# Whole-Exome Sequencing Association Study Implicates Genetic Effects on Viral Oncogenesis and Tumor Microenvironment in Nasopharyngeal Carcinoma

Yanni Zeng[1,2,3#], Chun-Ling Luo[1,4#], Guo-Wang Lin[5#], Fugui Li[6], Xiaomeng Bai[1,2,3], Josephine Mun-Yee Ko[7], Yang Liu[1,4], Shuai He[1,4], Jia-Xin Jiang[1,4], Enya Hui Wen ONG[8], Zheng Li[9], Ya-Qing Zhou[1,4], Yun-He Zhou[1,4], An-Yi Xu[1,4], Shu-Qiang Liu[1,4], Yun-Miao Guo[10], Jie-Rong Chen[11], Xi-Xi Cheng[1,4], Yu-Lu Cao[1], Xia Yu[6], Biaohua Wu[6], Pan-Pan Wei[1,4], Zhao-Hui Ruan[1,4], Qiu-Yan Chen[1], Lin-Quan Tang[1], James D. McKay[12], Wei-Hua Jia[1], Hai-Qiang Mai[1], Jian-Jun Liu[13], Dong-Xin Lin[1], Chiea Chuen Khor[9], Melvin Lee Kiang CHUA[8,14,15*], Mingfang Ji[6*], Maria Li Lung[7*], Yi-Xin Zeng[1*], Jin-Xin Bei[1,4,16*†]

1. State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center - Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, P.R. China
2. Faculty of Forensic Medicine, Guangdong Province Translational Forensic Medicine Engineering Technology Research Center, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, P.R. China.
3. Guangdong Province Key Laboratory of Brain Function and Disease, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, P.R. China.
4. Department of Experimental Research, Sun Yat-sen University Cancer Center, Guangzhou, P.R. China.
5. Department of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, P.R. China.
6. Cancer Research Institute of Zhongshan City, Zhongshan City People 's Hospital, Zhongshan, P.R. China.
7. Department of Clinical Oncology, School of Clinical Medicine, University of Hong Kong, Hong Kong SAR, P.R. China.
8. Precision Radiotherapeutics Oncology Programme, Division of Medical Sciences, National Cancer Centre Singapore, Singapore.
9. Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore.

10. Zhanjiang Institute of Clinical Medicine, Central People's Hospital of Zhanjiang, Guangdong Medical University Zhanjiang Central Hospital, Zhanjiang, P. R. China.
11. Department of Laboratory Medicine, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, P.R. China.
12. Genomic Epidemiology Branch, International Agency for Research on Cancer/World Health Organization (IARC/WHO), France.
13. Laboratory of Human Genomics, Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore.
14. Department of Head and Neck and Thoracic Radiation Oncology, National Cancer Centre Singapore, Singapore.
15. Oncology Academic Clinical Programme, Duke-NUS Medical School, Singapore.
16. Department of Medical Oncology, National Cancer Centre Singapore, Singapore.

# These authors contributed equally.
* Senior authors who jointly directed the study.
† Correspondence to:
Dr Jin-Xin Bei, State Key Laboratory of Oncology in South China, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
E-mail: beijx@sysucc.org.cn

**Abstract**

Nasopharyngeal carcinoma (NPC) poses a substantial clinical challenge with limited understanding of its genetic underpinnings. Here we conduct the largest-scale whole-exome sequencing association study of NPC to date, involving 6,969 NPC cases and 7,100 controls and revealing three novel germline genetic variants linked to NPC susceptibility: a common variant rs2276868 in *RPL14*, a rare variant rs5361 in *SELE*, and a common variant rs1050462 in *HLA-B*. Through a multiomics approach, which integrates both bulk (n=206) and single-cell RNA-sequencing (n=56) data along with experimental validations, we demonstrate that the *RPL14* variant modulates Epstein-Barr virus (EBV) life cycle and NPC pathogenesis. Additionally, we show that the *SELE* variant plays a role in modifying endothelial cell function, thus promoting NPC progression. Our study also underscores the critical impact of rare genetic variants on NPC heritability. We introduce a refined composite polygenic risk score (rcPRS) that outperforms existing models in predicting NPC risk. Notably, our findings reveal that the polygenic risk for NPC is mediated by EBV infection status. Overall, our study provides crucial insights into the intricate genetic architecture of NPC. It highlights the critical interplay between genetic variations and essential elements of the tumor microenvironment, such as EBV and endothelial cells, in predisposing to NPC. This work opens new avenues for personalized risk assessments, early diagnosis, and targeted therapeutic strategies for NPC.

**Keywords**

Nasopharyngeal carcinoma (NPC), whole-exome sequencing, genetic susceptibility, EBV, *RPL14, SELE,* HLA

**Highlights**

1. The largest whole-exome sequencing association study unveils three novel germline genetic variants linked to nasopharyngeal carcinoma (NPC) susceptibility, revealing the genetic architecture of NPC heritability with substantial contribution from rare variants.

2. *RPL14* variant modulates the Epstein-Barr virus life cycle and NPC pathogenesis, providing novel insights into genetic and viral interactions in cancer development.

3. *SELE* variant alters endothelial cell function and its favor to migration and metastasis of NPC, advancing our understanding of how genetic susceptibility variants influence tumor microenvironment components thereby contribute to cancer development.

4. A refined composite polygenic risk score (rcPRS) outperforms existing models in predicting NPC risk, emphasizing the influence of EBV infection status on NPC susceptibility.

108 **Introduction**

109 Nasopharyngeal carcinoma (NPC) is a highly lethal malignancy that exhibits a significant
110 geographic prevalence in specific regions of East and Southeast Asia[1]. These areas account for
111 approximately 70% of new cases worldwide[1], with the highest incidence rate (9.69/100,000)
112 reported in southern China area[2]. Familial clustering has been documented in a notable
113 proportion (3.64~19%) of NPC patients from diverse populations[3-6], suggesting a strong link
114 between family history and NPC risk[7]. Genetic susceptibility, Epstein-Barr virus (EBV)
115 infection, and environmental effects are believed to play essential roles in NPC development[1].

116 Efforts have been made to identify germline variations contributing to the genetic
117 susceptibility of NPC. Multiple linkage and association studies have reported that common
118 genetic variants in HLA region confer a significant risk for NPC[8-11]. Besides the HLA loci,
119 linkage studies have also identified risk loci in 4p12-p15, 3p21.31-21.2, and 5p13.1[9,12,13].
120 However, the replication of these linkage results has been challenging due to the diminishing
121 size of family pedigrees and the low resolution of reported loci, limiting their practical
122 application. Recently, genome-wide association studies (GWASs) have identified replicated
123 common single nucleotide polymorphism (SNPs) in multiple loci associated with NPC risk,
124 including *MECOM*, *CDKN2A/B*, *TNFRSF19*, *CIITA*, and *TERT/CLPTM1L*, in addition to the
125 HLA loci[8,14-17]. However, these GWAS loci explain only a small portion (2.84%) of the overall
126 NPC susceptibility[18], despite the relatively high heritability of NPC (10~61.3%)[19,20].

127 The issue of "missing heritability" (the gap between the heritability explained by GWAS
128 loci and the total heritability) in NPC, similar to other complex polygenic diseases, may be
129 attributed to the limited statistical power of studies targeting common variants (minor allele
130 frequency (MAF)>=0.01) and even poorer power of studies capturing rare variants
131 (MAF<0.01), such as those reported by utilizing whole-genome sequencing (WGS) or whole-
132 exome sequencing (WES) [21,22]. Currently, only a few studies have examined the contribution
133 of rare variants to NPC risk through WES, identifying NPC-associated genes such as *MST1R,*
134 *RAD54L* and *POLN* [23-26]. However, these findings either did not reach exome-wide statistical
135 significance or have not been independently validated, likely due to limited sample sizes or
136 genetic heterogeneity across population. Therefore, the genetic architecture involving both
137 common and rare variants in NPC risk remains unclear.

138 The role of EBV molecules in the screening and diagnosis of NPC is significant, as
139 evidenced by their presence in NPC tumor and peripheral blood[27,28]. This is particularly

140 remarkable given that over 90% of the adult population is infected with EBV[29]. However, the
141 exact mechanisms linking EBV to NPC, and the potential involvement of host genetic
142 variations in those processes, are still enigmatic. EBV is known for its ability to transform B
143 cells into lymphoblastoid cell lines and for encoding oncogenic genes such as LMP1 and
144 BARFs[30]. In NPC, EBV infects epithelial cells and encodes viral genes related to NPC
145 progression[31]. Furthermore, the interaction of EBV with malignant epithelial cells and various
146 stromal cell types, each exhibiting unique gene expression profiles, fosters a complex tumor
147 microenvironment (TME), leading to both intra- and inter-tumor heterogeneity in NPC[32,33].
148 Despite these insights, the influence of genetic variations on TME components, and how this
149 contributes to NPC predisposition remains elusive.

150 This study aims to deepen our understanding of the genetic architecture of NPC and to
151 translate genetic findings into insights about its pathological mechanisms. We first conducted
152 a comprehensive two-stage association study, utilizing WES and capture-sequencing (Cap-seq)
153 analyses across multiple levels—variant, gene, and pathway—involving 14,069 individuals of
154 southern Chinese descent, including those from Guangdong, Hong Kong in China, and
155 Singapore. Furthermore, we employed extensive bulk and single cell RNA sequencing analyses,
156 along with experimental assays, to elucidate the functional mechanisms of the identified risk
157 loci. Critically, we uncovered the profound impact of both common and rare genetic variants
158 on cancer heritability, particularly how these genetic variations distinctly affect EBV-related
159 pathogenesis and TME components, thereby contributing to varying susceptibilities to NPC.

160

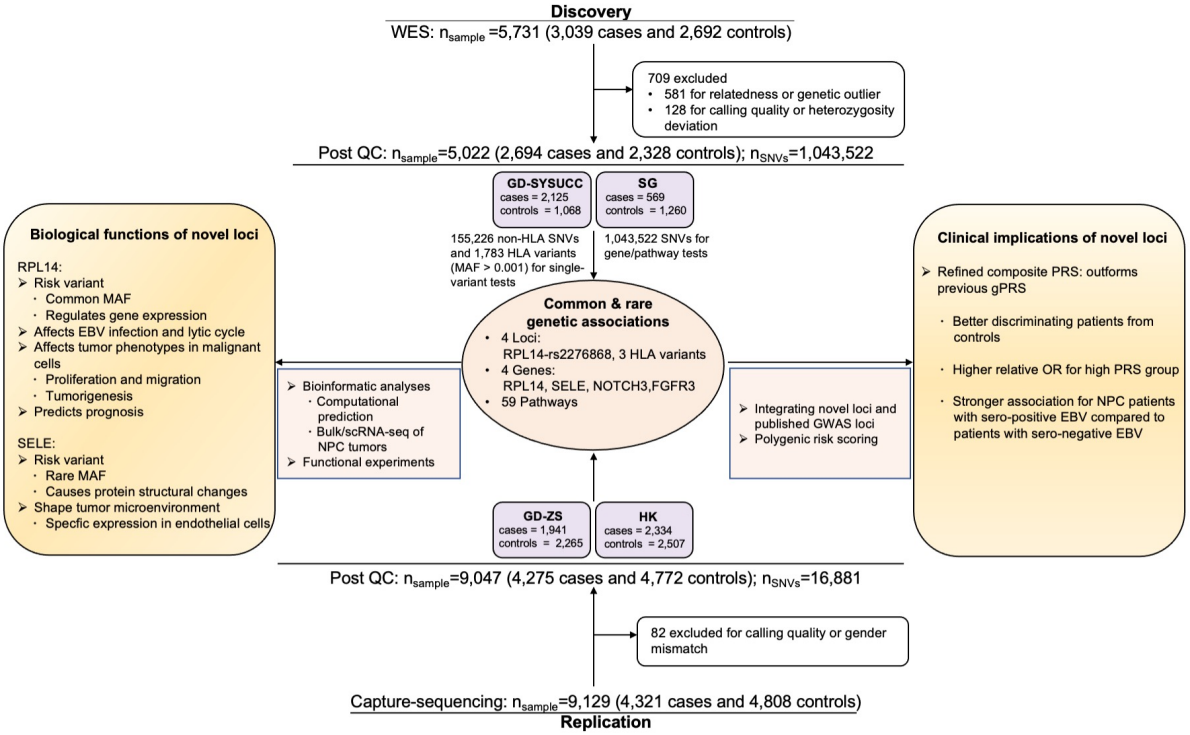



**Figure 1. Study overview.** A two-stage association study design was applied to investigate the genetic factors associated with NPC. In the discovery stage, a total of 5,022 samples, including the GD-SYSUCC cohort from Guangdong in China and the SG cohort from Singapore, were genotyped with WES and analyzed to identify four independent variants (two of which are novel), four novel genes, and 59 pathways associated with NPC. The associations were subsequently validated in the replication stage, which included 9,047 samples from the GD-ZS cohort in Guangdong and the HK cohort in Hong Kong, China. Bioinformatic analyses and functional experiments were conducted to explore the biological functions and the clinical implications of the novel loci.

## Results

### *Association of single variants with NPC susceptibility*

To identify susceptibility loci of NPC at single variant level, we first performed whole-exome sequencing (WES) on genomic DNA samples for the discovery stage, involving 2,694 NPC cases and 2,328 healthy controls of Han Chinese descent from southern China (GD-SYSUCC) and Singapore (SG; see Methods, **Figure 1**). After stringent quality controls, we obtained 1,043,522 single nucleotide variants (SNVs), including 362,993 potentially functional or pathogenic variants (Methods) and 335,163 novel variants which were not present in the 1000 genome phase 3, gnomAD exome, and genome database[34,35].

181 Subsequently, we performed a single-variant-based exome-wide association study (EWAS)
182 on 155,226 non-HLA SNVs and 1,783 HLA variants (including HLA alleles, amino acids [AA],
183 and SNVs) with a MAF >= 0.001 among 2,694 NPC cases (all NPC cases, ALLNPC) and
184 2,328 controls, as well as 409 NPC cases with a family history (FHNPC) and the same set of
185 controls. We observed neither evidence of genomic inflation of the associations ($\lambda_{ALLNPC}$=1.07,
186 $\lambda_{FHNPC}$=0.89, **Figure S1A**) nor significant genetic substructure between case and control
187 groups (**Figure S2**). The EWAS results revealed a total of 242 variants associated with NPC
188 risk in the ALLNPC group and 128 variants associated with NPC risk in the FHNPC group,
189 surpassing the Bonferroni significant threshold for multiple testing correction ($P < 3.2 \times 10^{-7}$;
190 **Figure S1B**). Notably, we discovered a novel non-HLA locus with several significant
191 associated SNVs (**Figure S1B**). The sentinel SNV rs2276868, located at the 5'-untranslated
192 region (5'-UTR) of the ribosomal gene *RPL14*, exhibited the strongest associations in both the
193 FHNPC ($P_{FHNPC}$=2.3 × $10^{-8}$, Odds Ratio (OR) $_{FHNPC}$ =1.575, 95% Confident Interval (CI)
194 =1.343-1.846) and the ALLNPC groups ($P_{ALLNPC}$=4.2 × $10^{-7}$, $OR_{ALLNPC}$ =1.242, 95%CI=1.142-
195 1.351; **Table 1, Figure S1B**). Furthermore, the association was stronger in the family cases
196 than in the sporadic cases ($OR_{sporadic_NPC}$ =1.200, 95%CI =1.097-1.313), indicating a robust
197 genetic predisposition to NPC. Conditional analysis revealed that all associations at the *RPL14*
198 locus vanished upon adjusting for rs2276868, suggesting that this variant explains all the
199 observed associations (**Table S1, S2**). Beyond the non-HLA locus, we also identified HLA
200 variants significantly associated with NPC (**Figure S1B**). Stepwise conditional analysis
201 localized three independently associated variants, including *HLA-DQB1*03:01:01*, *HLA-A:*
202 *A_62_Q*, and rs1050462 at HLA-B (**Tables 1, Table S1**). Except for the two known
203 associations at *HLA-A: A_62_Q* and *HLA-DQB1*03:01:01*[18,36], the association of rs1050462
204 with NPC is a novel discovery whose signal is not attributed to previously reported *HLA*
205 variants [8,10,11] (see Methods; $P \leq 3.23 \times 10^{-5}$; **Table S3**), according to an additional conditional
206 analysis using an extended GD-SYSUCC cohort consisting of 1,583 cases and 3,049 controls
207 with GWAS array data [17].

