse and visualize population structure. Molecular Ecology Resources 2017;17:27-32.

25. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 2004;20:289-90.

26. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 2009;25:1972-73.

27. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 2015;32:268-74.

28. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y, McInerny G. ggtree: anrpackage for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 2017;8:28-36.

29. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CCA, Iqbal Z, Clifton DA, Hopkins KL, Woodford N, Smith EG, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nature Microbiology 2016;1.

30. Browning Brian L, Browning Sharon R. Genotype Imputation with Millions of Reference Samples. The American Journal of Human Genetics 2016;98:116-26.

31. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics 2012;44:821-24.

32. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 2010;11.

33. Kwok H, Wu CW, Palser AL, Kellam P, Sham PC, Kwong DL, Chiang AK. Genomic diversity of Epstein-Barr virus genomes isolated from primary nasopharyngeal carcinoma biopsy samples. J Virol 2014;88:10662-72.

34. Kim SM, Kang SH, Lee WK. Identification of two types of naturally-occurring intertypic recombinants of Epstein-Barr virus. Mol Cells 2006;21:302-07.

35. Dawson CW, Port RJ, Young LS. The role of the EBV-encoded latent membrane proteins LMP1 and LMP2 in the pathogenesis of nasopharyngeal carcinoma (NPC). Seminars in Cancer Biology 2012;22:144-53.

36. Miller WE, Edwards RH, Walling DM, Raab-Traub N. Sequence variation in the Epstein--Barr virus latent membrane protein 1. Journal of General Virology 1994;75:2729-40.

37. Zhang X-S, Song K-H, Mai H-Q, Jia W-H, Feng B-J, Xia J-C, Zhang R-H, Huang L-X, Yu X-J, Feng Q-S, Huang P, Chen J-J, et al. The 30-bp deletion variant: a polymorphism of latent membrane protein 1 prevalent in endemic and non-endemic areas of nasopharyngeal carcinomas in China. Cancer Letters 2002;176:65-73.

38. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nature Reviews Genetics 2016;18:41-50.

39. Shen Z-c, Luo B, Chen J-n, Chao Y, Shao C-k, Liu Q-q, Wang Y. High Prevalence of the EBER Variant EB-8m in Endemic Nasopharyngeal Carcinomas. PLOS ONE 2015;10:e0121420.

40. Correia S, Bridges R, Wegner F, Venturini C, Palser A, Middeldorp JM, Cohen JI, Lorenzetti MA, Bassano I, White RE, Kellam P, Breuer J, et al. Sequence Variation of Epstein-Barr Virus: Viral Types, Geography, Codon Usage, and Diseases. J Virol 2018;92.

41. Tong JHM, Lo KW, Au FWL, Huang DP, To KF. Re: Discrete Alterations in the BZLF1 Promoter in Tumor and Non-Tumor-Associated Epstein-Barr Virus. JNCI Journal of the National Cancer Institute 2003;95:1008-09.

42. Feng F-T, Cui Q, Liu W-S, Guo Y-M, Feng Q-S, Chen L-Z, Xu M, Luo B, Li D-J, Hu L-F, Middeldorp JM, Ramayanti O, et al. A single nucleotide polymorphism in the Epstein-Barr virus genome is strongly associated with a high risk of nasopharyngeal carcinoma. Chinese Journal of Cancer 2015;34.

43. Liao H-M, Liu H, Lei H, Li B, Chin P-J, Tsai S, Bhatia K, Gutierrez M, Epelman S, Biggar R, Nkrumah F, Neequaye J, et al. Frequency of EBV LMP-1 Promoter and Coding

Variations in Burkitt Lymphoma Samples in Africa and South America and Peripheral Blood in Uganda. Cancers 2018;10.

44. Lee N, Moss Walter N, Yario Therese A, Steitz Joan A. EBV Noncoding RNA Binds Nascent RNA to Drive Host PAX5 to Viral DNA. Cell 2015;160:607-18.

45. Iwakiri D, Sheen T-S, Chen J-Y, Huang DP, Takada K. Epstein–Barr virus-encoded small RNA induces insulin-like growth factor 1 and supports growth of nasopharyngeal carcinoma-derived cell lines. Oncogene 2004;24:1767.

