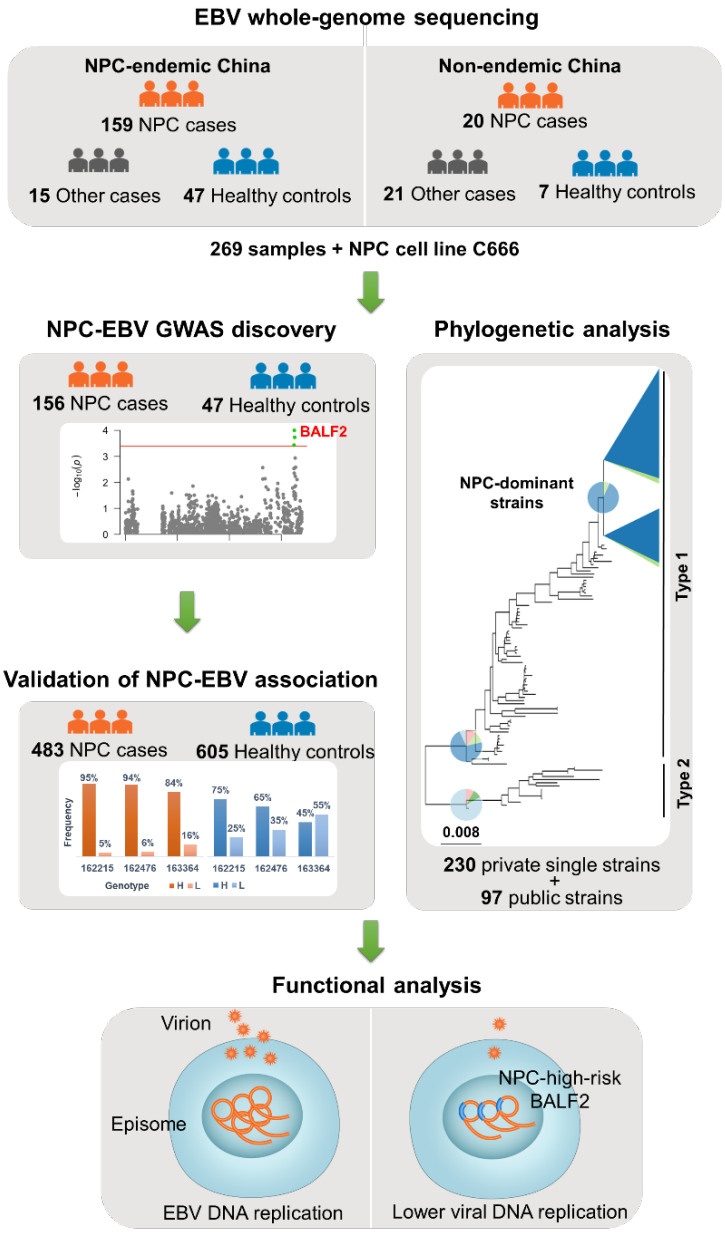
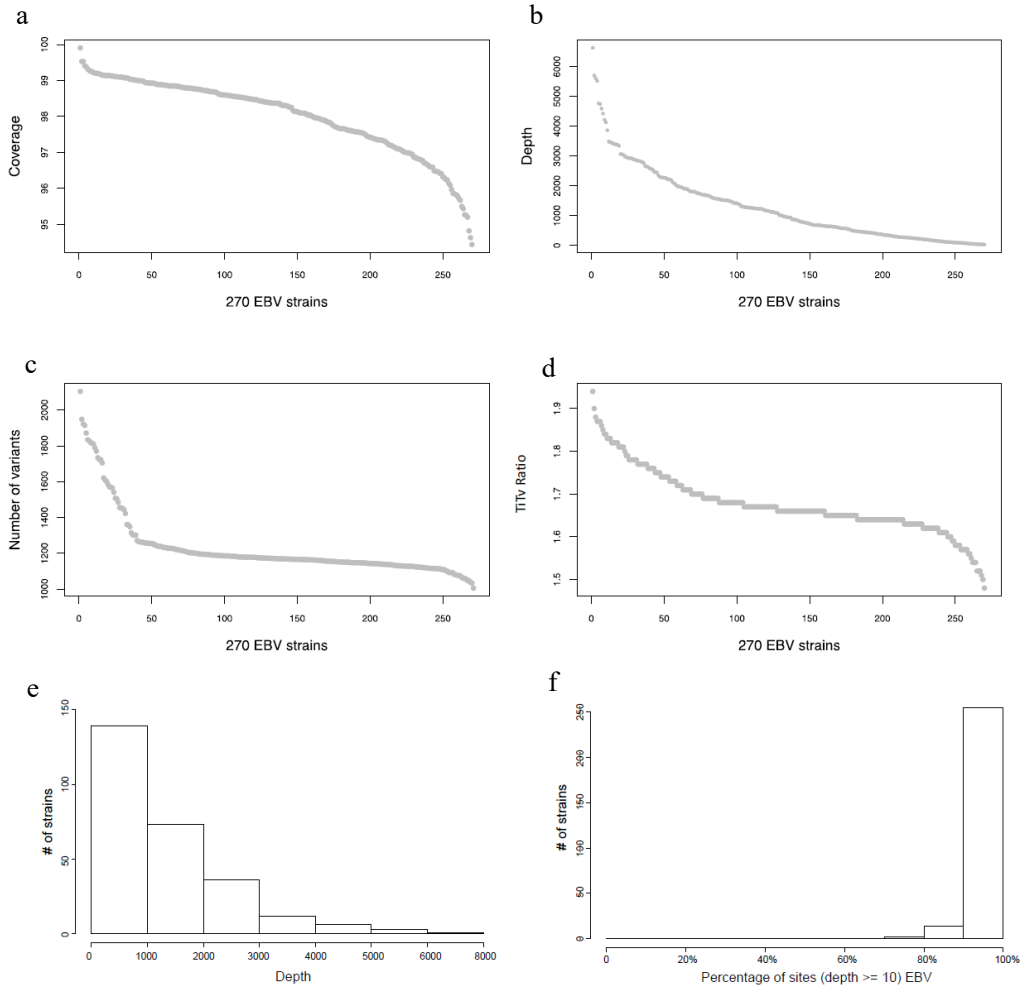
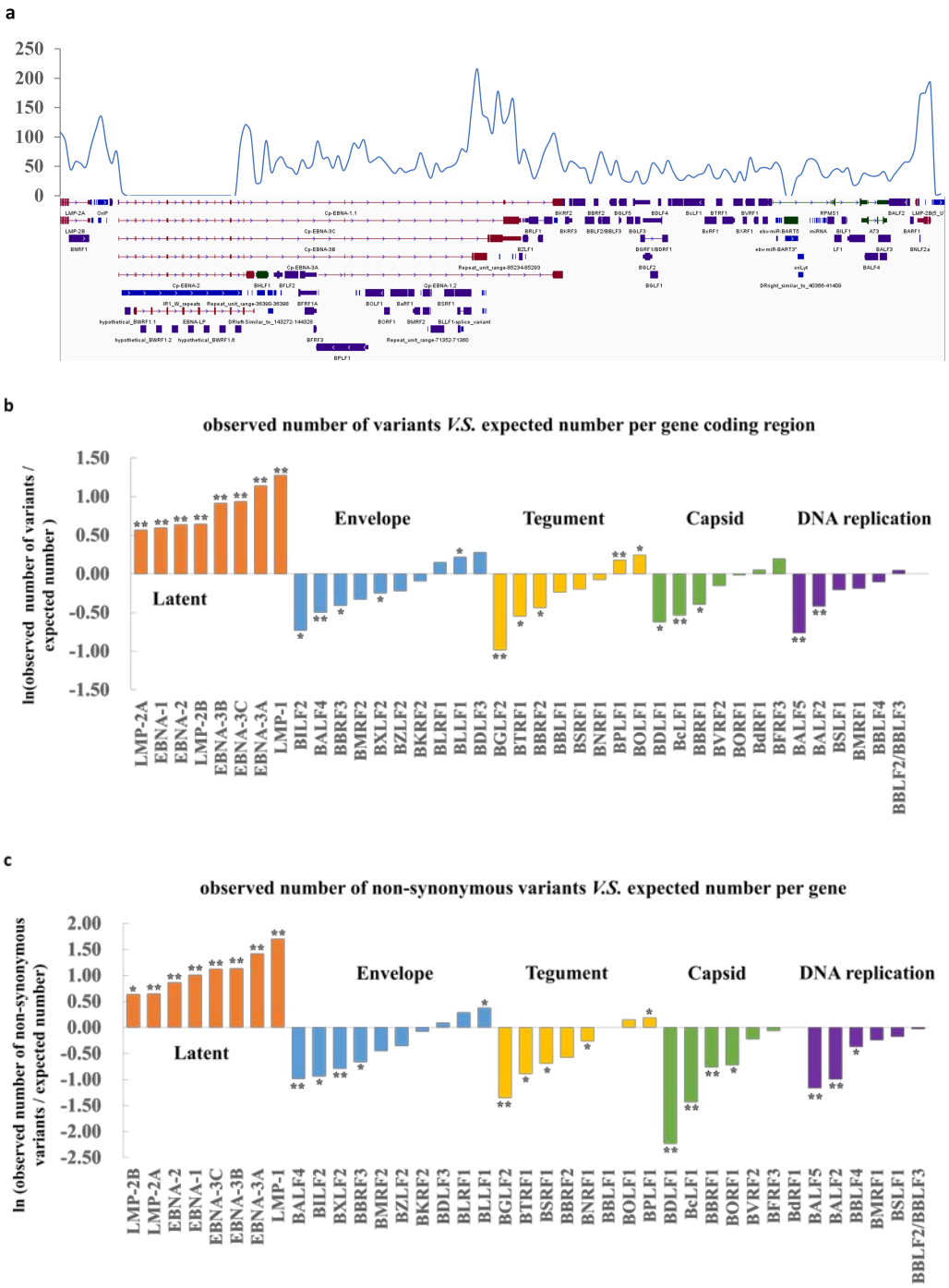