208 We next conducted validations of the identified associations using two independent
209 replication cohorts consisting of 9,047 samples of Southern Chinese recruited from Hong Kong
210 (HK; 2,334 cases and 2,507 controls) and Guangdong (GD-ZS; 1,941 cases and 2,265 controls)
211 in China (see Methods). We genotyped the samples using Cap-seq to target the genomic
212 regions spanning the above variants. Meta-analysis across the replication samples consistently
213 confirmed the associations of the four variants in *RPL14* and *HLA* loci ($P_{meta} < 0.05$; **Table 1**).

214 Moreover, when combining the discovery and replication samples, the meta-analysis revealed
215 that the four variants in *RPL14* and *HLA* loci were significantly associated with NPC risk,
216 surpassing the threshold for genome-wide significance ($P < 5 \times 10^{-8}$; **Table 1**).

217

218 ***Association of genes and pathways with NPC susceptibility***

219 To identify the cumulative genetic effects of both rare and common variants within individual
220 genes on the risks of NPC, we conducted a gene-based association analysis in the discovery
221 dataset using an ensemble method that incorporates four algorithms simultaneously (see
222 Methods). For each gene, we performed separate tests that use all SNVs within the genic region
223 or only consider coding-affecting SNVs of that gene. Our analysis revealed a significant
224 association between *RPL14* and NPC, with statistical significance surpassing Bonferroni
225 threshold for multiple testing correction ($P < 2.3 \times 10^{-6}$, **Table 2**). Additionally, we validated
226 previous associations in several genes (**Table S4**).

227 We further performed a pathway-based analysis to explore the cumulative genetic effects
228 of genes involved in 6,204 curated molecular pathways in the discovery dataset (see Methods).
229 This analysis identified a total of 59 pathways associated with NPC in either the ALLNPC or
230 FHNPC groups, adhering to the false discovery rate (FDR) threshold set for multiple testing
231 correction ($P < 3.09 \times 10^{-4}$; **Figure 2A, Table S5)**. Among these pathways, we observed that
232 11 sentinel pathway genes were involved in at least two NPC-associated pathways and
233 exhibited an association with NPC in gene-based test with a significant threshold of $P < 0.01$
234 (**Figure 2B**). Notably, *SELE, NOTCH3*, and *FGFR3* had strong associations with NPC ($P <$
235 $1\times10^{-4}$; **Figure 2B**). *SELE* was particularly outstanding, with the strongest association with
236 NPC ($P_{SELE} = 2.69 \times 10^{-6}$) and the involvement in the largest number of NPC-associated
237 pathways (**Figure 2B**).

238 To validate the associations for the significant gene *RPL14* and the top three sentinel
239 pathways genes (*SELE*, *FGFR3*, and *NOTCH3*), we re-sequenced exon regions of these four
240 candidate genes in the replication cohorts ($n_{GD\text{-}ZS}$=4,206 and $n_{HK}$=4,841). Gene-based
241 association analyses statistically validated all the candidate genes in the replication dataset
242 ($P_{RPL14}=1.23 \times 10^{-7}$**,** $P_{SELE}=4.05 \times 10^{-4}$**,** $P_{NOTCH3}=3.28 \times 10^{-6}$, and $P_{FGFR3}=8.96 \times 10^{-4}$; **Table 2**).
243 The association for *RPL14* was particularly robust across all replication cohorts ($P_{GD\text{-}ZS}=3.50$
244 $\times 10^{-6}$ and $P_{HK}=8.44 \times 10^{-5}$), whereas the associations for *SELE*, *NOTCH3*, and *FGFR3* were
245 more prominent in the GD-ZS cohort (**Table 2**).

246 **Table 1. Association of single variants with NPC susceptibility**

| Variants information | | | | Discovery (2,694 cases and 2,328 controls) | | | | Replication (4,275 cases and 4,772 controls) | | | | | Combined samples (6,969 cases and 7,100 controls) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variant ID | CHR | A1 | Gene | ALLNPC: OR (95% CI) | P | FHNPC: OR (95% CI) | P | OR (95% CI) HK (2,334 cases and 2,507 controls) | OR (95% CI) GD-ZS (1,941 cases and 2,265 controls) | $OR_{meta}$: (95% CI) | $P_{meta}$ | $I_{meta}^2$ | $OR_{meta}$ (95% CI) | $P_{meta}$ | $I_{meta}^2$ |
| rs2276868 | 3 | T | *RPL14* | 1.242 (1.142-1.351) | 4.16E-07 | 1.575 (1.343-1.846) | 2.30E-08 | 1.253 (1.154-1.36) | 1.193 (1.091-1.305) | 1.225 (1.153-1.302) | 5.26E-11 | 0.00 | 1.231 (1.172-1.293) | 1.22E-16 | 0.00 |
| DQB1*03:01:01 | 6 | Present | *HLA-DQB1* | 0.633 (0.568-0.705) | 1.00E-16 | 0.585 (0.467-0.734) | 3.64E-06 | 0.694 (0.622-0.774) | 0.589 (0.522-0.663) | 0.640 (0.545-0.752) | 5.56E-08 | 74.84 | 0.638 (0.582-0.700) | 9.43E-22 | 50.44 |
| A_62_Q | 6 | Present | *HLA-A* | 0.507 (0.462-0.556) | 4.61E-47 | 0.483 (0.401-0.581) | 1.20E-14 | 0.585 (0.532-0.642) | 0.560 (0.505-0.621) | 0.573 (0.535-0.615) | 1.73E-55 | 0.00 | 0.549 (0.504-0.598) | 4.69E-43 | 57.52 |
| rs1050462 | 6 | G | *HLA-B* | 0.646 (0.585-0.713) | 7.58E-18 | 0.637 (0.521-0.779) | 1.08E-05 | 0.568 (0.513-0.629) | 0.588 (0.526-0.658) | 0.577 (0.535-0.622) | 1.92E-46 | 0.00 | 0.600 (0.555-0.649) | 2.55E-37 | 40.86 |

247 Notes: CHR, chromosome; A1, target allele/amino acid; OR (95%CI), odds ratio estimate and the 95% confidence interval; $I^2$, the proportion of the variance
248 in observed effect is due to variance in true effects rather than sampling error; Meta, meta-analyzed using the random effects model.

249

250 **Table 2. Associations of individual genes with NPC susceptibility**

| Gene | Discovery (2,694 cases and 2,328 controls) | | | | Replication (4,275 cases and 4,772 controls) | | | | Combined samples (6,969 cases and 7,100 controls) |
|---|---|---|---|---|---|---|---|---|---|
| Name | Significant in tests | SNV group | Method | P | P (GD-ZS; 1,941 cases and 2,265 controls) | P (HK; 2,334 cases and 2,507 controls) | $P_{meta}$ | Adjusted $P_{meta}$ | $P_{meta}$ |
| *RPL14* | Gene association | ALLSNV | SKAT (C+R) | 2.17E-06 | 3.50E-06 | 8.44E-05 | 1.23E-07 | 6.16E-07 | 1.79E-10 |
| *SELE* | Sentinel pathway genes | ALLSNV | SKAT (C+R) | 2.69E-06 | 1.01E-03 | 3.94E-02 | 4.05E-04 | 2.02E-03 | 1.44E-07 |
| *SELE* | Sentinel pathway genes | coding affecting SNV | SKAT (C+R) | 3.61E-06 | 1.36E-03 | 1.20E-01 | 1.40E-03 | 7.00E-03 | 4.63E-07 |
| *FGFR3* | Sentinel pathway genes | ALLSNV | Burden | 1.24E-04 | 7.40E-04 | 1.28E-01 | 8.96E-04 | 4.48E-03 | 3.64E-06 |
| *NOTCH3* | Sentinel pathway genes | ALLSNV | SKAT (C+R) | 3.60E-04 | 6.00E-06 | 1.07E-02 | 3.28E-06 | 1.64E-05 | 5.41E-08 |

251 Notes: ALLSNV, all SNVs; SKAT (C+R), SNP-set (Sequence) Kernel Association Test for the combined effect of common (C) and rare (R) variants; Burden,
252 burden test; Meta, meta-analyzed using the summation of logits method; Adjusted P, the P value adjusted using the FDR method.

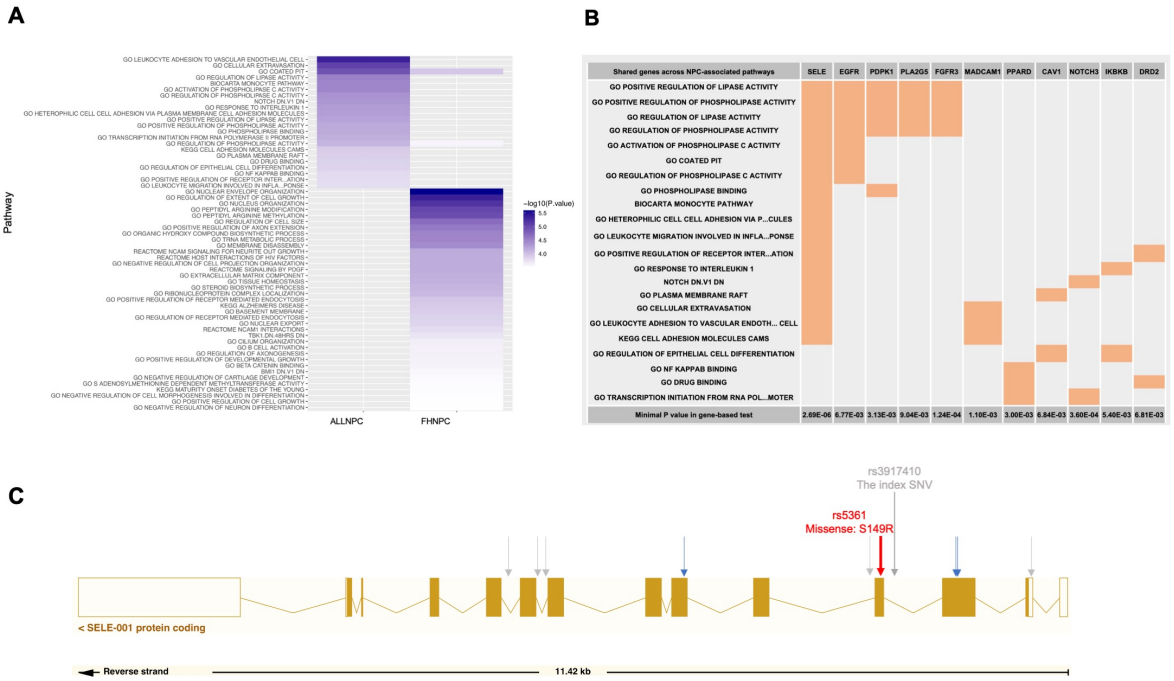



**Figure 2. Significant pathways and sentinel genes associated with NPC. A**. Significant pathways associated with NPC. ALLNPC denotes all NPC samples in the discovery stage, whereas FHNPC represents NPC cases with a familial NPC history. Color bars from white to purple indicate minus log-transformed P values. **B**. Sentinel genes for NPC-associated pathways. Genes indicated atop are highlighted in orange if they are part of a specific pathway. The listed genes are those belonging to at least two NPC-associated pathways and exhibiting a gene-based association P value below 0.01. **C**. Locations of the rare variants associated with NPC at significance level of $P < 1\times10^{-4}$ within the genic regions of SELE. Grey arrows denote non-coding variants, blue arrows represent synonymous coding variants, and the red arrow indicates non-synonymous coding variant predicted as "deleterious". The rs5361 minor allele introduces a missense mutation at position 149, resulting in an amino acid substitution from S to R. The index SNV rs3917410 showed the most significant P value in the association tests in the discovery stage.

We next fine-mapped the variants (MAF > 0.001) contributing to the gene-level associations for the four replicated genes using the discovery dataset (n=5,022). A constituent variant was considered to have a "major" contribution if its variant-level association P value was less than $1 \times 10^{-4}$. Using this criterion, we detected a major contribution from a common variant for *RPL14* (**Table S6**), as well as major contributions from multiple rare variants for *SELE* (**Table S7**). However, no major contributing variants were detected for *FGFR3* (**Table S8**) or *NOTCH3* (**Table S9**). Specifically, among the four SNVs (MAF > 0.001) in *RPL14*, the sentinel SNV rs2276868 (MAF=0.36, $P_{SNV}=4.16 \times 10^{-7}$) and another SNV rs2276869 (MAF=0.10, $P_{SNV}=0.001$) shared modest linkage disequilibrium (LD; $R^2=0.2$) and were the

277 only two *RPL14* variants associated with NPC ($P_{SNV}$ <0.05; **Table S6**). Removing these two
278 SNVs from the gene-based analysis diminished the gene-level association of *RPL14* with NPC
279 ($P_{gene}$=0.34). Controlling for rs2276868 also abolished the associations of rs2276869 (**Table**
280 **S6**). These findings suggest a predominant contribution of common variant rs2276868 to
281 *RPL14*'s association with NPC.

282 For *SELE*, a total of 23 SNVs (MAF > 0.001) were detected, including seven common
283 (MAF>=0.01) and 16 rare variants (MAF<0.01; **Table S7**). Of these, 10 SNVs were strongly
284 associated with NPC ($P_{SNV}=2.68 \times 10^{-5} \sim 4.98 \times 10^{-5}$; **Table S7**), with rs3917410 as the sentinel
285 SNV ($P=2.68 \times 10^{-5}$) and sharing nearly complete LD with the other nine SNVs ($R^2 > 0.99$;
286 **Table S7**). All the 10 SNVs are rare, including one non-synonymous coding (rs5361), three
287 synonymous coding, and six non-coding variants. Controlling for rs5361 variant abolished the
288 association signals with NPC for the other rare variants in *SELE* (Adjusted P>0.05; **Table S7**).
289 The rs5361 variant is a missense mutation (T>G) in the exon 4 of *SELE* (**Figure 2C**), resulting
290 in an amino acid substitution from the conservative Serine (S) (**Figure 2C, Figure S3A**) to
291 Arginine (R) at position 149. Furthermore, multiple algorithms predicted this S149R
292 substitution as "deleterious" (Polyphen-2 score=0.997; CADD score=23; SIFT score=0;
293 REVEL score=0.516; MetaLR score=0.772; **Figure S3B**). Collectively, these findings suggest
294 that the rare variant rs5361 (*SELE*-S149R) is likely a causal variant accounting for the NPC
295 associations observed in *SELE*.

