

1 **Genome sequencing analysis identifies high-risk Epstein-Barr** 2 **virus subtypes for nasopharyngeal carcinoma**

3
4 Miao Xu^{1,6,12}, Youyuan Yao^{1,11,12}, Hui Chen^{2,12}, Shanshan Zhang^{1,12}, Tong Xiang¹, Su-
5 -Mei Cao¹, Zhe Zhang³, Bing Luo⁴, Zhiwei Liu^{5,8}, Zilin Li⁶, Guiping He¹, Qi-Sheng Feng¹,
6 Li-Zhen Chen¹, Xiang Guo¹, Weihua Jia¹, Ming-Yuan Chen¹, Bingchun Zhao¹, Xiao Zhang¹,
7 Shang-Hang Xie¹, Roujun Peng¹, Ellen T. Chang⁷, Vincent Pedergnana², Lin Feng¹, Jin-Xin
8 Bei¹, Ruihua Xu¹, Mu Sheng Zeng¹, Weimin Ye⁸, Hans-Olov Adami⁸, Xihong Lin⁶,
9 Weiwei Zhai^{2,9,10}, Yi-Xin Zeng¹, Jianjun Liu²

10
11 ¹State Key Laboratory of Oncology in South China, Collaborative Innovation Center for
12 Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China. ² Human
13 Genetics, Genome Institute of Singapore, Agency for Science, Technology and Research
14 (A*STAR), Singapore. ³Department of Otolaryngology/Head and Neck Surgery, First
15 Affiliated Hospital of Guangxi Medical University, Nanning, China. ⁴Department of Medical
16 Microbiology, Qingdao University Medical College, Qingdao, China. ⁵Infections and
17 Immunoepidemiology Branch, National Cancer Institute, Rockville, MD, USA ⁶Department
18 of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. ⁷Division of
19 Epidemiology, Department of Health Research and Policy, Stanford University School of
20 Medicine, Stanford, CA, USA. ⁸Department of Medical Epidemiology and Biostatistics,
21 Karolinska Institutet, Stockholm, Sweden. ⁹Key Laboratory of Zoological Systematics and
22 Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. ¹⁰Center for
23 Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming,
24 China. ¹¹Department of Comprehensive Medical Oncology, Key Laboratory of Head & Neck
25 Cancer Translational Research of Zhejiang Province, Zhejiang Cancer Hospital, Hangzhou,
26 China. ¹²These authors contributed equally to the work. Correspondence should be addressed
27 to J.L. (liuj3@gis.a-star.edu.sg), Y.-X.Z. (zengyx@sysucc.org.cn) and W.Z.
28 (zhaiww1@gis.a-star.edu.sg).
29

30 **Summary (169 words)**

31 Epstein-Barr virus (EBV) infection is ubiquitous worldwide and associated with multiple
32 cancers including nasopharyngeal carcinoma (NPC). The role of EBV viral genomic
33 variation in NPC development and its striking endemicity in southern China has been poorly
34 explored. Through large-scale genome sequencing and association study of EBV isolates
35 from China, we identified two non-synonymous EBV variants within *BALF2* strongly
36 associated with NPC risk (conditional *P* value 1.75×10^{-6} for SNP162476_C and 3.23×10^{-13}
37 for SNP163364_T), whose cumulative effects contributed to 83% of the overall risk in
38 southern China. Phylogenetic analysis of the risk variants revealed a unique origin in
39 southern China followed by clonal expansion. EBV *BALF2* haplotype carrying the risk
40 variants were shown to reduce viral lytic DNA replication, as a result potentially promoting
41 viral latency. Our discovery has not only provided insight to the unique endemic pattern of
42 NPC occurrence in southern China, but also paved the way for the identification of
43 individuals at high risk of NPC and effective intervention program to reduce the disease
44 burden in southern China.

45

46 **Introduction**

47 Epstein-Barr virus (EBV) was discovered in 1964^{1,2}, and has been the first human virus to be
48 associated with cancers, including nasopharyngeal carcinoma (NPC), a subset of gastric
49 carcinoma, and several kinds of lymphomas³. Although EBV infection is ubiquitous in
50 human populations worldwide, its most closely associated malignancy, NPC, has unique
51 geographic distribution. NPC is rare in most part of the world, but very common in southern
52 China where the incidence rate can reach 20 - 40 per 100,000 individuals per year⁴. Multiple
53 human susceptibility loci, including *HLA*, *CDKN2A/2B*, *TNFRSF19*, *MECOM*, and *TERT*
54 loci, have been discovered for NPC, but these loci have limited contribution to overall risk⁵⁻⁸.
55 Moreover, the risk variants of these loci distributed widely in Chinese and could not explain
56 the unique endemics of NPC in southern China. Understanding the etiology of NPC,
57 commonly known as the Cantonese cancer, remains enigmatic.

58

59 Since the first EBV genome sequence, B95-8, was published in 1984⁹, more than a hundred
60 EBV genomes have been sequenced in spontaneous lymphoblastoid cell lines and patients
61 diagnosed with EBV-associated diseases, which revealed significant genomic variation
62 among EBV isolates from different geographic origins¹⁰⁻¹⁵. Even though the role of EBV
63 genome variation in the risk of EBV-associated diseases has been explored¹⁵⁻¹⁷, previous

64 studies suffered from the confounding effect of geographic distribution and insufficient
65 sample size. As a result, robust epidemiological and genetic evidence linking specific EBV
66 strains to the pathogenesis of NPC is yet to be uncovered.

67

68 In the current study, we performed large-scale EBV whole-genome sequencing (WGS) study
69 of 215 EBV isolates from patients diagnosed with EBV-associated cancers (including NPC,
70 gastric carcinoma, and lymphomas) and 54 isolates from healthy controls that were recruited
71 from both NPC-endemic and non-endemic regions of China. Through a comprehensive and
72 systematic association analysis of EBV genomic variation and subsequent replication
73 analysis in an independent sample, we identified two high-risk EBV variants for NPC.
74 Functional investigation uncovered their decreased ability to activate EBV lytic DNA
75 replication which potentially promotes EBV latency and NPC tumorigenesis. In addition,
76 phylogenetic analysis of the isolates from the current study and worldwide strains suggested a
77 unique evolutionary origin of the two NPC-high-risk variants in southern China. For the first
78 time, we have uncovered the high-risk EBV subtypes that contributed significantly to the
79 overall risk of NPC, as well as its unique endemics in southern China.

80

81 **EBV whole-genome sequencing**

82 Using a capture-based WGS protocol, we obtained EBV genome sequences from 215 tumor
83 tissue, saliva and plasma samples of EBV-associated cancer patients (NPC, gastric carcinoma,
84 and lymphomas) and 54 saliva samples of healthy donors, as well as one from NPC cell line
85 C666.1 (For details, see **Supplementary Table 1** and **Methods**). Of the 269 EBV isolates,
86 220 were obtained from the NPC-endemic region of southern China (Guangdong and
87 Guangxi Provinces), and 49 were from NPC-non-endemic regions of China. The average
88 sequencing depth of all the isolates was 1,282×, and on average 95% of EBV genome was
89 covered with at least 10× reads (**Supplementary Fig. 1**). Using B95-8 as the reference, we
90 identified a total of 8,469 variants (8015 SNPs, 454 INDELS) across the EBV genome (for
91 variant statistics, see **Supplementary Table 2** and **Supplementary Fig. 1**). The number of
92 variants identified in each sample ranged from 1,006 to 2,104, with EBNAs and LMPs genes
93 being the most polymorphic genes (**Supplementary Fig. 2a, b**), consistent with prior
94 reports¹⁴⁻¹⁶.

95

96 To explore the accuracy in sequencing and variant calling, we compared the re-sequenced
97 C666-1 EBV genome against the previous published record, and a high concordance rate of

98 97.9% was found¹⁸ (**Supplementary Table 3**). In addition, when subsets of variants
99 discovered by WGS were re-genotyped by Sanger sequencing and MassArray iPLEX assay;
100 97.55% and 99.99% of tested variants were confirmed, respectively (**Supplementary Tables**
101 **4 and 5**). Both of these evidences suggest a very high accuracy in our sequencing and variant
102 calling.

103

104 In order to understand intra-host polymorphism within an individual, two EBV fragments
105 were amplified and sequenced in 25 paired saliva and tumor samples from NPC cases. The
106 variant dissimilarity between the paired saliva and tumor samples (median 1.1%, 1st to 3rd
107 quartile: 0-3.4%) was substantially lower than the between-host dissimilarity (median 13.5%,
108 1st to 3rd quartile: 3.7-16.9%) (**Supplementary Fig. 3**). In addition, we sequenced the EBV
109 whole genomes from the same NPC patient in paired tumor and saliva samples, and we
110 observed that 99.27% of the variants were concordant between the EBV tumor and saliva
111 isolates (**Supplementary Table 6**). Taken together, these observations suggested that paired
112 saliva and tumor samples from the same subject contained the same EBV genome or strain.
113 Therefore, we combined the genome sequence information from tumor tissue and saliva
114 samples from NPC cases in subsequent analyses.