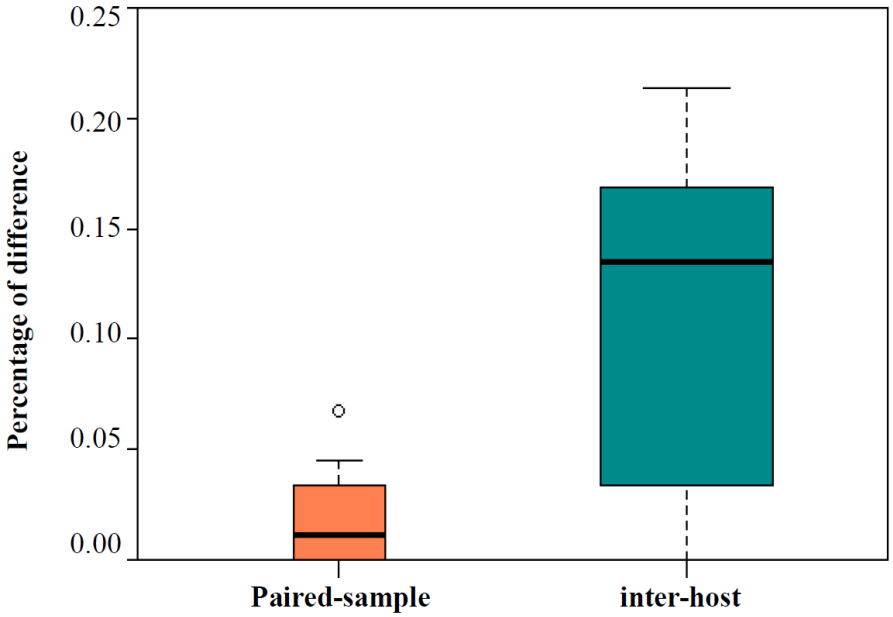
Supplementary Figure 1 Experimental and analytic overview. For EBV whole genome sequencing, we obtained 269 DNA samples from patients with NPC and other EBV-associated cancers and healthy controls from NPC-endemic and non-endemic regions in China, as well as one from EBV-positive NPC cell line C666.1. The EBV genome-wide association study (GWAS) with NPC used whole-genome variants derived from 156 NPC cases and 47 healthy controls from the NPC-endemic regions. The GWAS discovery phase and fine-mapping identified three non-synonymous variants in the *BALF2* gene region showing the strongest association with NPC. These associations were further validated in 483 independent cases and 605 controls from the NPC-endemic region. Phylogenetic analysis of whole genomes of 230 single EBV strains together with 97 published sequences revealed the unique origin of the risk variants in NPC-endemic China, followed by a clonal expansion. In *vitro* functional analyses compared the effects of EBV *BALF2* haplotypes carrying the risk variants on viral lytic DNA replication.



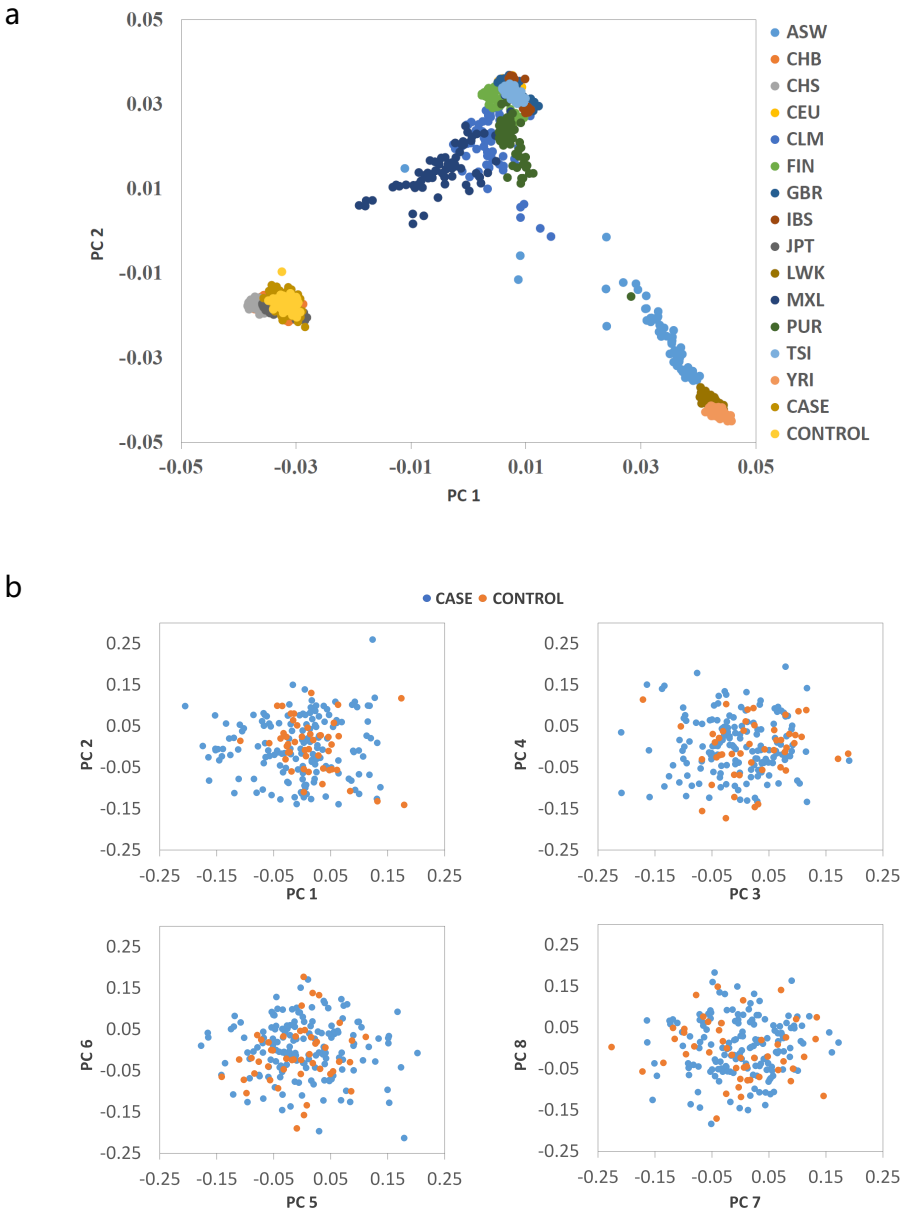
Supplementary Figure 2 Sequencing and variant statistics of each EBV genome isolates indicate no outliers among the 270 EBV isolates. (**a**) Sequencing coverage across EBV genome, ranging from 94% to 99%. (**b**) Average sequencing depth. (**c**) Number of variants. (**d**) Ratio of transition to transversion. (**e**) Frequency histogram of average sequencing depth per isolate. (**f**) Frequency histogram of percentage of reference genome that was covered by 10 or more reads.



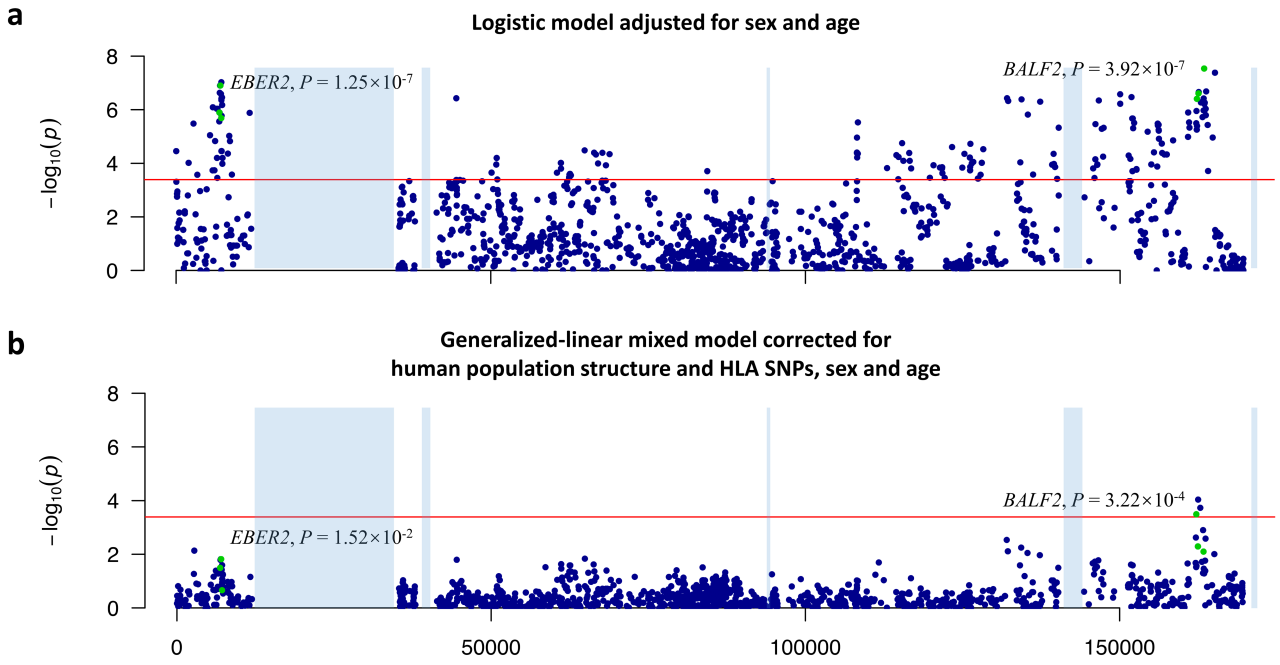
Supplementary Figure 3 Regions encoding latent proteins have highest diversity across EBV genomes. (**a**) Variant frequency across EBV genomes derived from 270 samples. The line graph is plotted across the genome showing the total number of variants in a sliding 1000-nt window.(**b**) Comparison of the observed and expected numbers of variants in gene coding regions. Expected numbers of variants were calculated by multiplying the length of gene coding regions by the number of variants per kilo base. *P* values were calculated by two-sided *Fisher's* exact tests. *, $P < 0.05$; **, $P < 0.001$.(**c**) Comparison of the observed and expected numbers of non-synonymous variants in gene coding regions. Expected numbers of non-synonymous variants were calculated by multiplying the length of gene coding regions by the number of non-synonymous variant per kilobase. *P* values were calculated by two-sided *Fisher's* exact tests. *, $P < 0.05$; **, $P < 0.001$.