46. Iwakiri D, Zhou L, Samanta M, Matsumoto M, Ebihara T, Seya T, Imai S, Fujieda M, Kawa K, Takada K. Epstein-Barr virus (EBV)–encoded small RNA is released from EBV-infected cells and activates signaling from toll-like receptor 3. The Journal of Experimental Medicine 2009;206:2091-99.

47. Tse K-P, Su W-H, Chang K-P, Tsang N-M, Yu C-J, Tang P, See L-C, Hsueh C, Yang M-L, Hao S-P, Li H-Y, Wang M-H, et al. Genome-wide Association Study Reveals Multiple Nasopharyngeal Carcinoma-Associated Loci within the HLA Region at Chromosome 6p21.3. The American Journal of Human Genetics 2009;85:194-203.

48. Bei J-X, Li Y, Jia W-H, Feng B-J, Zhou G, Chen L-Z, Feng Q-S, Low H-Q, Zhang H, He F, Tai ES, Kang T, et al. A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. Nature Genetics 2010;42:599-603.

49. Dai W, Zheng H, Cheung AKL, Tang CS-m, Ko JMY, Wong BWY, Leong MML, Sham PC, Cheung F, Kwong DL-W, Ngan RKC, Ng WT, et al. Whole-exome sequencing identifiesMST1Ras a genetic susceptibility gene in nasopharyngeal carcinoma. Proceedings of the National Academy of Sciences 2016;113:3317-22.

50. Li YY, Chung GTY, Lui VWY, To K-F, Ma BBY, Chow C, Woo JKS, Yip KY, Seo J, Hui EP, Mak MKF, Rusan M, et al. Exome and genome sequencing of nasopharynx cancer identifies NF-κB pathway activating mutations. Nature Communications 2017;8.

Coordinates	Risk allele	Non-risk allele	Frequency in NPC	Frequency in Control	FDR adjusted P	P-value	Annotation
5399*	A	G	0.93548	0.4085	1.19×10^{-04}	9.98x10^{-05}	p.Val1222Ile in BNRF1
5850*	T	A	0.93548	0.4085	1.19×10^{-04}	9.98x10^{-05}	BNRF1 3’ UTR
6484	T	C	0.95161	0.4085	1.19×10^{-04}	1.10x10^{-05}	Between BNRF1 and EBER1
6584*	G	A	0.93548	0.4085	1.19×10^{-04}	8.98x10^{-05}	
6866	A	G	0.96774	0.4014	1.19×10^{-04}	2.60x10^{-07}	
6884	G	A	0.96774	0.4014	1.19×10^{-04}	2.60x10^{-07}	
6886	T	G	0.96774	0.4014	1.19×10^{-04}	2.60x10^{-07}	
6911	A	G	0.96774	0.4085	1.19×10^{-04}	7.03x10^{-07}	
6944	G	A	0.96774	0.4085	1.19×10^{-04}	7.31x10^{-07}	
6999	G	T	0.96774	0.4085	1.19×10^{-04}	6.64x10^{-07}	EBER2 (44)
7001	T	A	0.96774	0.4085	1.19×10^{-04}	6.64x10^{-07}	EBER2 (46)
7012	G	A	0.96774	0.4155	1.19×10^{-04}	1.60x10^{-06}	EBER2 (57)
7016	T	A	0.96774	0.4085	1.19×10^{-04}	6.64x10^{-07}	EBER2 (61)
7048	C	A	0.96774	0.4085	1.19×10^{-04}	6.64x10^{-07}	EBER2 (93)
7121	C	CTA	0.96774	0.4085	1.19×10^{-04}	6.64x10^{-07}	EBER2 (167-168)
7125	G	T	0.96774	0.4085	1.19×10^{-04}	6.64x10^{-07}	EBER2 (170)
7134	C	G	0.96774	0.4085	1.19×10^{-04}	6.64x10^{-07}	Between EBER2 and OriP
7187	A	AAACT	0.96774	0.4014	1.19×10^{-04}	1.91x10^{-07}	
7198	T	C	0.96774	0.4085	2.40×10^{-04}	6.64x10^{-07}	
7206	A	T	0.96774	0.4085	2.40×10^{-04}	6.64x10^{-07}	
7213	C	G	0.96774	0.4085	1.47×10^{-03}	6.64x10^{-07}	
7233	A	G	0.95161	0.4085	1.47×10^{-03}	1.11x10^{-05}	
7262	A	G	0.96774	0.4085	7.58×10^{-03}	6.64x10^{-07}	
7297	T	C	0.96774	0.4155	1.09×10^{-02}	1.65x10^{-06}	
7327	C	T	0.96774	0.4085	1.12×10^{-02}	6.64x10^{-07}	OriP
8568*	T	A	0.93548	0.4043	1.12×10^{-02}	5.97x10^{-05}	
59515*	CCTCCTT	C	0.95161	0.4789	1.69×10^{-02}	2.79x10^{-04}	p.Gly1145_Gly1146insGluGly in BOLF1
59518*	CCTCCTA	C	0.95161	0.4789	2.91×10^{-02}	1.56x10^{-04}	p.Gly1144_Gly1145insValGly in BOLF1
137316*	C	A	0.91935	0.3944	3.97×10^{-02}	3.94x10^{-04}	p.His560Pro in BVRF2

