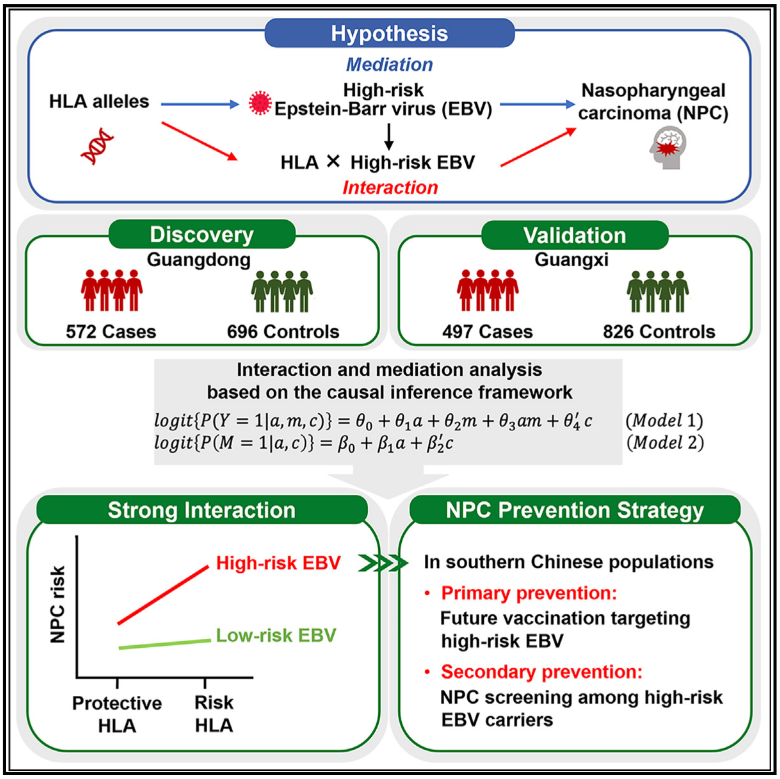
Cell Genomics

Article

Host genetic variants, Epstein-Barr virus subtypes, and the risk of nasopharyngeal carcinoma: Assessment of interaction and mediation

Graphical abstract



Highlights

- A causal inference framework applied to disentangle interaction and mediation effects
- The high-risk EBV subtype and HLA alleles jointly determine the majority of NPC risk
- The mediated effect through favoring the increased frequency of high-risk EBV is weak
- Targeting high-risk EBV could largely reduce NPC risk in southern China

Authors

Miao Xu, Ruimei Feng, Zhonghua Liu, ..., Hans-Olov Adami, Yi-Xin Zeng, Xihong Lin

Correspondence

ywm@fjmu.edu.cn (W.Y.),
hadami@hsph.harvard.edu (H.-O.A.),
zengyx@sysucc.org.cn (Y.-X.Z.),
xlin@hsph.harvard.edu (X.L.)

In brief

Xu et al. revealed that the risk of nasopharyngeal carcinoma (NPC) associated with HLA variants depended on infection of a high-risk EBV subtype, indicating that most NPC risk could theoretically be eliminated by intervening against infection of the high-risk EBV and by routine NPC screening among high-risk EBV carriers for early detection.

Xu et al., 2024, Cell Genomics *4*, 100474
February 14, 2024 © 2024 The Authors.
https://doi.org/10.1016/j.xgen.2023.100474

Cell Genomics

CellPress
OPEN ACCESS

Article

Host genetic variants, Epstein-Barr virus subtypes, and the risk of nasopharyngeal carcinoma: Assessment of interaction and mediation

Miao Xu,[1,2,19] Ruimei Feng,[3,19] Zhonghua Liu,[4,19] Xiang Zhou,[1,5,19] Yanhong Chen,[1] Yulu Cao,[1] Linda Valeri,[4,6] Zilin Li,[2,7] Zhiwei Liu,[8] Su-Mei Cao,[1] Qing Liu,[1] Shang-Hang Xie,[1] Ellen T. Chang,[9,10] Wei-Hua Jia,[1] Jincheng Shen,[11] Youyuan Yao,[12] Yong-Lin Cai,[13] Yuming Zheng,[13] Zhe Zhang,[14] Guangwu Huang,[14] Ingemar Ernberg,[15] Minzhong Tang,[14] Weimin Ye,[16,17,*] Hans-Olov Adami,[6,16,18,*] Yi-Xin Zeng,[1,*] and Xihong Lin[2,20,*]

[1]State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China
[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
[3]Department of Epidemiology, School of Public Health, Shanxi Medical University, Taiyuan 030012, Shanxi, China
[4]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA
[5]Prenatal Diagnostic Center, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangdong, China
[6]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
[7]School of Mathematics and Statistics, Northeast Normal University, Changchun, China
[8]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA
[9]Center for Health Sciences, Menlo Park, CA, USA
[10]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA
[11]Department of Population Health Sciences, University of Utah, Salt Lake City, UT, USA
[12]Department of Geriatric Oncology, Jiangsu Province Hospital, The First Affiliated Hospital with Nanjing Medical University, Nanjing, China
[13]Guangxi Health Commission Key Laboratory of Molecular Epidemiology of Nasopharyngeal Carcinoma, Wuzhou Red Cross Hospital, Wuzhou, China
[14]Department of Otolaryngology/Head and Neck Surgery, First Affiliated Hospital of Guangxi Medical University, Nanning, China
[15]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden
[16]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[17]Department of Epidemiology and Health Statistics & Key Laboratory of Ministry of Education for Gastrointestinal Cancer, Fujian Medical University, Fuzhou, China
[18]Clinical Effectiveness Group, Institute of Health and Society, University of Oslo, Oslo, Norway
[19]These authors contributed equally
[20]Lead contact
*Correspondence: ywm@fjmu.edu.cn (W.Y.), hadami@hsph.harvard.edu (H.-O.A.), zengyx@sysucc.org.cn (Y.-X.Z.), xlin@hsph.harvard.edu (X.L.)

https://doi.org/10.1016/j.xgen.2023.100474

SUMMARY

Epstein-Barr virus (EBV) and human leukocyte antigen (*HLA*) polymorphisms are well-known risk factors for nasopharyngeal carcinoma (NPC). However, the combined effects between *HLA* and EBV on the risk of NPC are unknown. We applied a causal inference framework to disentangle interaction and mediation effects between two host *HLA* SNPs, rs2860580 and rs2894207, and EBV variant 163364 with a population-based case-control study in NPC-endemic southern China. We discovered the strong interaction effects between the high-risk EBV subtype and both *HLA* SNPs on NPC risk (rs2860580, relative excess risk due to interaction [RERI] = 4.08, 95% confidence interval [CI] = 2.03–6.14; rs2894207, RERI = 3.37, 95% CI = 1.59–5.15), accounting for the majority of genetic risk effects. These results indicate that *HLA* genes and the high-risk EBV have joint effects on NPC risk. Prevention strategies targeting the high-risk EBV subtype would largely reduce NPC risk associated with EBV and host genetic susceptibility.

INTRODUCTION

Although nasopharyngeal carcinoma (NPC) is rare in most parts of the world, it is one of the most common cancers in southern China.[1] Epstein-Barr virus (EBV) has long been postulated to be a near-necessary factor for NPC development because it is present in the tumor cells of almost all patients with NPC, and it is the basis for serologic viral antibody and DNA tests that are widely used for screening and early diagnosis of NPC in high-risk populations.[2–4] Recent studies identified that the EBV subtype that carries the non-synonymous variant at position 163364, encoding the V317M mutation in EBV BALF2 protein, significantly contributes to the overall risk of NPC ($p = 2.40E-32$, odds ratio [OR] = 6.14), and its distribution is strongly associated with the unique epidemic of NPC in southern China.[5] However, because EBV infection is common and NPC is rare, it is widely accepted that other host genetic or environmental factors are also important determinants of NPC risk. Previous studies implicated that both host genetic, including the SNPs in the human leukocyte antigen (*HLA*), *TERT*, *CDKN2A/2B*, *TNFRSF19*, *MECOM*, *CIITA*, and *ITAG9* regions,[6–9] and environmental factors, including cigarette smoking, consumption of salt-preserved fish (a traditional Cantonese food suggested to be an NPC risk factor), and occupational exposures to wood dust,[1] may play a role in NPC development.

Among the genetic factors, *HLA* genes have the most consistent and prominent evidence for the association with NPC. Previous genome-wide association studies (GWASs) have identified two host *HLA* genetic variants, rs2860580 ($p_{GWAS} = 1.34E-28$, OR = 1.72) and rs2894207 ($p_{GWAS} = 1.22E-16$, OR = 1.64), to be strongly associated with NPC.[6–8] Given the *a priori* knowledge of the central role of *HLA* in host immune response against virus and the most prominent associations of *HLA* genes and the high-risk EBV subtypes with NPC, it is postulated that *HLA*-mediated pathways may cooperate with EBV in NPC development.[1] However, genetic and epidemiological evidence is lacking on how *HLA* genes and EBV act collaboratively to cause NPC. Without this knowledge, the clinical and public health utility of the genetic findings is limited.

Specifically, one possible causal pathway is that certain *HLA* genes may increase individuals' vulnerability to the oncogenic effect of high-risk EBV infection and that they synergistically influence the risk of NPC, a form of gene-EBV interaction. The other possible pathway is that the association of *HLA* genes with NPC risk may be mediated through increasing susceptibility to high-risk EBV infection, a form of mediation effect through EBV. Hence, a comprehensive and quantitative assessment of *HLA*-EBV genetic interplay is required to provide novel insights into the etiology of NPC and may inform the design of effective prevention strategies. Traditional methods to disentangle these possible pathways have been limited by the lack of adequate methods to accommodate mediation and interaction effects within a single framework and inadequate handling of case-control data and confounders.

Here, in a population-based case-control study of NPC conducted in NPC-endemic areas of southern China, we applied an advanced causal inference framework[10–13] to study whether the effects of two important host *HLA* variants (rs2860580 and rs2894207) on NPC are mediated by the EBV variant 163364, as well as their interactions. The causal inference method we applied accommodates gene-EBV interaction and mediation and allows us to disentangle mediation and interaction effects within one framework. To ensure the validity and reproducibility of this study,[14] we conducted mediation and interaction analyses using a two-phase design: an original study followed by a replication study, where we replicated the novel findings of the significant interactions between the two host *HLA* variants and the EBV variant from analyses of the original study with an independent, non-overlapping dataset. To highlight the public health implication, we further evaluated how much the genetic effects of the two *HLA* variants on NPC that are mediated by EBV, or are due to their interaction, can theoretically be eliminated by prevention strategies targeting individuals infected with the high-risk EBV subtype.

RESULTS

Study population characteristics and associations with host and EBV variants

We performed the causal inference analyses in a population-based case-control study conducted in two provinces, Guangdong and Guangxi, in NPC-endemic areas of southern China. The case-control data were frequency matched on the following variables, sex, 5-year age group, and area of residence, as described in the STAR Methods. Data exclusion criteria are outlined in Figure S1. Briefly, two human SNPs at the *HLA* locus, rs2860580 and rs2894207, and EBV variant 163364 were genotyped from saliva DNA. The success rate for genotyping both human SNPs was 98.7% (3,906/3,956, 1,683 affected individuals and 2,223 control subjects), whereas the genotyping rate for the EBV variant was 66.9% (2,648/3,956, 1,098 affected individuals and 1,550 control subjects). Hence, the success rate for EBV genotyping is likely dictated by the quantity of EBV DNA in saliva. As shown in the literature, periodic lytic EBV production in the oral epithelium is hypothesized to be the main source of the likely random fluctuation observed for saliva EBV DNA.[15–18] The missingness of EBV genotyping data did not differ by age, the consumption of salt-preserved fish, educational level, rural or urban area of residence, current occupation, selected environmental exposures, or a family history of NPC, and successful EBV genotyping was not correlated with increased NPC risk (Figure S2). A lower missing rate of EBV variant was found among smokers and men (who were substantially more likely than women to be smokers) (Figure S2B). This pattern is concordant with the observation that smoking stimulates EBV lytic production, which increases the chance of EBV being genotyped.[19,20] However, because the relative risk and attributable risk of NPC associated with smoking are relatively small, with a relative risk of only 1.1–1.5[19] compared to a relative risk of 6–7 associated with the high-risk EBV subtype,[5] any bias caused by smoking in our dataset would be small. Additionally, sex, age at interview, a family history of NPC, salt-preserved fish consumption, smoking, educational level, rural or urban area of residence, current occupation, and environmental exposure were included as covariates in the logistic regression models for the interaction and mediation analyses to control for confounding in the following causal inference analyses (STAR Methods).

Affected individuals and control subjects recruited from Guangdong were used in the original study, while those recruited from Guangxi were used for the replication study. Table 1 summarizes the demographic characteristics of the original and replication study subjects. The original and replication studies had similar distributions for the case-control status, sex, and age. In the pooled dataset, affected individuals were slightly younger than control subjects and were more likely to live in urban areas, to have a first-degree family history of NPC, to be less educated, to have blue-collar jobs, and to be exposed to selected hazardous agents (mostly inhalants; Table 1).