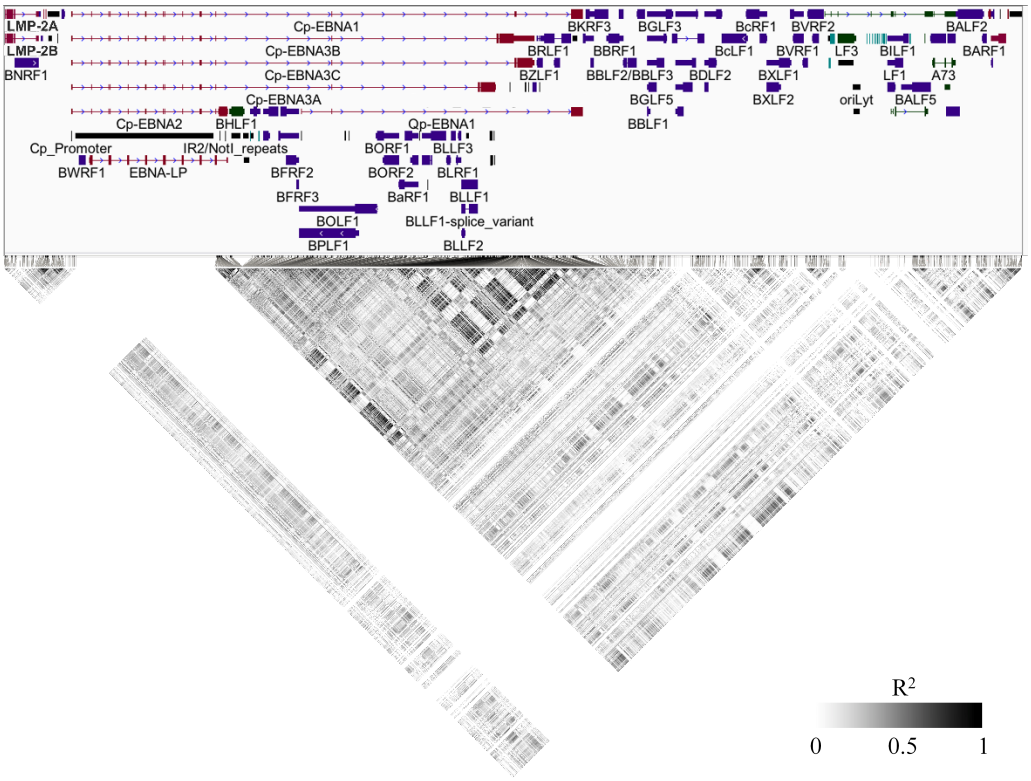
Supplementary Figure 4 The variant discordance rate between paired tumor and saliva samples (paired-sample) versus that between tumors from different patients (inter-host). EBV DNA fragments were sequenced from 25 pairs of NPC tumor and saliva samples. Median, 1st (Q1) and 3rd (Q3)quartiles were shown (box). The outlier is shown as black circle. Q3 + 1.5 × IQR and Q1 - 1.5 × IQR are shown as highest and lowest horizontal line (IQR = Q3 - Q1).



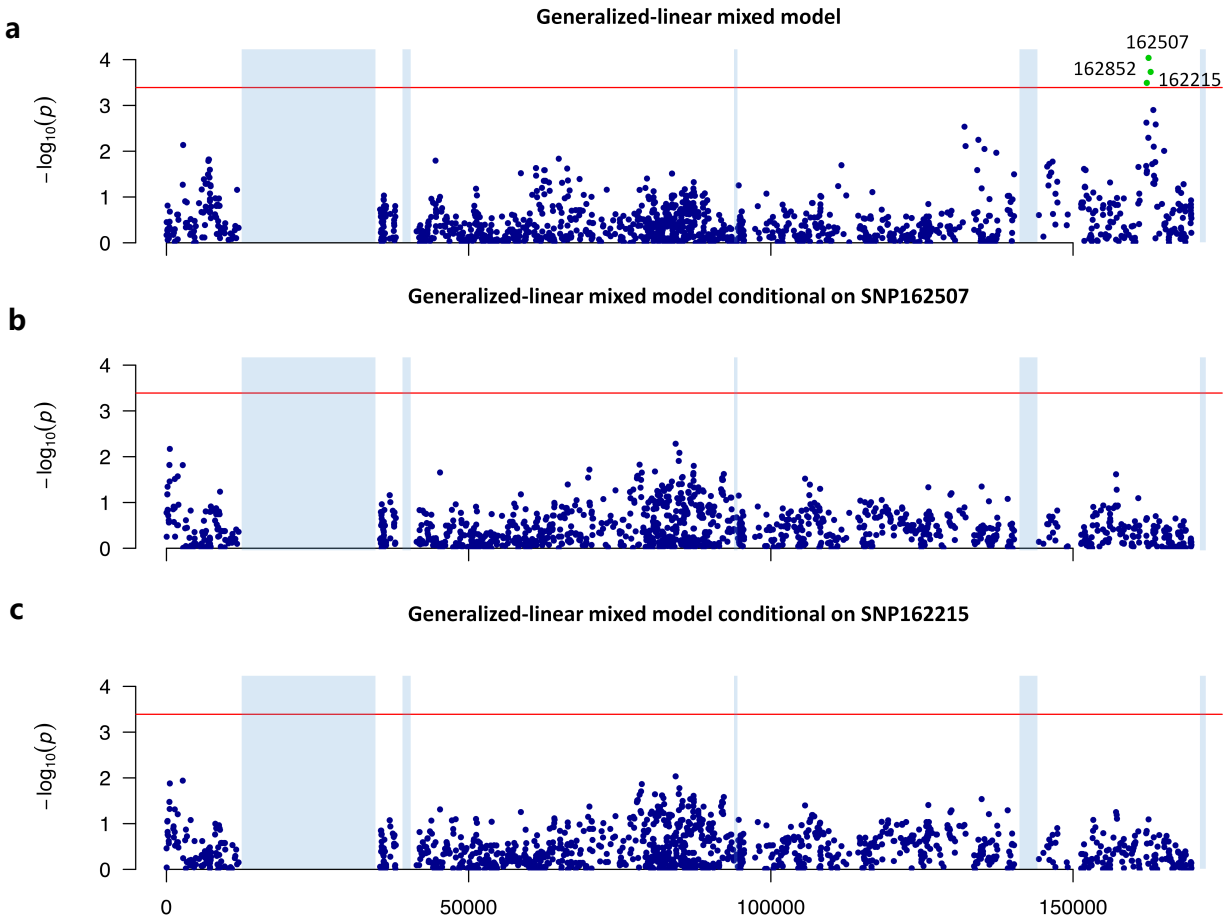
Supplementary Figure 5 Human principal component analysis of the samples used for EBV genome-wide association analysis with NPC. (**a**) The PC scores for each sample were plotted against the first two PCs (PC 1 and PC 2), together with 1000 genome project samples. No outlier was observed between our cases and controls using appropriate criteria in the paper of Price 2006, which defines individuals whose ancestry is at least 6 standard deviations from the mean of one of the top ten PC values as outlier [41]. Population codes and NPC cases and controls used for EBV GWAS were listed at the right panel. ASW, Americans of African ancestry in SW USA; CEU, Utah Residents with Northern and Western European Ancestry; CHB, Chinese Han in Beijing, China; CHS, Southern Han Chinese; CLM, Colombians from Medellin, Colombia; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian Population in Spain; JPT, Japanese in Tokyo, Japan; LWK, Luhya in Webuye, Kenya; MXL, Mexican Ancestry from Los Angeles USA; PUR, Puerto Ricans from Puerto Rico; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeria. (**b**) The PC scores for each NPC case and control were plotted against the first eight PCs (PC 1 to PC 8).