Table 1 | The most significant NPC-associated polymorphisms identified in the GWAS. The table shows the polymorphisms that pass false discovery rate (FDR) adjusted P of 0.05 in figure 3a. Genome-wide significant cut-off of P = 1.71×10^{-5} was used. *polymorphisms identified by controlling FDR at 0.05.

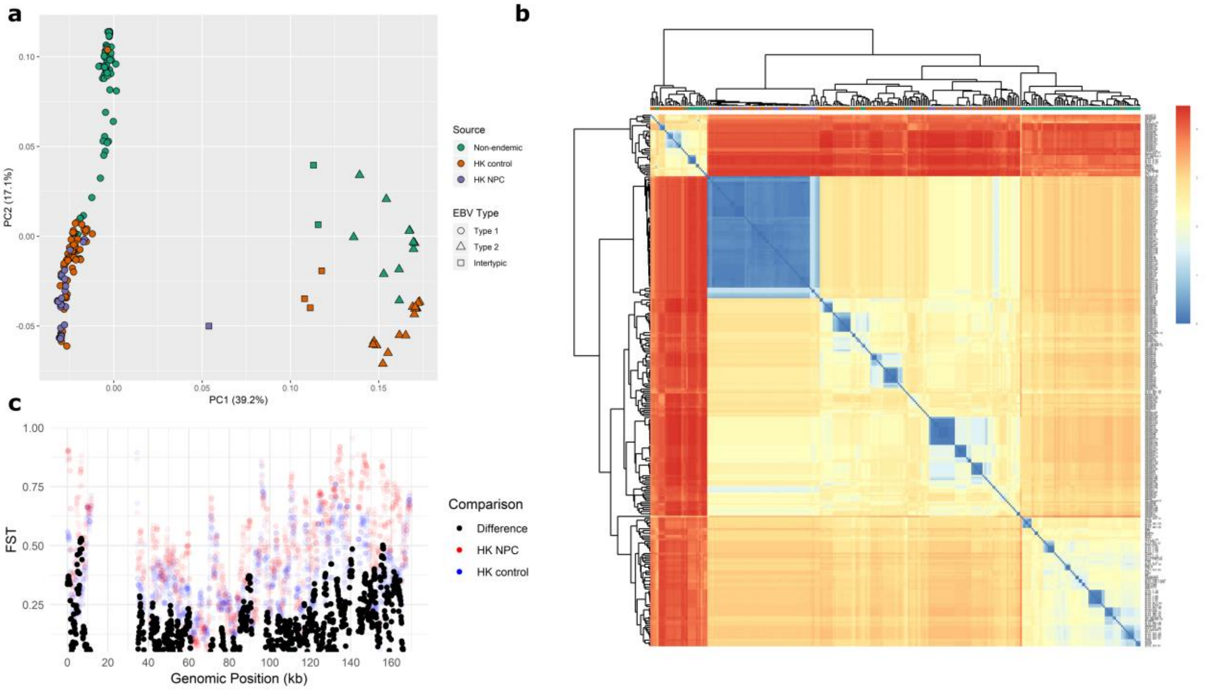
Figure legends

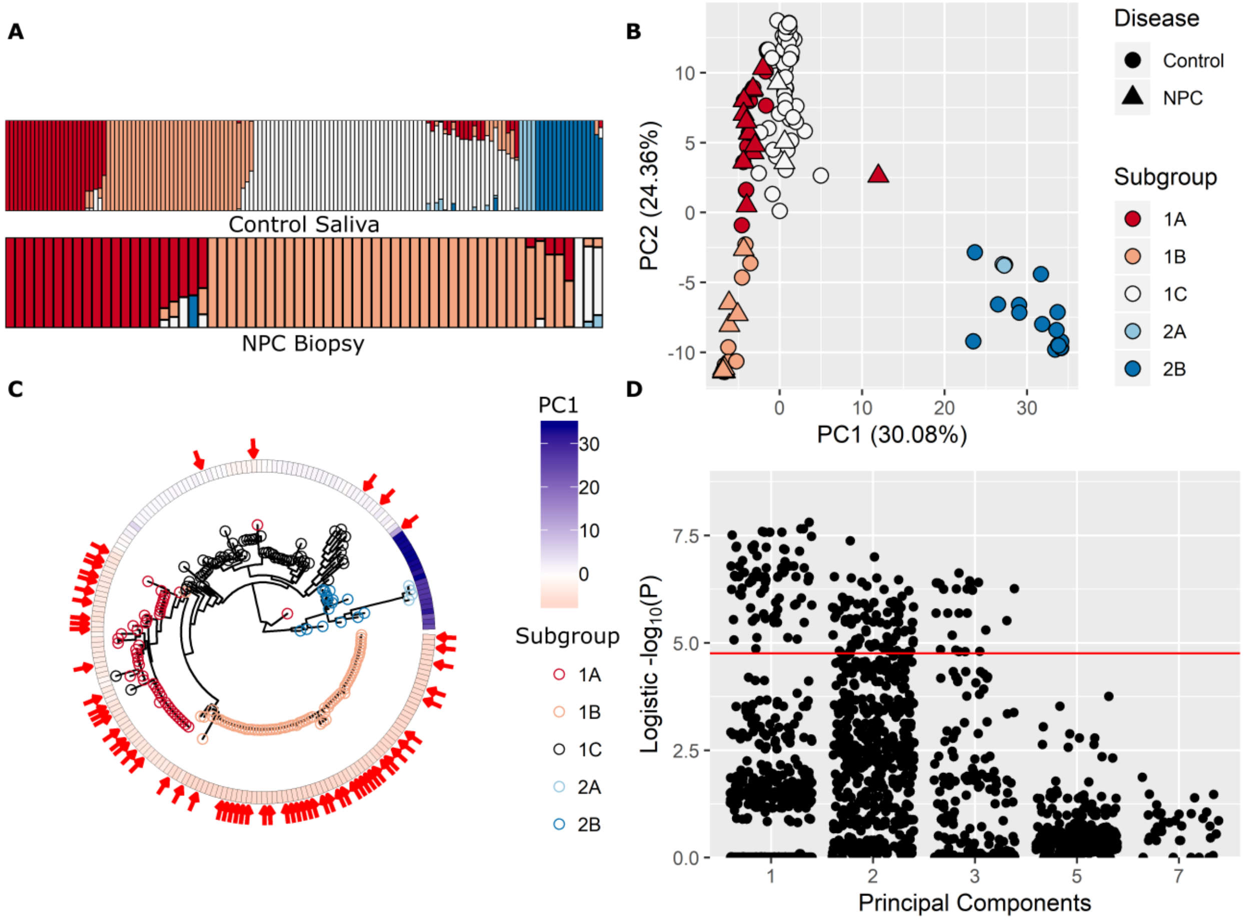
Figure 1 | Geographic differences between EBV variants isolated from individuals of Hong Kong and non-endemic regions. (a) Principal component analysis of EBV genomes isolated from 142 healthy carriers and 62 NPC patients of Hong Kong and 97 individuals of non-endemic regions. Percentages of variance explained are indicated in the axes. **(b)** Hierarchical clustering of pairwise distance amongst EBV genomes. The figure displays a sample by sample matrix where the distances are represented in colour scales from red (distant) to blue (close). The sources of the samples are indicated at the top of each column with the same colour codes used in Fig 1a. **(c)** Fixation index (F_{ST}) which signifies how different the type 1 EBV genomes are from the non-NPC endemic EBV genomes is shown. Each dot represents the F_{ST} in 1000 nucleotide region comparing type 1 Hong Kong NPC (HK NPC; red) or controls (HK control; blue) with type 1 non-endemic EBV. The difference (black) between the two F_{ST} values in each region is shown.