Table 1. Characteristics of nasopharyngeal carcinoma affected individuals and control subjects among the original, replication, and pooled studies

	Original study			Replication study			Pooled study		
Variables	Affected individuals n = 572 n (%)	Control subjects n = 696 n (%)	p	Affected individuals n = 497 n (%)	Control subjects n = 826 n (%)	p	Affected individuals n = 1,069 n (%)	Control subjects n = 1,522n (%)	p
Age, years	–	–	0.103	–	–	0.002	–	–	2.2E−4
Mean (SD)	48.6 (10.9)	49.6 (10.7)	0.108	49.2 (10.8)	51.1 (10.9)	0.003	48.9 (10.8)	50.4 (10.8)	3.9E−4
<35	58 (10.1)	52 (7.5)	–	45 (9.1)	53 (6.4)	–	103 (9.6)	105 (6.9)	–
35–59	424 (74.1)	512 (73.6)	–	366 (73.6)	568 (68.8)	–	790 (73.9)	1,080 (71.0)	–
>59	90 (15.7)	132 (19.0)	–	86 (17.3)	205 (24.8)	–	176 (16.5)	337 (22.1)	–
Sex	–	–	0.879	–	–	0.928	–	–	0.819
Male	431 (75.4)	527 (75.7)	–	381 (76.7)	635 (76.9)	–	812 (76.0)	1,162 (76.4)	–
Female	141 (24.7)	169 (24.3)	–	116 (23.3)	191 (23.1)	–	257 (24.0)	360 (23.7)	–
Education level, years	–	–	0.117	–	–	0.102	–	–	0.021
<7	229 (40.0)	247 (35.5)	–	216 (43.5)	310 (37.5)	–	445 (41.6)	557 (36.6)	–
7–9	233 (40.7)	286 (41.1)	–	179 (36.0)	327 (39.6)	–	412 (38.5)	613 (40.3)	–
≥10	110 (19.2)	163 (23.4)	–	102 (20.5)	189 (22.9)	–	212 (19.8)	352 (23.1)	–
Residential area	–	–	4.6E−6	–	–	0.251	–	–	2.9E−5
Urban	96 (16.8)	58 (8.3)	–	65 (13.1)	89 (10.8)	–	161 (13.1)	147 (10.8)	–
Rural	476 (83.2)	638 (91.7)	–	432 (86.9)	737 (89.2)	–	908 (86.9)	1,375 (89.2)	–
Salt-preserved fish consumption in 2000–2002	–	–	0.186	–	–	0.027	–	–	0.119
Yearly or less	390 (68.2)	450 (64.7)	–	431 (86.7)	678 (82.1)	–	821 (76.8)	1,128 (74.1)	–
Monthly or more	182 (31.8)	246 (35.3)	–	66 (13.3)	148 (17.9)	–	248 (23.2)	394 (25.9)	–
NPC history among first-degree relatives	–	–	1.9E−8	–	–	7.3E−5	–	–	2.4E−13
No	487 (85.1)	653 (93.8)	–	449 (90.3)	790 (95.6)	–	936 (87.6)	1,443 (94.8)	–
Yes	73 (12.8)	26 (3.7)	–	39 (7.9)	22 (2.7)	–	112 (10.5)	48 (3.2)	–
Unknown/missing	12 (2.1)	17 (2.4)	–	9 (1.8)	14 (1.7)	–	21 (2.0)	31 (2.0)	–
Smoking status	–	–	0.732	–	–	0.785	–	–	0.864
Never	224 (39.2)	266 (38.2)	–	217 (43.7)	367 (44.4)	–	441 (41.3)	633 (41.6)	–
Ever	348 (60.8)	430 (61.8)	–	280 (56.3)	459 (55.6)	–	628 (58.8)	889 (58.4)	–
Current occupation	–	–	3.9E−12	–	–	5.5E−4	–	–	1.0E−5
Farmer/Unemployment	208 (36.4)	232 (33.3)	–	192 (38.6)	388 (47.0)	–	400 (37.4)	620 (40.7)	–
Blue collar	236 (41.3)	213 (30.6)	–	190 (38.2)	265 (32.1)	–	426 (39.9)	478 (31.4)	–
White collar	89 (15.6)	102 (14.7)	–	59 (11.9)	118 (14.3)	–	148 (13.8)	220 (14.5)	–
Unknown	39 (6.8)	149 (21.4)	–	56 (11.3)	55 (6.7)	–	95 (8.9)	204 (13.4)	–

(*Continued on next page*)

Table 1. ***Continued***

Variables	Original study			Replication study			Pooled study		
	Affected individuals n = 572 n (%)	Control subjects n = 696 n (%)	p	Affected individuals n = 497 n (%)	Control subjects n = 826 n (%)	p	Affected individuals n = 1,069 n (%)	Control subjects n = 1,522n (%)	p
Selected environmental exposure[a]	–	–	6.5E−19	–	–	0.029	–	–	1.3E−16
None	58 (10.1)	82 (11.8)	–	35 (7.0)	85 (10.3)	–	93 (8.7)	167 (11.0)	–
Dust exposure	226 (39.5)	217 (31.2)	–	246 (49.5)	326 (39.5)	–	472 (44.2)	543 (35.7)	–
Smoke/exhaust exposure	107 (18.7)	91 (13.1)	–	125 (25.2)	167 (20.2)	–	232 (21.7)	258 (17.0)	–
Other exposure	175 (30.6)	193 (27.7)	–	85 (17.1)	245 (29.7)	–	260 (24.3)	438 (28.8)	–
Unknown/missing	6 (1.1)	113 (16.2)	–	6 (1.2)	3 (0.4)	–	12 (1.1)	116 (7.6)	–
EBV infection determined by variant 163364	–	–	0.693	–	–	0.692	–	–	0.890
Multiple strains	16 (2.8)	17 (2.4)	–	16 (3.2)	30 (3.6)	–	32 (3.0)	47 (3.1)	–
Single strain	556 (97.2)	679 (97.6)	–	481 (96.8)	796 (96.4)	–	1,037 (97.0)	1475 (96.9)	–

Multiple strains represent the genotype CT for EBV variant 163364; single strain represents the genotype C or T for the same variant.

[a] Dust exposure includes exposure to wood, metal, textile, leather, cement, and other types of non-soil dust. Smoke/exhaust exposure includes exposure to exhaust of diesel, gasoline, coal, firewood, asphalt/tar, nature gas, and other types of exhaust/smoke. Other environmental exposure includes exposure to wood preservatives, formaldehyde, organic solvents, pesticides, and other types of chemical vapor, as well as sulfuric acid, hydrochloride, and other types of acid/alkali.

Table 2. Association between two host genetic variants or EBV variant 163364 and risk of nasopharyngeal carcinoma

	Affected individuals n (%)	Control subjects n (%)	OR (95% CI)[a]
Original study in Guangdong			
rs2860580 (risk allele = G)			
AA/AG	242 (42.3)	402 (57.8)	reference
GG	330 (57.7)	294 (42.2)	1.93 (1.52, 2.44)
rs2894207 (risk allele = T)			
CC/CT	148 (25.9)	266 (38.2)	reference
TT	424 (74.1)	430 (61.8)	1.70 (1.32, 2.19)
Joint status of rs2860580 and rs2894207[b]			
Low risk	303 (53.0)	493 (70.8)	reference
High risk	269 (47.0)	203 (29.2)	2.24 (1.75, 2.85)
EBV 163364 (high-risk subtype = T)			
C	83 (14.5)	374 (53.7)	reference
CT/T	489 (85.5)	322 (46.3)	6.99 (5.25, 9.31)
Replication study in Guangxi			
rs2860580 (risk allele = G)			
AA/AG	213 (42.9)	459 (55.6)	reference
GG	284 (57.1)	367 (44.4)	1.73 (1.38, 2.18)
rs2894207 (risk allele = T)			
CC/CT	111 (22.3)	269 (32.6)	reference
TT	386 (77.7)	557 (67.4)	1.74 (1.34, 2.26)
Joint status of rs2860580 and rs2894207[b]			
Low risk	256 (51.5)	561 (67.9)	reference
High risk	241 (48.5)	265 (32.1)	2.12 (1.68, 2.69)
EBV 163364 (high-risk subtype = T)			
C	122 (24.6)	563 (68.2)	reference
CT/T	375 (75.5)	263 (31.8)	6.55 (5.07, 8.46)
Pooled study			
rs2860580 (risk allele = G)			
AA/AG	455 (42.6)	861 (56.6)	reference
GG	614 (57.4)	661 (43.4)	1.80 (1.53, 2.11)
rs2894207 (risk allele = T)			
CC/CT	259 (24.2)	535 (35.2)	reference
TT	810 (75.8)	987 (64.9)	1.68 (1.41, 2.01)
Joint status of rs2860580 and rs2894207[b]			
Low risk	559 (52.3)	1,054 (69.3)	reference
High risk	510 (47.7)	468 (30.8)	2.12 (1.79, 2.50)
EBV 163364 (high-risk subtype = T)			
C	205 (19.2)	937 (61.6)	reference
CT/T	864 (80.8)	585 (38.4)	6.86 (5.68, 8.27)

[a] Adjusted for age at interview, sex and smoking joint status, education level, salt-preserved fish consumption in 2000–2002, NPC history among first-degree relatives, rural or urban area of residence, current occupation, and environmental exposure.

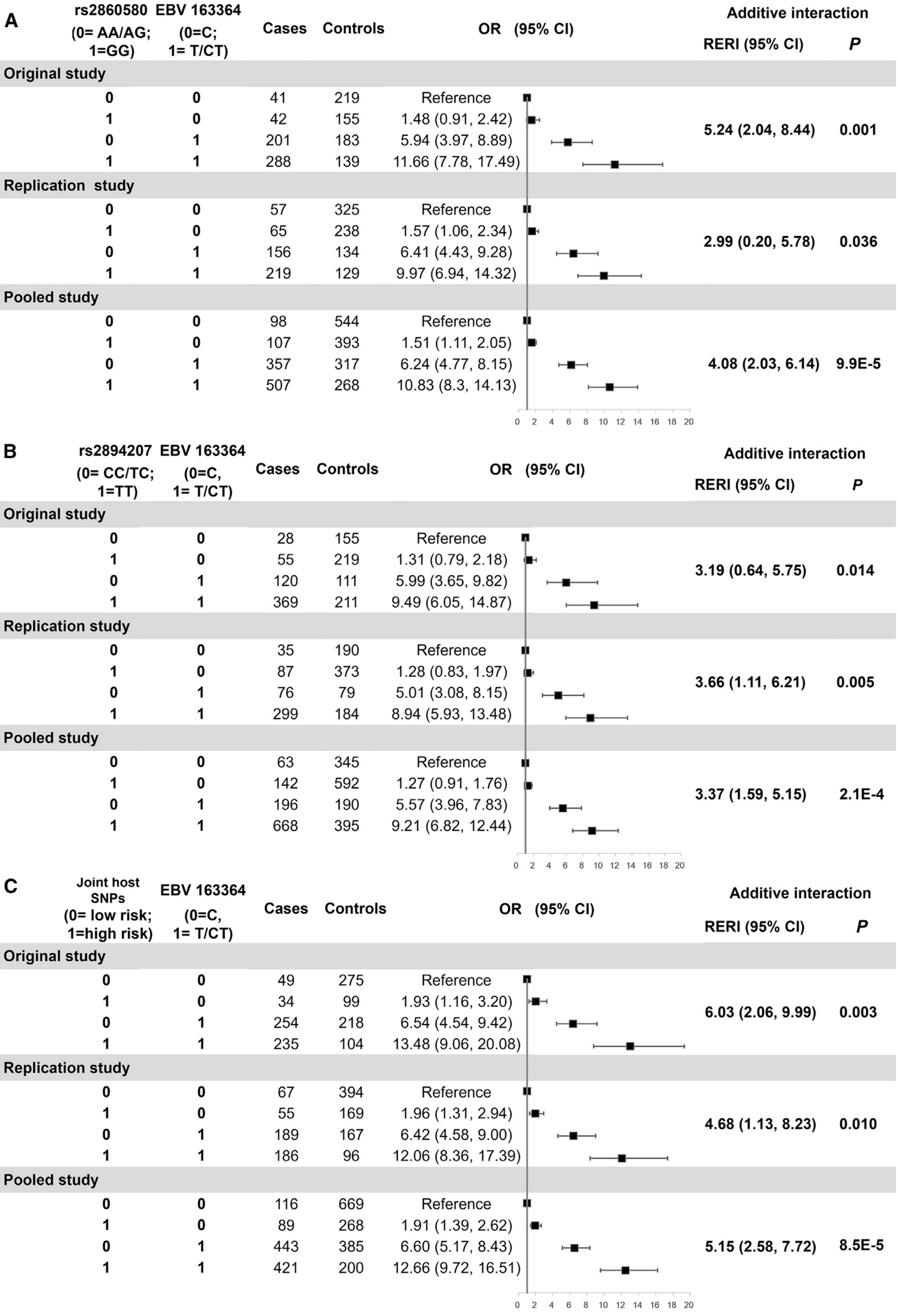
[b] The joint status of high-risk group: GG for rs2860580 and TT for rs2894207; the joint status of low-risk group: AA/AG for rs2860580 or CC/CT for rs2894207.

Table 2 displays the effects of the two host SNPs, rs2860580 and rs2894207, their joint status, and EBV variant 163364 on NPC risk in the original, replication, and pooled studies. No deviation from Hardy-Weinberg equilibrium was detected for the two human SNPs. The two human SNPs were independent of each other with weak linkage disequilibrium ($R^2 = 0.01$) among the control subjects in this study, consistent with previous GWASs.[6] In the pooled study, individuals carrying only the risk allele of rs2860580, only the risk allele of rs2894207, and only the risk alleles of both host SNPs had an increased NPC risk by 1.80-, 1.68-, and 2.12-fold, respectively, compared to the individuals in the reference group indicated in Table 2; the effect associated with per risk alleles was 1.68 (95% confidence interval [CI] = 1.47–1.91) and 1.65 (95% CI = 1.40–1.93) for rs2860580 and rs2894207, respectively, consistent with the published GWAS results (Table S1). EBV variant 163364 was associated with 6.86-fold increased risk of NPC in both datasets (Table 2).

Interaction effects between host SNPs and EBV variant

Figure 1 shows ORs, additive interactions, and their 95% CIs for the joint effects of host *HLA* SNPs rs2860580 and rs2894207 and EBV variant 163364 on NPC risk. Compared to the subjects carrying both protective alleles (AA/AG) of rs2860580 and the low-risk EBV variant (C), those carrying only the susceptible alleles (GG) of rs2860580, only the high-risk EBV variant (CT/T), or both had approximately 1.5-, 6-, and 11-fold increased risk, respectively, in the original and the replication studies (Figure 1A). Similarly, joint ORs and 95% CIs are shown for rs2894207 and EBV variant 163364 (Figure 1B). Importantly, significant additive interactions were observed between the two host SNPs, rs2860580 and rs2894207, and EBV variant 163364 in both the original and replication studies (Figure 1). In the pooled study, the total effect of carrying only host-susceptible *HLA* alleles and the high-risk EBV subtype due to interaction (relative excess risk due to interaction [RERI]) was 4.08 (95% CI = 2.03–6.14) for rs2860580 and 3.37 (95% CI = 1.59–5.15) for rs2894207 (Figure 1). Furthermore, in the interaction analyses, we combined the two host genetic variants into one categorical variable by their joint status, which divided the study subjects into two groups: one group carrying only risk alleles of both host SNPs and the other group carrying protective alleles of either host SNP. We identified significant and stronger interaction effects between the joint status of the two host SNPs and EBV subtypes in both the original study (RERI = 6.03, 95% CI = 2.06–9.99) and the replication study (RERI = 4.68, 95% CI = 1.13–8.23) than analyzing the two host SNPs separately (Figure 1C).