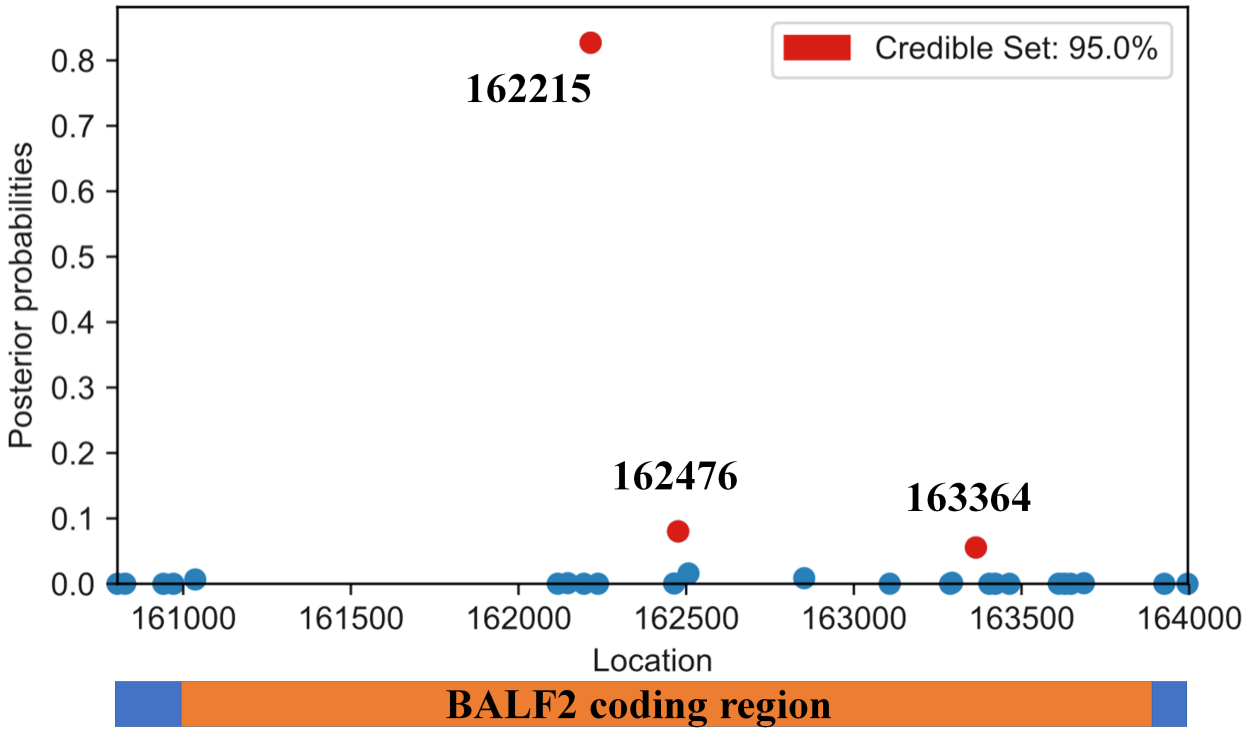
Supplementary Figure 6 Manhattan plot of genome-wide P values from association analyses. Genome-wide *P* values from association analyses without controlling for EBV and human population structure adjusted for age and sex (**a**) versus with controlling for EBV and human population structure using generalized-linear (logistic) mixed model adjusted for human PCs and HLA SNPs, sex and age (**b**) . The $-\log_{10}$-transformed *P* values (y axis) are presented according to their positions in the EBV genome. The red line is the suggestive genome-wide significance *P* value threshold of 4×10^{-4}. The *EBER2* SNPs 6886T>G, 7048A>C and deletion 7188_7191del reported by Hui et al. and our discoveries, *BALF2* non-synonymous variants 162215C>A, 162476T>C and 163364C>T are labeled as green. The association *P* values of *EBER2* SNP 7048A>C and *BALF2* SNP 162215C>A are shown. Repeat regions of the EBV genomes are labeled in light blue.



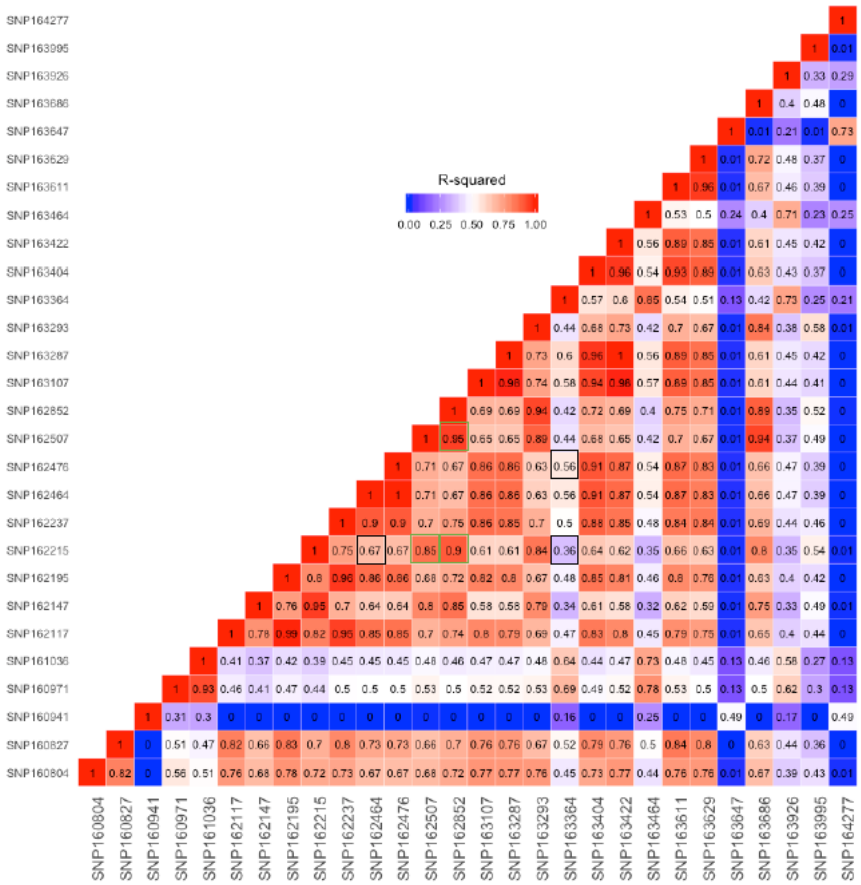
Supplementary Figure 7 EBV genome-wide linkage disequilibrium heatmap. Pair-wise R-squared values between 1545 variants with minor genotype frequency > 0.05 in 156 NPC cases and 47 controls were plotted in lower plot. Higher linkage disequilibrium is presented with darker blocks. The upper panel shows EBV genome annotation.



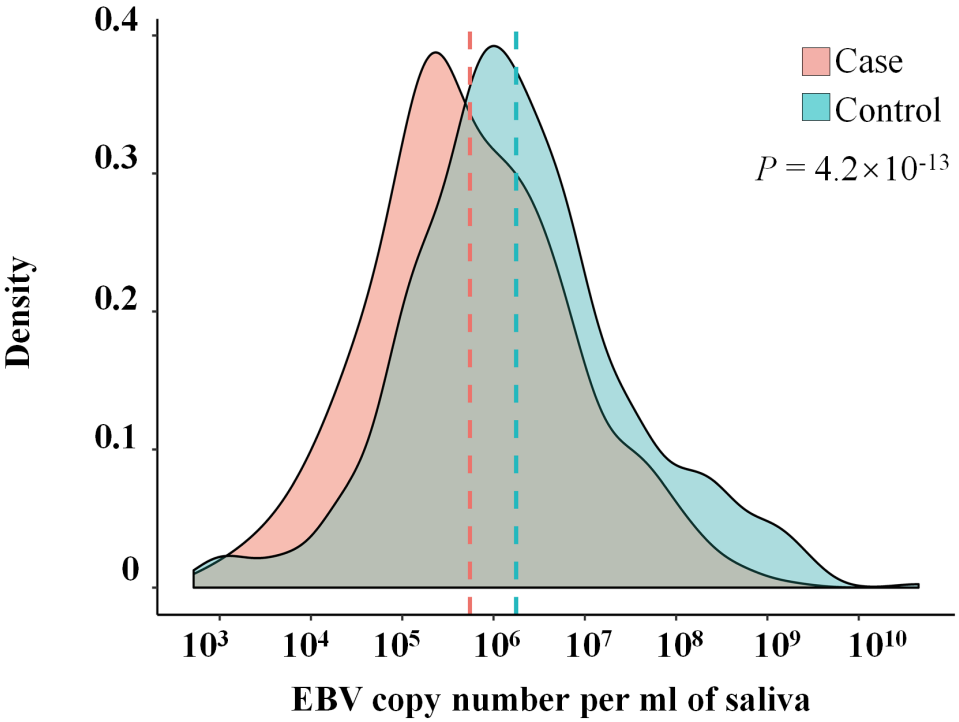
Supplementary Figure 8 EBV genome association study with NPC conditional on SNPs 162507 C>T and 162215C>A. (**a**) Manhattan plot of the genome-wide *P* values of association study. Association was assessed by generalized-linear mixed model (GLMM) with age, sex, status of single or multiple EBV infection, four human PCs and human *HLA* SNPs (rs2860580 and rs2894207) as fixed effects and genetic relatedness matrix as random effects. Genome-wide significant *P* value threshold (red line) was 4.07×10^{-4}. Associations reached genome-wide significance were highlighted by green. (**b-c**) Logistic regressions with GLMM were conditional on SNPs 162507C>T and 162215C>A as indicated. Repeat regions of the EBV genomes are labeled in light blue.



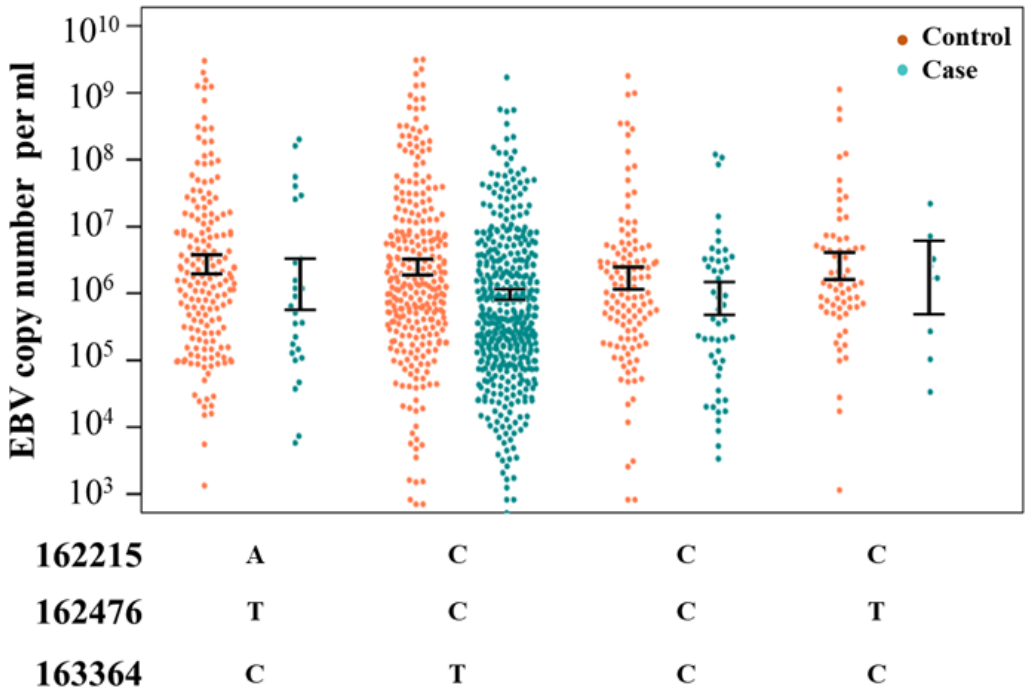
Supplementary Figure 9 Posterior probability association for 28 variants in *BALF2* gene region. Posterior probability was estimated by PAINTOR. The positions of variants in the EBV genome are indicated.