115

116 **Identification of high-risk EBV variants by two-stage genome-wide association analysis**

117 To investigate the impact of EBV genomic variations on NPC risk, we performed a two-stage
118 genome-wide association study of the EBV genome. The genome-wide discovery analysis
119 was performed by testing 1545 EBV variants in 156 NPC cases and 47 healthy controls from
120 Guangdong and Guangxi Provinces in the NPC-endemic region of Southern China
121 (**Discovery phase**). A principal component analysis (PCA) of the human genome variation
122 of all the cases and controls with the reference population samples from the 1000G project¹⁹
123 confirmed the Chinese origin of and genetic match between the cases and controls
124 (**Supplementary Fig. 4a, b**). We also performed a PCA of EBV viral genome variation by
125 using all the strains and C666-1 genome sequence from the current study with 97 publicly
126 available genomes. We observed a continuous distribution of the EBV strains along the first
127 principal component (PC) ranging from Africa and Europe to Asia (**Fig. 1a**). Within Asia, the
128 second PC showed a partial separation of the isolates from NPC-endemic region and the ones
129 from the non-endemic region of Asia (**Fig. 1a, d**).

130

131 To control for the potential impact of the population structures of both the human and EBV
132 genomes, the genome-wide discovery association analysis was performed using generalized-
133 linear mixed model with age, sex, the first four human PCs as fixed effect and the genetic
134 relatedness matrix of EBV genomes as random effect²⁰. The discovery analysis revealed
135 multiple association signals along the EBV genome, with the strongest association observed
136 within the *BALF2* region (SNP162507, $P = 9.99 \times 10^{-5}$) without any indication of inflation
137 (genomic control inflation factor $\lambda_{GC} = 1.01$) due to genetic structure (**Fig. 2a** and
138 **Supplementary Table 7**). In addition, we also performed a multi-SNP genome-wide
139 association analysis using Bayesian variable-selection regression by piMASS²¹, which
140 provided consistent and strong evidence for the association within the *BALF2* region
141 (posterior probability = 0.96) (**Fig. 2b**). We evaluated the statistical significance of
142 association using permutation (see **Methods**), and only the associations within the *BALF2*
143 region reached genome-wide significance (suggestive genome-wide significance, $P <$
144 4.07×10^{-4}). Consistent with the extensive linkage disequilibrium (LD) within the EBV
145 genome (**Supplementary Fig. 5**), conditioning on the genetic effects of the SNPs within the
146 *BALF2* region greatly reduced the extensive associations across the entire EBV genome
147 (**Supplementary Fig. 6**).

148
149 We performed a Bayesian fine-mapping analysis to prioritize potentially causal SNPs in the
150 *BALF2* gene region using PAINTOR and found that only the three non-synonymous coding
151 variants (SNPs 162215, 162476, and 163364) showed significant evidence for association
152 (**Supplementary Fig. 7 and Supplementary Table 8**). We genotyped the three non-
153 synonymous coding SNPs in an independent sample of 483 NPC cases and 605 age- and sex-
154 matched healthy population controls (**Validation phase**) (**Supplementary Table 9**). To
155 eliminate any potential impact of population stratification, all the cases and controls were
156 recruited from the single NPC-endemic region, Zhaoqing County in the Guangdong Province
157 of China. All three SNPs were significantly associated with NPC risk in the independent
158 sample ($P < 0.017, 0.05/3$), consistent with the discovery phase results (**Table 1**). The meta-
159 analysis of the combined discovery and validation phases confirmed the associations with the
160 three SNPs of *BALF2* with genome-wide significance according to both permutation analysis
161 and Bonferroni correction for multiple testing (SNP 162215_C, OR = 7.62, $P = 2.98 \times 10^{-19}$;
162 162476_C, OR = 8.79, $P = 5.90 \times 10^{-26}$ and 163364_T OR = 6.52, $P = 7.18 \times 10^{-36}$) (**Table 1**).
163 All the three SNPs showed significant LD (**Supplementary Fig. 8**), but conditional analysis
164 demonstrated that the associations with SNPs 162215 and 162476 were correlated, whereas

165 SNP 163364 showed an independent association that also reached genome-wide significance
166 (**Table 1**).

167

168 We further explored the association between the haplotypes (strains) composed of SNPs
169 162215, 162476 and 163364 and NPC risk. Taking the haplotype composed of the 3 low-risk
170 variants (A-T-C) as a reference, we did not observe association for the haplotype carrying the
171 high-risk variant for SNP162215 (C-T-C; odd ratio (OR) = 1.10, $P = 0.83$), although the
172 number of haplotypes for testing was limited (**Table 2** and **Supplementary Table 10**). Both
173 the haplotypes carrying the high-risk variants of either all the three SNPs or only SNPs
174 162215 and 162476 showed strong risk effect (haplotype C-C-T: OR = 12.10, $P = 1.17 \times 10^{-25}$;
175 haplotype C-C-C: OR = 3.30, $P = 2.20 \times 10^{-5}$) (**Table 2** and **Supplementary Table 10**), but
176 the haplotype C-C-T showed significantly stronger effect than the haplotype C-C-C ($P =$
177 1.85×10^{-12}), clearly indicating the additional risk effect of SNP163364. The haplotype
178 analysis further confirmed that NPC risk is primarily associated with SNPs 162476 and
179 163364, and SNP162215 needs to be further evaluated. We also performed pair-wise
180 interaction analysis showing no evidence for interaction between SNPs 162476 and 163364
181 ($P = 0.67$). Lastly, we performed a multiple regression analysis that yielded independent risk
182 effects (OR) of 3.15 for SNP 162476_C and 3.68 for SNP 163364_T (**Supplementary Table**
183 **11**), which were consistent with the risk effect of the haplotype carrying the two high-risk
184 variants (haplotype C-C-T: OR = 12.10) (**Table 2**).

185

186 **The population distribution and phylogenetic analysis of EBV risk haplotypes for NPC**

187 In China, the frequency of the two high-risk haplotypes (C-C-T and C-C-C) was very high in
188 the NPC-endemic region (93.27% in NPC cases and 63.04% in controls), but much lower in
189 non-endemic areas (55% in NPC cases, 14.29% in controls) (**Supplementary Table 12**).
190 The high-risk haplotypes were also observed in other EBV-associated cancers with a
191 frequency of about 40% in lymphoma patients from the NPC-endemic region and 8.3% in the
192 lymphoma and gastric carcinoma patients from the non-endemic regions. These frequencies
193 are comparable to those observed in healthy controls and much lower than those observed in
194 NPC patients (**Supplementary Table 12**). However, the number of samples from the non-
195 endemic region and other EBV-association cancers were much smaller than the NPC samples
196 from the endemic region. Interesting, the two risk haplotypes were absent or extremely rare
197 in non-Asian populations (African and western countries) (**Supplementary Table 12**),
198 suggesting the Asian origin of the EBV high-risk variants.

199

200 To further explore the origin of the EBV risk variants, we investigated the evolutionary
201 relationship among the EBV strains from the current study and the previously published ones.
202 By examining the frequency and distribution of heterozygous SNPs, we identified 229 EBV
203 single strains from the 269 genome isolates (see **Methods, Supplementary Fig. 9 and**
204 **Supplementary Table 13**). Using these 229 strains and C666-1 EBV genome sequence from
205 the current study as well as 97 publicly available genomes, we performed phylogenetic
206 inference analysis and found that the evolutionary relationship among all sequences was
207 highly unbalanced, with a deep split between Type 1 and Type 2 EBV isolates (**Fig. 1b**). All
208 Type 2 EBV isolates were geographically restricted to Africa, as previously observed^{14,15,22}.
209 The Type 1 EBV clade showed a continuous branching starting from Africa, Europe, and
210 Asia, matching the overall distribution along the first PC in the PCA analysis (**Fig. 1b-d**). As
211 shown in previous study^{17,23}, 97% of all 269 EBV isolates were found to be China 1 subtype,
212 and 2% were China 2 (defined by LMP1 classification) (**Supplementary Fig. 10**). Within
213 the Asian group, isolates from NPC-non-endemic areas clustered towards the basal position
214 of the lineage, similar to the pattern observed along the second PC in the PCA map (**Fig. 1b-**
215 **d**). The most striking pattern in the phylogenetic relationship was a rapid radiation of NPC-
216 dominant strains in the endemic population from southern China. EBV genomes from NPC
217 patients appeared to have expanded recently from a common ancestor, and more than half (22
218 of 37) of healthy controls from this region were also infected with NPC-dominant strains (**Fig.**
219 **1b, c**).

220

221 We also mapped the three SNPs of *BALF2* (SNP 162215, 162476, and 163364) onto the
222 phylogenetic tree of the EBV genomes. We observed that all the strains carrying the risk
223 variants of SNP162476, and 163364 were within the Asian subclade, whereas the carriers of
224 SNP 162215 had a much broader distribution (**Fig. 1b, e**). Within the Asian subclades, the
225 carriers of SNP162476 and 163364 were enriched in the strains from NPC patients (NPC-
226 dominant strains). These results provided strong evidence for the Asian origin of SNP162476
227 and 163364 and were consistent with their risk effect on NPC. The results also suggested that
228 SNP 162215 was less likely to be a risk variant for NPC, and its association effect was due to
229 its LD with SNP162476 (LD R-squared = 0.67).