To evaluate the impact of individual alleles of the two variants and produce OR for each risk allele, we performed the interaction analyses using the additive model that compares the effects of carrying homozygous versus heterozygous versus no risk alleles. The interaction effects with EBV variant 163364 on NPC were statistically significant in both the original and replication studies and were 3.08 (95% CI = 1.79–4.37) and 2.57 (95% CI = 1.24–3.90) per risk allele for rs2860580 and rs2894207, respectively, in the pooled study (Figure S3). These results further showed that the findings of highly significant interaction



A

rs2860580 (0= AA/AG; 1=GG)	EBV 163364 (0=C; 1= T/CT)	Cases	Controls	OR (95% CI)	Additive interaction RERI (95% CI)	*P*
Original study						
0	0	41	219	Reference	5.24 (2.04, 8.44)	0.001
1	0	42	155	1.48 (0.91, 2.42)		
0	1	201	183	5.94 (3.97, 8.89)		
1	1	288	139	11.66 (7.78, 17.49)		
Replication study						
0	0	57	325	Reference	2.99 (0.20, 5.78)	0.036
1	0	65	238	1.57 (1.06, 2.34)		
0	1	156	134	6.41 (4.43, 9.28)		
1	1	219	129	9.97 (6.94, 14.32)		
Pooled study						
0	0	98	544	Reference	4.08 (2.03, 6.14)	9.9E-5
1	0	107	393	1.51 (1.11, 2.05)		
0	1	357	317	6.24 (4.77, 8.15)		
1	1	507	268	10.83 (8.3, 14.13)		

B

rs2894207 (0= CC/TC; 1=TT)	EBV 163364 (0=C, 1= T/CT)	Cases	Controls	OR (95% CI)	Additive interaction RERI (95% CI)	*P*
Original study						
0	0	28	155	Reference	3.19 (0.64, 5.75)	0.014
1	0	55	219	1.31 (0.79, 2.18)		
0	1	120	111	5.99 (3.65, 9.82)		
1	1	369	211	9.49 (6.05, 14.87)		
Replication study						
0	0	35	190	Reference	3.66 (1.11, 6.21)	0.005
1	0	87	373	1.28 (0.83, 1.97)		
0	1	76	79	5.01 (3.08, 8.15)		
1	1	299	184	8.94 (5.93, 13.48)		
Pooled study						
0	0	63	345	Reference	3.37 (1.59, 5.15)	2.1E-4
1	0	142	592	1.27 (0.91, 1.76)		
0	1	196	190	5.57 (3.96, 7.83)		
1	1	668	395	9.21 (6.82, 12.44)		

C

Joint host SNPs (0= low risk; 1=high risk)	EBV 163364 (0=C, 1= T/CT)	Cases	Controls	OR (95% CI)	Additive interaction RERI (95% CI)	*P*
Original study						
0	0	49	275	Reference	6.03 (2.06, 9.99)	0.003
1	0	34	99	1.93 (1.16, 3.20)		
0	1	254	218	6.54 (4.54, 9.42)		
1	1	235	104	13.48 (9.06, 20.08)		
Replication study						
0	0	67	394	Reference	4.68 (1.13, 8.23)	0.010
1	0	55	169	1.96 (1.31, 2.94)		
0	1	189	167	6.42 (4.58, 9.00)		
1	1	186	96	12.06 (8.36, 17.39)		
Pooled study						
0	0	116	669	Reference	5.15 (2.58, 7.72)	8.5E-5
1	0	89	268	1.91 (1.39, 2.62)		
0	1	443	385	6.60 (5.17, 8.43)		
1	1	421	200	12.66 (9.72, 16.51)		

(legend on next page)

effects between EBV and the two human *HLA* genetic variants were robust to the assumed genetic models, i.e., both the additive model and the recessive model (comparing the individuals carrying only susceptible alleles versus those carrying protective alleles). Taken together, these coherent results indicate that host-susceptible *HLA* genes and the high-risk EBV subtype have synergistic interaction effects on the risk of NPC.

Genetic effects mediated by high-risk EBV subtype

The associations between the two host *HLA* SNPs, rs2860580 and rs2894207, and EBV variant 163364 (Table S2) indicate that the genetic effects might be mediated through increasing the frequency of high-risk EBV subtype. We applied causal mediation analyses, allowing for interaction between host SNPs and EBV (STAR Methods). The mediation effects (indirect effects) for NPC risk through EBV variant 163364 and direct effects are shown in Table 3. Both the original and replication studies, as well as the pooled study, revealed significant direct effects and small, statistically non-significant or weakly significant indirect effects. The direct effect in the pooled dataset was estimated as ORs 1.69 (95% CI = 1.40–2.03) for rs2860580 and 1.56 (95% CI = 1.27–1.90) for rs2894207, whereas the indirect effect was close to 1 (OR = 1.07, 95% CI = 0.98–1.17 for rs2860580; OR = 1.10, 95% CI = 1.00–1.21 for rs2894207). When we combined the two host SNPs as one categorical variable by their joint status and compared the individuals carrying only risk alleles of both host SNPs to those carrying protective alleles of either host SNP, the mediation effect through EBV subtypes (indirect effect) became significant, albeit weak, in the pooled analysis at a 5% significance level (indirect effect, OR = 1.12, 95% CI = 1.02–1.23; Table 3), possibly due to the stronger genetic effects of combining two SNPs together (Table 2).

We further used the additive model to assess the mediation effects per risk allele of each *HLA* SNP. Similarly, the direct effects of both host *HLA* SNPs were highly significant in both the original and replication studies, while the indirect effects through high-risk EBV variant 163364 were statistically non-significant for rs2860580 (indirect effect, OR = 1.05, 95% CI = 0.98–1.22) and became significant for rs2894207 at a 5% significance level (indirect effect, OR = 1.12, 95% CI = 1.03–1.21; Table S3) in the pooled analysis. In accordance with the results using the recessive model (comparing the individuals carrying only susceptible alleles versus those carrying protective alleles; Table 3), the effect sizes using the additive model per risk allele on NPC through increasing the frequency of the high-risk EBV subtype (Table S3; indirect effect, ORs = 1.05 and 1.12 for rs2860580 and rs2894207, respectively) were small and could not explain the majority of genetic effects for the two *HLA* SNPs or the EBV variant on NPC. Taken together, these results indicate that the majority of the effects of the host SNPs rs2894207 and rs2860580 on NPC risk might not be mediated by the high-risk EBV subtype.

NPC risk attributable to the high-risk EBV subtype

Evaluating the interaction between host genes and EBV subtypes enables quantification of attributable risk, that is, the potential beneficial impact of preventing infection with the high-risk EBV subtype. Therefore, we applied four-way decomposition to evaluate the proportion of NPC risk due to host-virus interaction and mediation that can be reduced or eliminated by intervention against high-risk EBV. The four-way decomposition method separates the excess relative risk of NPC due to host genetic effects into four parts involving interaction, mediation, both, or neither (Figure 2). The decomposition analysis showed that the excess relative risk of NPC due to pure interaction between *HLA* SNPs and EBV variant 163364 was significant in the pooled dataset (rs2860580: reference interaction = 0.47, 95% CI = 0.21–0.73; rs2894207: reference interaction = 0.40, 95% CI = 0.15–0.65) and accounted for the largest proportion for both SNPs (Figures 2A and 2B). In the pooled dataset, interaction with the high-risk EBV (reference interaction + mediated interaction) accounted for 66.0% of the total excess risk associated with SNP rs2860580 and 69.2% with rs2894207, comparing individuals carrying only susceptible *HLA* alleles with those carrying protective alleles. The association of NPC risk with both SNPs mediated through the high-risk EBV (mediated interaction + pure indirect effect) was non-significant in both the original and replication studies (Figures 2A and 2B). When we combined the two *HLA* SNPs in one model and compared the individuals carrying only risk alleles of both host SNPs to those carrying protective alleles of either host SNP, the excess relative risk of NPC due to the pure interaction effect between *HLA* SNPs and EBV variant 163364 became even stronger (reference interaction = 0.57, 95% CI = 0.26–0.87 in the pooled dataset; Figure 2C).

Furthermore, by combining the effects of host-virus interaction and mediation (reference interaction + mediated interaction + pure indirect effect) in the pooled dataset, we found that 74.5% and 82.7% of the total excess relative risk associated with carrying only the susceptible alleles of rs2860580 and rs2894207, respectively, can potentially be eliminated by preventing high-risk EBV infection (Figures 2A and 2B). Consistently, in the model using the joint status of the two SNPs, 69.9% of the total excess risk associated with carrying only susceptible alleles of both SNPs can potentially be eliminated by preventing high-risk EBV infection (Figure 2C). The remaining effects independent of high-risk EBV (i.e., the controlled direct effects) would be small (Figure 2).

Finally, the conclusion of the relatively small risk effects independent of high-risk EBV is robust under the additive model. Similarly, with the additive model (Figure S4), interaction with the high-risk EBV subtype accounted for 61.6% and 54.8%, the largest proportion, of the genetic effects of per risk allele on NPC risk for rs2860580 and rs2894207, respectively; 69.0% and 71.1% of the total excess relative risk associated per risk allele of rs2860580 and rs2894207, respectively, can potentially be eliminated by preventing high-risk EBV infection.

Figure 1. Joint effect and additive interaction between EBV variant 163364 and host HLA SNPs on the risk of nasopharyngeal carcinoma
(A) rs2860580, (B) rs2894207, and (C) their joint status. OR, odds ratio; CI, confidence interval; RERI, relative excess risk due to interaction. Two host SNPs were combined as one categorical variable in the models by their joint status, which divided the study subjects into two groups: one group at higher risk carrying only risk alleles of both host SNPs and the other group at lower risk carrying protective alleles of either host SNP.

Table 3. Direct and indirect effects on nasopharyngeal carcinoma between host SNPs rs2860580 and rs2894207 as well as their joint status and EBV variant 163364

	Effect	OR[a]	95% CI[a]	p
rs2860580 and EBV 163364				
Original study	natural direct effect	1.88	1.44, 2.46	3.4E−6
	natural indirect effect	1.02	0.90, 1.16	0.730
	marginal total effect	1.92	1.44, 2.58	1.2E−5
Replication study	natural direct effect	1.56	1.20, 2.02	8.0E−4
	natural indirect effect	1.13	0.99, 1.28	0.066
	marginal total effect	1.76	1.31, 2.35	1.5E−4
Pooled study	natural direct effect	1.69	1.40, 2.03	2.5E−8
	natural indirect effect	1.07	0.98, 1.17	0.157
	marginal total effect	1.80	1.47, 2.21	1.9E−8
rs2894270 and EBV 163364				
Original study	natural direct effect	1.53	1.16, 2.03	0.003
	natural indirect effect	1.13	0.99, 1.29	0.062
	marginal total effect	1.73	1.27, 2.36	5.0E−4
Replication study	natural direct effect	1.62	1.21, 2.17	0.001
	natural indirect effect	1.08	0.94, 1.25	0.294
	marginal total effect	1.75	1.27, 2.41	6.1E−4
Pooled study	natural direct effect	1.56	1.27, 1.90	1.5E−5
	natural indirect effect	1.10	1.00, 1.21	0.052
	marginal total effect	1.71	1.37, 2.14	1.7E−6
Joint status of host SNPs[b] and EBV 163364				
Original study	natural direct effect	2.04	1.55, 2.69	3.5E−7
	natural indirect effect	1.12	0.98, 1.27	0.088
	marginal total effect	2.28	1.68, 3.09	1.2E−7
Replication study	natural direct effect	1.90	1.46, 2.48	2.2E−6
	natural indirect effect	1.13	0.99, 1.29	0.067
	marginal total effect	2.15	1.59, 2.91	6.2E−7
Pooled study	natural direct effect	1.92	1.59, 2.31	1.3E−11
	natural indirect effect	1.12	1.02, 1.23	0.020
	marginal total effect	2.14	1.73, 2.64	2.2E−12

[a] Adjusted for age at interview, sex and smoking joint status, education level, salt-preserved fish consumption in 2000–2002, NPC history among first-degree relatives, rural or urban area of residence, current occupation, and environmental exposure.

[b] Two host SNPs were combined as one categorical variable in the models by their joint status, which divided the study subjects into two groups: one group at higher risk carrying only risk alleles of both host SNPs and the other group at lower risk carrying protective alleles of either host SNP.