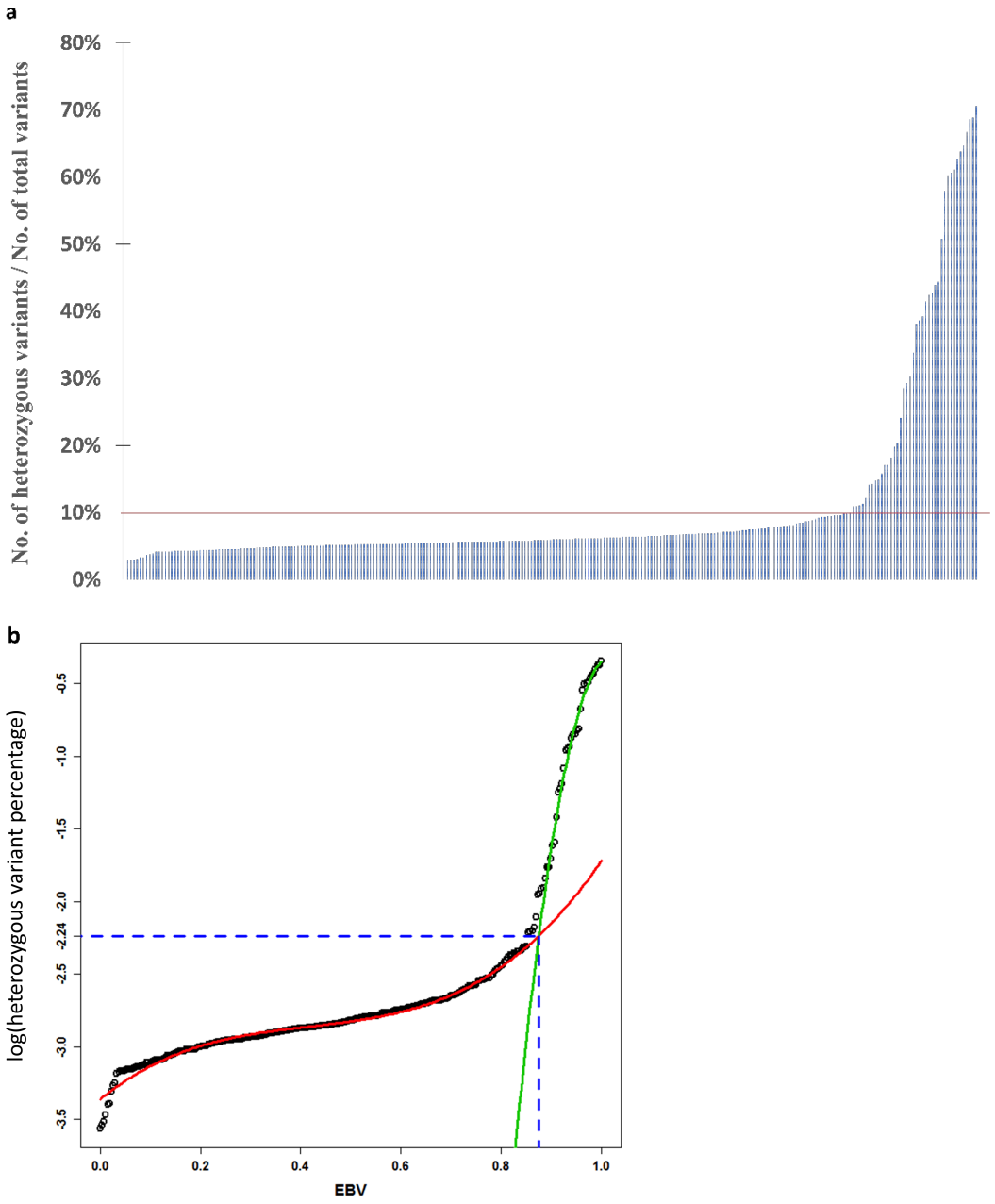
Supplementary Figure 10 Linkage disequilibrium structure of *BALF2* gene region. Pair-wise r-squared values of 28 SNPs in *BALF2* gene region are shown. The R-squared values of SNPs 162215C>A, 162476T>C and 163364C>T are highlighted with black squares. The R-squared values of SNPs 162215C>A and 162507C>T, 162852 G>T are highlighted with green squares.



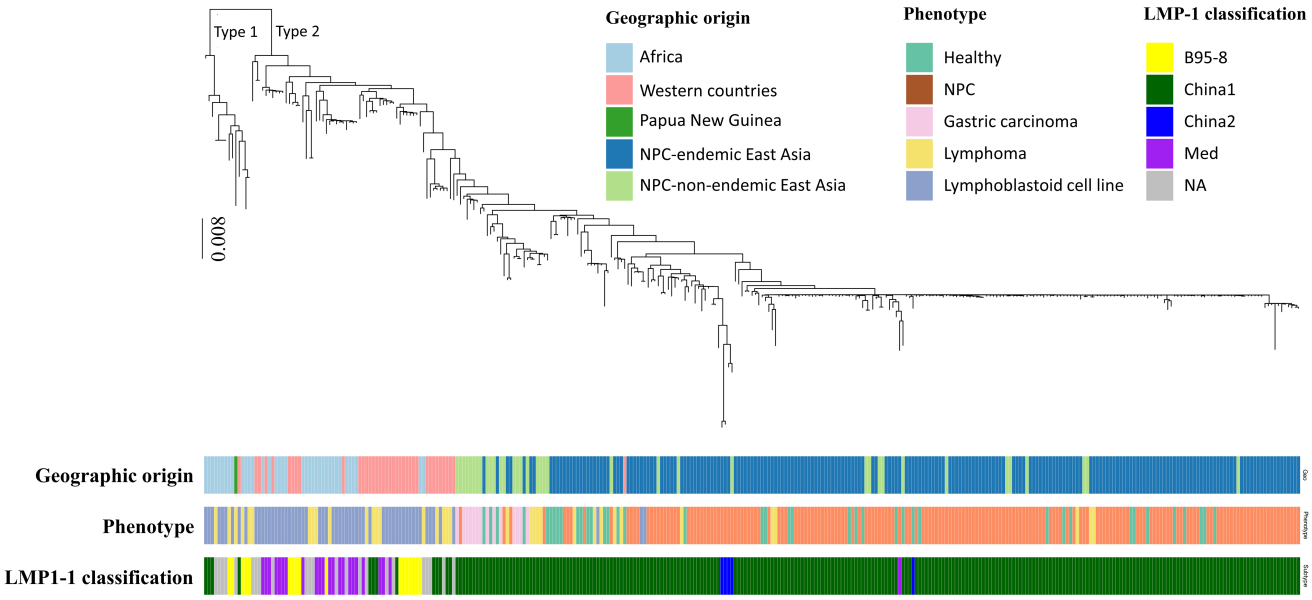
Supplementary Figure 11 Distribution of EBV DNA abundance in saliva samples of 533 NPC cases and 651 controls from the NPC-endemic Zhaoqing County. The distribution frequency density of log-scale of EBV DNA copy number is shown. The mean of EBV copy number per ml saliva of the case and control is 5.50×10^5 and 1.78×10^6 respectively (dashed line indicated). Of the 536 saliva samples from cases, three had missing values for EBV DNA amount. The *P* value was determined with a linear regression of the log of EBV DNA copy number versus case-control status adjusted for age and sex.



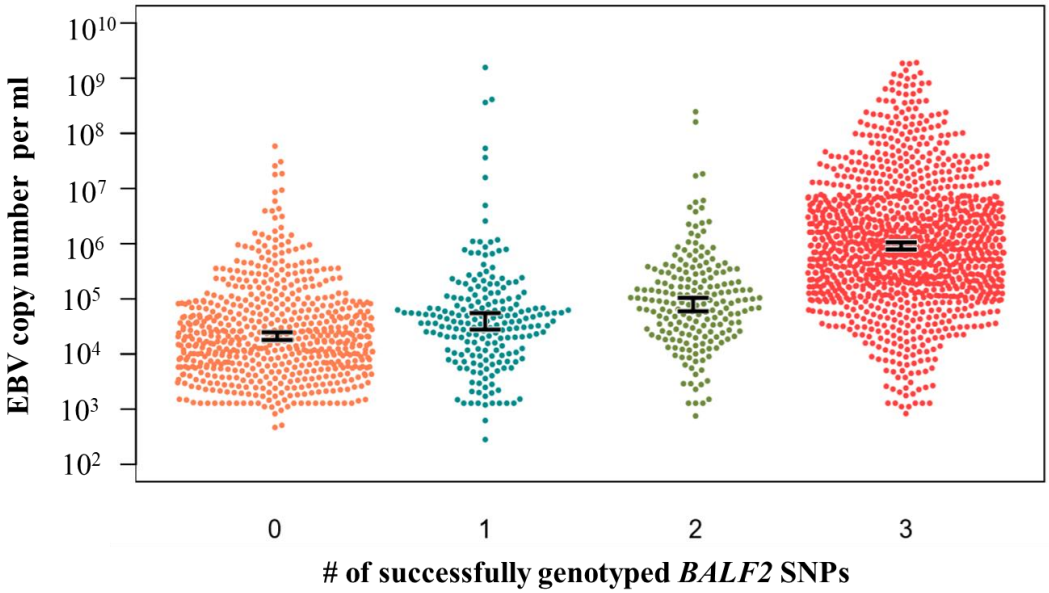
Supplementary Figure 12 Distribution of EBV DNA abundance in saliva samples from 533 NPC cases and 651 controls by each EBV *BALF2* haplotype. *BALF2* haplotypes: low-risk A-T-C, high-risk C-C-T and C-C-C, wild-type B95-8 C-T-C. Of the total 536 saliva samples from cases, three had missing EBV DNA amount values. The 95% confidence intervals of the mean are shown as black bars , respectively.