230

231 **Functional analysis of EBV *BALF2* nonsynonymous variants**

232 Since all three risk variants encode amino acid alterations within the *BALF2* gene that is
233 responsible for opening the viral DNA for lytic replication. In order to explore the
234 functional role of the three NPC-associated EBV variants, we investigated whether the three
235 viral SNPs influenced viral lytic DNA replication. First, we performed *in vitro* functional
236 analysis in EBV-positive NPC cell line TW03. After the stimulation of lytic cycle activation,
237 we measured viral DNA abundance within cells that were transfected with the reference
238 haplotype of B95-8 (C-T-C), the low-risk haplotype (A-T-C) and the high-risk haplotype (C-
239 C-T) of *BALF2* and the empty vector. We found that the viral DNA level was significantly
240 lower in cells carrying the high-risk haplotype compared to cells carrying the other two
241 haplotypes ($P < 0.05$). No difference was observed between the low-risk and reference
242 haplotypes, and between the high-risk haplotype and the empty vector (**Fig. 3a**). These
243 results indicated that both the reference and low-risk haplotypes of *BALF2* had similar ability,
244 whereas the high-risk haplotype had weaker ability in supporting EBV lytic DNA replication.
245 Furthermore, we performed an *in vivo* analysis of the viral DNA abundance in the saliva
246 samples from the 533 NPC cases and 651 healthy controls of the validation sample. We
247 observed a large variation of viral DNA load across the samples, and found that viral DNA
248 abundance in saliva was significantly lower in NPC patients than in healthy controls ($P =$
249 6.6×10^{-15}) (**Fig. 3b**). However, we did not observe the association of the high-risk haplotype
250 (C-C-T) of *BALF2* with saliva viral DNA abundance (**Supplementary Table 14**).

251

252 Discussion

253 Because of the ubiquity of EBV infection, the determinants of the distinctive geographical
254 distribution of NPC have long puzzled the scientific community. Using large-scale
255 sequencing and functional analyses, we discovered for the first time two EBV coding SNPs
256 162476 and 163364 showing a strong risk effect for NPC. The more than 6-fold increase in
257 NPC risk conferred by these two high-risk EBV variants is far greater than the effects of any
258 other known risk factors for this disease, including host genetic variants (**Table 1**). In
259 particular, with a population frequency of 45% and an OR of 12.10, the EBV haplotype C-T
260 of the two SNPs is the dominant NPC risk factor, contributing 71% (95% confidence interval:
261 67-74%) of the overall risk of NPC in the endemic population of southern China. The second
262 risk haplotype C-C also contributed about 10%, such that the two high-risk EBV haplotypes
263 combined accounted for 83% (95% confidence interval: 79-87%) of NPC risk in this
264 population (**Supplementary Table 15**). In non-endemic regions of China, the frequency of

265 these high-risk haplotypes is much lower (about 10%), but they still make a significant
266 contribution (~50%) to NPC risk. The frequency of the two high-risk EBV subtypes was not
267 associated with the risk of developing other EBV-related cancers in our study, suggesting that
268 their oncogenic effects might be specific to NPC. However, this observation would benefit
269 from further work since our study was only powered to explore NPC.

270

271 When we mapped these two causal variants onto the phylogenetic tree of EBV genomes, we
272 observed a distinct subclade of EBV subtypes carrying the two high-risk variants within Asia.
273 The carriers could only be found inside of Asia, which strongly demonstrates the Asian origin
274 of these two risk variants. Most interestingly, the phylogenetic analysis suggests a rapid
275 clonal expansion of these unique high-risk EBV subtypes. This is consistent with the current
276 distribution of these subtypes in China with very high frequency in the NPC-endemic region
277 (93.27% in NPC cases and 63.04% in controls), but much lower in the non-endemic areas (55%
278 in NPC cases, 9.68% in other samples) (**Supplementary Table 12**). It remains to be
279 investigated what kind of positive selection has driven their emergence. Taken together, the
280 strong risk effect, the confined geographic distribution, and the rapid clonal expansion and
281 consequently extremely enriched frequency of these two high-risk variants in the NPC-
282 endemic region strongly suggest that these two EBV risk factors are the driving factors for
283 the unique endemics of NPC in southern China.

284

285 Our genetic discovery has provided novel biological insight in the development of NPC by
286 highlighting the important role of suppressed viral lytic DNA replication in EBV-mediated
287 NPC tumorigenesis. Consistent with the role of *BALF2* as the core component in the viral
288 DNA replication complex, we showed that the high-risk haplotype of *BALF2* suppressed the
289 lytic DNA replication using the *in vitro* cell line analysis. Consistently, EBV DNA
290 abundance in saliva was found to be significantly lower in the NPC cases than in controls,
291 suggesting that EBV in buccal epithelium is less lytic in NPC patients. EBV latency promotes
292 the expression of oncogenes and is therefore indispensable for EBV-mediated
293 carcinogenesis^{24,25}, and the expansion of EBV latently-infected nasopharyngeal cells has
294 proven to be an early event in NPC tumorigenesis^{26,27}. Our results demonstrated that the
295 impairment of *BALF2* function due to EBV genetic variation potentially promotes viral
296 latency and fosters NPC development by suppressing viral lytic replication. The discovery of
297 these high-risk EBV variants also has major implications for public health efforts to reduce
298 the burden of NPC, particularly in the endemic region of southern China. Testing for these

299 high-risk EBV variants can enable the identification of high-risk individuals for targeted
300 implementation of routine clinical monitoring for early detection of NPC. Primary prevention
301 by developing vaccines against high-NPC-risk EBV strains is expected to lead to great
302 attenuation of the Cantonese Cancer in China.

303

304 **Methods**

305 **Study participants and samples.** Participants of the current study were enrolled through two
306 recruitments. The first one was a hospital-based study enrolling patients diagnosed with
307 EBV-related cancers, including NPC, Burkitt lymphoma, Hodgkin lymphoma, NK/T cell
308 lymphoma and gastric carcinomas as well as healthy controls from the Sun Yat-sen university
309 Cancer Center in Guangdong Province, the First Affiliated Hospital of Guangxi Medical
310 College in Guangxi Province, and the Affiliated Hospital of the Qingdao University in
311 Shandong Province of China. The geographical origin of the participants covers NPC-
312 endemic southern China (Guangdong and Guangxi Provinces where NPC has highest
313 incidence of 20-40/100,000 individuals per year) and non-endemic regions in China where
314 NPC is rare. After quantitative measurement of EBV DNA, 170 tumor tissue and/or saliva,
315 plasma samples were selected from the first recruitment for EBV whole-genome sequencing
316 (WGS). The second recruitment was a population-based NPC case-control study enrolling
317 NPC cases and population control subjects from Zhaoqing County, Guangdong Province of
318 China (NPC-endemic region). The population controls were matched to the cases by their age
319 and sex. Saliva samples were collected from all the subjects. After quantitative measurement
320 of EBV DNA of the second study, 99 saliva samples (53 NPC cases and 46 controls) were
321 selected for EBV WGS. Written informed consent was obtained from each participant before
322 undertaking any study-related procedures, and all studies was approved by institutional ethics
323 committee of Sun Yat-sen University Cancer Center.

324

325 Detailed sample information including the geographic origin of the 269 isolates used for
326 WGS was given and summarized in **Supplementary Table 1**. For discovery phase of EBV
327 whole-genome association study (GWAS) with NPC, we selected 156 cases and 47 controls
328 exclusively from the NPC-endemic region out of the 269 EBV WGS isolates. For the
329 validation phase, 990 NPC cases and 1105 healthy controls from the endemic population-
330 based case-control study were used by genotyping GWAS candidate SNPs (For details, see
331 **Supplementary Note and Supplementary Fig. 11**).

332

333 **Sample processing.** Saliva samples were collected into vials containing lysis buffer (50 mM
334 Tris, pH 8.0, 50 mM EDTA, 50 mM sucrose, 100 mM NaCl, 1% SDS). Tumor specimens
335 were obtained from biopsy samples collected during surgical treatment and confirmed by
336 histopathological examination. Both saliva and tumor specimens were stored at -80 °C. DNA
337 was extracted from the saliva using the Chemagic STAR (Hamilton Robotics, Sweden) and
338 from the tumor biopsy, plasma and NPC cell line C666-1 using the DNeasy blood and tissue
339 kit (Qiagen).

340

341 **EBV genome quantification, whole genome sequencing and variant calling.** Using real
342 time PCR targeting a DNA fragment at the *BALF5* gene (5' and 3' primers,
343 GGTCACAATCTCCACGCTGA and CAACGAGGCTGACCTGATCC), we quantified the
344 amount EBV DNA in each sample. The mean Ct values from three independent replicates
345 was used to select patient samples for viral whole genome sequencing (Ct value < 30,
346 detailed information can be found in the **Supplementary Note**).