Potential mechanistic interaction between EBV subtypes and *HLA* alleles

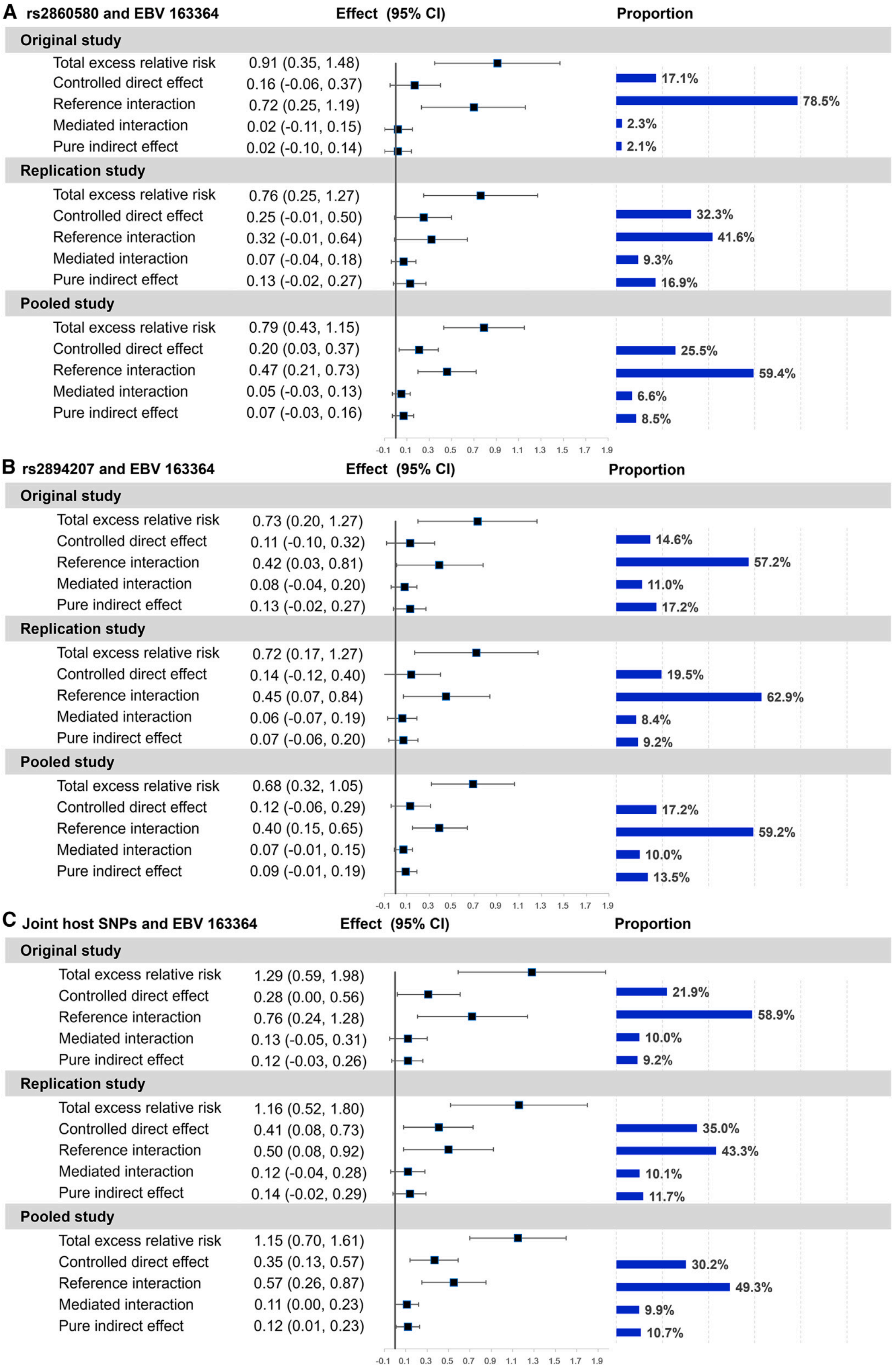
By contributing the majority of disease risk, the strong interaction effect we discovered indicates that NPC risk depends not only on EBV subtypes but also on the *HLA* alleles of the host. To explore the plausible mechanisms underlying the high-risk EBV subtype and its interaction with host *HLA* variants on NPC risk, we evaluated the impact of high-risk EBV variant 163364 on the protein function. We used AlphaFold2 to predict the structure of BALF2 proteins from the low-risk and high-risk EBV strains.[21,22] Interestingly, we observed that BALF2 amino acid 317, encoded by EBV SNP 163364, is located in the neck region, a presumed single-stranded DNA (ssDNA) binding pocket (Figure S5A). Within this pocket, two positively charged amino acids, D309 and E546 on the opposite side, form a gate structure, facilitating the movement of ssDNA through the binding pocket. Meanwhile, two negatively charged amino acids, R322 and K548, dock the ssDNA. With the presence of V317M mutation, encoded by the high-risk variant 163364, the long side chain of M drives a shift of α-helix (314–328) harboring M317 and its neighboring loop (308–313), narrowing the gate between D309 and E546, as well as the docking interface between R322 and K548 (Figures S5B–S5D). This narrowed gate and binding pocket due to the M317 high-risk variant may potentially affect DNA movement and viral DNA replication during the lytic cycle.

Furthermore, our findings underscore a substantial additional risk associated with the concurrent presence of the high-risk EBV subtype and susceptible *HLA* alleles. This association may be related to distinct *HLA*-mediated T cell immune responses to different EBV subtypes, suggesting a potential for immune evasion by the high-risk EBV subtype from susceptible *HLA* alleles. Among the *HLA* alleles associated with NPC in southern Chinese populations, *A*0207* is the most significantly associated risk allele.[9,23,24] Thus, we evaluated *HLA-A*0207* binding affinity for the nonamer peptide pairs from high-risk and low-risk EBV subtypes using NetMHCpan-4.1.[25] There is no difference in the predicted binding affinity for the high-risk peptide containing BALF2 variant 163364, but some other peptides associated with this variant and NPC risk have a lower binding affinity to *A*0207* in the high-risk strain (Figure S6). Therefore, EBV SNP 163364 may be a marker linked to yet still unidentified functional variation in the NPC high-risk strain.

DISCUSSION

Our analyses revealed significant additive interactions on NPC risk between the EBV subtype classified by variant 163364 and host SNPs rs2860580 at *HLA-A* and rs2894207 at *HLA-B/C* loci. The evidence was weaker for the EBV-mediated genetic effects associated with the two *HLA* SNPs. By decomposing the total effects of the host risk alleles into separate and joint effects of interaction and mediation by high-risk EBV subtype, we showed that interaction between host SNPs and EBV accounts for the majority of excess NPC risk conferred by the host genetic variants rs2860580 and rs2894207. Finally, we found that nearly three-quarters of the excess NPC risk attributable to both host *HLA* SNPs could in theory be eliminated by prevention of the infection with the high-risk EBV subtype through vaccination. Our study thus provides strong evidence of the critical interplay between human genetics and EBV subtype in the etiology of NPC and lays the groundwork for EBV-subtype-specific prevention for reducing the NPC burden in endemic populations.

A strong interaction between host genetic susceptibility and EBV subtypes indicates that the direct effects of the host variants on NPC risk may occur primarily among individuals carrying the high-risk EBV subtype. This risk model is supported by the fact that the ORs associated with rs2860580 and rs2894207 were 1–2 among individuals carrying the low-risk EBV subtype,



(legend on next page)

whereas the ORs associated with the two SNPs were 6–13 among individuals carrying the high-risk EBV subtype. The strong interaction highlights that the susceptible *HLA* alleles increase the risk associated with high-risk EBV infection for NPC, supporting potential immune evasion by the high-risk EBV subtype from susceptible *HLA* alleles. The association between *HLA* genes and NPC risk has been confirmed consistently in both candidate-gene studies and several independent GWASs, highlighting *HLA-A*1101* and *A*0207* as the most significantly associated genes in these investigations.[6,9,23,24,26] *HLA-A*1101* is the protective allele, while *A*0207* is the risk allele. The EBV EBNA-3B epitope IVTDFSVIK, restricted to *HLA-A*11*, has a high-frequency mutation (IVTDFSVIKN) among southern Chinese populations with a high *A*11* frequency. These mutations are thought to provide selective advantage in the highly *A*11*-positive populations.[27–29] Our findings, revealing an additional risk associated with the co-occurrence of NPC-high-risk EBV and *HLA-A*0207*, suggest that the high-risk EBV may carry the sequence variations correlated with reduced binding affinity for *A*0207*, potentially contributing to an elevated NPC risk among individuals carrying the *A*0207* allele. A recent study has also reported the trend of decreased *HLA-A*02* binding affinity with peptides harboring NPC-high-risk mutations.[30] Specifically, the LMP-1 YFLEILWRL mutant peptide, which shows association with NPC risk, has been reported to evade recognition by *A*02*-restricted epitope (YLLEMLWRL)-specific T cells (Figure S6; Table S4) and to fail to elicit T cell responses in patients with NPC.[31,32] Since EBV LMP-1 protein is among the few latent antigens expressed in NPC cells, it is plausible that NPC cells infected with the high-risk EBV subtype possess an enhanced ability to evade *HLA-A*0207*-mediated T cell immune surveillance. This scenario could further increase the NPC risk among individuals carrying the susceptible *HLA* allele and the high-risk EBV subtype. Extensive mapping of T cell epitopes of the high-risk EBV subtype is important for designing EBV vaccines and T cell therapies targeting NPC.

Furthermore, the NPC-derived EBV strain, M81, has been shown to exhibit an epitheliotropism and a high level of spontaneous replication in B cells.[33] These unique properties of M81 are thought to be related to its epithelial oncogenic potential and consistent with the observed increased viral replication preceding NPC onset. Polymorphisms within the NPC-high-risk EBV subtype, particularly in the transactivator protein BZLF1 and its promoter region, the non-coding RNA EBER2, as well as the gene structure of BALF5, have been reported to contribute to these properties.[33–38] Given the role of BALF2 as the ssDNA binding protein, an essential component of EBV DNA replication complex, the V317M mutation (variant 163364) in the BALF2 protein of high-risk EBV strains, could potentially influence the conformation of the ssDNA interaction surface, thereby altering its function during viral DNA replication. Functional analysis of these high-risk EBV variants would be indispensable to elucidate whether they might contribute the enhanced oncogenicity. Taken together, the distinct viral functional properties between NPC-high-risk and low-risk EBV, coupled with their interplay with *HLA* genetic factors, suggest that vaccine design aimed at NPC prevention should take into account the genetic variations within the high-risk EBV subtype.

In summary, our findings constitute strong epidemiological evidence for the joint interaction effect between host *HLA* genes and EBV subtypes on the risk of NPC, thereby providing an illuminating model of the interplay between critical host genetic factors and the virus in NPC carcinogenesis. Notably, the substantial contribution of the interaction with the high-risk EBV subtype to the genetic susceptibility associated with *HLA* SNPs in NPC suggests that a vaccine targeting high-risk EBV could significantly mitigate NPC risk associated with both viral and host genetic factors within the southern Chinese population. In this context, careful consideration of the genetic diversity specific to the NPC-high-risk EBV subtype is imperative for the development of vaccines aimed at NPC prevention. Moreover, given that both precancerous lesions and early NPC can be treated successfully, routine NPC screening would benefit early disease detection and treatment among individuals carrying the high-risk EBV subtype in endemic populations.

Limitations of the study

First, the weak mediation effect cannot be robustly detected with the current sample size. Using the simulation method proposed by Rudolph et al.,[39] we found that under the recessive genetic model, the necessary sample size with at least 80% power for detecting the mediation effect (natural indirect effect [NIE]) of 1.07 for the host SNP rs2860580 through EBV variant 163364 needs to be at least 2,390 affected individuals and 3,501 control subjects, more than double the current sample size (1,069 affected individuals and 1,522 control subjects). Under the additive model, the current sample size has the power to detect a modest mediation effect (NIE) of 1.12 for the host SNP rs2894207 but still lacks sufficient power to robustly detect a weak mediation effect (NIE) of 1.05 for rs2860580. For a mediation effect of 1.05 per risk allele of rs2860580 through EBV variant 163364, the necessary sample size with at least 80% power needs to be 1,999 affected individuals and 2,846 control subjects. However, the significant interaction effects between the two *HLA* SNPs and the EBV variant were robust to both the additive and recessive genetic models, indicating a relatively strong *HLA*-EBV interaction effect on NPC risk. Second, although we have adjusted for major NPC risk factors and several factors related to socioeconomic status, our estimates of interaction and mediation effects might be biased by unadjusted or incompletely adjusted confounders. Our case-control study bases were chosen because of the high incidence of NPC, the relatively stable population, the geographic contiguity, and the comparable industrialized level. The control subjects were completely randomly sampled (within strata of age and sex) from the same population in each study base using

Figure 2. Four-way decomposition of total excess relative risk for nasopharyngeal carcinoma associated with host SNPs

(A) rs2860580, (B) rs2894207, and (C) their joint status. CI, confidence interval; proportion, the proportion of total excess relative risk. Two host SNPs were combined as one categorical variable in the models by their joint status, which divided the study subjects into two groups: one group at higher risk carrying only risk alleles of both host SNPs and the other group at lower risk carrying protective alleles of either host SNP.

computerized population registries, and the participation rate was high (83%),[40] such that the control subjects are a representative sample of the underlying population where affected individuals were recruited, and they are genetically homogeneous with the affected individuals. Taken together, using a population-based study design and controlling for the confounders, including the rural or urban area of residence, current occupation, and environmental exposure, as well as educational level as a representative of socioeconomic status, the potential confounding related to population stratification, environmental exposure, and spatial dependence could be largely reduced. Moreover, because the more than 6-fold increased NPC risk conferred by the high-risk EBV subtype is far greater than that associated with other known or suspected risk factors, the potential unmeasured confounding bias, if any, is expected to be proportionally small and unlikely to change the direction and significance of the strong interaction effect between host *HLA* SNPs and EBV. Finally, the current datasets used in this study do not have sufficient power to detect the interaction and mediation effects between EBV and the non-*HLA* susceptibility SNPs, i.e., the SNPs from *TERT*, *CDKN2A/2B*, *TNFRSF19*, *MECOM*, *CIITA*, and *ITAG9* loci. To comprehensively understand the etiology of NPC, a genome-wide gene-EBV interaction analysis as well as the interplay between gene-EBV-environmental factors merit future studies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - ○ Genotyping of human and EBV variants
  - ○ Prediction of BALF2 structure with AlphaFold2
  - ○ Prediction of *HLA* binding affinity with EBV peptides
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100474.

ACKNOWLEDGMENTS

We sincerely thank Dr. Yi Zeng, a virologist who passed away in 2020, for his contributions as a founder member of NPC research and in this case-control study of NPC. We thank the members of the External Advisory Board of the NPC Genes, Environment, and Epstein-Barr Virus Study, including Dr. Curtis Harris (US National Cancer Institute), Dr. Allan Hildesheim (US National Cancer Institute), Dr. Wei-Cheng You (Peking University), and Dr. You-Lin Qiao (Chinese Academy of Medical Sciences), for their guidance of the overall case-control study. This work was supported by the National Key Research and Development Program of China (grant number 2022YFC2305400 to M.X.), the National Natural Science Foundation of China (grant number 82122050 to M.X.), the National Cancer Institute (grant numbers R35 CA197449 and P01 CA134294 to X.L.), the National Institutes of Health (grant number R01CA11587301 to H.-O.A. and Y.-X.Z.), the Swedish Research Council, and the National Natural Science Foundation of China (grant number 8171101281 to W.Y. and W.-H.J.).