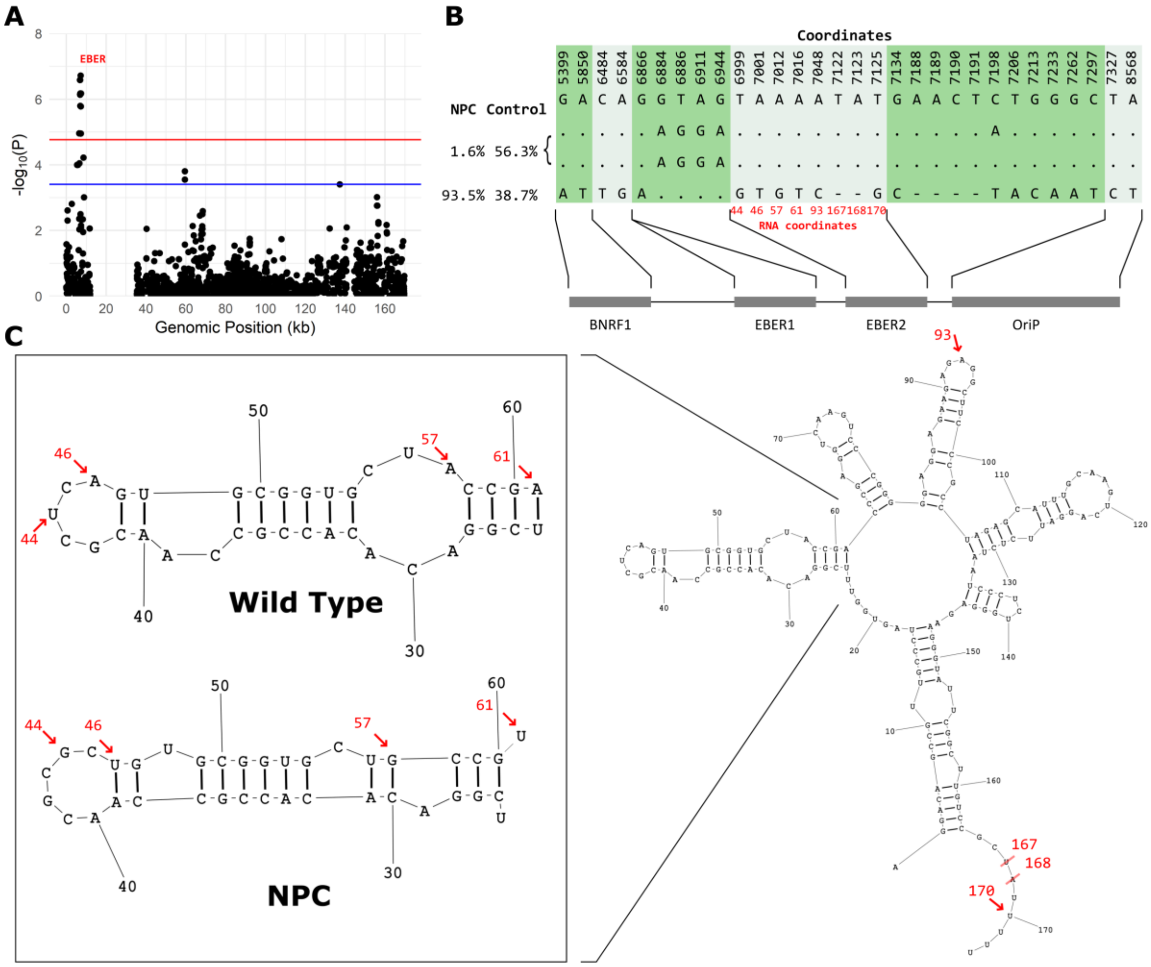
Figure 2 | Population structure of EBV variants isolated from population carriers and NPC patients of Hong Kong. **(a)** Admixture analysis output from Bayesian Analysis of Population structure (BAPS).[23] Each sample is represented by a column split into up five colours. Each colour represents a subgroup and the length shows the percentage of the ancestral source in an assembly. (b) Principal component analysis of 142 Hong Kong controls and 62 NPC samples. Percentages of variance explained are indicated in the axes. The PCs on the x-axis is sorted by their significance. (c) Maximum likelihood phylogenetic tree of the total 204 samples. Red arrow points to NPC cases. The colour intensities of the outer ring represent the values of samples in PC1 (d) The association tests for SNPs under logistic regression model adjusting for sex and age. The PCs on the x-axis are order by their significance to NPC. Only the top 5 PCs are shown. The red lines show the genome-wide significance cut-off.