347

348 EBV genomes were captured using the MyGenostics GenCap Target Enrichment Protocol
349 (GenCap Enrichment, MyGenostics, USA). After capture enrichment, DNA libraries were
350 prepared and sequenced using the Illumina HiSeq 2000 platform according to standard
351 protocols (Illumina Inc., San Diego, CA, USA). After raw sequence processing and quality
352 control, paired-end reads were aligned to the EBV B95-8 reference genome (NC_007605.1)
353 using the Burrows-Wheeler Aligner (BWA, version 0.7.5a)^{28 29}. The average sequencing
354 depth was 1,282 (range, 32 to 6,629). High coverage (average, 98.02%; range, 94.44% to
355 99.91%;) was achieved. (**Supplementary Fig. 1**).

356

357 Following GATK's best practice (version 3.2-2), an initial set of 8,469 variants were first
358 called after base and variant recalibration and filter³⁰. In order to avoid inaccurate calling, we
359 further filtered out variants that has (i) low coverage support (depth < 10×), (ii) in repetitive
360 elements, (iii) within 5 bp of an indel, and 7,962 variants were retained for subsequent EBV
361 phylogenetic, principal component and association analyses. The functional annotation of the
362 EBV variants was performed using the SNPEff package according to the reference genome
363 (NC_007605.1, NCBI annotation, Nov 2013)³¹. A complete description of the sequencing
364 and variant calling is presented in the **Supplementary Note**. Sequencing and variant statistics
365 did not find outlier of EBV isolates sequenced in current study (**Supplementary Fig. 1**).

366

367 To evaluate the accuracy of our sequencing and variant calling, subsets of EBV variants were
368 validated using either the Sanger sequencing or MassArray iPLEX assay (Agena
369 Bioscience). Two independent technologies can provide orthogonal evaluations of the
370 sequencing accuracy. 299 PCR fragments were amplified from 53 randomly selected EBV
371 isolates and re-sequenced using the Sanger sequencing. Comparing the SNPs called by WGS
372 and the Sanger sequencing revealed a concordance rate of 97.55% (**Supplementary Table 4**).
373 Similarly, a high concordance rate of 99.988% between the WGS and MassArray iPLEX
374 assay was found when genotyping 37 variants in 239 selected samples (**Supplementary**
375 **Table 5**). In addition, when comparing the re-sequenced C666-1 EBV genome against the
376 publicly available sequence¹⁸, a high concordance rate of 97.93% was found
377 (**Supplementary Table 3**).

378

379 In order to understand viral genomes from multiple sample types from the same patient, two
380 EBV fragments (position 80,089 to 80,875 and position 81,092 to 81,829) containing 89
381 SNPs were resequenced using the Sanger method from paired saliva and tumor samples from
382 the same set of patients. Across 25 NPC patients with paired tumor and saliva samples,
383 pairwise difference (defined as the genotype discordance rate at the 89 SNPs) between the
384 tumor samples of the 25 patients (inter-host difference) as well as between the paired tumor
385 and saliva samples of the same patient (intra-host difference) were calculated and compared
386 (**Supplementary Fig. 3**). The median inter-patient difference was 13.5% (1st to 3rd quartile:
387 3.7-16.9%), and the median intra-host difference was only 1.1% (1st to 3rd quartile: 0-3.4%).
388 High concordance rate between saliva and tumor tissue suggests that paired saliva and tumor
389 sample from the same patient are highly similar.

390

391 **Genotyping analysis of EBV variants by MassArray iPLEX.** In order to genotype the
392 EBV variants in the Zhaoqing 990 cases and 1105 controls, genotyping was conducted using
393 customized primers and following the recommended protocol by the Agena Bioscience
394 MassArray iPLEX platform. A fixed position within the human albumin gene was used as a
395 positive control. Since the genotyping success rate strongly correlates with the EBV DNA
396 abundance (**Supplementary Fig. 12**), about half of the validation samples (483 of cases and
397 605 of controls) could be successfully genotyped for all the three GWAS candidate markers
398 (i.e. SNP 162215, 162476 and 163364). The slightly lower success rate in cases is consistent

399 with the fact that the EBV DNA abundance was lower in the saliva from patients than the
400 healthy controls. For detailed information, see **Supplementary Note**.

401

402 **Determining single versus multiple EBV infections.** Previous studies have found that EBV
403 genome usually underwent clonal expansion in NPC tumors or other malignancies^{26,27,32}. In
404 the scenario of clonal expansion, EBV genome is stable and intra-host mutation rate is often
405 low, and heterozygous variants as a result of quasi-species evolution within a host are not
406 frequent^{12,18,33}. On the contrary, EBV isolates from specimens with multiple infections will
407 have a higher number of heterozygous variants. We plotted the percentage of heterozygous
408 variants across all the 270 samples from the WGS analysis and observed that heterozygosity
409 (defined as percentage of heterozygous variants) across all the samples showed two different
410 distributions with low and high numbers of heterozygous variants. By fitting two curves to
411 the lower and higher quantiles of the empirical distribution, we defined the reflection point
412 (i.e. the intersection of the two distributions) as the cutoff value (**Supplementary Fig. 9**).
413 Samples with the proportion of heterozygous variants lower than the cutoff value were
414 identified as single-infection samples, whereas samples above this threshold were identified
415 as multi-infection samples. For the validation cohort, samples with the homozygous calls at
416 all the three EBV SNPs were regarded as patients infected by single EBV haplotypes. For
417 samples with multiple EBV haplotype infection, haplotypes of the three SNPs were inferred
418 by Beagle 4.1³⁴. For details, see **Supplementary Note**.

419

420 **Phylogenetic and principal component analyses of EBV genome sequences.** The
421 phylogenetic and principal component analyses were performed using 230 EBV single-
422 infection isolates and 97 publicly accessible EBV genomes. For the phylogenetic analysis, we
423 first created the fasta sequence for each resequenced isolate using the variant data extracted
424 from the variant calling. The 230 whole genomes were subsequently combined with the 97
425 public genomes and multiple sequence alignment was carried out using the Multiple
426 Alignment using Fast Fourier Transform (MAFFT)³⁵. After masking the regions of repetitive
427 sequences and poor coverage in resequencing maximum likelihood inference of the
428 phylogenetic relationship was conducted using the Randomized Axelerated Maximum
429 Likelihood (RAxML) assuming a General Time Reversible (GTR) model³⁶. The inferred
430 phylogeny was subsequently rooted using the Evolutionary Placement Algorithm (EPA)
431 algorithm³⁷ from RAxML using a Macacine herpesvirus 4 genome sequence (NC_006146) as
432 the outgroup.

433

434 In the PCA analysis, genomic variation from the 97 public genomes was generated by global
435 pairwise sequence alignment of published genome sequences against the B95-8 reference
436 genome (NC_007605) using the EMBOSS Stretcher³⁸. The variant set is then combined with
437 the polymorphism data extracted from WGS. A combined set of 12,182 SNPs from the 270
438 newly-sequenced isolates and 97 published ones were then used for the PCA analyses.
439 During the PCA analysis, SNPs were first filtered based on allele frequency (minor genotype
440 frequency > 0.05) and linkage disequilibrium (LD pruning with pairwise correlation r^2 value >
441 0.6 within a 1000-bp sliding window). 495 SNPs were included in the PCA analysis using the
442 R package “SNPRelate”³⁹.

443

444 **Principal component analysis of the cases and controls.** To assess the human population
445 structure of the 156 cases and 47 healthy controls used for the EBV GWAS discovery phase,
446 the human DNAs of these samples were genotyped using the OmniZhongHua-8 Chip
447 (Illumina). After sample filtering based on a series of criteria including (i) the calling rate
448 (above 95%), (ii) SNP filtering by minor allele frequency (above 5%), (iii) Hardy-Weinberg
449 equilibrium ($P > 1 \times 10^{-6}$), (iv) LD-based SNP pruning ($r^2 < 0.1$ and not within the five high-
450 LD regions⁵), PCA analysis was performed using the PLINK (Version 1.9) based on the
451 discovery samples alone or by combining them with reference samples from the 1000
452 Genome project¹⁹.

453

454 **Association analysis.** Genetic association analysis of EBV variants was performed by testing
455 either single or multiple variants. Single variant association analysis was performed using
456 generalized-linear mixed model with EBV genetic relatedness matrix as random effect²⁰. Sex
457 and age were included as fixed effect, as well as four human PCs to correct for any potential
458 impact of host population structures on the association results. Both single- and multiple-
459 infection samples were included in the association analysis with the status of single- or
460 multiple-infection being included as a covariate to correct for any potential confounding
461 effect due to multiple infection. The genome-wide discovery analysis was performed by
462 testing 1,545 EBV variants (with missing rate < 10%, minor genotype frequency > 0.05 and
463 heterozygosity < 0.1) in 156 cases and 47 healthy controls. The validation analysis was
464 performed by testing three EBV non-synonymous coding SNPs (162215, 162476 and 163364)
465 of *BALF2* in additional 483 cases and 605 population healthy controls that were matched to

466 the cases in term of age and sex from the case-control study in Zhaoqing County, Guangdong
467 China. The logistic regression model was used for validation phase with the adjustment of
468 age, sex and status of single- or multiple-infection of EBV strains. The meta-analysis of the
469 discovery and validation phases was performed with zscore pooling method. Considering the
470 extensive LD across the EBV genome, to obtained a suggestive genome-wide significance of
471 association, we used permutation of a logistic model with adjustment of age, sex, status of
472 single- or multiple-infection and host and EBV population structures. The genome-wide
473 significance (4.07×10^{-4}) was determined by 5% quantile of the empirical distribution of
474 minimum P -values from 10, 000 permutations, as the data-driven threshold to control family-
475 wise error rate under multiple correlated testing.