AUTHOR CONTRIBUTIONS

X.L., Y.-X.Z., and M.X. conceived and designed the study; H.-O.A., Y.-X.Z., W.Y., and E.T.C. supervised the design and implementation of the population-based case-control study; M.X., X.Z., and Y.Y. performed sample preparation, quality control, and genotyping; M.X., R.F., Zhonghua Liu, X.Z., L.V., Z. Li, Y. Chen, and J.S. conducted statistical analyses; Y. Cao conducted protein structural analysis; S.-M.C., Q.L., W.-H.J., S.-H.X., Zhiwei Liu, Y.-L. Cai, Y.Z., Z.Z., I.E., M.T., and G.H. participated in the case-control study design, subject recruitment, and sample collection; and the manuscript was drafted by M.X., Zhonghua Liu, and R.F. under the supervision of X.L. and was edited by L.V., E.T.C., H.-O.A., W.Y., and X.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 30, 2023
Revised: September 29, 2023
Accepted: December 6, 2023
Published: January 9, 2024

REFERENCES

1. Chang, E.T., Ye, W., Zeng, Y.-X., and Adami, H.-O. (2021). The Evolving Epidemiology of Nasopharyngeal Carcinoma. Cancer Epidemiol. Biomarkers Prev. *30*, 1035–1047.
2. Chien, Y.C., Chen, J.Y., Liu, M.Y., Yang, H.I., Hsu, M.M., Chen, C.J., and Yang, C.S. (2001). Serologic markers of Epstein-Barr virus infection and nasopharyngeal carcinoma in Taiwanese men. N. Engl. J. Med. *345*, 1877–1882.
3. Liu, Y., Huang, Q., Liu, W., Liu, Q., Jia, W., Chang, E., Chen, F., Liu, Z., Guo, X., Mo, H., et al. (2012). Establishment of VCA and EBNA1 IgA-based combination by enzyme-linked immunosorbent assay as preferred screening method for nasopharyngeal carcinoma: a two-stage design with a preliminary performance study and a mass screening in southern China. Int. J. Cancer *131*, 406–416.
4. Chan, K.C.A., Woo, J.K.S., King, A., Zee, B.C.Y., Lam, W.K.J., Chan, S.L., Chu, S.W.I., Mak, C., Tse, I.O.L., Leung, S.Y.M., et al. (2017). Analysis of Plasma Epstein-Barr Virus DNA to Screen for Nasopharyngeal Cancer. N. Engl. J. Med. *377*, 513–522.
5. Xu, M., Yao, Y., Chen, H., Zhang, S., Cao, S.-M., Zhang, Z., Luo, B., Liu, Z., Li, Z., Xiang, T., et al. (2019). Genome sequencing analysis identifies Epstein–Barr virus subtypes associated with high risk of nasopharyngeal carcinoma. Nat. Genet. *51*, 1131–1136.
6. 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., et al. (2010). A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. Nat. Genet. *42*, 599–603.
7. Bei, J.X., Su, W.H., Ng, C.C., Yu, K., Chin, Y.M., Lou, P.J., Hsu, W.L., McKay, J.D., Chen, C.J., Chang, Y.S., et al. (2016). 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. Biomarkers Prev. *25*, 188–192.
8. Cui, Q., Feng, Q.-S., Mo, H.-Y., Sun, J., Xia, Y.-F., Zhang, H., Foo, J.N., Guo, Y.-M., Chen, L.-Z., Li, M., et al. (2016). An extended genome-wide association study identifies novel susceptibility loci for nasopharyngeal carcinoma. Hum. Mol. Genet. *25*, 3626–3634.

9. Hildesheim, A., Apple, R.J., Chen, C.-J., Wang, S.S., Cheng, Y.-J., Klitz, W., Mack, S.J., Chen, I.-H., Hsu, M.-M., Yang, C.-S., et al. (2002). Association of HLA Class I and II Alleles and Extended Haplotypes With Nasopharyngeal Carcinoma in Taiwan. J. Natl. Cancer Inst. *94*, 1780–1789.

10. VanderWeele, T.J., and Knol, M.J. (2014). A tutorial on interaction. Epidemiol. Methods *3*, 33–72.

11. Valeri, L., and VanderWeele, T.J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. Psychol. Methods *18*, 137–150.

12. VanderWeele, T.J. (2014). A Unification of Mediation and Interaction: A 4-Way Decomposition. Epidemiology *25*, 749–761.

13. VanderWeele, T.J. (2015). Explanation in Causal Inference: Methods for Mediation and Interaction (Oxford University Press).

14. National Academies of Sciences, E., and Medicine (2019). Reproducibility and Replicability in Science (The National Academies Press).

15. Kieff, E.D., and Rickinson, A.B. (2007). Epstein-Barr Virus and Its Replication. In Fields' virology, D.M. Knipe and P.M. Howley, eds. (Lippincott Williams & Wilkins, Wolters Kluwer)), pp. 2603–2654.

16. Borza, C.M., and Hutt-Fletcher, L.M. (2002). Alternate replication in B cells and epithelial cells switches tropism of Epstein-Barr virus. Nat. Med. *8*, 594–599.

17. Frangou, P., Buettner, M., and Niedobitek, G. (2005). Epstein-Barr virus (EBV) infection in epithelial cells in vivo: rare detection of EBV replication in tongue mucosa but not in salivary glands. J. Infect. Dis. *191*, 238–242.

18. Hadinoto, V., Shapiro, M., Sun, C.C., and Thorley-Lawson, D.A. (2009). The Dynamics of EBV Shedding Implicate a Central Role for Epithelial Cells in Amplifying Viral Output. PLoS Pathog. *5*, e1000496.

19. Xu, F.-H., Xiong, D., Xu, Y.-F., Cao, S.-M., Xue, W.-Q., Qin, H.-D., Liu, W.-S., Cao, J.-Y., Zhang, Y., Feng, Q.-S., et al. (2012). An Epidemiological and Molecular Study of the Relationship Between Smoking, Risk of Nasopharyngeal Carcinoma, and Epstein–Barr Virus Activation. J. Natl. Cancer Inst. *104*, 1396–1410.

20. He, Y.-Q., Liao, X.-Y., Xue, W.-Q., Xu, Y.-F., Xu, F.-H., Li, F.-F., Li, X.-Z., Zhang, J.-B., Wang, T.-M., Wang, F., et al. (2019). Association Between Environmental Factors and Oral Epstein-Barr Virus DNA Loads: A Multicenter Cross-sectional Study in China. J. Infect. Dis. *219*, 400–409.

21. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

22. Mapelli, M., Panjikar, S., and Tucker, P.A. (2005). The crystal structure of the herpes simplex virus 1 ssDNA-binding protein suggests the structural basis for flexible, cooperative single-stranded DNA binding. J. Biol. Chem. *280*, 2990–2997.

23. Hildesheim, A., and Wang, C.P. (2012). Genetic predisposition factors and nasopharyngeal carcinoma risk: a review of epidemiological association studies, 2000-2011: Rosetta Stone for NPC: genetics, viral infection, and other environmental factors. Semin. Cancer Biol. *22*, 107–116.

24. Tang, M., Lautenberger, J.A., Gao, X., Sezgin, E., Hendrickson, S.L., Troyer, J.L., David, V.A., Guan, L., McIntosh, C.E., Guo, X., et al. (2012). The principal genetic determinants for nasopharyngeal carcinoma in China involve the HLA class I antigen recognition groove. PLoS Genet. *8*, e1003103.

25. Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. *48*, W449–W454.

26. 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., et al. (2009). Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the HLA region at chromosome 6p21.3. Am. J. Hum. Genet. *85*, 194–203.

27. Midgley, R.S., Bell, A.I., McGeoch, D.J., and Rickinson, A.B. (2003). Latent gene sequencing reveals familial relationships among Chinese Epstein-Barr virus strains and evidence for positive selection of A11 epitope changes. J. Virol. *77*, 11517–11530.

28. de Campos-Lima, P.-O., Gavioli, R., Zhang, Q.-J., Wallace, L.E., Dolcetti, R., Rowe, M., Rickinson, A.B., and Masucci, M.G. (1993). HLA-A11 epitope loss isolates of Epstein-Barr virus from a highly A11+ population. Science *260*, 98–100.

29. Midgley, R.S., Bell, A.I., Yao, Q.-Y., Croom-Carter, D., Hislop, A.D., Whitney, B.M., Chan, A.T.C., Johnson, P.J., and Rickinson, A.B. (2003). HLA-A11-restricted epitope polymorphism among Epstein-Barr virus strains in the highly HLA-A11-positive Chinese population: incidence and immunogenicity of variant epitope sequences. J. Virol. *77*, 11507–11516.

30. Deng, C.M., Wang, T.M., He, Y.Q., Zhang, W.L., Xue, W.Q., Li, D.H., Yang, D.W., Wang, Q.L., Liao, Y., Diao, H., et al. (2023). Peptidome-wide association analysis of Epstein− Barr virus identifies epitope repertoires associated with nasopharyngeal carcinoma. J. Med. Virol. *95*, e28860.

31. Lin, J.C., Cherng, J.M., Lin, H.J., Tsang, C.W., Liu, Y.X., and Lee, S.P. (2004). Amino acid changes in functional domains of latent membrane protein 1 of Epstein-Barr virus in nasopharyngeal carcinoma of southern China and Taiwan: prevalence of an HLA A2-restricted 'epitope-loss variant. J. Gen. Virol. *85*, 2023–2034.

32. Duraiswamy, J., Burrows, J.M., Bharadwaj, M., Burrows, S.R., Cooper, L., Pimtanothai, N., and Khanna, R. (2003). Ex vivo analysis of T-cell responses to Epstein-Barr virus-encoded oncogene latent membrane protein 1 reveals highly conserved epitope sequences in virus isolates from diverse geographic regions. J. Virol. *77*, 7401–7410.

33. Tsai, M.H., Raykova, A., Klinke, O., Bernhardt, K., Gärtner, K., Leung, C.S., Geletneky, K., Sertel, S., Münz, C., Feederle, R., and Delecluse, H.J. (2013). Spontaneous lytic replication and epitheliotropism define an Epstein-Barr virus strain found in carcinomas. Cell Rep. *5*, 458–470.

34. Bristol, J.A., Djavadian, R., Albright, E.R., Coleman, C.B., Ohashi, M., Hayes, M., Romero-Masters, J.C., Barlow, E.A., Farrell, P.J., Rochford, R., et al. (2018). A cancer-associated Epstein-Barr virus BZLF1 promoter variant enhances lytic infection. PLoS Pathog. *14*, e1007179.

35. Li, Z., Baccianti, F., Delecluse, S., Tsai, M.-H., Shumilov, A., Cheng, X., Ma, S., Hoffmann, I., Poirey, R., and Delecluse, H.-J. (2021). The Epstein–Barr virus noncoding RNA EBER2 transactivates the UCHL1 deubiquitinase to accelerate cell growth. Proc. Natl. Acad. Sci. USA *118*, e2115508118.

36. Li, Z., Tsai, M.-H., Shumilov, A., Baccianti, F., Tsao, S.W., Poirey, R., and Delecluse, H.-J. (2019). Epstein–Barr virus ncRNA from a nasopharyngeal carcinoma induces an inflammatory response that promotes virus production. Nat. Microbiol. *4*, 2475–2486.

37. Wang, Y., Ungerleider, N., Hoffman, B.A., Kara, M., Farrell, P.J., Flemington, E.K., Lee, N., and Tibbetts, S.A. (2022). A Polymorphism in the Epstein-Barr Virus EBER2 Noncoding RNA Drives In Vivo Expansion of Latently Infected B Cells. mBio *13*, e0083622.

38. Church, T.M., Verma, D., Thompson, J., and Swaminathan, S. (2018). Efficient Translation of Epstein-Barr Virus (EBV) DNA Polymerase Contributes to the Enhanced Lytic Replication Phenotype of M81 EBV. J. Virol. *92*, e01794-17.

39. Rudolph, K.E., Goin, D.E., and Stuart, E.A. (2020). The Peril of Power: A Tutorial on Using Simulation to Better Understand When and How We Can Estimate Mediating Effects. Am. J. Epidemiol. *189*, 1559–1567.

40. Ye, W., Chang, E.T., Liu, Z., Liu, Q., Cai, Y., Zhang, Z., Chen, G., Huang, Q.-H., Xie, S.-H., Cao, S.-M., et al. (2017). Development of a population-based cancer case-control study in southern china. Oncotarget *8*, 87073–87085.

41. Knol, M.J., and VanderWeele, T.J. (2012). Recommendations for presenting analyses of effect modification and interaction. Int. J. Epidemiol. *41*, 514–520.

42. Zhou, X., Cao, S.-M., Cai, Y.-L., Zhang, X., Zhang, S., Feng, G.-F., Chen, Y., Feng, Q.-S., Chen, Y., Chang, E.T., et al. (2021). A comprehensive risk score for effective risk stratification and screening of nasopharyngeal carcinoma. Nat. Commun. *12*, 5189.

43. Schrodinger, L.L.C. (2015). The PyMOL Molecular Graphics System Version 0.99. .

44. Paul, S., Croft, N.P., Purcell, A.W., Tscharke, D.C., Sette, A., Nielsen, M., and Peters, B. (2020). Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. PLoS Comput. Biol. *16*, e1007757.

45. Ferlay J, E.M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., and Bray, F. (2018). Global Cancer Observatory: Cancer Today (International Agency for Research on Cancer). https://gco.iarc.fr/today.

46. Ji, M.F., Sheng, W., Cheng, W.M., Ng, M.H., Wu, B.H., Yu, X., Wei, K.R., Li, F.G., Lian, S.F., Wang, P.P., et al. (2019). Incidence and mortality of nasopharyngeal carcinoma: interim analysis of a cluster randomized controlled screening trial (PRO-NPC-001) in southern China. Ann. Oncol. *30*, 1630–1637.

47. VanderWeele, T.J., and Vansteelandt, S. (2010). Odds Ratios for Mediation Analysis for a Dichotomous Outcome. Am. J. Epidemiol. *172*, 1339–1348.

48. Hosmer, D.W., and Lemeshow, S. (1992). Confidence interval estimation of interaction. Epidemiology *3*, 452–456.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human saliva DNA	Biological repository of the NPC Genes, Environment, and EBV (NPCGEE) study (Ye et al.[40])	https://doi.org/10.18632/oncotarget.19692
Deposited data		
Human genotype data	This paper	National Genomics Data Center (NGDC: GVM000648, https://bigd.big.ac.cn/gvm/getProjectDetail?Project=GVM000648)
EBV genotype data	This paper	National Genomics Data Center (NGDC: GVM000647, https://bigd.big.ac.cn/gvm/getProjectDetail?Project=GVM000647)
Software and algorithms		
SAS code: interaction effects	A tutorial on interaction (VanderWeele et al.[10]); Recommendations for presenting analyses of effect modification and interaction (Knol et al.[41])	https://www.hsph.harvard.edu/tyler-vanderweele/tools-and-tutorials/; https://doi.org/10.1093/ije/dyr218
SAS macro: direct and indirect effect	Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros (Valeri et at.[11])	https://doi.org/10.1037/a0031034 (https://www.hsph.harvard.edu/tyler-vanderweele/tools-and-tutorials/)
SAS code: interaction, mediation, and four-way decomposition analyses	A Unification of Mediation and Interaction: A 4-Way Decomposition (VanderWeele et al.[12])	https://doi.org/10.1097/EDE.0000000000000121

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Xihong Lin (xlin@hsph.harvard.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The EBV genotype and human genotype data are available at the National Genomics Data Center (NGDC: GVM000647 and NGDC: GVM000648). This study did not generate any new code. Data web links and code/software used in this paper are also listed in the key resources table.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

We assessed interaction and mediation effects in a population-based case-control study of NPC based in the Zhaoqing area (including 7 cities/counties) of Guangdong Province and the Wuzhou and Guiping/Pingnan areas (including 6 cities/counties) of Guangxi Zhuang Autonomous Region (Guangxi), between 2010 and 2014. These bases were chosen because of the high incidence of NPC, the relatively stable population, the geographic contiguity, and the comparable industrialized level. This case-control study of NPC followed a population-based and frequency-matched study design. To ensure a high identification rate of NPC cases in the study bases, a rapid case ascertainment system involving a network of local physicians in each study base was built to recruit cases before case enrollment. Control subjects, who were frequency matched to the expected five-year age and sex distribution of the cases, were randomly selected every six months from total population registries in each study base. The overall participation rate was high, 83.8% for cases and 82.7% for controls, respectively. A total of 85.6% of participants enrolled from Zhaoqing area of Guangdong and 85.8% from Wuzhou and Guiping/Pingnan areas of Guangxi were from the rural area. The study design and subject enrollment have been previously described in detail.[40]

Overall, 1306 and 1248 eligible newly diagnosed NPC cases were recruited in Guangdong and Guangxi, respectively. Through random selection from total population registries, 1356 population-based control subjects in Guangdong and 1292 in Guangxi were identified and enrolled with frequency matching to the cases by sex, 5-year age group, and area of residence. All cases and controls were aged between 20 and 74 and did not have a history of cancer, or congenital or acquired immune deficiency. Each participant completed an in-person, structured interview conducted by a trained interviewer. This study was approved by the institutional ethics committees of all the collaborative institutes. Written informed consent was obtained from each study participant.