296

297 ***Functional characterization of the novel loci***

298 ***Distinct cellular expression patterns of NPC-associated genes***

299 To understand the functional consequences of the novel loci in NPC, we first investigated their
300 cellular expression patterns along with the known NPC-associated genes (**Table S10**). We
301 harnessed single-cell transcriptomic analyses of tumor samples from 56 NPC patients and
302 various non-tumor tissues from 15 donors without cancer[37] (see Methods). *RPL14* exhibited a
303 universal expression across all cell types, notably in the malignant epithelial cells (**Figure 3**).
304 Similarly, most HLA-I genes also showed broad expression in all cell types. By contrast, *SELE*
305 exhibited specific expression in endothelial cells across both tumor and non-tumor tissues
306 (**Figure 3C, Figure S4**). We also noted specific expression of some genes in cell types other
307 than the malignant epithelial cells in NPC tumors (**Figure 3**). For example, *NOTCH3* was
308 specifically expressed in cancer associated fibroblasts, while HLA-II genes were mainly

309 expressed in various stromal cell types in the tumor microenvironment (TME; **Figure 3**).
310 Additionally, by performing single-cell transcriptomic analyses of tumor samples from other
311 epithelial cancers, including lung, gastric, and colorectal cancers (see Methods), we observed
312 that gene associated with those cancers exhibited similar expression patterns observed in NPC,
313 with some expressed predominantly in epithelial cells and some predominantly in stromal cells
314 in the TME (**Figure S5**). These findings strongly indicate that genes conferring susceptibility
315 to cancers are not necessarily confined to expression in malignant cells. Instead, these genes
316 often exhibit predominant expression in diverse cell populations within the TME, suggesting a
317 crucial role of genetic impact on stromal cells in cancer development.

318

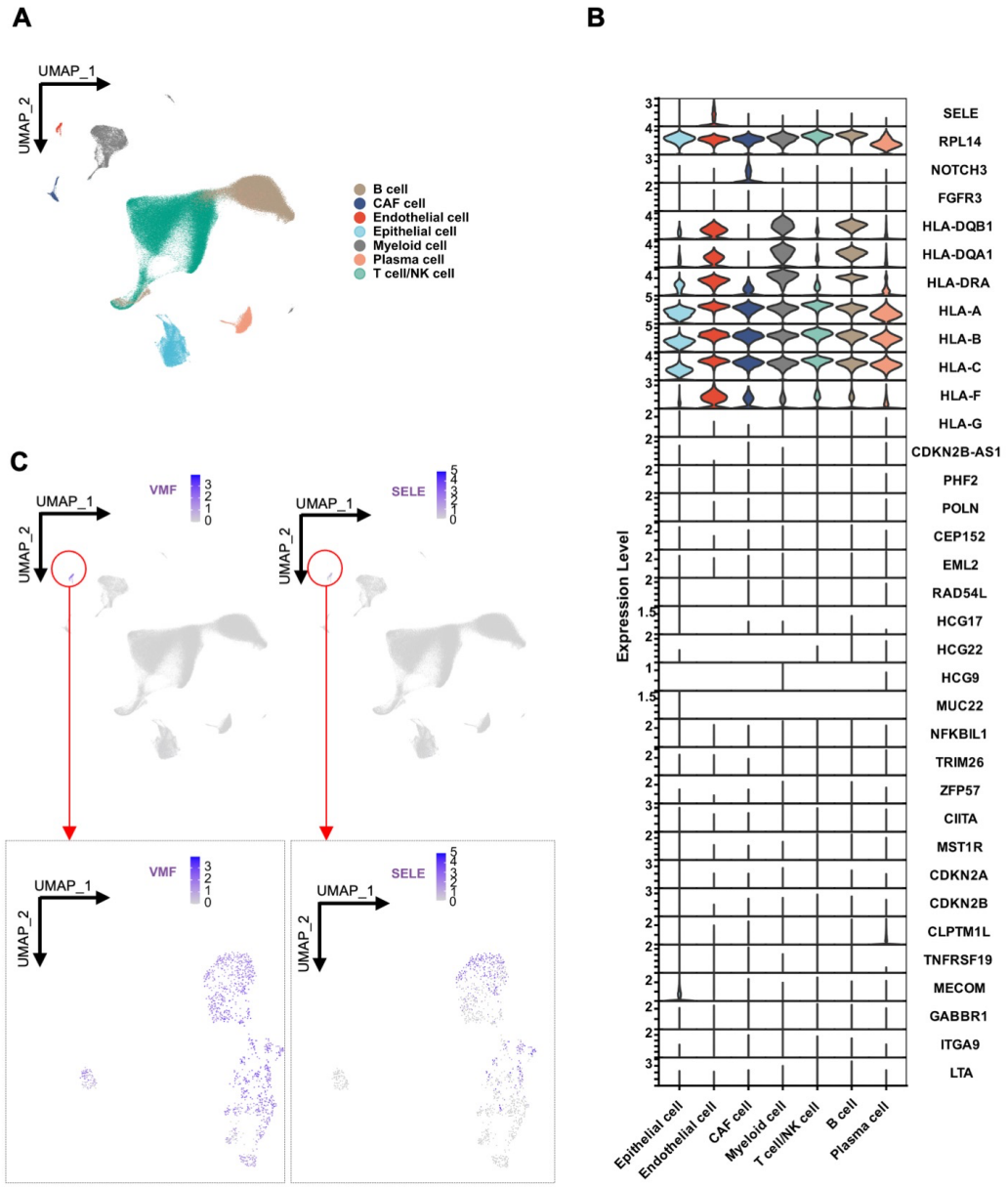



**Figure 3. Expression patterns of novel and known NPC-associated genes across diverse cell types in NPC tumor tissues.** Single-cell transcriptomic analyses of 223,593 cells derived from NPC tumor tissues in 56 patients. **A**. UMAP plot of 223,593 single cells grouped into seven major cell clusters. **B**. Violin plot illustrating normalized expression of NPC-associated genes across these major cell clusters. All epithelial cells captured in NPC tumor were malignant (see Methods). **C**. The expression of the marker gene (*VWF*) for endothelial cells and alongside the novel NPC-associated gene *SELE*; top panel: initial UMAP plot, bottom panel: re-normalized UMAP emphasizing cells highlighted by the red circles in the initial UMAP plot.

332 ***rs2276868 regulates RPL14 expression through the NKRF transcription factor***

333 Given the location of rs2276868 at the RPL14 promoter [38] (**Figure S6, S7A**) and the stable
334 expression of *RPL14* in malignant epithelial cells (**Figure 3**), we aimed to investigate the
335 potential regulatory role of rs2276868 in *RPL14* mRNA expression in epithelial cells. We first
336 performed a single-cell-transcriptome-based eQTL analysis using 15,623 malignant epithelial
337 cells from tumor tissues of 35 NPC patients. We observed that CC or CT genotype of
338 rs2276868 was significantly associated with a higher expression of *RPL14* in malignant NPC
339 cells compared to the TT genotype (**Figure 4A**). Subsequently, luciferase reporter assay
340 demonstrated that the rs2276868-[T] construct had significantly lower luciferase activity
341 compared to the rs2276868-[C] construct in human 293T cells (**Figure 4B**). These findings
342 strongly suggest the influence of rs2276868 on *RPL14* transcription, which may involve
343 modulating the binding affinity of certain transcription factor (TF).

344 To identify such potential TF, we conducted a comprehensive screening and identified 24
345 candidate TFs, which showed a positive correlation with the mRNA expression of *RPL14* in
346 two independent bulk transcriptomic datasets of NPC tissues ($n_{Bei}$=87 and $n_{Zhang}$=113; **Figure**
347 **4B, Figure S7B**). Subsequent siRNA screening assays confirmed that five of the 24 TFs were
348 capable of modulating *RPL14* expression (**Figure S7C**). Among them, *NKRF* knockdown
349 significantly reduced *RPL14* transcription in NPC cells (**Figure 4B, Figure S7C-E**).
350 Furthermore, knockdown or overexpression of *NKRF* exhibited a preferential effect on the
351 transcription activity of the rs2276868-[C] construct over the rs2276868-[T] construct in
352 luciferase reporter assays (**Figure 4E, F**). Consistently, single-cell transcriptome analysis
353 revealed a stronger positive correlation between the expression of *NKRF* and *RPL14* in
354 malignant NPC cells carrying the rs2276868-[C] compared to the rs2276868-[T] carriers
355 (**Figure 4J**). Moreover, chromatin immunoprecipitation (ChIP) assay demonstrated that NKRF
356 specifically binds to a refined genomic region encompassing rs2276868 (**Figure 4G-I, S7F**).
357 Taken together, these data strongly suggest that rs2276868 influences *RPL14* expression
358 through its interaction with *NKRF*, which preferentially binds to the promoter region housing
359 rs2276868-[C] to foster enhanced transcription of *RPL14* in NPC tumors.

360

361

362

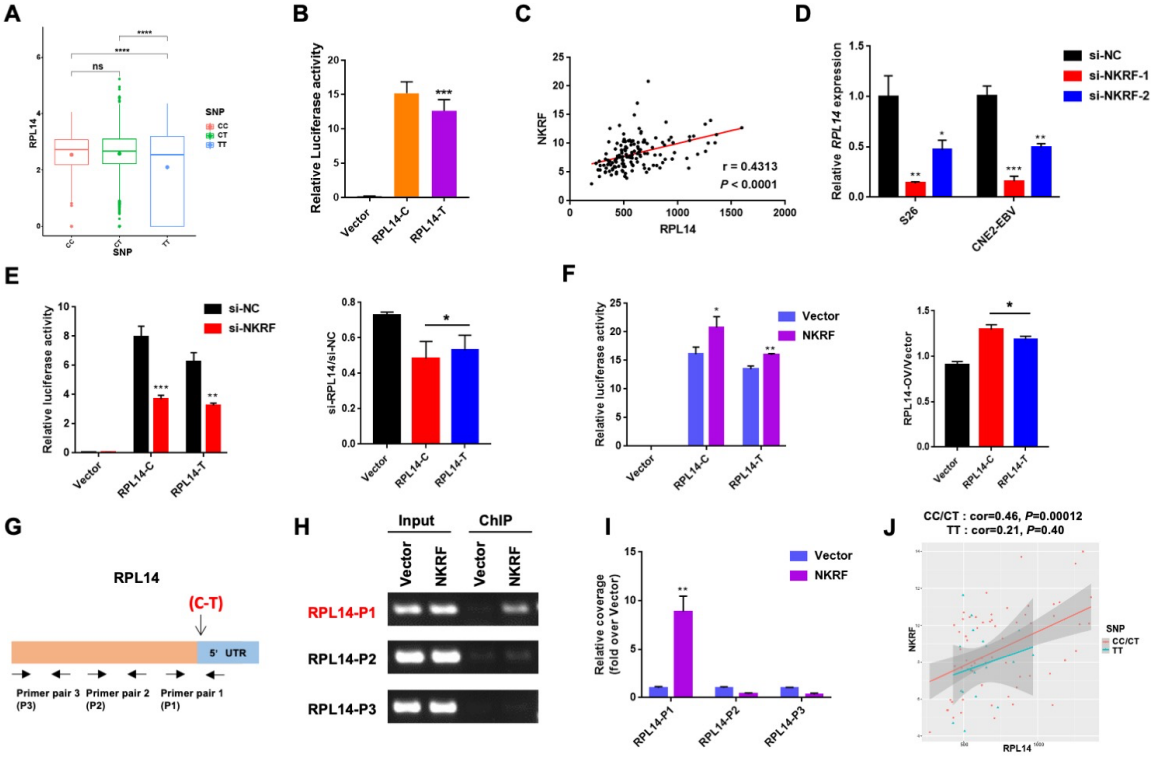



**Figure 4. rs2276868 regulates the expression of *RPL14*. A**. Single-cell transcriptome analysis showed the mRNA expression of RPL14 in 15,623 malignant epithelial cells from 35 NPC samples with different rs2276868 genotypes. **B**. Relative luciferase activity changes in 293T cells transfected with plasmids containing rs2276868 -[C], -[T], or control vectors. **C**. Pearson correlation analysis indicates the relationship between *RPL14* and *NKRF* expression (measured as transcript per million, TPM) in transcriptome data of NPC patients (n=87). **D**. RT-qPCR illustrates the mRNA expression of RPL14 in NPC cells transfected with NKRF siRNAs or control siRNA. **E-F**. Relative luciferase activity in 293T cells co-transfected with the rs2276868 -[C] or -[T] plasmids and NKRF siRNA (F) or NKRF overexpression vectors (G). Corresponding statistics are presented at the right. **G**. Schematic diagram indicates primer pairs used for PCR amplification of RPL14 fragments. **H-I**. ChIP assay in S26 cells transfected with Flag-NKRF and control vectors. ChIP PCR (H) and qPCR (I) showed RPL14 expression level in cells. P1-3 denote primer pairs targeting genomic regions shown in G. **J**. Pearson correlation analysis indicates the relationship between NKRF and RPL14 expression in single-cell transcriptome data are different for NPC patients with different genotypes. rs2276868 -[CC/CT] patients have a stronger correlation than rs2276868-[TT] patients. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

### *RPL14 suppresses EBV infection and lytic cycle in NPC*

To investigate the role of *RPL14* in NPC development, we first examined the relationship between *RPL14* expression and the transcriptional activity scores of various molecular pathways in NPC (see details in Methods). Through REACTOME and GO analysis of malignant epithelial cells in NPC scRNA-seq data ($n_{sample}$=35 and $n_{cell}$=15,623), we observed

387 a significant association between *RPL14* and a pathway related to viral translation
388 (GO:0019081: viral translation. $P_{sc}=1.63 \times 10^{-6}$; **Figure 5A, Table S11**). Additionally,
389 REACTOME and GO analysis of bulk RNA-seq data from NPC tissues ($n_{bulk_Bei}=93$ and
390 $n_{bulk_Zhang}=113$) revealed significant associations between *RPL14* and multiple pathways
391 related to viral infection. These pathways include SARS-COV-2 modulation of host translation
392 machinery ($P_{bulk_Bei}=8.48 \times 10^{-16}$, $P_{bulk_Zhang}=9.11 \times 10^{-14}$) and influenza infection
393 ($P_{bulk_Bei}=2.00 \times 10^{-14}$, $P_{bulk_Zhang}=1.66 \times 10^{-13}$; **Figure S8; Table S12, S13**). Considering the
394 well-known involvement of EBV in NPC etiology[39], we hypothesized a potential link between
395 *RPL14* and EBV processes. Supportively, in EBV-positive malignant NPC cells, we observed
396 a significant negative association ($R= -0.27$, $P < 2.2\times10^{-16}$) between *RPL14* expression and the
397 EBV activity score, a metric developed by consolidating expression levels of multiple EBV
398 genes (see Methods; **Figure 5B**). Furthermore, rs2276868-[T] carriers exhibited decreased
399 *RPL14* expression but increased expression of EBV latent genes *LMP1* and *LMP2* as compared
400 to rs2276868-[C] carriers in the same cell group (**Figure 5C**). These results corroborate the
401 connection between *RPL14* and EBV-driven processes.