Figure 3 | Identification of the strongest NPC-associated polymorphisms in the EBER region by GWAS. (a) Manhattan plot from the GWAS. The results are based on 2919 SNPs/indels with MAF > 0.05 in a dataset of EBV derived from 62 biopsies of NPC patients and 142 saliva samples of controls. The red line shows the genome-wide cut-off $p=1.71\times10^{-5}$ after Bonferroni correction. The blue line shows the cut-off when controlling false discovery rate at 0.05. **(b)** Three major haplotypes of polymorphisms near EBERs (refer to Table 1) in NPC cases and controls. Dots represent reference genotypes. Rare haplotypes, which are found in only 1 sample, are not shown (refer to supplementary Table 8). **(c)** Predicted RNA secondary structures of EBER2 wild-type and the mutant commonly found in NPC cases. Left, the stem loop that is structurally different between wild type and NPC dominant EBER2. Right, the predicted RNA secondary structure of EBER2. Red arrows denote the significant variations identified in GWAS.

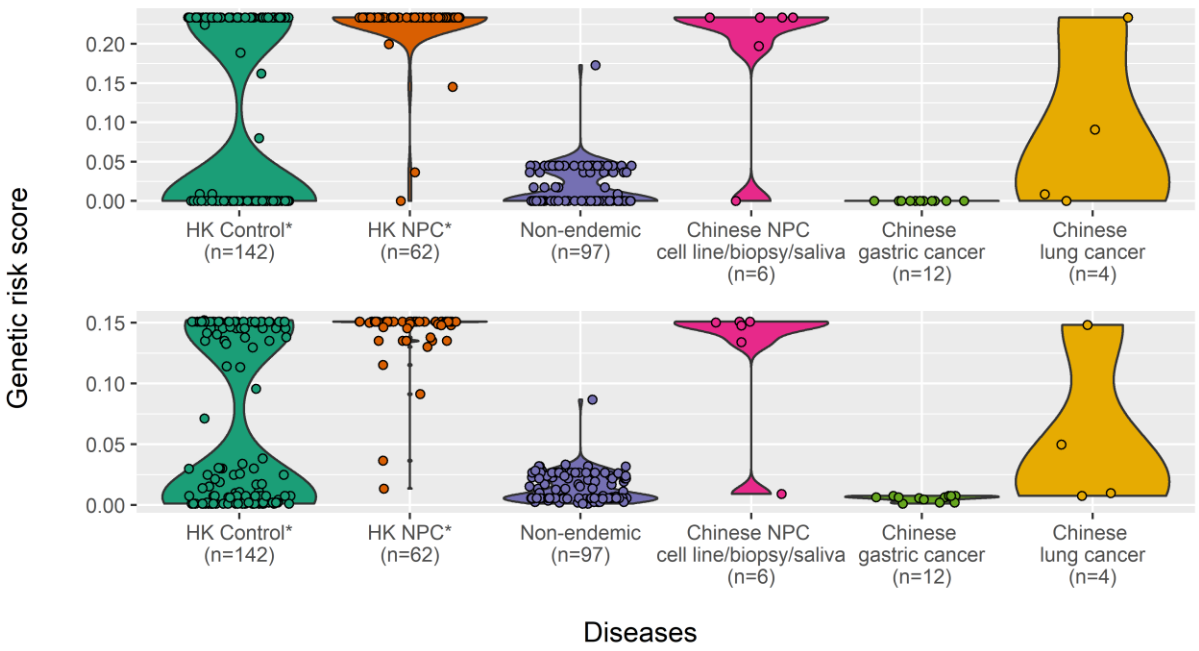
Figure 4 | Genetic risk scores (GRS) of EBV variants isolated from different types of samples and geographic regions. The GRS is calculated based on the polymorphisms in **(a)** EBER locus (Table 1) and **(b)** EBER locus + PC1 SNPs that are significant in logistic regression model (refer to table 1 and supplementary table 5). The distributions of GRS are shown with violin plots. The width represents the density of points. The maximum widths are normalised across categories for clarity. *data used in GWAS.