476

477 The genome-wide multi-variant-based association analysis was performed by testing 1477 bi-
478 allelic EBV variants in Bayesian variable selection regression implemented in piMASS²¹.
479 Age, sex, four human PCs and two EBV PCs were included as covariates. The analysis was
480 performed by partitioning EBV genome into the regions of 20-SNP sliding window with 10
481 overlapping SNPs. The sum of the posterior probabilities of the SNPs (being associated)
482 within a window was calculated as the “region statistic” indicating the strength of the
483 evidence for genetic association in that region.

484

485 To further prioritize potentially casual SNPs in the top hit *BALF2* gene region for validation,
486 we applied further fine-mapping analysis using Bayesian multiple variable selection by
487 PAINTOR3.1⁴⁰. Functional annotation of SNPs was used as a prior to compute the
488 probability of being causal SNPs for each variant in the region. We assumed a single causal
489 variant in *BALF2* genes and calculated 90% credible set which contains the minimum set of
490 variants that jointly have at least 90% probability of including the causal variant.

491

492 **Estimation of population attributable fraction of risk.** The proportion of NPC risk
493 explained by the effect of the two high-risk haplotypes of SNPs 162476 and 163364 (C-T and
494 C-C) was estimated in the validation sample. The attributable fraction of risk and 95%
495 confidence interval were estimated in a logistic regression model with adjustment for age and
496 sex by R package ‘AF’⁴¹. As NPC is not a common disease (prevalence < 40/100,000), the
497 risk ratio can be approximated by OR. Thus, the population attributable fraction can be
498 approximated by $AF \approx 1 - E_z\{OR^{-X}(Z) | Y = 1\}$.

499

500 **Functional analysis of NPC-associated BLAF2 SNPs in cell line.** The DNA fragment of
501 the *BALF2* reference haplotype C-T-C was obtained from B95-8 EBV-BAC plasmid p2089⁴²
502 using PCR. The DNA fragments carrying one of the three non-synonymous SNPs 162215,
503 162476 and 163364 were constructed by site-directed mutagenesis through PCR. The
504 reference and three mutated *BALF2* haplotype PCR fragments was subsequently cloned into
505 the vector (pCDH-CMV-MCS-EF1-Puro), and sequences were verified by Sanger
506 sequencing. The four *BALF2* constructs and the empty vector were transfected into 293T
507 cells using polyethylenimine (PEI) for lentivirus production. Lentivirus infection with EBV-
508 positive TW03 cells and selection of stable cell line were performed as previously described⁴³.
509 Overexpression of the *BALF2* construct was confirmed by Western blot.

510

511 EBV lytic cycle was induced in TW03 cells using TPA (phorbol-12-myristate-13-acetate) (20
512 ng/ml) and SB (sodium butyrate) (2.5mM) for 12 hours. After 12 to 48 hours' culture, total
513 viral DNA within the cells and in the supernatant of culture were extracted using Qiagen
514 DNeasy Blood & Tissue Kits and Qiagen QIAamp DNA Blood Mini Kit, respectively. Three
515 biological replicates were conducted. EBV DNA copy number in the supernatant or cells was
516 measured in triplicate relative to a standard curve by quantitative PCR, and the measurements
517 were normalized by the total number of cells in culture.

518

519 **Acknowledgements**

520 We want to thank all the participants for their generous support for the current study. We
521 would also like to thank Dr. Cui Jie from the Wuhan Institute of Virology, Chinese Academy
522 of Sciences for helpful discussions on viral evolution and phylogenetic analysis, Dr. Zhen Lin
523 from Tulane University for kindly sharing EBV genome annotation files, Mr. Wen-Sheng Liu
524 and Dr. Xiaoyu Zuo from Sun Yat-sen University Cancer Center for providing code supports.
525 This work was supported by the National Natural Science Foundation of China (key project,
526 81430059), the National Key R&D Program of China (No. 2016YF0902000), the National
527 Cancer Institute at the US National Institutes of Health (R01CA115873-01), 1000-talent
528 award (to W.Z.) of China, and the Agency of Science, Technology and Research (A*STAR),
529 Singapore.

530

531 **Author contributions**

532 Y.-X.Z., J.L. and W.Z. were the principal investigators who conceived the study. Y.-X.Z and
533 M.X. obtained financial support. Y.-X.Z., J.L., W.Z. and M.X. designed and oversaw the
534 study. J.L. and X.L. provided supervision over viral genome-wide association study. M.X.,
535 Y.Y. and L.Z. performed sample preparation, quality control, sequencing and genetic
536 statistical analysis. H.C. and W.Z. performed the phylogenetic analysis. M.X, S.Z. and Y.Y.
537 performed genotyping by MassArray iPlex assay. M.X., S.Z. and G.H. performed functional
538 experiments. H.-O.A., W.Y. and Y.-X.Z. supervised the design and implementation of the
539 population-based case-control study in Zhaoqing. W.Y., E.T.C., S.-M.C., W.J., S.-H.X., Z.L.
540 participated in the case-control design, study recruitment, and sample storage and preparation.
541 Z.Z. was responsible for the collection of NPC tissue samples from Guangxi. B.L. was
542 responsible for the collection of NPC and EBV-GC tissue samples from North China. X.G.,
543 M.-Y.C. and R.-J.P. were responsible for the collection of NPC and EBV-lymphoma tissue
544 samples from Guangdong. The manuscript was drafted by M.X., J.L., W.Z. and Y.-X.Z. and
545 revised by V.P. and E.T.C.. All the authors critically reviewed the article and approved the
546 final manuscript.

547

548 **References**

- 549 1 Epstein, M. A., Achong, B. G. & Barr, Y. M. Virus Particles in Cultured Lymphoblasts
550 from Burkitt's Lymphoma. *Lancet* **1**, 702-703 (1964).
- 551 2 Epstein, A. Why and How Epstein-Barr Virus Was Discovered 50 Years Ago. *Current*
552 *topics in microbiology and immunology* **390**, 3-15, doi:10.1007/978-3-319-22822-8_1
553 (2015).
- 554 3 Kieff, E. *Epstein-Barr Virus and Its Replication*.
- 555 4 Zhang, L. F. *et al.* Incidence trend of nasopharyngeal carcinoma from 1987 to 2011 in
556 Sihui County, Guangdong Province, South China: an age-period-cohort analysis. *Chin*
557 *J Cancer* **34**, 350-357, doi:10.1186/s40880-015-0018-6 (2015).
- 558 5 Bei, J. X. *et al.* A genome-wide association study of nasopharyngeal carcinoma
559 identifies three new susceptibility loci. *Nature genetics* **42**, 599-603,
560 doi:10.1038/ng.601 (2010).
- 561 6 Bei, J. X. *et al.* A GWAS Meta-analysis and Replication Study Identifies a Novel
562 Locus within CLPTM1L/TERT Associated with Nasopharyngeal Carcinoma in
563 Individuals of Chinese Ancestry. *Cancer Epidemiol Biomarkers Prev* **25**, 188-192,
564 doi:10.1158/1055-9965.EPI-15-0144 (2016).
- 565 7 Cui, Q. *et al.* An extended genome-wide association study identifies novel
566 susceptibility loci for nasopharyngeal carcinoma. *Human molecular genetics*,
567 doi:10.1093/hmg/ddw200 (2016).
- 568 8 Tang, M. *et al.* The principal genetic determinants for nasopharyngeal carcinoma in
569 China involve the HLA class I antigen recognition groove. *PLoS genetics* **8**,
570 e1003103, doi:10.1371/journal.pgen.1003103 (2012).
- 571 9 Baer, R. *et al.* DNA sequence and expression of the B95-8 Epstein-Barr virus
572 genome. *Nature* **310**, 207-211 (1984).