Supplementary Figure 13 Distribution of genome-wide heterozygous variants in 270 EBV genome isolates. (**a**) Heterozygous variant percentage in 270 EBV genome isolates. Y axis is the percentage of heterozygous variants out of total variants per isolate. Red represents the cut-off of 10.7%. (**b**) Heterozygosity of 10.7% was determined as the cut-off for single infection. Two curves of heterozygosity of single infections (EBV isolates with heterozygous variant proportion < 8% , red curve) and multiple infections (isolates with heterozygous variant proportion > 15% , green curve) were fitted with cubic model. Since the distribution of heterozygous variant proportion is markedly skewed, we applied log-transformation and used the transformed heterozygous variant proportion in the analysis. The two fitted curves intersected at heterozygosity of 10.7% (blue dashed line), which was used as the cut-off value of single infection determination.



Supplementary Figure 14 Classification of 230 newly-sequenced EBV isolates and 97 published EBV isolates based on *LMP-1* C-terminal signatures. Phylogeny of 327 EBV strains. Macacine herpesvirus 4 genome sequence (NC_006146) was used as the outgroup to root the tree. *LMP-1* classifications, geographical origins and phenotypes from which EBV strains were sequenced are shown with colors as indicated.



Supplementary Figure 15 Distribution of EBV DNA abundance in saliva samples from 990 NPC cases and 1105 controls by the number of successfully genotyped *BALF2* SNPs. Of the total 990 saliva samples from cases, three had missing EBV DNA amount values. The 95% confidence intervals of the mean are shown as black bars , respectively.

Supplementary Table 2 Summary of geographic origins and phenotypes of 367 EBV isolates

	Africa	Western countries	NPC-endemic China	NPC-non-endemic East Asia	Total
NPC			**175**	**20**	**195**
Gastric carcinoma				**17**	**17**
Healthy control		**1**	**47**	**7**	**55**
Lymphoma					**43**
Hodgkin		8	8	3	
Burkkit	13		2	2	
NK/T cell			5	2	
LCL	**24**	**5**		**3**	**32**
sLCL-IM		5			**5**
sLCL-PTLD		**19**			**19**
Total	**37**	**38**	**237**	**54**	**366**

One single published strain was sequenced from Burkkit's lymphoma cell line with Papua New Guinea origin and is not listed in this table

Supplementary Table 3 Summary of geographic origins, phenotypes and sample types of 270 EBV isolates sequenced in current study

	NPC-endemic China	NPC-non-endemic China
NPC	**160**	**20**
Tumor tissue	105	19
Saliva	54	
Plasma		1
Cell line	1	
Healthy control saliva	**47**	**7**
Lymphoma biopsy	**15**	**5**
Hodgkin	8	3
Burkkit	2	
NK/T cell	5	2
Gastric carcinoma tissue		**16**
Total	**222**	**48**

Supplementary Table 4 Summary of age and sex of study participants for EBV whole genome sequencing

	Male				Female			
	Age			Male total	Age			Female total
	0-27	28-59	60-		0-27	28-59	60-	
Hodgkin lymphoma				**11**				
NPC-endemic China		6	2					
non-endemic China	1		2					
Burkitt lymphoma				**2**				
NPC-endemic China		2						
NPC				**129**				**50**
NPC-endemic China	1	94	21		3	30	10	
non-endemic China	1	10	2			5	2	
Gastric carcinoma				**14**				**2**
non-endemic China	1	9	4			2		
Healthy control				**41**				**13**
NPC-endemic China		31	7		1	5	3	
non-endemic China	2	1			1	3		
NK/T cell lymphoma				**6**				**1**
NPC-endemic China	1	4						
non-endemic China	1					1		
Total				**203**				**66**

One EBV genome was sequenced from NPC cell line C666.1 and is not included in this table.

Supplementary Table 6 Concordance rate between SNPs from the C666-1 EBV genome sequenced in current study and in published study

C666-1		Current study		Total
		nVariants	nReferences	
Published*	nVariants	1021[a]	117[b]	1138
	nReferences	51[c]	6943[d]	6994
Total		1072	7060	8132

Concordance rate was 97.93%, calculated by (a+d)/(a+b+c+d).

*GenBank accession number: KC617875.1

Supplementary Table 7 Concordance rate between variants discovered by targeted EBV whole-genome sequencing (EBV-WGS) and Sanger sequencing

		EBV-WGS		Total
		nVariants	nReferences	
Sanger	nVariants	153[a]	3[b]	156
	nReferences	5[c]	165[d]	170
Total		158	168	326

Concordance rate was 97.55%, calculated by (a+d)/(a+b+c+d).

Supplementary Table 8 Concordance rate between variants discovered by targeted EBV whole-genome sequencing (EBV-WGS) and MassArray iPlex assay

		EBV-WGS		Total
		nVariants	nReferences	
MassArray iPlex assay	nVariants	4328[a]	0[b]	4328
	nReferences	1[c]	4229[d]	4230
Total		4329	4229	8558

Concordance rate was 99.99%, calculated by (a+d)/(a+b+c+d).

Supplementary Table 9 Variant comparison between EBV isolates from paired saliva and NPC tumor samples from the same NPC patients

		tumor		Total
		nReference	nVariants	
saliva	nReference	7155[a]	48[b]	7203
	nVariants	13[c]	1136[d]	1149
Total		7168	1184	8352

Concordance rate was 99.27%, calculated by (a+d)/(a+b+c+d).

Supplementary Table 10 Association results of seven previously reported human SNPs in our current study

SNP	Chr	Locus	Minor Allele	Current study		Previous GWAS		
				Odds ratio	*P*	Odds ratio	*P*	
rs6774494	3	MDS1-EVI1	G	0.88	1.52E-01	0.84	5.05E-08	[5]
rs31489	5	CLPTM1L/TERT	A	0.81	3.56E-02	0.81	6.30E-13	[6]
rs2860580	**6**	**HLA-A**	**A**	**0.55**	**2.05E-10**	**0.58**	**3.65E-65**	**[5]**
rs2894207	**6**	**HLA-B/C**	**G**	**0.55**	**2.58E-07**	**0.61**	**1.83E-31**	**[5]**
rs28421666	6	HLA-DQ/DR	G	0.74	4.18E-02	0.67	1.40E-18	[5]
rs1412829	9	CDKN2A/2B	G	0.75	3.75E-02	0.78	3.51E-07	[5]
rs9510787	13	TNFRSF19	A	1.22	1.88E-02	1.20	9.57E-09	[5]

The association of seven human SNPs with NPC was assessed in the combined discovery and validation samples (639 cases *versus* 652 controls) using meta-analysis. The associations of seven SNPs reported by previou GWAS[5] (5,090 cases *versus* 4,957 controls) and GWAS[6] (6,868 cases *versus* 9,119 controls) are shown.

Supplementary Table 11 The top three associated SNPs in GWAS discovery phase reaching suggestive genome-wide significance ($P < 4.07\times10^{-4}$)

POS	Reference/ alternative genotypes	Alt Freq in cases	Alt Freq in controls	*P* _GWAS*	Zscore_GWAS*	LD r-squared with SNP		Annotation
						162215	162507	
162215	C/A	3.85%	40.43%	3.22E-04	-3.60		0.85	BALF2, non-synonymous, V700L
162507	C/T	2.56%	42.55%	9.17E-05	-3.91	0.85		BALF2, synonymous
162852	G/T	2.56%	40.43%	1.86E-04	-3.74	0.90	0.95	BALF2, synonymous

*Alternative genotypes were tested against reference genotypes in the mixed model in GWAS discovery phase.

Supplementary Table 12 Fine-mapping for casual SNPs associated with NPC risk in *BALF2* gene region

Position	Reference/ alternative genotypes	Alternative genotype frequency in cases	Alternative genotype frequency in controls	*P* _GWAS*	Zscore_GWAS*	Posterior probability by PAINTOR	Annotation
160804	C/T	7.24%	39.13%	1.38E-01	-1.48	0.00	BALF2, synonymous
160827	G/T	94.12%	65.22%	2.20E-02	2.29	0.00	BALF2, synonymous
160941	G/A	5.84%	13.33%	3.88E-01	0.86	0.00	BALF2, synonymous
160971	T/C	90.26%	59.09%	8.41E-02	1.73	0.00	BALF2, synonymous
161036	T/C	88.82%	56.82%	1.20E-01	1.56	0.01	BALF2, non-synonymous, S1093G
162117	A/G	93.59%	65.96%	2.09E-02	2.31	0.00	BALF2, synonymous
162147	G/A	3.85%	38.30%	2.37E-03	-3.04	0.00	BALF2, synonymous
162195	A/C	92.95%	65.96%	3.00E-02	2.17	0.00	BALF2, synonymous
162215	C/A	3.85%	40.43%	3.22E-04	-3.60	0.83	BALF2, non-synonymous, V700L
162237	C/G	93.51%	65.22%	2.70E-02	2.21	0.00	BALF2, synonymous
162464	G/A	93.59%	61.70%	5.09E-03	2.80	0.00	BALF2, synonymous
162476	T/C	93.59%	61.70%	5.09E-03	2.80	0.08	BALF2, non-synonymous, I613V
162507	C/T	2.56%	42.55%	9.17E-05	-3.91	0.02	BALF2, synonymous
162852	G/T	2.56%	40.43%	1.86E-04	-3.74	0.01	BALF2, synonymous
163107	A/C	93.59%	65.22%	1.91E-02	2.34	0.00	BALF2, synonymous
163287	G/A	93.59%	63.83%	5.02E-02	1.96	0.00	BALF2, synonymous
163293	G/A	3.85%	40.43%	1.26E-03	-3.23	0.00	BALF2, synonymous
163364	C/T	88.46%	48.94%	7.95E-03	2.65	0.06	BALF2, non-synonymous, V317M
163404	C/A	94.19%	63.04%	3.42E-02	2.12	0.00	BALF2, synonymous
163422	G/T	93.55%	63.04%	5.02E-02	1.96	0.00	BALF2, synonymous
163464	G/A	87.10%	52.17%	5.18E-02	1.94	0.00	BALF2, synonymous
163611	C/T	94.77%	65.22%	1.72E-02	2.38	0.00	BALF2, synonymous
163629	T/C	94.12%	65.22%	4.14E-02	2.04	0.00	BALF2, synonymous
163647	C/T	8.44%	15.22%	7.93E-01	0.26	0.00	BALF2, synonymous
163686	G/A	3.92%	42.55%	2.61E-03	-3.01	0.00	BALF2, synonymous
163926	C/T	83.97%	51.06%	1.63E-01	1.40	0.00	BALF2, synonymous
163995	C/T	5.77%	29.79%	1.63E-01	-1.39	0.00	BALF2, synonymous
164277	G/T	7.69%	17.02%	6.30E-01	-0.48	0.00	BALF2, synonymous

*Alternative genotypes were tested against reference genotypes in the mixed model in GWAS discovery phase.