a
0.10
0.05
0.00
-0.05
PC2 (17.1%)
0.00
0.05
0.10
0.15
PC1 (39.2%)
Source
Non-endemic
HK control
HK NPC
EBV Type
Type 1
Type 2
Intertypic
b
c
1.00
0.75
0.50
0.25
FST
0
20
40
60
80
100
120
140
160
Genomic Position (kb)
Comparison
Difference
HK NPC
HK control


A
Control Saliva
NPC Biopsy
B
Disease
Control
NPC
Subgroup
1A
1B
1C
2A
2B
PC2 (24.36%)
PC1 (30.08%)
10
5
0
-5
-10
0
10
20
30
C
PC1
30
20
10
0
Subgroup
1A
1B
1C
2A
2B
D
Logistic -log$_{10}$(P)
7.5
5.0
2.5
0.0
1
2
3
5
7
Principal Components



	Coordinates																													
	5399	5850	6484	6584	6866	6884	6886	6911	6944	6999	7001	7012	7016	7048	7122	7123	7125	7134	7188	7189	7190	7191	7198	7206	7213	7233	7262	7297	7327	8568
NPC Control	G	A	C	A	G	G	T	A	G	T	A	A	A	A	T	A	T	G	A	A	C	T	C	T	G	G	G	C	T	A
1.6% 56.3%	.	.	.	.	.	A	G	G	A	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.
	.	.	.	.	.	A	G	G	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
93.5% 38.7%	A	T	T	G	A	.	.	.	.	G	T	G	T	C	-	-	G	C	-	-	-	-	T	A	C	A	A	T	C	T


Genetic risk score
0.20
0.15
0.10
0.05
0.00
HK Control* (n=142)
HK NPC* (n=62)
Non-endemic (n=97)
Chinese NPC cell line/biopsy/saliva (n=6)
Chinese gastric cancer (n=12)
Chinese lung cancer (n=4)
0.15
0.10
0.05
0.00
HK Control* (n=142)
HK NPC* (n=62)
Non-endemic (n=97)
Chinese NPC cell line/biopsy/saliva (n=6)
Chinese gastric cancer (n=12)
Chinese lung cancer (n=4)
Diseases

Genetic variation in Epstein-Barr virus appears to play an important role in nasopharyngeal cancer (NPC). These authors conducted a case-control study in Hong Kong, where NPC incidence is high. They compared EBV genomes from NPC patients with those from population carriers. Population carriers harbored five different EBV types, they found, while only two showed up in tumor samples. A genome-wide association study identified a frameshift deletion in the EBER locus, which occurred in 97% of NPC cases and 40% of population carriers. More research into the geographic distribution of EBV variants could help explain why NPC incidence varies among populations.