- 573 10 Zeng, M. S. *et al.* Genomic sequence analysis of Epstein-Barr virus strain GD1 from a
574 nasopharyngeal carcinoma patient. *J Virol* **79**, 15323-15330,
575 doi:10.1128/JVI.79.24.15323-15330.2005 (2005).
- 576 11 Dolan, A., Addison, C., Gatherer, D., Davison, A. J. & McGeoch, D. J. The genome of
577 Epstein-Barr virus type 2 strain AG876. *Virology* **350**, 164-170,
578 doi:10.1016/j.virol.2006.01.015 (2006).
- 579 12 Liu, P. *et al.* Direct sequencing and characterization of a clinical isolate of Epstein-
580 Barr virus from nasopharyngeal carcinoma tissue by using next-generation
581 sequencing technology. *J Virol* **85**, 11291-11299, doi:10.1128/JVI.00823-11 (2011).
- 582 13 Lin, Z. *et al.* Whole-genome sequencing of the Akata and Mutu Epstein-Barr virus
583 strains. *J Virol* **87**, 1172-1182, doi:10.1128/JVI.02517-12 (2013).
- 584 14 Palser, A. L. *et al.* Genome diversity of Epstein-Barr virus from multiple tumor types
585 and normal infection. *J Virol* **89**, 5222-5237, doi:10.1128/JVI.03614-14 (2015).
- 586 15 Correia, S. *et al.* Natural Variation of Epstein-Barr Virus Genes, Proteins, and Primary
587 MicroRNA. *J Virol* **91**, doi:10.1128/JVI.00375-17 (2017).
- 588 16 Kwok, H. *et al.* Genomic diversity of Epstein-Barr virus genomes isolated from
589 primary nasopharyngeal carcinoma biopsy samples. *J Virol* **88**, 10662-10672,
590 doi:10.1128/JVI.01665-14 (2014).
- 591 17 Edwards, R. H., Seillier-Moiseiwitsch, F. & Raab-Traub, N. Signature amino acid
592 changes in latent membrane protein 1 distinguish Epstein-Barr virus strains. *Virology*
593 **261**, 79-95, doi:10.1006/viro.1999.9855 (1999).
- 594 18 Tso, K. K. *et al.* Complete genomic sequence of Epstein-Barr virus in nasopharyngeal
595 carcinoma cell line C666-1. *Infectious agents and cancer* **8**, 29, doi:10.1186/1750-
596 9378-8-29 (2013).
- 597 19 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature*
598 **526**, 68-74, doi:10.1038/nature15393 (2015).
- 599 20 Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in
600 Genetic Association Studies via Logistic Mixed Models. *American journal of human*
601 *genetics* **98**, 653-666, doi:10.1016/j.ajhg.2016.02.012 (2016).
- 602 21 Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide
603 association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780-1815,
604 doi:10.1214/11-AOAS455 (2011).
- 605 22 Rowe, M. *et al.* Distinction between Epstein-Barr virus type A (EBNA 2A) and type B
606 (EBNA 2B) isolates extends to the EBNA 3 family of nuclear proteins. *J Virol* **63**,
607 1031-1039 (1989).
- 608 23 Li, D. J. *et al.* The dominance of China 1 in the spectrum of Epstein-Barr virus strains
609 from Cantonese patients with nasopharyngeal carcinoma. *Journal of medical virology*
610 **81**, 1253-1260, doi:10.1002/jmv.21503 (2009).
- 611 24 Young, L. S. & Rickinson, A. B. Epstein-Barr virus: 40 years on. *Nat Rev Cancer* **4**,
612 757-768, doi:10.1038/nrc1452 (2004).
- 613 25 Tsao, S. W. *et al.* The biology of EBV infection in human epithelial cells. *Seminars in*
614 *cancer biology* **22**, 137-143 (2012).
- 615 26 Raab-Traub, N. & Flynn, K. The structure of the termini of the Epstein-Barr virus as a
616 marker of clonal cellular proliferation. *Cell* **47**, 883-889 (1986).
- 617 27 Pathmanathan, R., Prasad, U., Sadler, R., Flynn, K. & Raab-Traub, N. Clonal
618 proliferations of cells infected with Epstein-Barr virus in preinvasive lesions related to
619 nasopharyngeal carcinoma. *N Engl J Med* **333**, 693-698,
620 doi:10.1056/NEJM199509143331103 (1995).
- 621 28 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
622 transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

- 623 29 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
624 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
- 625 30 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-
626 generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806
627 (2011).
- 628 31 Cingolani, P. *et al.* A program for annotating and predicting the effects of single
629 nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*
630 strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92, doi:10.4161/fly.19695 (2012).
- 631 32 Neri, A. *et al.* Epstein-Barr virus infection precedes clonal expansion in Burkitt's and
632 acquired immunodeficiency syndrome-associated lymphoma. *Blood* **77**, 1092-1095
633 (1991).
- 634 33 Weiss, E. R. *et al.* Early Epstein-Barr Virus Genomic Diversity and Convergence
635 toward the B95.8 Genome in Primary Infection. *J Virol* **92**, doi:10.1128/JVI.01466-17
636 (2018).
- 637 34 Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and
638 missing-data inference for whole-genome association studies by use of localized
639 haplotype clustering. *American journal of human genetics* **81**, 1084-1097,
640 doi:10.1086/521987 (2007).
- 641 35 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid
642 multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*
643 **30**, 3059-3066 (2002).
- 644 36 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses
645 with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690,
646 doi:10.1093/bioinformatics/btl446 (2006).
- 647 37 Berger, S. A., Krompass, D. & Stamatakis, A. Performance, accuracy, and Web server
648 for evolutionary placement of short sequence reads under maximum likelihood. *Syst*
649 *Biol* **60**, 291-302, doi:10.1093/sysbio/syr010 (2011).
- 650 38 Li, W. *et al.* The EMBL-EBI bioinformatics web and programmatic tools framework.
651 *Nucleic acids research* **43**, W580-584, doi:10.1093/nar/gkv279 (2015).
- 652 39 Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal
653 component analysis of SNP data. *Bioinformatics* **28**, 3326-3328,
654 doi:10.1093/bioinformatics/bts606 (2012).
- 655 40 Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical
656 fine-mapping studies. *PLoS genetics* **10**, e1004722,
657 doi:10.1371/journal.pgen.1004722 (2014).
- 658 41 Dahlqvist, E., Zetterqvist, J., Pawitan, Y. & Sjolander, A. Model-based estimation of
659 the attributable fraction for cross-sectional, case-control and cohort studies using the
660 R package AF. *Eur J Epidemiol* **31**, 575-582, doi:10.1007/s10654-016-0137-7 (2016).
- 661 42 Delecluse, H. J., Pich, D., Hilsendegen, T., Baum, C. & Hammerschmidt, W. A first-
662 generation packaging cell line for Epstein-Barr virus-derived vectors. *Proc Natl Acad*
663 *Sci U S A* **96**, 5188-5193 (1999).
- 664 43 Deng, C. *et al.* TNFRSF19 Inhibits TGFbeta Signaling through Interaction with
665 TGFbeta Receptor Type I to Promote Tumorigenesis. *Cancer Res* **78**, 3469-3483,
666 doi:10.1158/0008-5472.CAN-17-3205 (2018).
- 667

668

669 **Figure legends**

670 **Figure 1 Principal component and phylogenetic analyses of EBV genomes.** (a) Principal
671 component analysis of 270 EBV isolates sequenced in current study and 97 published isolates.
672 PC1 and PC2 scores are shown. Explaining 26.9% of the total genomic variance, PC1
673 discriminates between East Asian and Western/African strains. PC2 explains 15.3% of the
674 total variance. (b) Phylogeny of 230 EBV single strains sequenced in current study and 97
675 published strains. Macacine herpesvirus 4 genome sequence (NC_006146) was used as the
676 outgroup to root the tree. (c) Geographical origins and phenotypes of samples from which
677 EBV strains were sequenced are shown with colors as indicated. (d) The normalized values
678 of the first two principal-component scores (PC1 and PC2) are shown by colors from blue to
679 red. (e) The genotypes of SNPs 162215, 162476 and 163364 in each isolate. Dashed lines in
680 (a) and (b) indicate the separation between East Asian and Western/African strains. Red dot
681 on the phylogeny indicates the lineage of NPC-dominant EBV strains, where 22 of 37 strains
682 from healthy controls from NPC-endemic southern China were located.

683

684 **Figure 2 Genome-wide association analysis of EBV variants in 156 NPC cases and 47**
685 **controls.** (a) Manhattan plot of genome-wide P values from the association analysis using
686 generalized-linear mixed model. The $-\log_{10}$ -transformed P values (y axis) of 1545 variants in
687 156 NPC cases and 47 controls are presented according to their positions in the EBV genome.
688 The minimum P value (SNP162507) is 9.99×10^{-5} . Suggestive genome-wide significance P
689 value threshold of 4.07×10^{-4} (red line) was shown. (b) The regional plot of posterior
690 probability of association. EBV genome was partitioned into overlapping 20-variant bins
691 with 10-variant overlap between adjacent bins. The sum of the posterior probability for
692 variants was assigned to each region. One region with strong evidence (> 0.90) for
693 association with NPC risk is shown in green. (c) Schematic of EBV genes. Repetitive regions
694 in EBV genomes are masked by light blue.

695

696 **Figure 3 Functional analysis of high-risk genotypes of SNPs 162215, 162476, and 163364**
697 **in terms of regulation of EBV lytic DNA replication.** (a) Effect of the *BALF2* haplotypes
698 composed of three SNPs on EBV lytic DNA replication. EBV DNA inside cells was
699 measured by quantitative PCR in the NPC TW03 cell line 48 hours after induction of viral
700 lytic cycle activation. The cells were transfected with the following haplotypes: reference
701 haplotype H-L-L (C-T-C), low-risk haplotype L-L-L (A-T-C) and high-risk haplotype H-H-H

702 (C-C-T). The cells transfected with empty vector were used as controls. L: low-risk genotype;
703 H: high-risk genotype. Significant differences in EBV DNA amount were calculated by
704 Student's *t* test: * $P = 0.024$, ** $P = 0.004$. Means and s.d. of three independent experiments
705 are shown. **(b)** Distribution of EBV DNA abundance in saliva samples from 533 NPC cases
706 and 651 controls relative to the average abundance in controls. Of the total 536 saliva
707 samples from cases, three had missing EBV DNA amount values. Distribution frequency
708 density of relative EBV DNA amount is shown. P value was determined using linear
709 regression adjusted for age and sex.