402 To further explore the interaction between *RPL14* and EBV, we monitored the protein
403 expression of RPL14 in NPC cells post EBV infection over different timepoints. We observed
404 a significant downtrend in RPL14 expression concurrent with EBV infection (**Figure 5D**). By
405 contrast, RPL14 knockdown resulted in a significant rise in EBV infection in NPC cells, while
406 RPL14 overexpression led to a marked reduction in EBV infection (**Figure 5E-H**). We also
407 noticed a decline in RPL14 expression when EBV lytic cycle was induced using TPA and NaB
408 in NPC cells (**Figure 5I, J; Figure S9**). Furthermore, RPL14 knockdown remarkably
409 augmented the expression of EBV lytic genes in NPC cells, while RPL14 overexpression
410 significantly reduced their expression (**Figure 5K, L**). These observations strongly suggest that
411 RPL14 acts as a suppressor of EBV infection and lytic cycle in NPC.

412

413

414

415

416

417

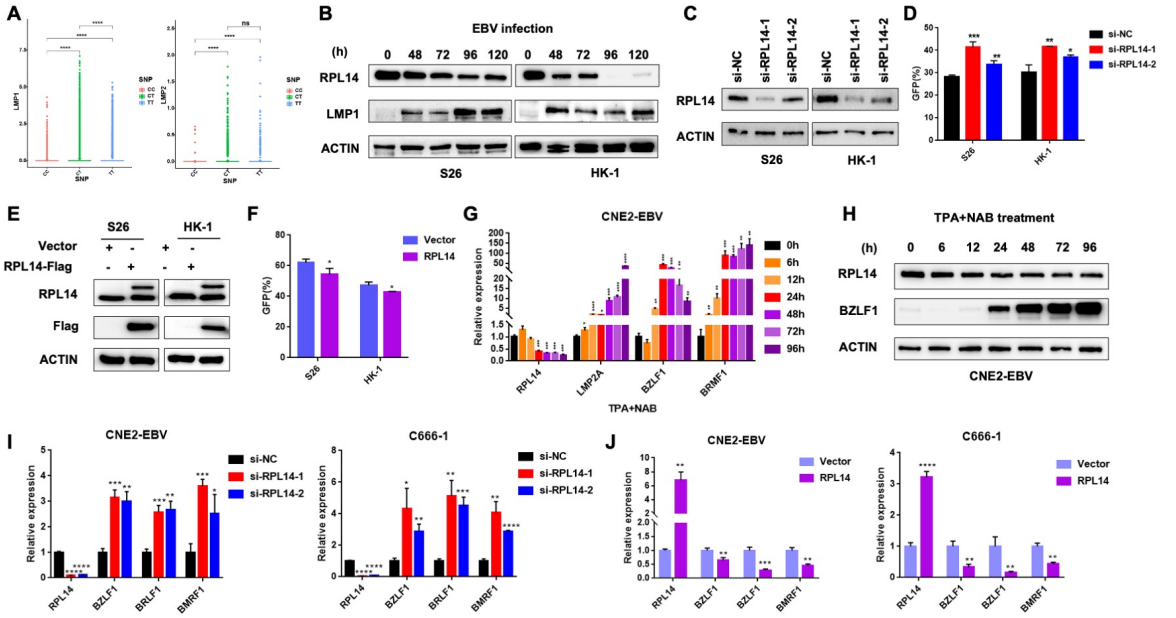


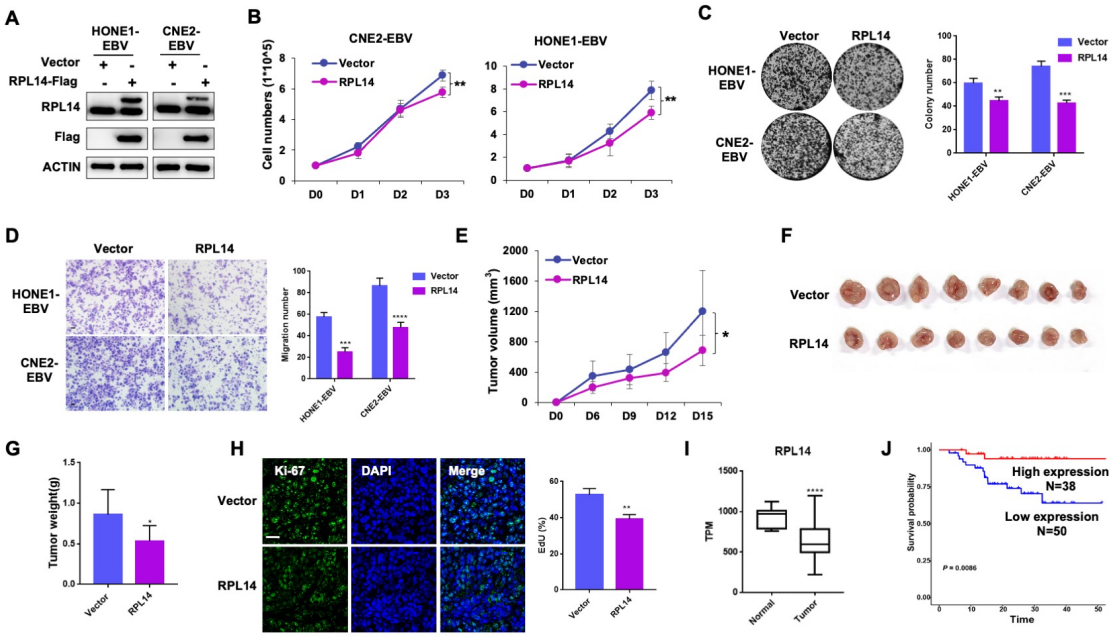

**Figure 5. *RPL14* inhibits EBV infection and lytic cycle activation in NPC cells. A.** The top 20 pathways significantly associated with RPL14 expression in malignant epithelial cells from NPC tumor. **B.** Correlation analysis between EBV-activity scores and *RPL14* expression within malignant epithelial cells. **C.** Single-cell transcriptome analysis showing *LMP1* and *LMP2* expression in NPC samples with either rs2276868 -[C] or -[T] genotypes. **D.** Western blot examination of RPL14 and LMP1 protein levels in NPC cell lines S26 and HK-1 cells infected with EBV at different timepoints. **E**. Western blot assessment of the knockdown efficiency of RPL14 siRNAs or control siRNA in S26 and HK-1 cells. **F.** Flow cytometry quantification of GFP intensity for the EBV infection efficiency in the NPC cells described in E. **G**. Western blot assay showing RPL14 protein expression in S26 and HK-1 cells infected with lentivirus stably expressing RPL14. ACTIN served as a loading control. **H**. Flow cytometry assessment of EBV infection efficiency in the cells described in G. **I-J**. Changes of gene expression in CNE2-EBV cells subjected to EBV lytic induction with TPA and NAB. RT-qPCR (I) and western blotting (J) were performed to detect the mRNA and protein expression, respectively, of RPL14 and EBV genes at different timepoints of EBV lytic induction. **K-L**. RT-qPCR analysis of EBV lytic gene expression in CNE2-EBV and C666-1 cells transfected with RPL14 siRNAs or control siRNA (K), or infected with lentivirus stably expressing RPL14 vector or control vector (L). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ***$p < 0.0001$.

### *RPL14 impedes tumorigenesis in NPC cells*

We next explored the function of *RPL14* in NPC tumorigenesis. We first manipulated *RPL14* expression in an *in vitro* NPC cell model (**Figure 6A; S10A**). Cell growth curves and colony formation assays revealed that the stable overexpression of *RPL14* significantly inhibited NPC cell proliferation. Conversely, stable knockdown of *RPL14* had an opposite effect, promoting proliferation (**Figure 6B-C, Figure S10B-C**). Transwell assays revealed that *RPL14*

445 overexpression suppressed NPC cell migration, while its knockdown enhanced the migration
446 (**Figure 6D; Figure S10D**). These observations strongly suggest that *RPL14* inhibits the
447 tumorigenesis of NPC cells.

448




449

450 **Figure 6. Tumor suppressive function of *RPL14* in NPC. A**. Western blot analysis showing
451 RPL14 protein levels in NPC cells infected with RPL14 overexpressing lentivirus. ACTIN
452 serves as the loading control. **B**. Cell growth curves of the cells described in A. **C**. Colony
453 formation assay for the cells described in A. Corresponding statistical analysis is shown to the
454 right. **D**. Transwell migration assay evaluating the migration ability of the cells described in A.
455 The statistical analysis is presented on the right. **E-H**. Tumor growth evaluation in a nude mice
456 model with subcutaneous injection of NPC cells described in A. Tumor volumes were
457 measured every three days. Visual presentation of tumor post-sacrifice (F) and weight (G) were
458 presented. IF detection of Ki-67 expression (H) in the tumors described in F, and the
459 corresponding statistical analysis is shown on the right. **I**. Transcriptomic analysis showcasing
460 mRNA levels of RPL14 in NPC (n=87) versus control samples (n=10). **J**. Kaplan-Meier
461 survival analysis linking RPL14 expression to survival time in NPC patients using bulk RNA-
462 seq data (n=88). $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, $****p < 0.0001$.

463

464 To corroborate the tumor-suppressive function of *RPL14 in vivo*, we established a mouse
465 model with subcutaneous injection of NPC cells stably expressing *RPL14*. We observed a
466 significant inhibition on tumorigenesis as reflected by the diminished tumor volume and weight
467 in mice with *RPL14* overexpression compared to the control group (**Figure 6E-G**).
468 Supportively, immunofluorescence staining of Ki-67, a proliferation marker, revealed a
469 substantial decrease in Ki-67 positive cells in the *RPL14*-overexpressing tumors compared to

470 the control tumors (**Figure 6H**). Furthermore, we observed a significant lower expression of
471 *RPL14* in NPC tissues compared to nasopharyngeal tissues from individuals with rhinitis
472 (**Figure 6I**). Kaplan-Meier survival analysis revealed that low expression of *RPL14* was
473 significantly correlated with a poor prognosis of NPC patient (**Figure 6I**). These findings
474 strongly suggest *RPL14* as a tumor suppressor in NPC.

475

476 ***SELE-S149R promotes NPC development through gain of function in endothelial cells***

477 *SELE* encodes E-selectin, a selectin cell adhesion molecule specifically expressed on the
478 surface of endothelial cells (**Figure 3C**). E-selectin plays a pivotal role in attracting cancer
479 cells by binding to ligands like glycoproteins[40]. Structural modeling of E-selectin paired with
480 glycomimetic antagonist ligand revealed that the Ser149 residue of SELE, implicated in the
481 rs5361 variant, is in close proximity to its ligand, and the S149R substitution (**Figure 2C**) leads
482 to an extension of E-selectin's side chain, potentially altering its interaction with glycoproteins
483 (**Figure 7A**). This aligns with a previous finding that *SELE*-S149R introduces a gain-of-
484 function effect on the recruiting capability of E-Selectin[41].

485 To probe the functional implications of the *SELE*-S149R mutation in the context of NPC,
486 we developed a model using human umbilical vein endothelial cells (HUVECs) stably
487 overexpressing either the wild-type SELE (*SELE*-Ser149, SELE-WT) or its S149R mutant
488 variant (SELE-MUT, **Figure 7B**). Transwell assays revealed that overexpression of SELE,
489 especially the SELE-MUT, significantly boosted the migration ability of HUVEC cells (**Figure**
490 **7C**). Tube formation assay further demonstrated that SELE, particularly the mutant form,
491 markedly enhanced the endothelial cell's ability to form capillary-like structures, (**Figure 7D**),
492 indicating a gain-of-function effect of the S149R mutation in promoting angiogenesis. We next
493 employed a co-culture model involving HUVECs (expressing either SELE-WT or MUT) and
494 NPC cell lines (HK-1 and S26). Strikingly, HUVECs overexpressing SELE, especially the
495 SELE-MUT variant, substantially increased the migration of NPC cells (**Figure 7E**). This was
496 further substantiated by *in vivo* xenograft assays, which showed increased tumorigenesis in
497 NPC cells co-cultured with HUVEC-SELE-MUT, evident from the augmented tumor volume
498 and weight (**Figure 7F-G**). Moreover, lymph node metastasis assays reinforced these findings,
499 demonstrating that co-culture with HUVEC-SELE-MUT cells substantially promoted the
500 metastasis of NPC cells (**Figure 7I-K**). These results strongly suggest that *SELE*-S149R

mutation in endothelial cells plays a crucial role in escalating both tumorigenesis and metastasis in NPC cells.