METHOD DETAILS

Genotyping of human and EBV variants

We selected two host single nucleotide polymorphisms (SNPs) rs2860580 and rs2896207 and the EBV variant at position 163364, based on published GWASs showing the most consistent statistically significant associations with NPC risk.[5–8] Human and EBV variant genotyping was performed using saliva DNA available from 1710 cases and 2246 controls with the Agena Bioscience MassArray platform and was previously described in detail.[42]

To assess potential selection bias, the association of missingness of EBV genotyping data with case-control status and a set of covariates was evaluated. Based on the analyses, which suggested limited bias (described in results and Figure S2), participants with missing EBV and host genotype data were excluded. We further excluded the study subjects with missing data on smoking, salted-preserved fish consumption, and educational level. Finally, 572 NPC cases and 696 controls from Guangdong were included in the original study, and a non-overlapping set of 497 NPC cases and 826 controls from Guangxi were included in the replication study (Figure S1). The current original study has been included in the validation phase of the initial study describing the EBV variant 163364 by Xu et al.,[5] while the replication dataset of the current study is completely independent of this prior study. Both the original and replication datasets of the current study are completely independent of the initial studies describing the two *HLA* genetic variants.[6–8]

Prediction of BALF2 structure with AlphaFold2

The structures of BALF2 proteins from the low-risk Akata EBV and the high-risk M81 EBV were predicted using AlphaFold2.[21] The conformational prediction was based on the crystal structure of single-stranded DNA-binding protein ICP8 from HSV-1 (PDB code: 1URJ), a homologous protein of BALF2, resolved at 3.0 Å resolution.[22] Given ∼25% amino acid sequence similarity between BALF2 and ICP8, we achieved a high confidence in the predicted structure of BALF2. For visualization and in-depth analysis of protein structure, we employed the PyMOL Molecular Graphics System (version 0.99, Schrödinger LLC; http://www.pymol.org/).[43] Among the variations characterizing the low-risk Akata and high-risk M81 BALF2 proteins, the high-risk EBV variant 163364 has the most pronounced influence on protein structure.

Prediction of *HLA* binding affinity with EBV peptides

We predicted *HLA-A*0207* binding affinity for peptide pairs derived from high-risk and low-risk EBV strains. Our analysis focused on the nonamer peptides encompassing these variants: BALF2 variant 163364, 24 non-synonymous NPC-associated variants that were in moderate to high linkage disequilibrium with 163364 (R2 > 0.25), and three non-synonymous variants within two previously reported NPC-associated *A*02* epitopes. Next, we evaluated their binding affinity to *HLA-A*0207* using NetMHCpan-4.1[25]. We used the sliding window to consider all unique peptide nonamers containing the NPC-risk-associated amino acid as input for predicting binding affinity. Peptide pairs from high-risk and low-risk EBV strains were retained if either member could be confidently assigned to *A*0207* with an affinity ranking <2%. We identified nine peptide pairs as candidate *A*0207* epitopes, covering 13 variants, including the two reported *A*02* epitopes (Table S4). Comprehensive MHC-I benchmarking suggests that this threshold captured approximately 90% of the epitopes that elicit T cell response *in vivo.*[44]

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were conducted separately in the original study, replication study, and pooled study. Characteristics between NPC cases and controls were compared using the Chi-square test or t-test. Individual or joint odds ratios (ORs) and their 95% confidence intervals (CIs) for the associations between NPC and the two host SNPs, rs2860580 (0 = AA/AG; 1 = GG) and rs2894207 (0 = CC/CT; 1 = TT), the joint status of two host SNPs (low risk, 0 = AA/AG for rs2860580 or CC/CT for rs2894207; high risk, 1 = GG for rs2860580 and TT for rs2894207) and EBV variant 163364 (0 = C, 1 = CT/T) were evaluated using logistic regression models where the recessive genetic model was employed. Alternatively, the effects associated with per risk allele were also evaluated using logistic regression models that employed the additive genetic model for the two *HLA* SNPs (risk allele = G for rs2860580 and T for rs2894207). In all the logistic models for the interaction and mediation analyses, we adjusted the same set of covariates, including age at interview (<35, 35–59, or >59 years old), cross-classified sex and smoking status (male and never smokers, male and ever smokers and female), education level (<7, 7–9 or ≥10 years), salt-preserved fish consumption in 2000–2002 (yearly or less, or monthly or more), family history among first-class relatives (yes or no), area of residence (rural or urban), current occupation (unemployment/farmer, blue-collar, white-collar), and selected environmental exposures (none, dust exposures including exposures to wood, metal, textile, leather, cement and other types dust but excluding the soil dust, exhaust/smoke exposures including exposures to diesel, gasoline, coal, firewood, asphalt/tar, nature gas and other types of exhaust/smoke but excluding dust exposures, and other exposures including exposures to chemical vapor and acid/alkali but excluding dust and smoke/exhaust exposures). Because NPC is a rare disease, with a prevalence of approximately 0.16% in endemic regions in Southern China in 2008–2013,[45,46] based on rare event assumption, we used OR as an approximation of the relative risk in the interaction and mediation analyses. The odds ratio (OR) of NPC (Y) associated with the host genotype (A = *a*) and the EBV variant (M = *m*) was defined as OR_{am}. C = *c* denotes the set of covariates.

We estimated the additive interaction effect between the two host genetic variants and the EBV variant on NPC risk as the relative excess risk due to interaction (RERI). We first fit the following logistic regression model for NPC:[10]

$$log\,it\{P(Y\ =\ 1|a,m,c)\}\ =\ \theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4'\,c\ (Model\ 1)$$

Under the rare outcome assumption, we have[10]

$$\begin{aligned} RERI &= OR_{11} - OR_{10} - OR_{01} + 1 \\ &= exp(\theta_1 + \theta_2 + \theta_3) - exp(\theta_1) - exp(\theta_2) + 1 \end{aligned}$$

The analyses of interaction effects were conducted using a published SAS code.[10,41]

Causal mediation analysis was used to investigate the mediation effect of EBV variant 163364 on the two host SNPs in relation to the risk of NPC (Y). Mediation analysis was conducted to decompose the effect of a total effect into a direct and an indirect effect, and these effects on the odds ratio (OR) scale were evaluated in a case-control study design setting.[11,47] In addition to the logistic regression for the outcome (Model 1), a second logistic regression model for the mediator (EBV subtypes, Model 2) was applied only on controls:[11,13]

$$log\,it\{P(M\ =\ 1|a,c)\}\ =\ \beta_0+\beta_1 a+\beta_2' c\ (Model\ 2)$$

Under the rare outcome assumption, we have

$$OR^{NDE}\cong\frac{\exp(\theta_1 a)\left\{1+\exp\left(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c\right)\right\}}{\exp(\theta_1 a^*)\left\{1+\exp\left(\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2' c\right)\right\}},$$

$$OR^{NIE}\cong\frac{\left\{1+\exp\left(\beta_0+\beta_1 a^*+\beta_2' c\right)\right\}\left\{1+\exp\left(\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2' c\right)\right\}}{\left\{1+\exp\left(\beta_0+\beta_1 a+\beta_2' c\right)\right\}\left\{1+\exp\left(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c\right)\right\}}$$

where a = 1 and a* = 0 denoting the host high-risk and low-risk genotype, respectively.[11] The natural direct effect (NDE) can be interpreted as the effect on odds ratio scale of the host SNPs (exposure variable) on NPC (outcome variable) not mediated by EBV variant 163364; while the natural indirect effect (NIE) can be interpreted as the causal effect on odds ratio scale of the host SNPs on NPC mediated by the high-risk EBV subtype. The analyses of mediation effects were conducted using a published SAS macro with the setting "casecontrol = true".[11]

In order to concurrently investigate the potential interaction and mediation together, we further conducted a joint causal mediation and interaction analysis by decomposing the total effect (TE) of each host genetic variant on NPC risk into the following four components[12]: (1) controlled direct effect (CDE), due to high-risk host genotype in the absence of high-risk EBV genotype; (2) reference interaction (INT^{ref}), only interaction effect between high-risk host genotype and high-risk EBV genotype; (3) mediated interaction (INT^{med}), both interaction and mediation effect with high-risk host genotype acting only via high-risk EBV genotype selected by high-risk host genotype; and (4) pure indirect effect (PIE), only mediation effect operating exclusively through high-risk EBV subtype, but suppressing the interaction. The total effect is a sum of the four components:[12]

$$TE\ =\ CDE+INT^{ref}+INT^{med}+PIE$$

More detailed calculation formulas for the four components are provided in the Section 2.3 of the eAppendix in the paper we cite.[12] For this analysis, the two logistic regression models, Model 1 and Model 2, were used, allowing interaction between host SNPs and EBV variant in Model 1.

The sum of reference interaction and mediated interaction is the total risk attributable to interaction, and the sum of pure indirect effect and mediated interaction is the total risk attributable to mediation. The sum of reference interaction, mediated interaction, and pure indirect effect is the total risk due to interaction and mediation, which can be potentially eliminated by intervention in high-risk EBV. The analyses of the four-way decomposition of total effect were conducted using the published SAS code.[12]

For details of the sample size estimation, see the section limitations of the study. All statistical tests used are two-sided. The 95% CIs and p values in the interaction, mediation, and four-way decomposition analyses were calculated using the delta method.[10–12,48] Analyses were implemented with SAS 9.4.

Cell Genomics, Volume *4*

Supplemental information

Host genetic variants, Epstein-Barr virus subtypes, and the risk of nasopharyngeal carcinoma: Assessment of interaction and mediation

Miao Xu, Ruimei Feng, Zhonghua Liu, Xiang Zhou, Yanhong Chen, Yulu Cao, Linda Valeri, Zilin Li, Zhiwei Liu, Su-Mei Cao, Qing Liu, Shang-Hang Xie, Ellen T. Chang, Wei-Hua Jia, Jincheng Shen, Youyuan Yao, Yong-Lin Cai, Yuming Zheng, Zhe Zhang, Guangwu Huang, Ingemar Ernberg, Minzhong Tang, Weimin Ye, Hans-Olov Adami, Yi-Xin Zeng, and Xihong Lin

Table S1. Association between two host genetic variants per risk allele and nasopharyngeal carcinoma, related to Table 2.

	Allele frequency in cases N (%)	Allele frequency in controls N (%)	OR per risk allele* (95% CI)	*P* value
Original study in Guangdong				
rs2860580 (risk allele = G)				
A	273 (23.9)	496 (35.6)	Reference	6.00E-10
G	871 (76.1)	896 (64.4)	1.79 (1.49, 2.15)	
rs2894207 (risk allele = T)				
C	155 (13.5)	299 (21.5)	Reference	2.74E-06
T	989 (86.5)	1093 (78.5)	1.71 (1.37, 2.14)	
Replication study in Guangxi				
rs2860580 (risk allele = G)				
A	240 (24.1)	547 (33.1)	Reference	3.62E-07
G	754 (75.9)	1105 (66.9)	1.62 (1.34, 1.95)	
rs2894207 (risk allele = T)				
C	120 (12.1)	300 (18.2)	Reference	3.64E-05
T	874 (87.9)	1352 (81.8)	1.63 (1.29, 2.06)	
Pooled study				
rs2860580 (risk allele = G)				
A	513 (24.0)	1043 (34.3)	Reference	3.76E-15
G	1625 (76.0)	2001 (65.7)	1.68 (1.47, 1.91)	
rs2894207 (risk allele = T)				
C	275 (12.9)	599 (19.7)	Reference	7.48E-10
T	1863 (87.1)	2445 (80.3)	1.65 (1.40, 1.93)	

Abbreviation: OR, odds ratio; CI, confidence interval.

* The OR per risk allele was estimated with the additive genetic model using logistic regression and adjusted for age at interview, sex and smoking joint status, education level, salt-preserved fish consumption in 2000-2002, nasopharyngeal carcinoma history among first-degree relatives, rural or urban area of residence, current occupation, and environmental exposure.

Table S2. Association between host genetic variants, rs2860580 and rs2894207, and their joint status and EBV variant 163364 †, related to Table 3.

	OR *	95% CI *	*P* value
rs2860580 and EBV 163364			
Original study	1.46	1.16, 1.84	0.002
Replication study	1.55	1.24, 1.92	1.0E-4
Pooled study	1.48	1.27, 1.74	8.7E-7
rs2894207 and EBV 163364			
Original study	1.62	1.27, 2.07	1.0E-4
Replication study	1.55	1.22, 1.98	4.0E-4
Pooled study	1.54	1.30, 1.83	5.2E-7
Joint status of host SNPs and EBV 163364			
Original study	1.74	1.36, 2.23	1.2E-5
Replication study	1.68	1.34, 2.11	7.4E-6
Pooled study	1.67	1.42, 1.97	1.0E-9

Abbreviation: OR, odds ratio; CI, confidence interval.

***** Adjusted for age at interview, sex and smoking joint status, education level, residential area, salt-preserved fish consumption in 2000-2002, nasopharyngeal carcinoma history among first-degree relatives, current occupation, and environmental exposure.

† Coding in logistic regression for rs2860580 (0=AA/AG; 1=GG), rs2894207 (0=CC/CT; 1=TT), the joint status of two host SNPs (low risk, 0 = AA/AG for rs2860580 or CC/CT for rs2894207; high risk, 1 = GG for rs2860580 and TT for rs2894207) and EBV variant 163364 (0=C, 1=CT/T).