Figure 1

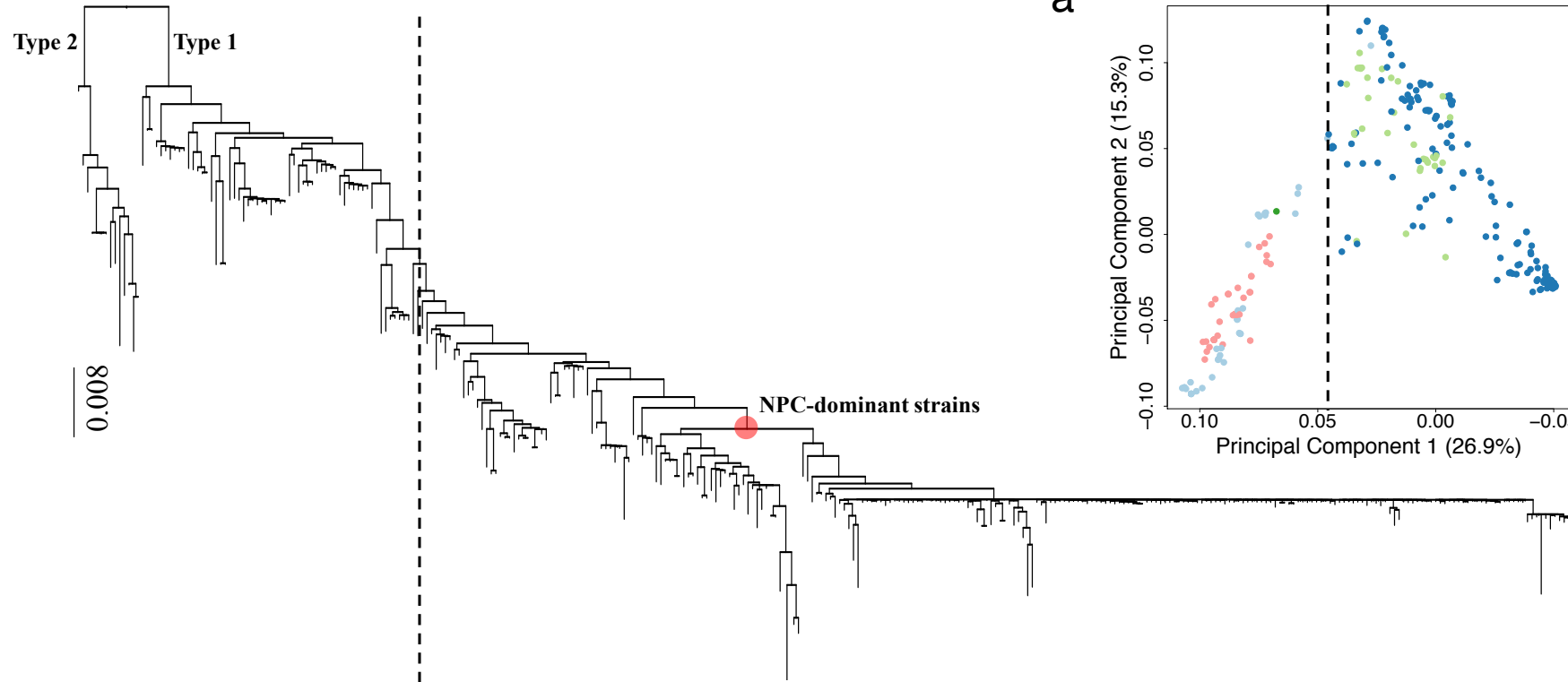
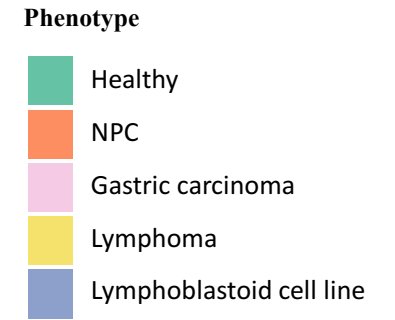
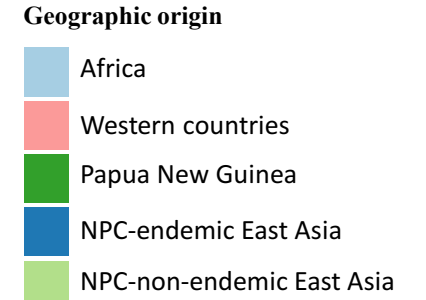
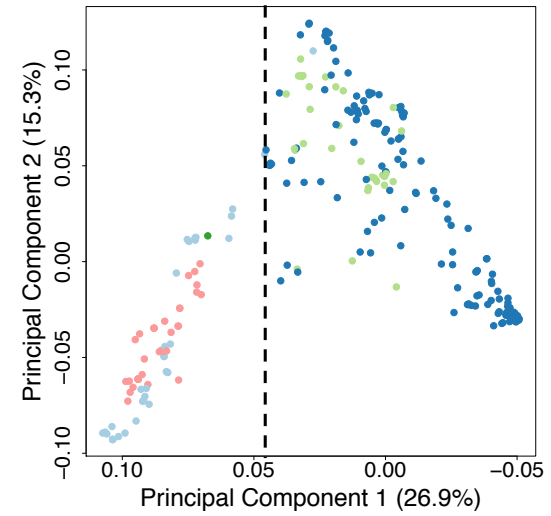
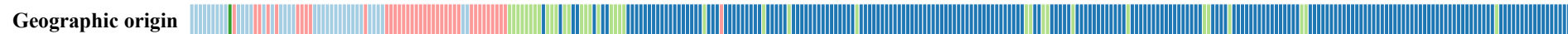
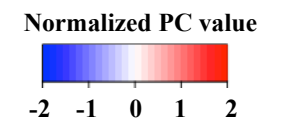
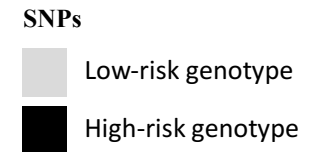
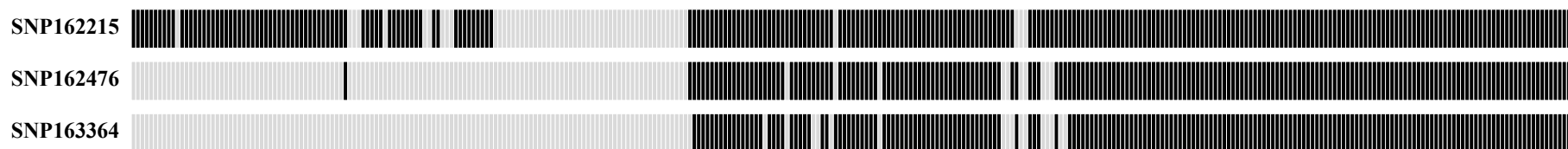
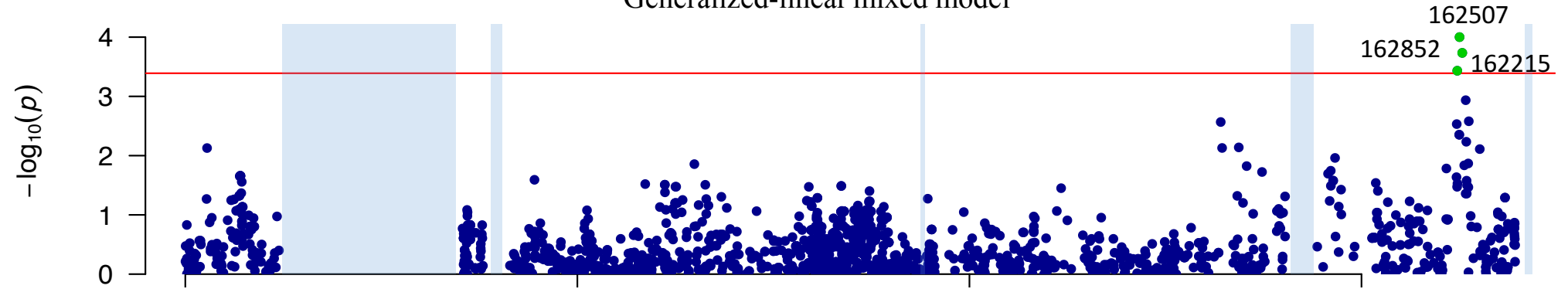
b**a****c****d****e**