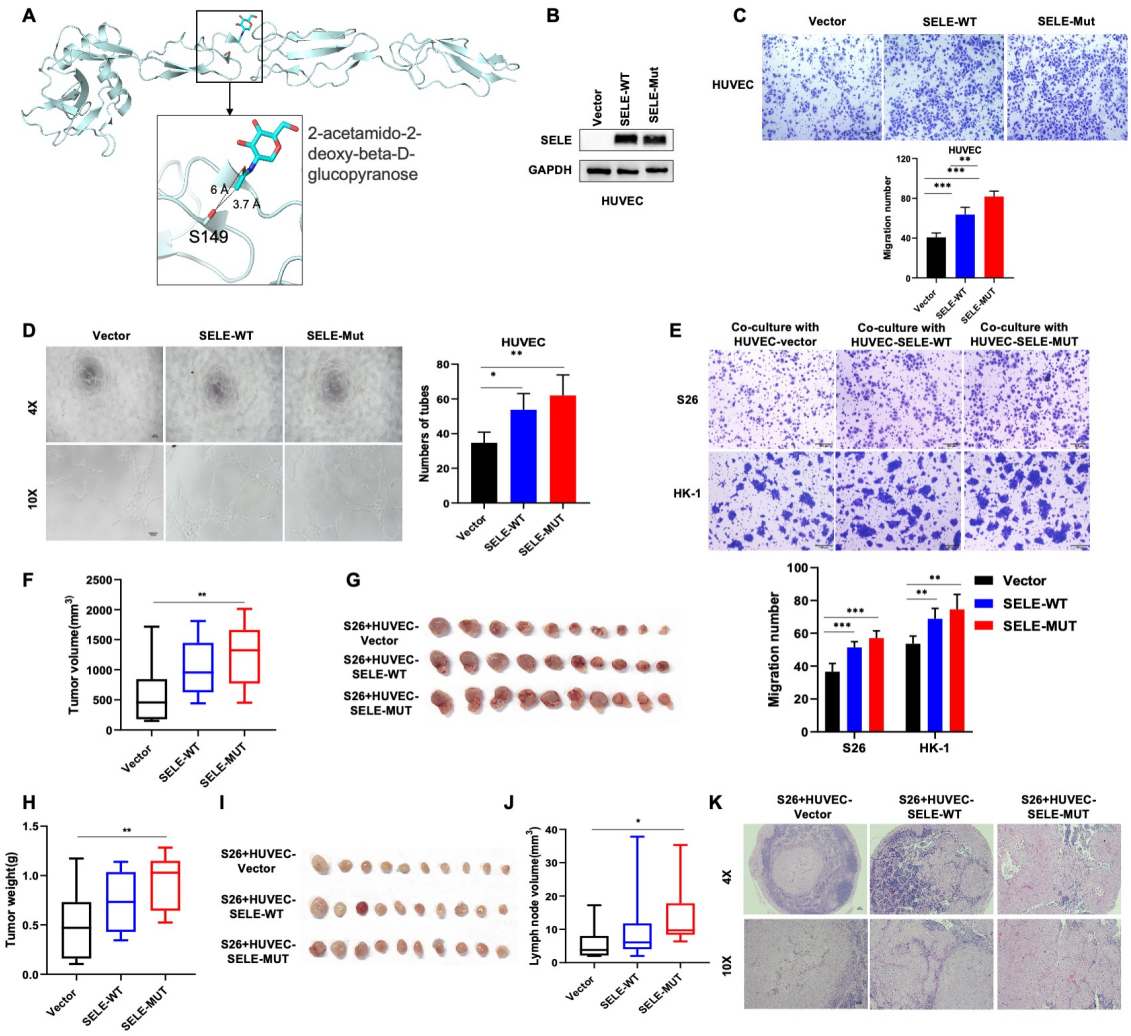



**Figure 7. SELE-S149R mutation in endothelial cells promotes the tumorigenesis and metastasis of NPC cells. A.** Structure prediction of E-selection paired with glycomimetic antagonist ligand 2-acetamido-2-deoxy-beta-D-glucopyranose (PDB code 4C16). The spatial proximity between Ser149 and the ligand is highlighted. Both the Ser149 side chain and the ligand are shown as sticks, with the distances between the OG atom of Ser149 and the C8 or O7 atoms of the ligand specified. **B.** Western blot examination of SELE protein level in HUVEC cells infected with lentivirus overexpressing SELE-WT, S149R mutant or control vectors. **C.** Transwell migration assay evaluating the migration ability of the cells described in A, with statistical analysis presented at the bottom. Scale bar, 100 μm. **D.** Tube formation assay with cells described in A. The statistical analysis is presented on the right. Scale bar, 100 μm. **E.** Transwell migration assay assessing the migration ability of the NPC cell lines (S26 and HK-1) co-cultured with HUVEC cells from A (S26/HK-1: HUVEC = 10:1), accompanied by statistical analysis presented at the bottom. Scale bar, 200 μm. **F-H.** Tumor growth evaluation in xenograft model with subcutaneous injection of S26 cells described in E. Tumor volumes. (F) are measured and visual presentation of tumor post-sacrifice (G) and weight (H)

are presented. **I-K.** Tumor lymphatic metastasis of S26 cells described in E. Lymph nodes are visualized (I) and measured (J), with HE staining conducted to assess metastasis in these lymph nodes (K). Scale bar, 100 μm. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

### *Genetic contributions of common and rare variants to NPC susceptibility*

We next undertook a comprehensive analysis to quantify the overall heritability of NPC explained by SNVs derived from WES, which was further broken down based on the MAF and LD score classifications of these SNVs, using GREML-LDMS model (see Methods). We observed that a total of 15.2% of NPC susceptibility (with a standard error of 3.6%) could be explained by the genetic effects of all WES-derived SNVs, including 4.4% and 10.8% attributed to variants within the HLA and non-HLA regions, respectively. Notably, rare variants (MAF < 0.01) in the non-HLA regions explained a substantial 82.4% of the WES SNV-heritability (**Figure 8A**). Furthermore, variants with the lowest LD scores, meaning relative independence, contributed the largest proportion (30.6%) to the WES SNV-heritability (**Figure 8A**).

Through a comparative analysis using a joint model that considers multiple genetic risk factors simultaneously (see Methods), we further revealed that the minor allele G of the rare variant *SELE*-rs5361 conferred a substantial disease risk (OR=2.20~2.22, **Figure S11**) in the two Guangdong cohorts (GD-SYSUCC and GD-ZS). This risk level is comparable to that associated with the highest 20% of the population's polygenic risk score (PRS) derived from HLA variants (OR=2.14~2.30, **Figure S11**). Intriguingly, the risk also overshadowed that conferred by the minor allele T of the common variant *RPL14*-rs2276868 (OR = 1.22~1.45) or the highest 20% of the population's PRS for NPC from non-HLA GWAS variants (OR = 1.31-1.80, **Figure S11**). Additionally, by assessing the phenotypic variance ascribed to each locus[42], we also revealed that HLA loci are predominantly responsible for the NPC phenotypic variance, while the *RPL14* common variant significantly explain a greater portion of disease variance than the *SELE* rare variant (**Figure S11**).

### *Polygenic risk prediction improved by novel loci and mediated by EBV infection*

To assess the efficacy of the novel loci in predicting NPC risk, we developed a refined composite polygenic risk score (rcPRS) that simultaneously considers NPC-associated SNVs and the leading SNVs in NPC-associated genes and pathways identified in our study and previous GWASs (see Methods; **Table S15**). Notably, the rsPRS demonstrated a consistent

553 right-skewed distribution for NPC cases compared to controls across all discovery and
554 replication datasets (**Figure 8B**), suggesting a higher predictive value of rsPRS in identifying
555 individuals at risk. Furthermore, the rcPRS achieved an area under curve (AUC) for risk
556 prediction ranging between 0.659 and 0.666 in the GD-SYSUCC discovery and replication
557 samples, outperforming the published GWAS-based PRS in the GD-SYSUCC sample (gPRS;
558 AUC = 0.649; **Figure 8C; see Methods**)[10]. This improvement in predictive accuracy is largely
559 attributable to the inclusion of two novel common variants, rs1050462 in *HLA-B* and
560 rs2276868 in *RPL14*, as evidenced by our leave-one-out analysis (**Table S16**). Further
561 stratification of individuals based on their NPC PRSs, either rcPRSs or gPRSs, revealed a more
562 pronounced rising trend in the relative disease risk with the escalating rcPRS compared to the
563 gPRS (**Figure 8D**). Individuals in the highest percentile rage (95-100%) of the rcPRS exhibited
564 approximately double the relative risk ($OR_{rcPRS_discov}$=15.1) than those with the gPRS
565 ($OR_{gPRS_discov}$=8.4; **Figure 8D**). These findings suggest that incorporating novel loci into the
566 rcPRS not only improves its accuracy in distinguishing disease risk but also enhances its
567 capability to stratify individuals based on their likelihood to develop NPC.

568 Noteworthily, in a subset of 1,018 cases and 774 controls from the GD-SYSUCC cohort
569 that had data on serological antibody against the EBV-encoded virus capsule antigen (EBV
570 VCA-IgA), we observed a stronger association between rcPRS and NPC susceptibility in
571 individuals tested seropositive for EBV VCA-IgA compared to those tested seronegative
572 **(Figure 8E)**. Interaction analysis statistically further confirmed that the association between
573 the rcPRS and NPC disease status was significantly mediated by EBV status ($P_{interaction}$=0.03).
574 These findings suggest that the polygenic risk of NPC is mediated by serum EBV activity, and
575 that rcPRS may have a greater utility in predicting NPC risk among individuals with serological
576 evidence of EBV infection.

577

578

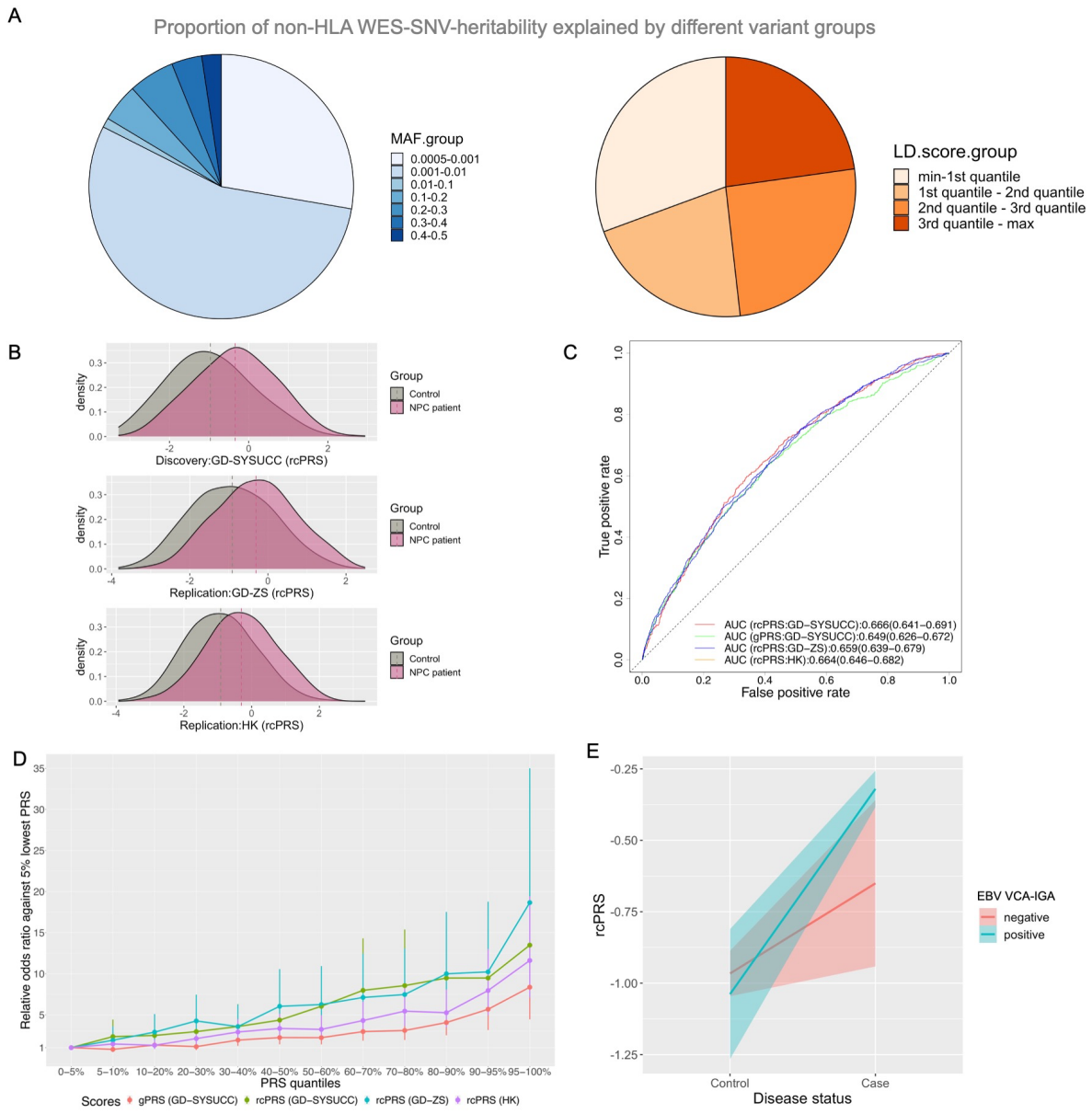



**Figure 8. Contribution of common and rare variants to NPC susceptibility. A**. Fractional representation of NPC heritability attributable to non-HLA WES-SNVs, categorized by minor allele frequency (MAF) and linkage disequilibrium (LD). For each variant, MAF was calculated using the discovery samples, and LD score represented the aggregated $R^2$ with adjacent variants spanning a 200kb window. **B**. Density plots illustrating the PRS incorporating the novel loci and previously known GWAS loci (refined composite PRS, rcPRS) for cases and controls in both the discovery and replication cohorts. **C**. Receiver operating characteristic curves comparing the rcPRS and previously reported GWAS PRS (gPRS) for NPC across different cohorts. **D**. Relative odds ratio comparing the rcPRS or gPRS bins and the 5% lowest quantile group in different cohorts. **E**. Correlation of rcPRS with disease status in individuals categorized by their EBV VCA-IgA status.

593 **Discussion**

594 Here we conducted the largest whole-exome sequencing association study on NPC to date and
595 unveiled three novel variants linked to NPC susceptibility, including the common variant
596 rs2276868 in the promoter (5'-UTR) region of *RPL14*, the rare coding variant rs5361 in *SELE*,
597 and another common variant rs1050462 in HLA-B locus[8,11]. Notably, the identification of the
598 *RPL14* locus underscores the significant role of familial history in NPC. This mirrors patterns
599 observed in other complex diseases[42], suggesting both genetic or environmental factors related
600 to family history might amplify the genetic impact of individual variants on NPC risk. A
601 possible explanation for this locus remaining undetected earlier could be its pronounced genetic
602 risk in familial cases ($OR_{FHNPC}$ =1.575) compared to sporadic cases ($OR_{sporadic_NPC}$ = 1.200),
603 coupled with the vast sample size of our study. Additionally, the successful replication of
604 associations with *SELE* and other two sentinel pathway genes underlines an efficient approach
605 to identify and decompose disease associations at the pathway level.

606 Our study sheds light on the functional roles of the novel genes, offering a better
607 understanding of the pathogenic mechanisms underlying NPC. We uncovered that the variant
608 rs2276868 modulates *RPL14* expression through the transcription factor *NKRF*, which is an
609 inhibitor to the NF-κB pathway involved in NPC development[43,44]. Single-cell transcriptomic
610 analyses revealed an inverse association between *RPL14* expression and EBV activity in
611 malignant NPC cells, which is further corroborated by the suppressive effect of RPL14 on EBV
612 infection and lytic cycle in NPC cells. Given the involvement of RPL22[45] and RPL4[46], both
613 of which belong to the same RPL family as RPL14[47], in EBV infection, these findings highlight
614 the crucial role of RPL family members in EBV life cycle. Furthermore, our study revealed
615 that *RPL14* inhibits NPC tumorigenesis, consistent with its tumor-suppressive function
616 reported in an *in vitro* NPC study[48] and in other cancer types[49]. Reinforcing this, a high *RPL14*
617 expression is associated with a favorable survival of NPC. Taken together, our findings suggest
618 that the polymorphism at rs2276868 predisposes individuals to varied *RPL14* expression,
619 which in turn influences EBV-related tumorigenesis in NPC.

620 Besides *RPL14* in our study, multiple NPC-associated genes such as *HLA-I, -II*, and *CIITA*
621 have been implicated in modulating EBV life cycle [29,30]. This reinforces that genetic variations
622 affect EBV to mediate varied susceptibility to NPC. Significantly, our findings reveals that the
623 effectiveness of the NPC polygenic risk score is influenced by the serological status of EBV
624 activation. Given that both the EBV DNA load and antibodies against EBV are established
625 serological markers for early NPC diagnosis[60,61], incorporating the rcPRS and EBV infection

626 markers could significantly enhance personalized risk assessments and lead to the development
627 of more effective screening strategies for NPC.