Supplemenary Table 13 Basic characteristics of 483 cases and 605 control individuals used for validation phase by age and sex

Variables	Cases	Controls	*P* (chisq)*
Sex			0.481
Male	364 (75.4%)	467 (77.2%)	
Female	119 (24.6%)	138 (22.8%)	
Age			0.369
Mean	48.7	49.3	
Standard Deviation	11.3	10.6	
< 37	17 (3.5%)	13 (2.1%)	
37-59	387 (80.1%)	487 (80.5%)	
>59	79 (16.4%)	105 (17.4%)	
Total	483	605	

* The *p* values were obtained from $\chi 2$ tests

Supplementary Table 14 EBV haplotypes and the risk for NPC in 536 cases and 651 controls from the NPC-endemic Zhaoqing County

EBV subtype (162215-162476-163364)	536 cases		651 controls		Odds Ratio *	95% CI	*P*
	no.	%	no.	%			
L-L-L (A-T-C)	22	4.10%	171	26.27%			
H-H-H (C-C-T)	451	84.14%	292	44.85%	11.81	7.49 - 19.49	2.82E-24
H-H-L (C-C-C)	51	9.51%	118	18.13%	3.52	2.02 - 6.28	1.24E-05
H-L-L (C-T-C)	9	1.68%	65	9.98%	1.10	0.45 - 2.47	8.31E-01
other subtypes	3	0.56%	5	0.77%	4.23	0.78 - 19.62	7.07E-02

* Odds ratio for individual EBV subtypes were estimated with logistic model by categorizing each subtype as a single variable and adjusted for age, sex, status of single or multiple infection and the human *HLA* SNPs (rs2860580 and rs 2894207). Subjects with EBV subtype L-L-L, a common low-risk subtype were used as the reference category. H represents high-risk genotypes; L represents low-risk genotypes.

Supplementary Table 15 Estimation of odds ratios of SNP 162476T>C and 163364C>T for NPC risk

	Beta	Standard Error	*P*	Odds ratio	95% Confidence interval
SNP162476	1.20	0.25	1.10E-06	3.31	2.06-5.40
SNP163364	1.21	0.18	4.84E-11	3.35	2.35-4.83

Odds ratios were estimated in the combined discovery and validation samples (639 cases and 652 controls) using a logistic regression model containing SNPs 162476T>C and 163364C>T. The logistic regression model was adjusted for age, sex and the human *HLA* SNPs (rs2860580 and rs 2894207).

Supplementary Table 16 The association of EBV haplotypes with EBV DNA abundance in saliva of 533 cases and 651 controls from the NPC-endemic Zhaoqing County

	10^5 EBV copies/ml case saliva		10^5 EBV copies/ml control saliva		EBV DNA load in saliva of 533 cases and 651 controls			EBV DNA load in saliva of 533 cases			EBV DNA load in saliva of 651 controls		
	Mean	**95% CI**	**Mean**	**95% CI**	**Adjusted fold change**	**95% CI**	***P* value**	**Ajusted fold change**	**95% CI**	***P* value**	**Adjusted fold change**	**95% CI**	***P* value**
L-L-L (A-T-C)	8.59	2.84 - 26.10	21.26	14.11 - 32.05	Ref			Ref			Ref		
H-H-H (C-C-T)	5.5	4.34 - 6.97	18.92	13.40 - 26.61	0.83	0.51 - 1.35	4.57E-01	0.66	0.22 - 1.97	3.75E-01	0.97	0.55 - 1.71	9.27E-01
H-H-L (C-C-C)	4.64	2.28 - 9.48	11.75	7.29 - 18.94	0.56	0.31 - 1.02	5.62E-02	0.54	0.15 - 1.97	3.08E-01	0.57	0.28 - 1.15	1.16E-01
H-L-L (C-T-C)	11.50	2.34 - 56.70	19.83	11.06 - 35.51	0.97	0.45 - 2.09	9.36E-01	1.33	0.16 - 10.76	7.65E-01	1.00	0.43 - 2.35	9.97E-01

*The adjusted fold changes in saliva EBV DNA load from individuals infected by EBV with the high-risk BALF2 haplotypes (C-C-T amd C-C-C) and wild-type B95-8 haplotype (C-T-C) versus low-risk BALF2 haplotype (A-T-C) were assessed using multiple linear regression against log-scale of EBV DNA copy numbers with EBV haplotypes and adjusted for single-multiple infection status, sex, age and case-control status. The adjusted fold change values were yielded by 2^(-regression coefficient). H and L represent high- and low- risk genotype, respectively. H-H-H represents the high-risk haplotype carrying risk genotypes of SNPs 162215C>A, 162476T>C and 163364C>T.

Supplementary Table 17 Frequency of high-risk EBV haplotypes in different regions

Geographic origin	Total frequency of high-risk haplotypes C-C-T and C-C-C				
	NPC cases			non-NPC samples	
Africa				0.00%	0/37
Western countries				2.63%	1/38
NPC-endmic China	**93.27%**	**596/639**		**62.54%**	**419/670**
			Healthy control	63.04%	411/652
			Lymphoma	40.00%	6/15
			Lymphoblastoid cell line	66.67%	2/3
NPC-non-endemic East Asia	**55.00%**	**11/20**		**9.68%**	**3/31**
			Healthy control	14.29%	1/7
			Lymphoma	28.57%	2/7
			Gastric carcinoma	0.00%	0/17

Geographic origin	Frequency of high-risk haplotype C-C-T				
	NPC cases			non-NPC samples	
Africa				**0.00%**	**0/37**
Western countries				**0.00%**	**0/38**
NPC-endmic China	**84.35%**	**539/639**		**44.93%**	**301/670**
			Healthy control	44.94%	293/652
			Lymphoma	40.00%	6/15
			Lymphoblastoid cell line	66.67%	2/3
NPC-non-endemic East Asia	**55.00%**	**11/20**		**6.45%**	**2/31**
			Healthy control	14.29%	1/7
			Lymphoma	14.29%	1/7
			Gastric carcinoma	0.00%	0/17

Supplementary Table 19 Estimation of the proportion of NPC population risk attributable to high-risk EBV haplotypes in 536 cases and 651 controls from the NPC-endemic Zhaoqing County

High-risk haplotype	Population attributable risk fraction	95% confidence interval
C-C-T	70.76%	64.08%-77.45%
C-C-T and C-C-C	82.84%	75.64%-90.04%

The attributable fraction of risk and 95% confidence interval were estimated in a logistic regression model with adjustment for age, sex, status of single- or multiple-infection and the human *HLA* SNPs (rs2860580 and rs 2894207). For details, see methods.

Supplementary Table 20 Comparison of the association results of EBV SNPs and NPC with human SNPs *versus* without human SNPs as covariates

Position	Genotypes	High-risk genotype	without HLA SNPs		with HLA SNPs		*P* value	
			Odds ratio	95% CI	Odds ratio	95% CI	without *HLA* SNPs	with *HLA* SNPs
162215	C/A	C	7.62	5.02 - 11.57	7.6	4.97 - 11.62	2.98×10^{-19}	1.42×10^{-18}
162476	T/C	C	8.79	5.90 - 13.09	8.69	5.79 - 13.03	5.90×10^{-26}	9.69×10^{-25}
163364	C/T	T	6.52	4.90 - 8.68	6.14	4.59 - 8.22	7.18×10^{-36}	2.40×10^{-32}

The association of three EBV SNPs with NPC risk was assessed with and without human *HLA* SNPs rs2860580 and rs2894207 as covariates. Odds ratios conferred by high-risk genotypes and the 95% confidence intervals (CI) were estimated by meta-analysis of combined discovery and validation phases.

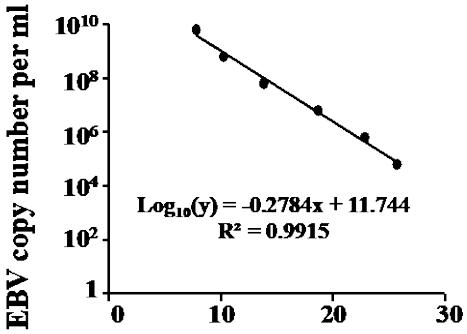
Supplementary Note

Patient recruitment in the population-based case-control study

The study design of the population-based case-control study has been described in detail in our previous publication[1]. To accommodate available resources, the present analysis was confined to NPC cases and controls enrolled from Zhaoqing County between January 2010 and October 2014, using the following eligibility criteria: (i) histological confirmation of NPC, (ii) age less than 80 years old, (iii) no treatment for NPC, and (iv) residence in Zhaoqing city. Among 1,306 eligible patients with NPC recruited into the study through a rapid case ascertainment system involving a network of physicians, 1,043 (79.9%) had saliva samples which were sequenced or genotyped. Controls, selected at random from the total population registry of Zhaoqing County, consisted of 1,151 population control subjects with no history of malignancy (84.3% of 1,365 eligible controls enrolled with frequency-matching to cases by 5-year age and sex) had saliva samples that were sequenced or genotyped.