Figure 2**a**

Manhattan plot of association study of EBV genome with NPC

156 cases, 47 controls

Generalized-linear mixed model

**b**

Sum of posterior probability of association

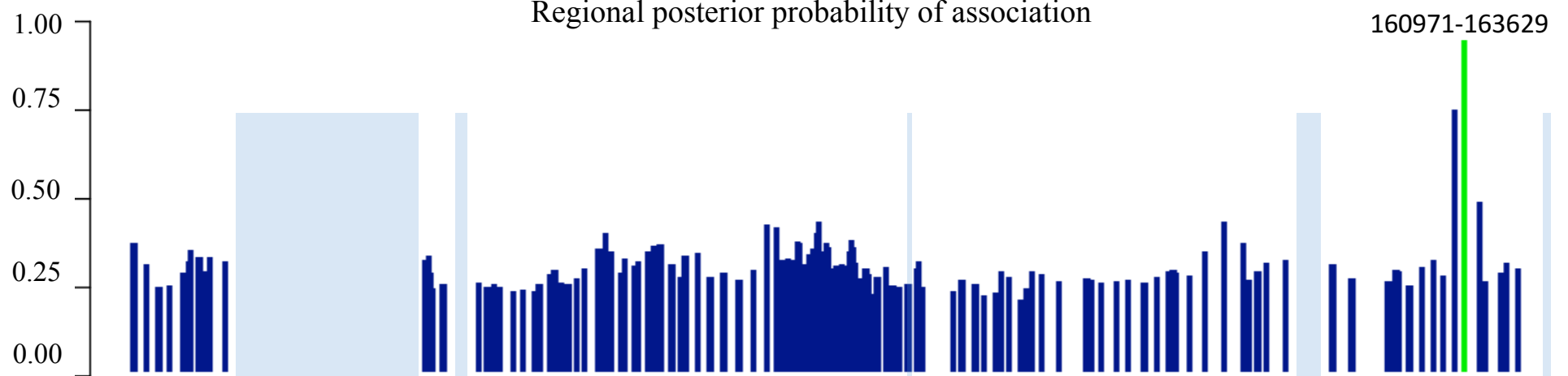
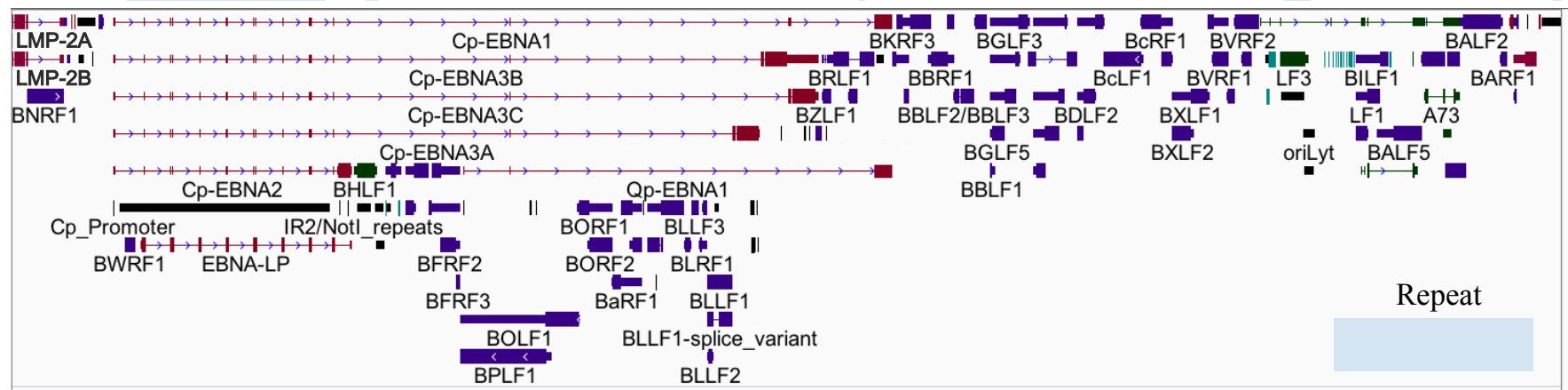
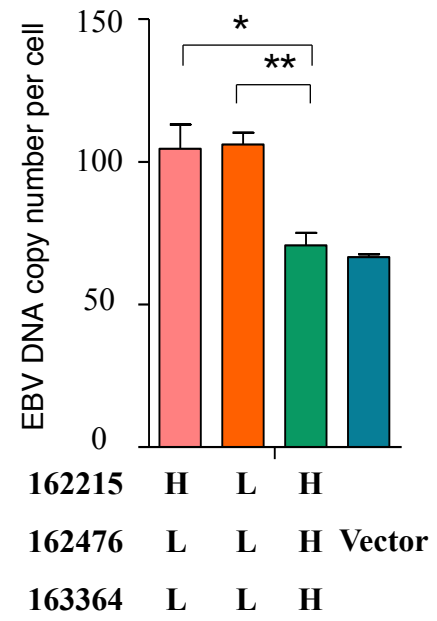
**c**

Figure 3

a



b

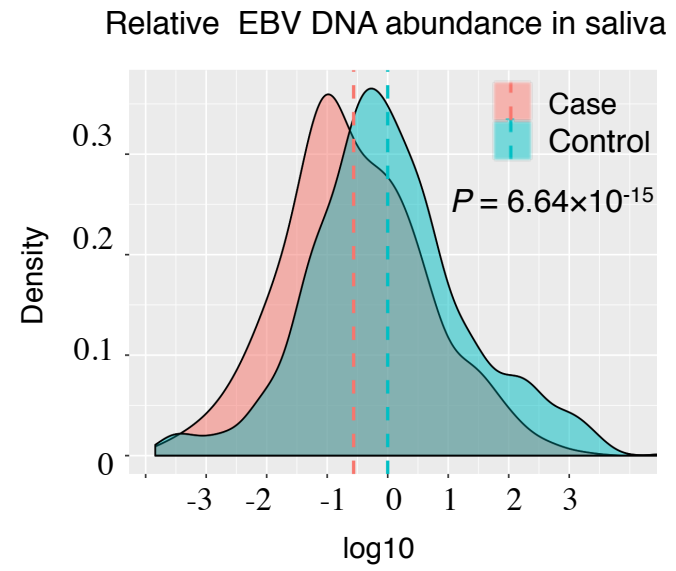


Table 1 The association of three non-synonymous SNPs in *BALF2* gene and their odds ratios for NPC risk.

| Position | Genotypes | High-risk genotype | Discovery | | | Validation | | | Combined | | | Odds ratio (95% CI) | <i>P</i> value conditional on SNPs | | Annotation |
|----------|-----------|--------------------|-----------|-------------|----------------|------------|--------------|----------------|-----------|--------------|----------------|---------------------|------------------------------------|----------|----------------------|
| | | | 156 cases | 47 controls | <i>P</i> value | 483 cases | 605 controls | <i>P</i> value | 639 cases | 652 controls | <i>P</i> value | | 163364 | 162476 | |
| 162215 | C/A | C | 96.15% | 65.96% | 3.69E-04 | 95.03% | 74.71% | 1.84E-16 | 95.31% | 74.08% | 2.98E-19 | 7.62 (5.02-11.57) | 1.30E-04 | 2.39E-01 | <i>BALF2</i> , V700L |
| 162476 | T/C | C | 93.59% | 61.70% | 4.44E-03 | 94.00% | 65.12% | 1.21E-24 | 93.90% | 64.88% | 5.90E-26 | 8.79 (5.90-13.09) | 1.75E-06 | | <i>BALF2</i> , I613V |
| 163364 | C/T | T | 88.46% | 48.94% | 5.83E-03 | 83.85% | 45.45% | 1.83E-35 | 84.98% | 45.71% | 7.18E-36 | 6.52 (4.90-8.68) | | 3.23E-13 | <i>BALF2</i> , V317M |

Frequencies of high-risk genotypes in discovery, validation and combined analyses are indicated. The association of three EBV SNPs with NPC risk was tested by meta-analysis of the combined discovery and validation phases. Conditional regression analyses were performed in combined data sets and *P* values of SNP association are listed. Odds ratios conferred by high-risk genotypes and the 95% confidence intervals (CI) were estimated by meta-analysis of combined discovery and validation phases.

Table 2 EBV haplotypes composed of SNPs 162215, 162476 and 163364 and their odds ratios for NPC risk in 639 cases and 652 controls.

| EBV subtype (162215-162476-163364) | 639 cases | | 652 controls | | Odds ratio (95% CI)* | P value |
|---------------------------------------|-----------|--------|--------------|--------|-------------------------|----------|
| | no. | % | no. | % | | |
| L-L-L (A-T-C) | 25 | 3.91% | 171 | 26.23% | | |
| H-H-H (C-C-T) | 539 | 84.35% | 293 | 44.94% | 12.10 (7.59 - 19.30) | 1.17E-25 |
| H-H-L (C-C-C) | 57 | 8.92% | 118 | 18.10% | 3.30 (1.90 - 5.72) | 2.20E-05 |
| H-L-L (C-T-C) | 13 | 2.03% | 65 | 9.97% | 1.10 (0.49 - 2.47) | 8.27E-01 |
| other subtypes | 5 | 0.78% | 5 | 0.77% | 4.63 (1.04 - 20.54) | 4.40E-02 |

* Odds ratios of individual EBV subtypes and 95% confidence intervals (CI) were estimated with logistic model by categorizing each subtype as a single variable and adjusted for age, sex and the status of single- or multiple-infection in combined discovery and validation data sets. Subjects with EBV subtype A-T-C, a common low-risk subtype were used as the reference category. H represents high-risk genotypes; L represents low-risk genotypes.