628 Our study has highlighted *SELE* as a novel NPC-associated gene. This association is likely
629 attributed to the rare germline mutation rs5361 in *SELE*, which encodes an adhesion protein E-
630 Selectin, known for its interactions with various cells, including leukocytes and malignant
631 epithelial cells[36,50]. A previous study has demonstrated that the deleterious S149R encoded by
632 rs5361 enhances the functionality of E-Selectin in endothelial cells, promoting their rolling and
633 adhesion capabilities[37]. Noteworthily, our study extends these findings by demonstrating that
634 *SELE* expression is confined to endothelial cells in tumor tissues. Our structural modelling and
635 functional experiments reveal that S149R mutation extends the SELE's side chain on the
636 surface of endothelial cells, thereby augmenting their migration and angiogenesis abilities.
637 Furthermore, we observed that endothelial cells expressing SELE, particularly those with the
638 S149R mutation, significantly render stronger migration and tumorigenesis to malignant NPC
639 cells. This supports the crucial role of E-selectin in interacting with cancer cells and modulating
640 metastasis across various cancers[36,51-53]. Therefore, these findings suggest that the *SELE*-
641 S149R may enhance the adhesion function of E-Selectin in endothelial cells, reshaping the key
642 TME component to favor the progression of cancer cells[53], consequently altering individual
643 susceptibility to NPC.

644 We further delved deeper into the cellular expression of various NPC-associated genes,
645 revealing versatile patterns. For example, certain genes, like *RPL14*, are expressed in most cell
646 types, including epithelial cells that give rise to malignant NPC cells. In contrast, genes like
647 *HLA-II* are found mainly in diverse stromal cells, which typically lack malignant potential in
648 NPC. Such patterns are also observed in other cancers. For instance, the well-known
649 susceptibility-associated genes *NOTCH4* in lung cancer and *SMAD1* in colorectal cancer[54,55]
650 are exclusively expressed in endothelial cells (not typically developing into malignancy) in
651 tumors. These findings highlight the multifaceted cellular implications of germline variations
652 in cancer susceptibility, resonating with a growing literature suggesting the engagement of
653 various cell types in the development of complex disease[56-58]. Given that the validation of the
654 functional relevance of genetic susceptibility genes often predominantly emphasizes the cells
655 of cancer origin (e.g., epithelial cells in NPC), our findings call attention to the importance of
656 exploring the roles of cancer risk genes across diverse TME cell types in cancer development
657 (such as endothelial cells)[54,55].

In summary, our study fills the missing heritability of NPC and enhances our understanding of the genetic architecture of NPC by unveiling novel genetic associations from variants of both common and rare prevalence. Furthermore, our study elucidates diverse pathogenic mechanisms, including the EBV-mediated tumorigenesis in malignant epithelial cells through *RPL14* and the adhesion component in endothelial cells via *SELE*. These findings highlight the substantial impacts of genetic variants on EBV life cycle and various cell types within the TME, which jointly contribute to NPC development. Moreover, our innovative rcPRS model, incorporating these novel discoveries, reflects the modifier effect of EBV status on polygenic risk and provides promising implications for NPC screening. Additionally, we acknowledge limitations, including sample size constraints, especially for the FHNPC group and rarer variants, and the need for further studies to delve into the mechanistic links between these genetic variants and NPC pathogenesis. Future work should explore the potential clinical benefits of these findings in NPC management.

**Acknowledgement**

We acknowledge supports from the National Natural Science Foundation of China (NSFC; 82261160657, 81971270), the National Key R&D Program of China (NKRDPC; 2022YFC3400901 and 2021ZD0202000), the Sci-Tech Project Foundation of Guangzhou City (201707020039), the Guangdong Innovative and Entrepreneurial Research Team Program (2016ZT06S638), the Chang Jiang Scholars Program (J.X.B.), Hong Kong Research Grant Council (RGC) Theme-based Research Scheme Funds (T12-703/22-R and T12-703/23-N), National Medical Research Council Singapore Clinician Scientist Award (NMRC/CSAINV20nov-0021), the Duke-NUS Oncology Academic Program Goh Foundation Proton Research Programme (M.L.C), NCCS Cancer Fund (M.L.C), the Kua Hong Pak Head and Neck Cancer Research Programme (M.L.C), the Research Grants Council Area of Excellence (AoE) Hong Kong NPC Research (AoE/M-06/08), and the Hong Kong Cancer Fund (M.L.L.). We thank all participants in the study, staffs at the biobank of SYSUCC for sample preparation, staffs at the Tissue Bank for providing NPC blood samples and the patient demographics of the HK cohort, Tan Kah Min's support for technical support and sample processing in SG cohort, and staffs at the High-Throughput Analysis Platform (HTAP) of SYSUCC for data generation and processing.

690

691 **Author contributions**

692 Designed the study: J.X.B, Y.X.Z; procured financial support: J.X.B, M.L.C, M.L.L, M.J;
693 sample recruitment, preparation, and data collection: M.J, F.G.L, C.C.K, Z.L, M.L.L, J.M.K,
694 M.L.C, E.H.O, Y.M.G, J.R.C, S.H, S.Q.L, X.X.C, X.Y, B.W, Y.H.Z, A.Y.X, P.P.W, Q.Y.C,
695 L.Q.T, W.H.J, H.Q.M; analyzed and interpreted data: Y.Z, J.X.B, C.L.L, G.W.L, X.B, Y.L,
696 S.H, Z.H.R, Y.L.C, C.C.K, J.J.L, M.L.C, J.M.K, E.H.O, J.D.M, D.X.L; performed functional
697 experiments: C.L.L, J.X.J, Y.Q.Z; wrote the paper (original draft): Y.Z, C.L.L, J.X.B; all
698 authors approved the final report.

699

700 **Availability of data and materials**

701 Key data were deposited in the Research Data Deposit public platform (RDD;
702 http://www.researchdata.org.cn ).

703

704 **Declaration of interests**

705 The authors declare no competing interests.

706

707 **Materials and Methods**

708 **Ethics**

709 The study was approved by the Sun Yat-sen University Cancer Center (SYSUCC) Ethics
710 Committee (reference no. SL-B2021-032-03), the SingHealth Institutional Review Board (IRB
711 protocol no. 2019/2177), and the Institutional Review Board (IRB) of the University of Hong
712 Kong. Informed consent was obtained from all participants.

713

714 **Patient recruitment and sample preparation**

715 ***Discovery stage***

716 A total of 3,039 NPC patients of Chinese Han ancestry was recruited from Guangzhou in
717 Guangdong province of China (GD-SYSUCC; n= 2,196) by SYSUCC and from Singapore
718 (SG; n=843) by the National Cancer Center (Singapore, collected from 1987-2018). A total of
719 2,692 healthy individuals of Chinese Han ancestry was recruited as controls from local
720 community in Guangdong (GD-SYSUCC; n=1,355) and from Singapore (SG; n =1,337) by

721 Genome Institute of Singapore (Singapore, details have been described previously[63]). All NPC
722 cases were histologically diagnosed. All healthy controls declared free of cancer through
723 questionnaires. After stringent quality control, the discovery dataset consisted of 2,694 NPC
724 cases (2,125 from GD-SYSUCC and 569 from SG) and 2,328 healthy controls (1,068 from
725 GD-SYSUCC and 1,260 from SG). Serum immunoglobulin A (IgA) antibodies to Epstein-Barr
726 virus (EBV) capsid antigen (IgA-VCA) were available for 1,646 NPC patients and 918 controls
727 in GD-SYSUCC subset. Individuals with IgA-VCA >=1:20 was categorized as EBV positive.

728

729 ***Replication stage***

730 The replication dataset included two cohorts: a Zhongshan cohort (GD-ZS) consisting of 1,941
731 cases and 2,265 controls recruited from the Zhongshan city in the Guangdong province by
732 Zhongshan City People's Hospital, and a Hong Kong cohort (HK) consisting of 2,334 cases
733 and 2,507 controls recruited from Hong Kong by the University of Hong Kong. Specifically
734 for the HK cohort, NPC cases and controls were collected by the Tissue Bank established under
735 the Area of Excellence (AoE) program from five Hong Kong public hospitals including Queen
736 Mary Hospital (QMH), Queen Elizabeth Hospital (QEH), Tuen Mun Hospital (TMH), Pamela
737 Youde Nethersole Eastern Hospital (PYNEH), and Princess Margaret Hospital (PMH) from
738 2010 to 2017, details have been described previously[11]. NPC cases were confirmed by
739 histopathology and enrolled with routine staging procedures according to the American Joint
740 Committee on Cancer (AJCC) TNM system, physical examination, and imaging tests. The
741 control population was recruited from the Red Cross and hospital cancer-free individuals from
742 QMH, QEH, TMH, PYNEH, and PMH. We considered GD-ZS as a Guangdong indigenous
743 cohort that shared regional demographic features with the discovery GD-SYSUCC cohort
744 collected from the same province, whereas the HK cohort shared ancestral genetics with other
745 cohorts but was relatively demographically different.

746

747 ***Bulk RNA-Seq samples***

748 Tumor biopsy samples from 93 NPC patients were collected for bulk RNA-seq (labelled as
749 "the Bei-lab collection"), 87 of which had prognosis data available. Additional bulk RNA-seq
750 data of tumor tissues from 113 independent NPC patients with prognosis data available were
751 collected from a published study (labelled as "the Zhang-lab collection")[64]. We further
752 collected nasopharynx samples from 10 cancer-free individuals with inflammations in

753 nasopharynx at SYSUCC, which were used as controls. Biopsies were collected before any
754 treatment and immersed in RNAlater.

755

756 ***NPC samples for Single-cell RNA-seq***

757 Tumor biopsies of 10 NPC patients were collected for single cell RNA sequencing (scRNA-
758 seq) in our previous study[32]. Among them, eight had tumor epithelial cells successfully
759 captured. By extracting data from four additional studies[33,65-67], we further collected scRNA-
760 seq data of 46 NPC tumor samples, 27 of which captured tumor epithelial cells[33,65-67]. In total,
761 scRNA-seq datasets consisting of 56 samples (of which 35 have epithelial cells captured) were
762 used in downstream analyses.

763

764 **Data generation, quality control and annotation**

765 ***Whole exome sequencing and variant calling***

766 Whole blood DNA sample was extracted from all participants in the discovery dataset for
767 whole exome sequencing. Library was constructed using one of the three products: Agilent
768 SureSelect Human All Exon V6+UTR kit, Agilent SureSelect Human All Exon V5+UTR kit,
769 and Roche SeqCap EZ Exome + UTR Target Enrichment Kit. This was followed by sequencing
770 using a paired-end 2 x 150 bp protocol on an Illumina HiSeq instrument. A custom optimized
771 version of Sentieon[68] was applied to preprocess sequencings reads, align reads to human
772 genome using GRCh37 (hg19) assembly as reference, and conduct single nucleotide variant
773 (SNV) calling using algorithms following the GATK best-practice pipeline[69,70]. After GATK's
774 VQSR quality control, hard filtering was applied to remove the variants with 1) Read depth
775 (DP) <10; 2) Allelic Balance (AB) > 0.8 or < 0.2 for heterozygous genotype; 3) Genotype
776 quality (GQ) < 30. For rare variants (minor allele frequency < 0.01), we applied additional
777 stringent quality control based on per-read mutation rate, allelic balance, and strand bias
778 (defined as proportion of allele called by reads from one strand over that by any of the two
779 strands). We used 1000 genome phase 1 high confident SNP set as golden set to establish
780 parameter threshold range where high liability variants can be called. This led to the following
781 filtering for rare variants: 1) refining alignment by removing any read with a mismatch distance
782 against the reference larger than 8 bp (any indel event counts for 1 bp) and re-apply hard
783 filtering as described above; 2) setting variant-sample elements with strand bias > 0.95 or <
784 0.05 for heterozygous genotype, or >0.97 or <0.03 for homozygous genotype to missing; 3)

785 setting variant-sample elements with allelic Balance (AB) > 0.75 or < 0.22 for heterozygous
786 genotype , or with allelic Balance (AB) < 0.95 and > 0.04 for homozygous genotype to missing.

787

788 ***Additional quality control for association test***

789 For both common and rare variants, we further removed variants that are 1) with a missing
790 rate >= 0.02; 2) with a Hardy-Weinberg equilibrium P value <= 1 x $10^{-6}$ in healthy controls; 3)
791 with a significant differential missing rate between case vs controls (Bonferroni corrected); 4)
792 variants in super duplicates or low complexity regions. Samples were removed if they are 1)
793 with a missing rate > 0.05; 2) with autosomal heterozygosity deviation outside 3 standard
794 deviations of the mean; 3) genetic outliers through eye check or with any the of top four
795 principal component value located outside 6 standard deviations of the mean. These filtering
796 steps were performed using PLINK[71]. We also removed one of each related pair (remained
797 sample t<0.05) using GCTA[72].

798

799 ***HLA genotyping***

800 High resolution WES-based HLA typing was performed on the bam files using HLAScan
801 (v2.1.3) with default parameters and implemented HLA database[73]. PCR amplification HLA
802 typing was also performed on 135 WES samples to cross-validate the WES-based typing
803 procedures. The concordance of the HLA genotypes was high (> 94%) at four-digit resolution
804 level for most HLA genes, except for HLA-DQA1 and -DPB1 with a relatively lower
805 concordance being observed (> 80%). The HLA amino acid (AA) polymorphism was obtained
806 using HLA AA alignments in IPD-IMGT database[74]. The genotypes of HLA SNPs were
807 inferred by the SNP2HLA software[75].

808

809 ***Capture sequencing for targeted genomic regions***

810 DNA library preparation was performed as described in the NadPrep DNA Library Preparation
811 Kit (#1002222; Nanodigmbio, Nanjing, China) for the MGI sequencing platform on the
812 MGISP-960 automated workstation (MGI Tech Co., Ltd., Shenzhen, China), in accordance
813 with the manufacturers' instructions. Hybridization capture-based target enrichment was
814 performed as described in the NadPrep Hybrid Capture Reagents kit (#1005101; Nanodigmbio)
815 on the MGISP-100 automated workstation (MGI Tech Co., Ltd., Shenzhen, China), in

816 accordance with the manufacturers' instructions. The library distribution was analyzed by
817 Qsep100 (BiOptic Inc, New Taipei City, China) and quantified by Qubit dsDNA HS Assay Kit
818 (Invitrogen, Termo Fisher Scientifc Inc, Waltham, USA). Final libraries were sequenced on
819 MGI DNBSEQ-T7 (MGI Tech Co., Ltd., Shenzhen, China) using a pair-end of 150 bp.

820

821 ***Whole transcriptomic sequencing***

822 For bulk RNA-seq, total RNA was extracted from tissue or cell line using Rneasy Mini Kit
823 (Qiagen), with ribosomal RNAs removed by Ribo-Zero Magnetic kit (Illumina). TruSeq RNA
824 Library Prep Kit (Illumina) was used to construct the library, followed by RNA sequencing.
825 RNA-seq was conducted with a pair-end of 150 bp protocol on an Illumina Hiseq X sequencer.
826 Bowtie was used to map post-Qced reads to human reference genome (hg19) and Htseq was
827 used to quantify reads counts[76,77]. Per-gene expression level was normalized using transcripts
828 per million (TPM). scRNA-seq was performed following the manufacturers' instructions (10x
829 Genomics single-cell 5' sequencing and 3' sequencing kits). Details were described in previous
830 studies[32,33,65-67].