Table S3. Direct and indirect effects on nasopharyngeal carcinoma between per risk allele of the two host SNPs and EBV variant 163364 under the additive genetic model, related to Table 3.

		OR *	95% CI *	*P* value
rs2860580 and EBV 163364				
	Natural direct effect	1.76	1.44, 2.15	3.9E-8
Original study	Natural indirect effect	1.01	0.93, 1.09	0.883
	Marginal total effect	1.77	1.43, 2.18	1.3E-7
	Natural direct effect	1.46	1.18, 1.80	4.0E-4
Replication study	Natural indirect effect	1.12	1.02, 1.24	0.023
	Marginal total effect	1.64	1.30, 2.07	3.0E-5
	Natural direct effect	1.58	1.37, 1.82	2.4E-10
Pooled study	Natural indirect effect	1.05	0.98, 1.12	0.151
	Marginal total effect	1.66	1.42, 1.94	1.2E-10
rs2894270 and EBV 163364				
	Natural direct effect	1.53	1.20, 1.96	7.1E-4
Original study	Natural indirect effect	1.15	1.03, 1.30	0.017
	Marginal total effect	1.77	1.35, 2.31	3.2E-5
	Natural direct effect	1.48	1.15, 1.90	0.002
Replication study	Natural indirect effect	1.09	0.98, 1.22	0.128
	Marginal total effect	1.61	1.23, 2.10	4.9E-4
	Natural direct effect	1.49	1.25, 1.77	6.6E-6
Pooled study	Natural indirect effect	1.12	1.03, 1.21	0.008
	Marginal total effect	1.66	1.38, 2.01	9.3E-8

Abbreviation: OR, odds ratio; CI, confidence interval.

* Adjusted for age at interview, sex and smoking joint status, education level, salt-preserved fish consumption in 2000-2002, nasopharyngeal carcinoma history among first-degree relatives, rural or urban area of residence, current occupation, and environmental exposure.

Table S4. HLA-A*0207 binding affinity with the peptides of NPC-low-risk and high-risk EBV subtypes, related to STAR Methods.

Gene_Amino acid change	NPC-low-risk peptide		NPC-high-risk peptide		Reported	OR (95% CI)*
	Peptide	Binding rank %	Peptide[+]	Binding rank %		
LMP2A_V254L/L255V	*FLACVLVLI*	0.10	*FLACLVVLI*	0.14		2.7 (1.7, 4.2)
LMP2A_C426S	*CLGGLLTMV*	0.55	*SLGGLLTMV*	0.30	Yes	2.2 (1.4, 3.5)
EBNA3B_AA_36E[*]	*GSDPISPEI*	1.88	*ESEPISPEI*	19.58		1.4 (1.0, 2.0)
EBNA3A_AA_814G[*]	*ALGYPLHAL*	0.91	*ALGYALHGL*	1.04		1.8 (1.2, 2.8)
BALF4_A743V	*LVAGVVILV*	0.84	*LVVGVVILV*	2.07		2.8 (1.8, 4.4)
BALF2_L700V	*RLYGRRLPV*	0.69	*RVYGRRLPV*	2.11		4.1 (2.4, 6.9)
BNRF1_V1222I	*FTNLGMPYV*	0.16	*FTNLGMPYI*	0.53		2.3 (1.6, 3.4)
BPLF1_L610I	*QLPPSATTL*	0.36	*QIPPSATTL*	1.39		1.8 (1.2, 2.5)
LMP1_L126F/M129I	*YLLEMLWRL*	0.01	*YFLEILWRL*	0.34	Yes	2.2 (1.1, 4.3)

Abbreviation: OR, odds ratio; CI, confidence interval; NPC, nasopharyngeal carcinoma.

[+] Amino acid changes in the high-risk peptides are highlighted in red.

*Multiple peptide haplotypes are present. The high-risk peptide and a major low-risk peptide are shown. The OR indicates the NPC risk associated with the high-risk peptide compared to the other peptide variants.

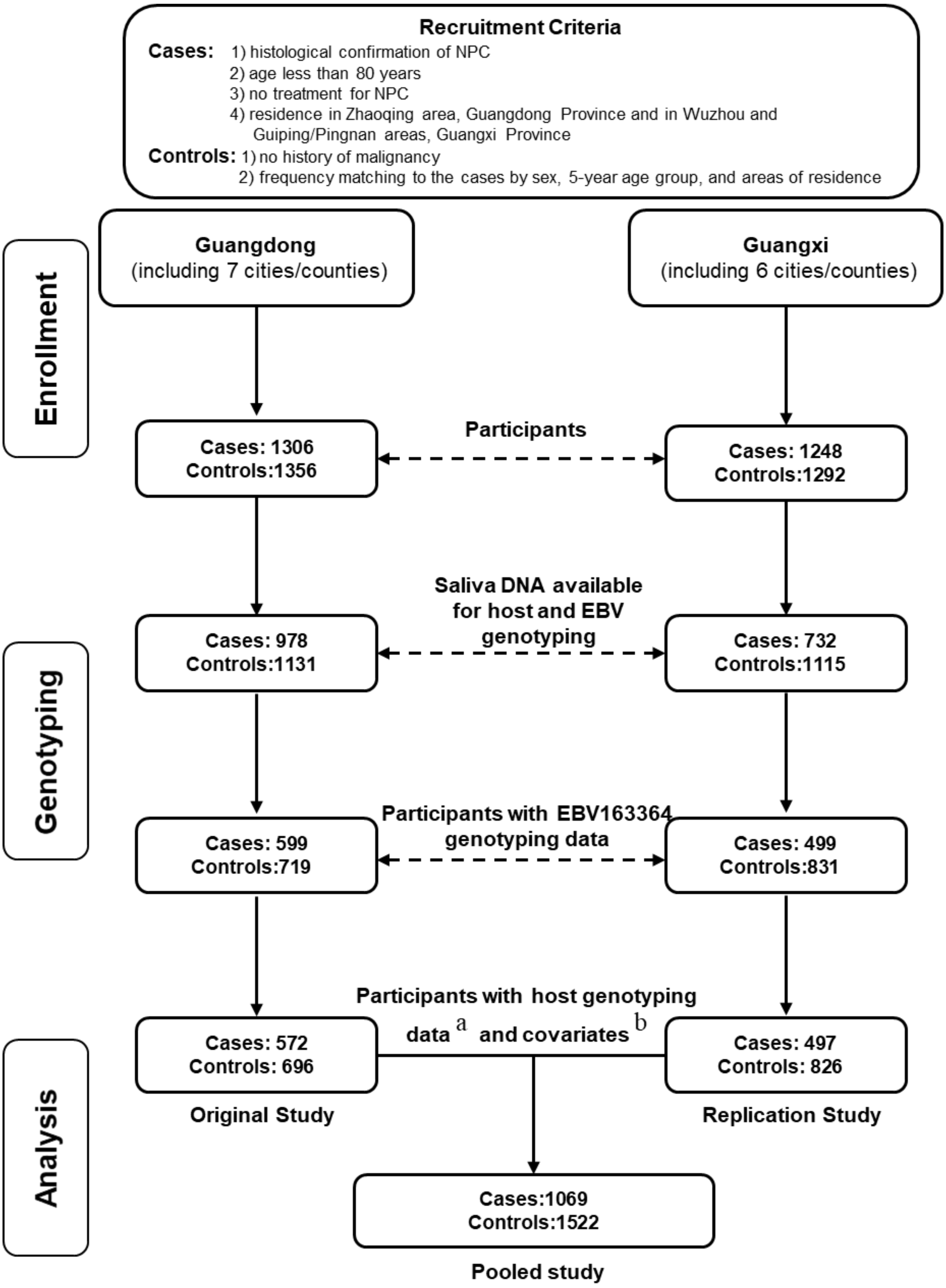

Recruitment Criteria
Cases: 1) histological confirmation of NPC
2) age less than 80 years
3) no treatment for NPC
4) residence in Zhaoqing area, Guangdong Province and in Wuzhou and Guiping/Pingnan areas, Guangxi Province
Controls: 1) no history of malignancy
2) frequency matching to the cases by sex, 5-year age group, and areas of residence
Enrollment
Guangdong (including 7 cities/counties)
Guangxi (including 6 cities/counties)
Participants
Cases: 1306 Controls:1356
Cases: 1248 Controls:1292
Genotyping
Saliva DNA available for host and EBV genotyping
Cases: 978 Controls:1131
Cases: 732 Controls:1115
Participants with EBV163364 genotyping data
Cases: 599 Controls:719
Cases: 499 Controls:831
Analysis
Participants with host genotyping data a and covariates b
Cases: 572 Controls: 696
Original Study
Cases: 497 Controls: 826
Replication Study
Cases:1069 Controls:1522
Pooled study

Figure S1: Flowchart for the study design, related to STAR Methods.

Abbreviation: NPC, nasopharyngeal carcinoma.

[a] Two host genetic SNPs: rs2860580 and rs2894207

[b] Covariates: sex, age, smoking, education level, salt-preserved fish consumption in 2000-2002, nasopharyngeal carcinoma history among first-degree relatives, rural or urban area of residence, current occupation, and environmental exposure.

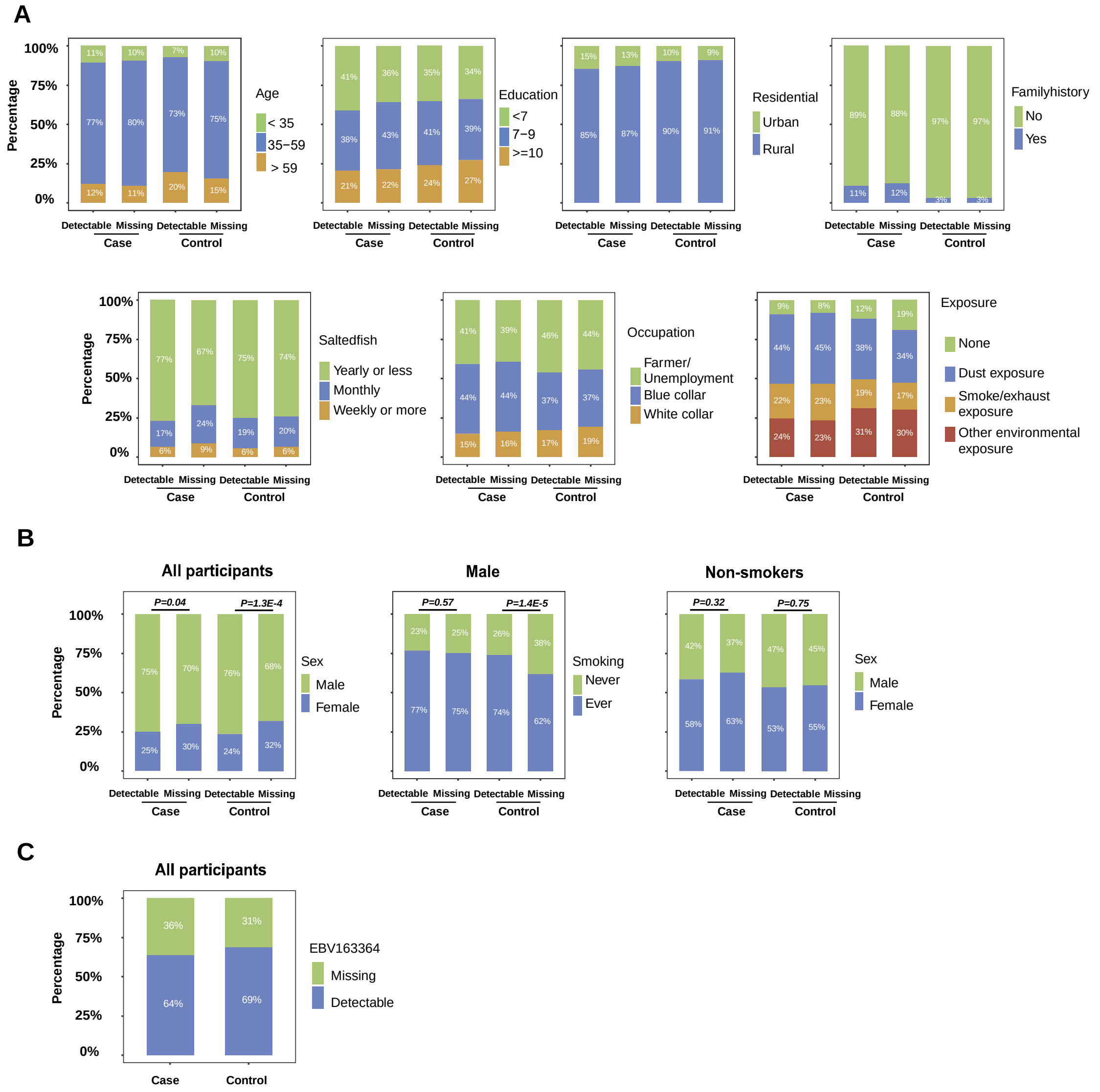


Figure S2. Distribution of genotyping success or failure for EBV variant 163364 in the study participants available for saliva DNA and variable information, related to STAR Methods. (A) Stacked bar plots of the distribution of genotyping success or failure for EBV SNP (163364). Variables: age at interview, education level, rural or urban area of residence, nasopharyngeal carcinoma history among first-degree relatives, salt-preserved fish consumption in 2000-2002, current occupation, and selected environmental exposures. (B) Stacked bar plots of the distribution of EBV genotyping success or failure by sex in all participants, by smoking status in men, and by sex among non-smokers. The values were calculated using χ² tests. (C) Stacked bar plots of EBV SNP genotyping failure among nasopharyngeal carcinoma cases and controls which were not associated with increased risk of nasopharyngeal carcinoma.