Evaluation of EBV DNA abundance in saliva and its correlation with genotyping success rate

The abundance of EBV DNA in saliva samples from the 1043 cases and 1151 controls in Zhaoqing County was measured in triplicate for each sample by fluorescence quantitative PCR (qPCR) using a DNA fragment of *BALF5*. The DNA copy number in saliva was deduced from qPCR standard curve shown below.



EBV DNA quantification standard curve derived by real-time PCR. The linear regression line plots the log of EBV concentrations *vs.* Ct values of each PCR reaction of standard samples.

Because EBV in the buccal mucosa periodically undergoes a lytic cycle[2-5], in a large proportion of both cases and controls, EBV DNA abundance was quite low and did not allow for EBV WGS or successful genotyping. From 53 cases and 46 controls, saliva samples with EBV DNA load higher than 4.6×10^5 (a Ct value < 30) copies per ml of saliva were used for EBV WGS. The remaining 990 cases and 1105 controls were used for genotyping the three GWAS candidate markers used for validation.

The amount of EBV DNA in saliva is highly correlated with the genotyping success rate (**Supplementary Fig. 15**). When the EBV copy number per ml of saliva was lower than 4.6×10^5 (a Ct value < 30), the genotyping success rate reached 87%. However, this rate dropped markedly as the EBV DNA load decreased. Therefore, all three SNP genotypes were only obtained in saliva samples from 483 cases (48.79%) and 605 controls (54.75%). The EBV genotyping success rate in saliva from controls was slightly higher than from cases. Consistently, in these saliva samples, we found that EBV DNA abundance from cases was significantly lower than from controls (**Supplementary Fig. 11**).

EBV whole-genome sequencing, variant calling, and filtering

Targeted EBV WGS. Genomic DNA extracted from tumor, saliva, plasma, and cell line was subject to hybrid capture by an EBV-targeting, single-stranded DNA probe developed by MyGenostics. Sequencing libraries were constructed by shearing genomic DNA into 150 to 200 bp DNA fragments, DNA purification, end blunting, and adaptor ligation, according to instructions from Illumina. Library concentrations were evaluated by Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). EBV DNA was captured from genomic DNA following the MyGenostics GenCap Target Enrichment Protocol (GenCap Enrichment, MyGenostics, USA). Libraries were hybridized with EBV probes at 65 ℃ for 24 h and then washed to remove uncaptured DNA. The eluted DNA fragments were amplified by 18 PCR cycles to generate libraries for sequencing. Libraries were quantified and subjected to paired-end sequencing on an Illumina HiSeq 2000 sequencer according to the manufacturer's instructions (Illumina Inc., San Diego, CA, USA).

Read Mapping. The quality of the raw reads was assessed with Trim-galore to remove adaptor sequences and low-quality reads. High-quality reads were aligned to the reference EBV genome (NC_007605.1) using Burrows-Wheeler Aligner (BWA, version 0.7.5a)[6]. Alignments were converted from the sequence alignment map format to sorted, indexed binary alignment map (BAM) files[7]. Duplicate reads were removed with the Picard tool. The depth and coverage were calculated for each sample. The average sequencing depth for EBV genomes was 1282 (range, 32 to 6629), and on average, 95.28% of the genome was covered with at least 10× reads (**Supplementary Fig. 2**).

Variant calling and filtering. GATK software tools (version 3.2-2) were used to improve alignments and genotype calling following GATK's Best Practice[8]. Briefly, BAM files were realigned with the GATK IndelRealigner. The base quality of the mapped reads was recalibrated with the GATK base quality recalibration tool BQSR. Because BQSR requires a genuine SNP database for recalibration, we generated our own high-quality EBV SNP database. Raw variants were first called by GATK UnifiedGenotyper and HaplotypeCaller separately against the reference EBV genome (NC_007605.1). Common variants identified by the two callers were selected and filtered to generate the SNP database for BQSR. Analysis-ready reads were prepared after three cycles of BQSR when before-and-after BQSR plots converged and the recalibration reached saturation. Subsequently, variants were called by GATK UnifiedGenotyper using analysis-ready reads. Because EBV has a small genome and a small number of variants relative to the human genome, hard-filtering was recommended by GATK developer to exclude the low-quality variant due to (i) low variant confidence, (ii) low read-mapping quality, (iii) strand bias (the variation being seen on only the forward or only the reverse strand) in the reads, and (iv) reads that were aligned to multiple positions in the EBV genome. In particular, SNPs and INDELs were filtered separately by GATK VariantFiltration with the parameter "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) 0.1)", "QUAL < 50.0", "QD < 2.0", "MQ < 40.0", "FS > 250.0" for SNPs and "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) 0.1)", "QUAL < 50.0", "FS > 200.0" for INDELs. We identified an initial set of high-quality 8,469 variants from 269 samples and the C666-1 cell line.

To avoid inaccurate calling, we further filtered out variants that had (i) low coverage support (depth < 10×), (ii) in repetitive elements (NCBI annotation of reference NC_007605.1), and (iii) within 5 bp of an indel; 7,962 variants were retained for subsequent EBV phylogenetic, principal component and association analyses. Metrics including the number of filtered SNP counts, the concordance of variants among samples, and the ratios of heterozygous to single variants were evaluated using GATK VariantEval. Variants were annotated and summarized by SNPEff[9], according to the annotation of NC_007605.1 (NCBI annotation, NOV 2013).

Determining single *versus* multiple EBV infections.

The EBV genome is stable, and intra-host mutation rate is often low[10], so heterozygous variants caused by intra-host mutation occur at a low frequency. We sequenced EBV genomes in quadruple replicates from the NPC cell line C666-1[11]. EBV genomes in cell lines and EBV-associated tumors usually undergo clonal expansion[12-14], and the heterozygous variants come from low-level genomic evolution during cell proliferation over decades. The proportions of heterozygous variants ranged from 6.7% to 9.5%, as determined from quadruple replicates of C666-1 EBV whole-genome sequencing. In contrast, the EBV isolates with multiple infections tend to have higher numbers of heterozygous variants. Therefore, we first extracted the empirical distribution of the percentage of heterozygous variants across all the samples. By fitting two different curves to the lower (< 8%) and higher quantiles (> 15%) of the empirical distribution, we identified a reflection point separating the two tails of the distribution (**Supplementary Fig. 13**). The reflection point (10.7%) was then used to define a threshold of the number of heterozygous SNPs constituting a single- or multiple- infection. The 230 samples with the number of heterozygous SNPs lower than the threshold were identified as single-infection samples for subsequent phylogenetic analysis.

In 483 NPC cases and 605 controls from the Zhaoqing case-control study, we genotyped all three EBV GWAS candidate markers. In saliva from 464 cases (96.07%) and 570 controls (94.25%), we detected EBV infection with single haplotypes where all three markers were homozygous, whereas saliva from only 19

cases (3.93%) and 35 controls (5.75%) contained EBV infection with multiple EBV haplotypes with heterozygous markers. The multiple-infection EBV haplotypes were deduced by phasing using Beagle 4.1[15]. In the association study, we included cases and controls carrying both single infection and multiple infection with EBV haplotypes. Because we adjusted for multiple-infection, which accounted only for a small proportion of cases and controls in the association study, our association results are not confounded by multiple-infection.

1 Ye, W. *et al.* Development of a population-based cancer case-control study in southern china. *Oncotarget* **8**, 87073-87085, doi:10.18632/oncotarget.19692 (2017).

2 Kieff, E. D. & Rickinson, A. B. in *Fields' virology* Vol. 68A (eds D.M. Knipe & P.M. Howley) 2603-2654 (Lippincott Williams & Wilkins, Wolters Kluwer, 2007).

3 Borza, C. M. & Hutt-Fletcher, L. M. Alternate replication in B cells and epithelial cells switches tropism of Epstein-Barr virus. *Nat Med* **8**, 594-599, doi:10.1038/nm0602-594 (2002).

4 Frangou, P., Buettner, M. & Niedobitek, G. 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, doi:10.1086/426823 (2005).

5 Hadinoto, V., Shapiro, M., Sun, C. C. & Thorley-Lawson, D. A. The dynamics of EBV shedding implicate a central role for epithelial cells in amplifying viral output. *PLoS Pathog* **5**, e1000496, doi:10.1371/journal.ppat.1000496 (2009).

6 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

7 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

8 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).

9 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92, doi:10.4161/fly.19695 (2012).

10 Weiss, E. R. *et al.* Early Epstein-Barr Virus Genomic Diversity and Convergence toward the B95.8 Genome in Primary Infection. *J Virol* **92**, doi:10.1128/JVI.01466-17 (2018).

11 Cheung, S. T. *et al.* Nasopharyngeal carcinoma cell line (C666-1) consistently harbouring Epstein-Barr virus. *Int J Cancer* **83**, 121-126 (1999).

12 Raab-Traub, N. & Flynn, K. The structure of the termini of the Epstein-Barr virus as a marker of clonal cellular proliferation. *Cell* **47**, 883-889 (1986).

13 Pathmanathan, R., Prasad, U., Sadler, R., Flynn, K. & Raab-Traub, N. Clonal proliferations of cells infected with Epstein-Barr virus in preinvasive lesions related to nasopharyngeal carcinoma. *N Engl J Med* **333**, 693-698, doi:10.1056/NEJM199509143331103 (1995).

14 Neri, A. *et al.* Epstein-Barr virus infection precedes clonal expansion in Burkitt's and acquired immunodeficiency syndrome-associated lymphoma. *Blood* **77**, 1092-1095 (1991).

15 Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics* **81**, 1084-1097, doi:10.1086/521987 (2007).