831

832 ***Variant annotation***

833 ANNOVAR was used to annotate SNV with following categories[78]:
834 1) SNPs in dbSNP150 database.
835 2) Variants in population genetics database: 1000 genome phase 3, exome and genome subset
836 of the Genome Aggregation Database (EXAC, gnomAD).
837 3) within genes in NCBI RefSeq Gene database.
838 4) within genes of pathways in curated canonical pathways, GO Ontology items and oncogenic
839 signature gene sets as downloaded from Molecular Signatures Database v6.2[79].
840 5) variants that have potential functional consequences as they ranked at top 5% of the rank
841 score of CADD (Combined Annotation Dependent Depletion), or DANN (deleterious
842 annotation of genetic variants using neural networks) or fitCons (fitness consequence score) [80-
843 82].
844 6) 'pathogenic' or 'likely pathogenic' variants as annotated in InterVar database (clinical
845 interpretation of missense variants)[83];

846

847 **Statistical analysis**

848 ***Single-variant-based association test***

849 For variants in non-HLA region, SNVs with a MAF >= 0.001 were included in a logistic
850 regression in which NPC disease status was regressed on the genotype of individual SNV. For
851 variants in HLA region, four types of variants were included in analysis: SNVs, four-digits
852 HLA alleles, six-digits HLA alleles, and amino acids. For each variant, two tests were
853 performed for three binary phenotypes: all NPC cases (ALLNPC) vs controls, and NPC cases
854 with family history (FHNPC) vs controls, respectively. Other covariates in the model includes
855 top five genetic principal components and gender. Bonferroni method was applied in multiple
856 testing correction (N=157,024), with significance threshold P value of $3.2 \times 10^{-7}$.

857

858 ***Conditional analysis for single-variant-based associations***

859 To test whether the single-variant association from target SNV was independent of the
860 associations of other SNVs, conditional analysis was performed by jointly fitting each of the
861 SNVs to be conditioned as covariate variable in the regression model.

862

863 ***SNV-set-based association analysis 1: gene-based test***

864 The gene boundary was defined by the transcription start and end sites. These was no MAF-
865 based filtering for this analysis (singletons were also included). Apart from testing using all
866 SNVs (SNV type 1), we also performed test that only consider coding-affecting SNVs. For
867 each SNV type, we tested 4 algorithms that are available on the "SKAT" package in R:

868 1) The original SNP-set (Sequence) Kernel Association Test (SKAT)[84].

869 2) The original Burden test[85].

870 3) SKAT for the combined effect of common and rare variants using the sum test as
871 described previously[86].

872 4) Burden test for the combined effect of common and rare variants using the sum test as
873 described previously[86].

874 For each gene tested using each SNV type for each phenotype, the result from the algorithm
875 with the most significant P value was remained. Top ten genetic principal components and

876 gender were included in the model as covariates. Bonferroni method was applied in multiple
877 testing correction (N=22,228), with significance threshold P value of $2.3 \times 10^{-6}$.

878

879 ***SNV-set-based association analysis 2: pathway-based test***

880 The pathway information was downloaded from Molecular Signature Database, including
881 curated canonical pathways, GO Ontology items and oncogenic signature gene sets[79]. After
882 annotating SNVs to the pathways, we performed pathway-level association analysis to test
883 cumulative effect from both rare and common variants within a pathway on NPC risk. To avoid
884 false positive results driven by very small or very large gene set, we excluded pathways with a
885 size larger than 200 genes or smaller than 5 genes. The same grouping and algorithms applied
886 in the gene-based test were also applied here. Multiple testing correction was performed using
887 the FDR method ($N_{pathway}$=6,204).

888

889 ***Meta-analysis for single-variant-based and gene-based associations.***

890 The random effect model (REML) was applied to meta-analyze the association between
891 individual variants and NPC using the R package "metafor"[87]. The summation of logits
892 method[88] was applied to meta-analyze the association between individual genes and NPC using
893 the R package "metap"[89].

894

895 ***Structural modelling of E-selectin***

896 The molecular structure of the E-selectin complexed with glycomimetic antagonist 2-
897 acetamido-2-deoxy-beta-D-glucopyranose (PDB code 4C16) was modelled as previously
898 described[90]. PyMOL with the Adaptive Poisson-Boltzmann Solver plugin was used in the
899 evaluation of the electrostatic surface and the visualization.

900

901 ***Assessing cell-type specific gene expression of NPC and other cancer-associated genes using***
902 ***scRNA-seq data of tumor and non-tumor tissues***

903 Susceptibility gene list of NPC, lung, colorectal and gastric cancers included those identified
904 in present study (for NPC only) and those collected from published GWAS studies (**Table S10)**.
905 Expression profiles were derived from multiple scRNA-seq datasets: the NPC tumor dataset

906 ($n_{sample}$=56, $n_{cell}$= 223,593) as described in "NPC samples for Single-cell RNA-seq" section;
907 the lung cancer tumor dataset ($n_{sample}$=19, $n_{cell}$=90,477) downloaded from the GEO database
908 (GSE131907, GSE139555) and collected from a previous pan-cancer study[91]; the gastric
909 cancer tumor dataset ($n_{sample}$=9, $n_{cell}$=28,299) provided by a published study[92]; the colorectal
910 cancer tumor dataset ($n_{sample}$=37, $n_{cell}$=70,015) downloaded from the GEO database
911 (GSE132465, GSE144735 and GSE139555); and the non-cancer tissue dataset consisting of
912 483,152 cells (for computational consideration, 100,000 cells were randomly selected in
913 downstream analyses) from 10 tissues of 15 donors provided by a published study[37]. The
914 quantification of gene expression and determination of cell types were performed separately
915 for each dataset.

916 Specifically, for NPC data we first used DoubletFinder to detect doublets[93] by inspecting
917 abnormal clusters that simultaneously express two or more cell type markers. For each sample,
918 doublets were removed with the expected doublet rate of 0.05. Cells with $n_{Feature_RNA} > 500$,
919 $n_{Count_RNA} > 1{,}000$ and percent.mt < 25% were retained for downstream analyses. Expression
920 Matrix of all cells were converted to a Seurat object and managed by Seurat's pipeline[94]:
921 normalization with NormalizeData and ScaleData, batch effect correction with Harmony[95],
922 dimensionality reduction with RunUMAP, clustering with FindNeighbors (k.param = 30) and
923 FindClusters. Clusters were annotated by classic markers as following: NK cells (KLRF1), T
924 cells (CD3D), Plasma cells (MZB1), B cells (MS4A1), Myeloid cells (AIF1), Cancer
925 Associated Fibroblasts (COL1A1), Endothelial cells (VWF) and Epithelial cells (EPCAM).
926 Normalized expression of marker genes for each cell type in each tumor datasets were shown
927 in **Figure S12**.

928 Lung, gastric, colorectal cancer tissue datasets were merged and analyzed. We first used
929 Scrublet[96] to remove estimated doublets and SoupX[97] to remove cell-free mRNA
930 contamination. Cells with $n_{Feature_RNA} > 500$, $n_{Count_RNA} > 1000$ and percent.mt < 25% through
931 quality control were kept for downstream analysis. NormalizeData and ScaleData in Seurat
932 was used to normalize data[98]; FindVariableFeatures with $n_{feature}$ = 3000 was used to select
933 hypervariable genes. We used Harmony to correct for batch effect caused by different patients[95],
934 and used RunUMAP with dims = 1:30 and FindClusters with resolution = 0.3 for dimension
935 reduction and clustering. The clusters were annotated by classic markers of major cell types.
936 Normalized expression of marker genes for each cell type in each tumor datasets were shown
937 in **Figure S12**.

938 For the non-cancer tissue dataset, cell clusters were determined by the original study[37].

939

940 ***Inference of malignant status of epithelial cells in scRNA-seq data in NPC tumor tissues***

941 Malignant cells have substantially increased copy number variation (CNV), such as large-scale
942 gain or deletions of chromatin. Therefore, we use inferCNV to identify malignant epithelial
943 cell clusters[99]. Comparing with control normal epithelial cells downloaded from GEO database
944 (GSE121600), all epithelial cell clusters identified in NPC scRNA-seq data in our study were
945 considered malignant.

946

947 ***Estimate of WES-SNV heritability for NPC***

948 GREML-LDMS (the LD and MAF stratified GREML) approach was applied to estimate the
949 proportion of NPC risk that can be explained by WES SNVs [100]. In brief, when calculating the
950 total WES SNV-heritability and the proportion explained by variants from HLA and non-HLA
951 regions, SNVs were stratified into four groups based on MAF and regions: MAF ≥ 0.01 in
952 HLA region, MAF<0.01 HLA region, MAF ≥ 0.01 in non-HLA region and MAF < 0.01 in
953 non-HLA region; when further decomposing the WES SNV-heritability in non-HLA region,
954 SNVs were stratified into 28 groups based on LD scores (four bins based on quartiles. LD score
955 was calculated as the sum of $R^2$ with variants in a 200kb window) and MAF (7 bins: 0.0005 <
956 MAF <= 0.001 ,0.001 < MAF <=0.01, 0.01 < MAF <=0.1, 0.1 < MAF <= 0.2, 0.2 < MAF
957 <=0.3, 0.3 < MAF <=0.4, and 0.4 < MAF <=0.5). For each SNV group, genomic relationship
958 matrix (GRM) was created. In each model (the four components and 28 components ones),
959 GRMs from all groups were then simultaneously fitted in the GREML model for NPC, where
960 the total NPC variation explained by all genetic components and the variation explained by
961 individual genetic component as represented by each GRM can by estimated.

962

963 ***Joint model for a comparative analysis to estimate the relative risks by RPL14, SELE, HLA,***
964 ***and non-HLA GWAS loci***

965 In discovery samples, a subset (1,382 cases and 912 controls) of the GD-SYSUCC cohort
966 simultaneously have imputed GWAS array[17] and WES data available, allowing the
967 comparisons of the relative disease risks conferred by the common variant rs2276868 in *RPL14*
968 and the rare variant *rs5361* in SELE with the polygenic risk from the HLA and other known
969 GWAS loci. In replication samples, the cap-seq designed to target at both published GWAS

970 loci and the candidate loci identified by the presented WES study also enabled this comparison.
971 In brief, for each cohort we applied a joint regression model that simultaneously accounts for
972 genetic effects from different variants: the effects from *RPL14* and *SELE* were presented by
973 the genotypes of rs2276868 (*RPL14*) and rs5361(*SELE*), respectively; the combined genetic
974 effects from published non-HLA GWAS loci and that from HLA loci were represented as two
975 polygenic risk scores (one for non-HLA GWAS loci and one for HLA loci) constructed using
976 odds ratios estimated from original publication or our single-variant analysis using discovery
977 cohorts (**Table S14**).

978

979 ***Construction and evaluation of polygenic risk scores (PRS) for NPC***

980 ***1. Construction of a refined composite PRS (rcPRS) for NPC risk***

981 The rcPRS was calculated for all individuals with non-missing genotypes in the discovery
982 sample (the subset of 1,382 cases and 912 controls with both GWAS array and WES data
983 available[17] in GD-SYSUCC cohort) and in all replication samples (the capture sequencing
984 allowed the genotyping for constituent variants of combined PRS). The following SNVs were
985 used to create rcPRS to simultaneously account for genetic risk factors identified in our WES
986 analyses and published GWASs[8,17,101]:

987 1) In WES data, significant variants in single-variant-based association test (including HLA
988 variants identified in conditional analysis).
989 2) In WES data, leading variants in gene-based test if they:
990 a. were in genes significant in SKAT-gene-based association test.
991 b. reached $P<1\times10^{-4}$ in single-SNV-based association test.
992 3) In WES data, leading variants in pathway-based test as they:
993 a. were in a gene of significant pathways and that gene has a $P<0.01$ in gene-based test.
994 b. reached $P<1\times10^{-4}$ in single-SNV-based association test.
995 4) Published GWAS loci [8,17,101]. Only non-HLA GWAS loci was extracted from this source
996 as majority of HLA SNPs identified in published GWAS were in non-coding regions,
997 which are beyond the regions captured by WES and that signals from HLA regions have
998 already been included in source 1.

999 To avoid duplication due to overlapped genes among pathways, each variant can be used
1000 only once. Clumping was performed so that in each cluster of LD partners with $R^2>=0.1$ (within
1001 250kb extended windows) only one variant could be remained. PRS was created as the sum of

1002 weighted genetic risk from all variants. For each variant, the weighted genetic risk was
1003 calculated as the product of allele dosage and the coefficient of variant effect estimated in
1004 single-variant-based test. For variants from published GWAS, the coefficient was extracted
1005 from original publications [8,17,101]. Details of the constituent variants were shown in **Table S15**.

1006

1007 ***2. Construction of GWAS-only PRS (gPRS) for NPC risk***

1008 For a subset sample (n=2, 294) of the discovery GD-SYSUCC cohort, we produced imputed
1009 genotypes for BeadChips array data. In brief, SNP array data was obtained from a previous
1010 study[8]. To prepare the data for imputation analysis, we first used Liftover to convert the SNP
1011 array data from hg18 to hg19 encoding strand. Next, we performed quality control at the SNP
1012 level, removing SNPs with a minor allele frequency (MAF) less than 0.01, genotype missing
1013 rate greater than 0.05, and those that failed the Hardy-Weinberg equilibrium test (HWE) with
1014 a P-value less than $10^{-5}$. A total of 463,270 SNPs passed quality control and were retained for
1015 further analysis. For the quality-controlled genotype data, we used SHAPEIT2 with default
1016 parameters to estimate haplotypes of the study samples[102]. We then used IMPUTE2 with
1017 default parameters and the haplotype data from the 1000 Genomes Project to impute the
1018 missing genotype of each individual's haplotype[103]. After imputation, we removed SNPs with
1019 an INFO value less than 0.8 and set loci with genotype probability (GP) values less than 0.9 to
1020 missing. Ultimately, we retained 6,095,964 SNPs for analysis. This allows the construction of
1021 a previous published PRS using NPC-associated loci reported by GWAS[10], following the
1022 instruction from the original study[10]. To note, for three SNPs (rs2106123, rs6475604,
1023 rs9507124) not genotyped and not imputed in GD-SYSUCC, we used three surrogate SNPs in
1024 complete LD ($R^2$=1 in Chinese Southern Han population in 1000 genome phase 3 dataset) with
1025 them (rs6774494 for rs2106123, rs1412829 for rs6475604, and rs9510790 for rs9507124) in
1026 the model.

1027

1028 ***Construction of composite scores for pathway and EBV activity***

1029 We applied the AddModuleScore function in R package Seurat [98] to profile each pathway from
1030 the REACTOME and the GO Ontology database in each sample, where bulk RNA-seq or
1031 scRNA-seq data was available. Based on AddModuleScore, the pathway score is obtained by
1032 subtracting the average expression of the corresponding control gene set from the average
1033 expression of the target gene set. For samples with scRNA-seq, we took each patient as a

1034 datapoint, meaning that the average expression of a given gene across all epithelial cells from
1035 the patient was treated as the expression of that gene for the patient. We additionally created
1036 composite scores for EBV activity. In brief, we downloaded the EBV gene list from the EBV
1037 reference sequence (Akata; https://github.com/flemingtonlab/public/tree/master/annotation)
1038 and mapped the scRNA-seq data of EBV-positive malignant epithelial cells onto them. We
1039 then used the annotated EBV gene data to construct an EBV-composite score for each NPC
1040 patient with scRNA-seq data derived from EBV-positive epithelial cells available, following
1041 the same steps as described above.