A

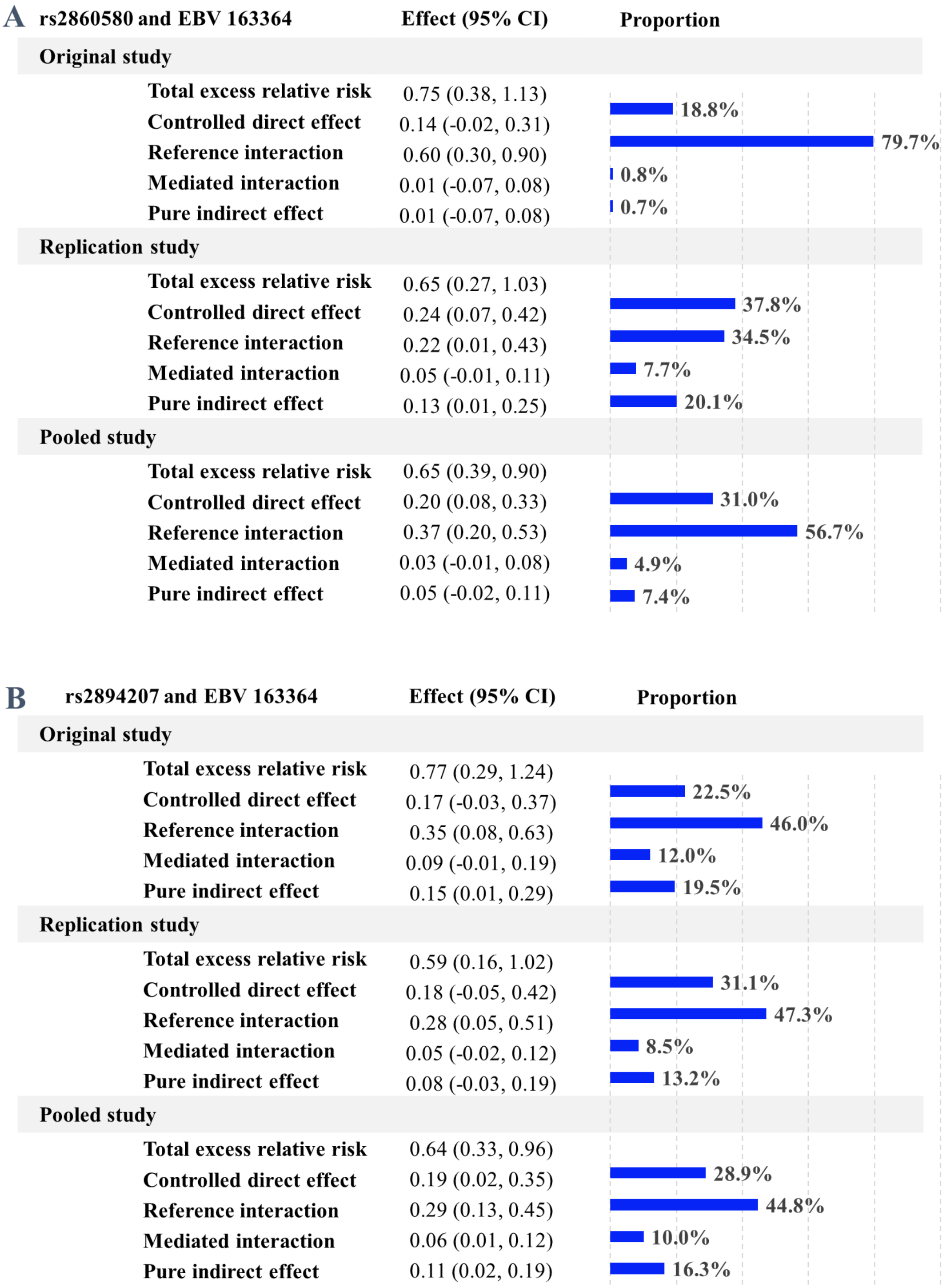
rs2860580	EBV 163364	Cases	Controls	OR (95% CI)*	Additive interaction	
(0= AA; 1=AG; 2=GG)	(0=C; 1= T/CT)				RERI (95% CI)	*P*
Original study						
0	**0**	9	50	0.88 (0.39, 1.99)	3.34 (1.50, 5.19)	3.8E-4
1	**0**	32	169	Reference		
2	**0**	42	155	1.44 (0.85, 2.43)		
0	**1**	22	44	2.71 (1.41, 5.24)		
1	**1**	179	139	6.76 (4.29, 10.64)		
2	**1**	288	139	11.32 (7.26, 17.64)		
Replication study						
0	**0**	6	68	0.43 (0.18, 1.05)	2.48 (0.50, 4.46)	0.014
1	**0**	51	257	Reference		
2	**0**	65	238	1.38 (0.92, 2.09)		
0	**1**	21	20	5.11 (2.54, 10.28)		
1	**1**	135	114	5.72 (3.84, 8.50)		
2	**1**	219	129	8.77 (6.01, 12.79)		
Pooled study						
0	**0**	15	118	0.62 (0.34, 1.12)	3.08 (1.79, 4.37)	2.9E-6
1	**0**	83	426	Reference		
2	**0**	107	393	1.38 (1.00, 1.91)		
0	**1**	43	64	3.59 (2.26, 5.69)		
1	**1**	314	253	6.23 (4.65, 8.35)		
2	**1**	507	268	9.90 (7.46, 13.14)		

B

rs2894207	EBV 163364	Cases	Controls	OR (95% CI)*	Additive interaction	
(0= CC; 1=CT; 2=TT)	(0=C; 1= T/CT)				RERI (95% CI)	*P*
Original study						
0	**0**	1	25	0.21 (0.03, 1.68)	2.81 (0.63, 4.99)	0.012
1	**0**	27	130	Reference		
2	**0**	55	219	1.23 (0.73, 2.06)		
0	**1**	6	8	2.63 (0.82, 8.47)		
1	**1**	114	103	5.54 (3.32, 9.23)		
2	**1**	369	211	8.28 (5.22, 13.14)		
Replication study						
0	**0**	2	26	0.40 (0.09, 1.79)	2.39 (0.71, 4.07)	0.005
1	**0**	33	164	Reference		
2	**0**	87	373	1.18 (0.75, 1.84)		
0	**1**	7	5	7.54 (2.21, 25.72)		
1	**1**	69	74	4.42 (2.66, 7.34)		
2	**1**	299	184	8.24 (5.40, 12.58)		
Pooled study						
0	**0**	3	51	0.29 (0.09, 0.95)	2.57 (1.24, 3.90)	1.5E-4
1	**0**	60	294	Reference		
2	**0**	142	592	1.16 (0.83, 1.62)		
0	**1**	13	13	4.41 (1.91, 10.14)		
1	**1**	183	177	5.02 (3.53, 7.15)		
2	**1**	668	395	8.18 (6.00, 11.14)		

Figure S3. Joint effect and additive interaction between EBV variant 163364 and per risk allele of the two host SNPs rs2860580 (A) and rs2894207 (B) on the risk of the nasopharyngeal carcinoma under the additive genetic model, related to Figure 1. The

analyses were adjusted for age at interview, sex and smoking joint status, education level, salt-preserved fish consumption in 2000-2002, NPC history among first-degree relatives, rural or urban area of residence, current occupation, and environmental exposure. Abbreviation: OR, odds ratio; CI, confidence interval; RERI, relative excess risk due to interaction.



A rs2860580 and EBV 163364	Effect (95% CI)	Proportion
Original study		
Total excess relative risk	0.75 (0.38, 1.13)	
Controlled direct effect	0.14 (-0.02, 0.31)	18.8%
Reference interaction	0.60 (0.30, 0.90)	79.7%
Mediated interaction	0.01 (-0.07, 0.08)	0.8%
Pure indirect effect	0.01 (-0.07, 0.08)	0.7%
Replication study		
Total excess relative risk	0.65 (0.27, 1.03)	
Controlled direct effect	0.24 (0.07, 0.42)	37.8%
Reference interaction	0.22 (0.01, 0.43)	34.5%
Mediated interaction	0.05 (-0.01, 0.11)	7.7%
Pure indirect effect	0.13 (0.01, 0.25)	20.1%
Pooled study		
Total excess relative risk	0.65 (0.39, 0.90)	
Controlled direct effect	0.20 (0.08, 0.33)	31.0%
Reference interaction	0.37 (0.20, 0.53)	56.7%
Mediated interaction	0.03 (-0.01, 0.08)	4.9%
Pure indirect effect	0.05 (-0.02, 0.11)	7.4%

B rs2894207 and EBV 163364	Effect (95% CI)	Proportion
Original study		
Total excess relative risk	0.77 (0.29, 1.24)	
Controlled direct effect	0.17 (-0.03, 0.37)	22.5%
Reference interaction	0.35 (0.08, 0.63)	46.0%
Mediated interaction	0.09 (-0.01, 0.19)	12.0%
Pure indirect effect	0.15 (0.01, 0.29)	19.5%
Replication study		
Total excess relative risk	0.59 (0.16, 1.02)	
Controlled direct effect	0.18 (-0.05, 0.42)	31.1%
Reference interaction	0.28 (0.05, 0.51)	47.3%
Mediated interaction	0.05 (-0.02, 0.12)	8.5%
Pure indirect effect	0.08 (-0.03, 0.19)	13.2%
Pooled study		
Total excess relative risk	0.64 (0.33, 0.96)	
Controlled direct effect	0.19 (0.02, 0.35)	28.9%
Reference interaction	0.29 (0.13, 0.45)	44.8%
Mediated interaction	0.06 (0.01, 0.12)	10.0%
Pure indirect effect	0.11 (0.02, 0.19)	16.3%

Figure S4. Four-way decomposition of total excess relative risk for nasopharyngeal carcinoma associated with per risk allele of the two host SNPs rs2860580 (A) and rs2894207 (B) using additive genetic model, related to Figure 2. The analyses were adjusted for age at interview, sex and smoking joint status, education level, salt-preserved fish consumption in 2000-2002, nasopharyngeal carcinoma history among first-degree relatives, rural or urban area of residence, current occupation, and environmental exposure.

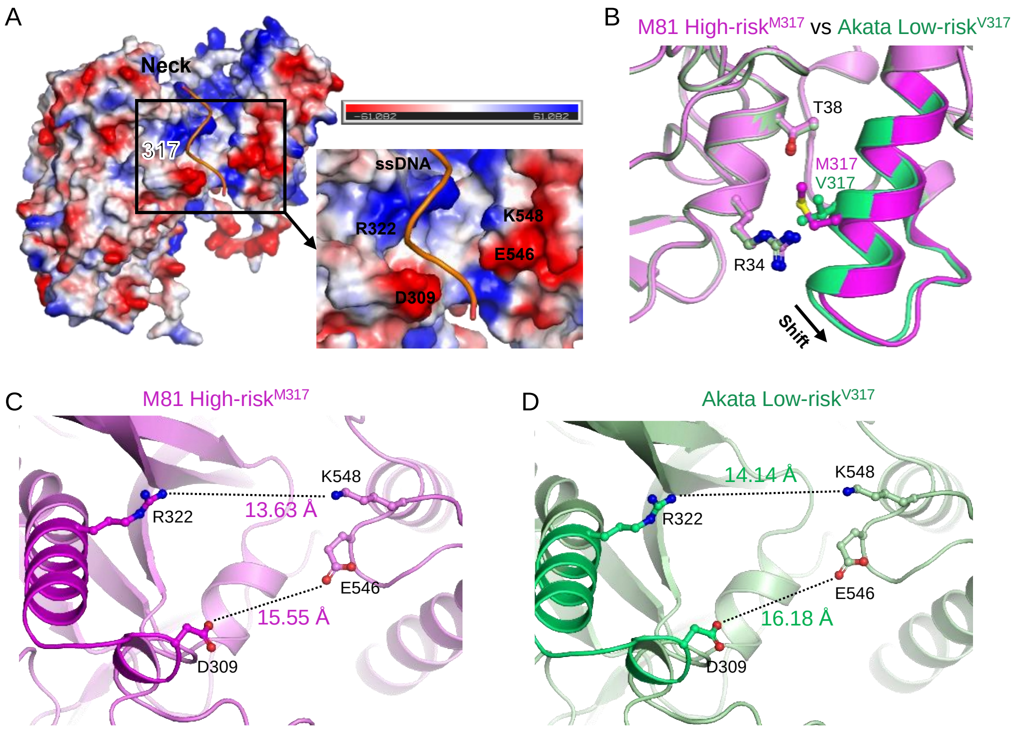


Figure S5. Predicted structures of BALF2 protein from high-risk M81 EBV and low-risk Akata EBV, related to STAR Methods. (A) Predicted protein conformation of BALF2 in complex with single-stranded DNA (ssDNA, orange). The amino acid 317, encoded by the high-risk variant 163364, and the key amino acids interacting with ssDNA are indicated. (B) The V317M mutation induces an alpha-helix shift. Magenta and green indicate regional structures of BALF2 protein from high-risk M81 EBV and low-risk Akata EBV, respectively. Other two amino acids (R34 and T38) that retain their position, in contrast to V317M, are highlighted. (C-D) Spatial distances between amino acids interacting with ssDNA are indicated for high-risk M81 EBV (C) and low-risk Akata EBV (D) BALF2 proteins, respectively.

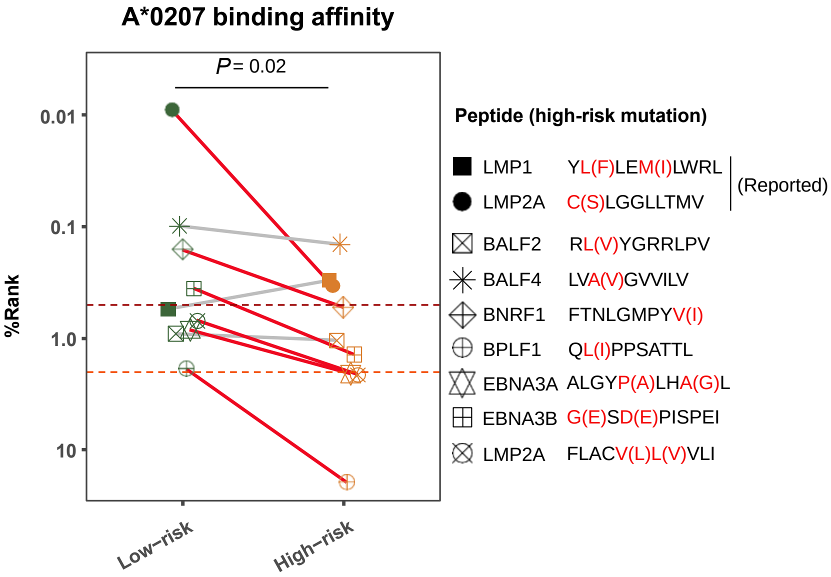


Figure S6. HLA-A*0207 binding affinity with the EBV peptides of nasopharyngeal carcinoma-low-risk and high-risk subtypes, related to STAR Methods. The 9-mer peptides are indicated on the right, and mutations in the high-risk EBV subtype are highlighted in red. The LMP1 and LMP2A peptides have been verified with functional T cell response assays in previous studies, indicating that the mutant LMP2A peptide failed to elicit T cell responses in patients with nasopharyngeal carcinoma. The affinity is shown as the binding ranking percentile predicted with NetMHCpan-4.1. The dark red dashed line represents a ranking percentile of 0.5%, indicative of strong binding affinity. The red dashed line represents a ranking percentile of 2%, indicative of weak binding affinity.