1042

1043 ***The correlation between the pathway composite scores and RPL14 expression***

1044 Pearson correlations between *RPL14* expression and composite scores were tested using the
1045 cor.test function in R. Multiple testing correction was performed using the FDR method. The
1046 Pearson correlations of individual EBV genes and *RPL14* were shown in **Table S17.**

1047

1048 **Functional characterization of NPC-associated loci**

1049 ***Cell culture***

1050 Human NPC cell lines S26, CNE2-EBV, and HONE1-EBV were kindly gifted by Professor
1051 Mu-Sheng Zeng at the SYSUCC. Human NPC cell lines HK-1 and C666-1 were kindly gifted
1052 by Professor KW Lo at the Chinese University of Hong Kong. Human embryonic kidney cell
1053 293T cell was purchased from Cell Bank of Type Culture Collection of Chinese Academy of
1054 Sciences, Chinese Academy of Sciences. All above cells were cultivated in DMEM medium
1055 containing 10% fetal bovine serum (FBS, Gibco, NY, USA) and 1% penicillin–streptomycin
1056 solution (FBS, Gibco, NY, USA). Human umbilical vein endothelial cells (HUVECs) were
1057 isolated and cultured in ECM medium (ScienCell, CA, USA) supplemented with 15% FBS and
1058 0.03 mg/mL of endothelial cell growth supplement (ECGS, ScienCell, CA, USA). All cells
1059 were cultured at 37°C under a humidified atmosphere with 5% CO2. For all cell lines,
1060 mycoplasma was routinely determined using PCR with specific primers and no contamination
1061 was detected (**Table S18**).

1062

1063 ***Cell proliferation***

1064 For cell growth curves, a total of $1\times10^5$ NPC cells was plated in a 12-well plate with three
1065 replicates. The number of cells was measured by a cell counter with trypan blue staining on 24,
1066 48 and 72h. For colony formation assays, NPC cells were seeded at a density of 3000 cells per
1067 replicate into 6-well plate for 7-10 days. Then, colonies were fixed with 4% paraformaldehyde
1068 for 30min at room temperature. After stained with crystal violet for 10min, cells were
1069 photographed with the Bio-Rad ChemiDoc Touch (Hercules, CA, USA).

1070

1071 ***Cell migration***

1072 Transwell assay was performed as described previously[104]. Briefly, NPC cells in serum-free
1073 medium were plated at a density of $6\times10^4$ cells into the transwell chamber (8 μm pores, Corning,
1074 NY, USA), which then placed at 24-well plates containing 600 μL DMEM suppled with 10%
1075 FBS. After 16-24h, cells on the membrane were fixed with 4% paraformaldehyde for 30min
1076 and stained with trypan blue solution for 15min at room temperature. Then the cells were
1077 imaged under a microscope.

1078

1079 ***Luciferase reporter assay***

1080 Luciferase reporter assay was performed as described previously[105]. Briefly, the promoter
1081 region of RPL14 containing the rs2276868-C or -T was cloned into a luciferase reporter pGL3-
1082 basic plasmid. Afterwards, 293T cells were co-transfected with 200 ng of pGL3-RPL14-C or
1083 -T constructs and 10 ng of pRL (Renilla luciferase) plasmid together with respective siRNAs
1084 or overexpression plasmids. After 48h, cells were harvested to measure the luciferase activity
1085 through a Dual-Luciferase Assay Kit (Promega, USA) according to the manufacturer's
1086 instructions.

1087

1088 ***Immunofluorescence staining (IF)***

1089 Paraffin slides of xenograft tumors were deparaffinized with xylene and rehydrated through a
1090 gradient ethanol series. After microwave antigen retrieval in sodium citrate solution (pH 6),
1091 slides were incubated in 0.3% hydrogen peroxide for 15 min at room temperature to quench
1092 the endogenous peroxidase activity. Afterwards, the samples were incubated with the primary
1093 anti-Ki-67 (AB_393778, BD Pharmingen™, USA) antibodies overnight at 4 °C. Next day,
1094 Alexa Fluor-conjugated secondary antibody (Thermo Fisher Scientific, USA) was used for 1 h

1095 at room temperature, followed by counterstaining with antifade mounting medium and 4′ ,6-
1096 diamidino-2-phenylindole (DAPI) for 10 min. Images were then captured with a confocal laser
1097 scanning microscope (Carl Zeiss, Microscope 880, Germany).

1098

1099 ***Manipulating gene expression using siRNA and lentivirus***

1100 For transient knockdown of target genes, siRNAs specific against RPL14 or NKRF were
1101 transfected into cells using Lipofectamine RNAiMAX (Invitrogen, USA) according to the
1102 manufacturer's protocol. After 48-72 hours, cells were harvested to detect the knockdown
1103 efficiency of target genes through RT-qPCR as described below. Full length cDNA of NKRF
1104 was cloned into the pcDNA3.1 vectors. RPL14, SELE-WT and MUT were cloned into the
1105 pCDH-puro lentiviral vectors. For lentivirus preparation, 293T cells were transfected with
1106 above lentivirus plasmids or control vector using Lipofectamine 2000 (Invitrogen, USA)
1107 according to the manufacture's procedures. Afterwards, the supernatant media containing
1108 lentivirus were collected two times at 24 and 48h post-transfection to infect NPC cell lines,
1109 following selection with puromycin (2 μg/mL) at least for one week.

1110

1111 ***RNA isolation, RT-qPCR and ChIP PCR***

1112 Total RNA was isolated using Trizol reagent (Invitrogen, USA) following the manufacturer's
1113 protocols. After determination of RNA concentration using a Nanodrop Spectrophotometer
1114 (Thermo Fisher Scientific, USA), cDNA was synthesized using oligo (dT) primers and M-
1115 MLV Reverse Transcriptase (Promega, USA) according to the manufacturer's procedures. RT-
1116 qPCR was performed with SYBR Premix Ex Taq kit (Takara, Japan) on a LightCycler 96 (Bio-
1117 rad, USA). GAPDH or 18S RNA was used as an internal control. The primer sequences are
1118 shown in **Table S18**. For ChIP assay, S26 cells transfected with Flag-NKRF plasmids or
1119 control vectors were grown in complete DMEM media to 80%–90% confluency. The media
1120 were removed and replaced with media containing 1% formaldehyde and crosslinked for
1121 10 min at 37°C, followed by SimpleChIP® Plus Enzymatic Chromatin IP Kit (Magnetic Beads;
1122 CST, USA) according to the manufacturer's instructions. Afterwards, PCR or qPCR was
1123 performed with specific detection primers following the standard procedures (**Table S18**).

1124

1125 ***EBV preparation and infection***

1126 AKATA cells carrying EBV recombinant viruses were treated with 0.75 (v/v) of goat anti-
1127 human immunoglobulin G IgG (H0111-6, Tianfun Xinqu Zhenglong Biochem, China) for 6h
1128 at 37℃ to induce the viral productive cycle. Virus supernatants were harvested three days post-
1129 induction, filtered through a 0.45 μm filter and concentrated at 20000g for 3h. Afterwards,
1130 virus concentrate was added into NPC cells and concentrated at 2000g in 37℃ for 1h. Cell
1131 supernatants were removed and replaced with normal media 6-8h post-infection. The infection
1132 efficiency was measured using flow cytometry analysis of GFP-positive cells.

1133

1134 ***Western blotting***

1135 NPC cells were lysed in ice-cold cell lysis buffer (CST, USA) containing 1X protease inhibitors
1136 (Beyotime, China). Lysate was cleared by centrifugation at 14,000g for 10 min at 4°C and
1137 boiled in 1X LDS loading buffer at 100°C for 10 min. Proteins were separated by SDS-PAGE
1138 and transferred to 0.2 μm polyvinylidene difluoride (PVDF) membrane (Merck Millipore,
1139 USA). Membranes were blocked with 5% bovine serum albumin (BSA; Sangon Biotech, China)
1140 in tris-buffered saline and Tween 20 (TBST) at room temperature for 1h. Afterwards, the
1141 membranes were incubated with target-specific antibodies in blocking buffer at 4°C overnight.
1142 Next day, horseradish peroxidase–conjugated (HRP) secondary antibodies were used at room
1143 temperature for 1h. Proteins were then detected with Fdbio-Dura ECL kit (Fdbio science, China)
1144 and quantified by Bio-Rad ChemiDoc Touch (Hercules, USA). Primary antibodies were
1145 commercially available, including RPL14 and LMP1 (ab181200 and ab78113, Abcam, USA),
1146 SELE (20894-1-AP, Proteintech, China), Flag (F1804, Sigma-Aldrich, USA), BZLF1(sc53904,
1147 Santa Cruz, USA) and ACTIN (AC004, Abclonal, China).

1148

1149 ***In vivo tumor xenograft***

1150 Six-week-old male BALB/c nude mice (Beijing Vital River Laboratory Animal Technology,
1151 China) were grown in the Specific Pathogen Free (SPF) environment. For subcutaneous
1152 xenograft assay, $5\times10^6$ CNE2-EBV cells or $2\times10^6$ S26 cells co-culturing with HUVEC cells
1153 (S26: HUVEC = 10:1) were suspended in 100μL ice-old PBS mixed with Matrigel (0.20 v/v,
1154 Corning Incorporated, USA) and subcutaneously injected to the two flanks of the mice.
1155 Macroscopic observation and tumor volume measurement using a caliper were performed
1156 every 3 days. After 15 days, all mice were sacrificed, and tumor tissues were carefully dissected
1157 and weighed. Tumor volume was calculated following the formula: tumor volume (mm3) =

1158 length (mm) × (width (mm)) 2 /2. For tumor lymphatic metastasis assay, $1\times10^6$ S26 cells co-
1159 culturing with HUVEC cells (S26: HUVEC = 10:1) were injected into the foot pads of the node
1160 mice. After one month, the lymph nodes were harvested, volumed and paraffin embedded
1161 further HE staining. For animal studies *in vivo*, all experiments were performed in strict
1162 accordance with the instructions approved by the Institutional Animal Care and Use Committee
1163 of Sun Yat-sen University.

1164

1165 ***Tube formation assays***

1166 For tube formation assays, HUVEC cells stably expressing SELE wild-type (SELE-WT), the
1167 S149R mutant or control vectors were resuspended in the ECM medium. Then the cells were
1168 seeded into 96-well plates pre-coated with growth-factor-reduced Matrigel (Corning
1169 Incorporated, USA) and incubated at 37°C for 5 hours. The formation of capillary-like
1170 structures was subsequently observed using a microscope. Quantitative analysis was conducted
1171 by counting the number of tube networks formed across the entire field, serving as an indicator
1172 of *in vitro* angiogenesis capability.

1173

1174 **References**


1175 1 Chen, Y.-P. *et al.* Nasopharyngeal carcinoma. *The Lancet* **394**, 64-80,
1176 doi:10.1016/s0140-6736(19)30956-0 (2019).
1177 2 Wei, K. R. *et al.* Nasopharyngeal carcinoma incidence and mortality in China, 2013.
1178 *Chin J Cancer* **36**, 90, doi:10.1186/s40880-017-0257-9 (2017).
1179 3 Cao, S. M., Simons, M. J. & Qian, C. N. The prevalence and prevention of
1180 nasopharyngeal carcinoma in China. *Chin J Cancer* **30**, 114-119,
1181 doi:10.5732/cjc.010.10377 (2011).
1182 4 Chen, C. *et al.* Multiple risk factors of nasopharyngeal carcinoma: Epstein-Barr virus,
1183 malarial infection, cigarette smoking and familial tendency. **10**, 547-553 (1990).
1184 5 IRELAND, B., LANIER, A. P., KNUTSON, L., CLIFT, S. E. & HARPSTER, A. J. I.
1185 j. o. e. Increased risk of cancer in siblings of Alaskan native patients with
1186 nasopharyngeal carcinoma. **17**, 509-511 (1988).
1187 6 Loh, K. S. *et al.* Familial nasopharyngeal carcinoma in a cohort of 200 patients. **132**,
1188 82-85 (2006).
1189 7 Liu, Z. *et al.* Quantification of familial risk of nasopharyngeal carcinoma in a high-
1190 incidence area. **123**, 2716-2725 (2017).
1191 8 Bei, J. X. *et al.* A genome-wide association study of nasopharyngeal carcinoma
1192 identifies three new susceptibility loci. *Nature genetics* **42**, 599-603,
1193 doi:10.1038/ng.601 (2010).
1194 9 Lu, S.-j. *et al.* Linkage of a nasopharyngeal carcinoma susceptibility locus to the HLA
1195 regions. *Nature* **346**, 470-471 (1990).
1196 10 He, Y. Q. *et al.* A polygenic risk score for nasopharyngeal carcinoma shows potential
1197 for risk stratification and personalized screening. *Nature communications* **13**, 1966,
1198 doi:10.1038/s41467-022-29570-4 (2022).

1199 11 Ning, L. *et al.* Nasopharyngeal carcinoma MHC region deep sequencing identifies
1200 HLA and novel non-HLA TRIM31 and TRIM 39 loci. **3**, 759 (2020).
1201 12 Xiong, W. *et al.* A susceptibility locus at chromosome 3p21 linked to familial
1202 nasopharyngeal carcinoma. **64**, 1972-1974 (2004).
1203 13 Hu, L.-F. *et al.* A genome-wide scan suggests a susceptibility locus on 5p13 for
1204 nasopharyngeal carcinoma. **16**, 343-349 (2008).
1205 14 Fachiroh, J. *et al.* Tobacco consumption and genetic susceptibility to nasopharyngeal
1206 carcinoma (NPC) in Thailand. *Cancer Causes & Control* **23**, 1995-2002,
1207 doi:10.1007/s10552-012-0077-9 (2012).
1208 15 Ng, C. C. *et al.* A genome-wide association study identifies ITGA9 conferring risk of
1209 nasopharyngeal carcinoma. *Journal of human genetics* **54**, 392-397,
1210 doi:10.1038/jhg.2009.49 (2009).
1211 16 Bei, J.-X. *et al.* A GWAS meta-analysis and replication study identifies a novel locus
1212 within CLPTM1L/TERT associated with nasopharyngeal carcinoma in individuals of
1213 Chinese ancestry. **25**, 188-192 (2016).
1214 17 Cui, Q. *et al.* An extended genome-wide association study identifies novel
1215 susceptibility loci for nasopharyngeal carcinoma. **25**, 3626-3634 (2016).
1216 18 Wang, T. M. *et al.* Fine-mapping of HLA class I and class II genes identified two
1217 independent novel variants associated with nasopharyngeal carcinoma susceptibility.
1218 **7**, 6308-6316 (2018).
1219 19 Dai, J. *et al.* Estimation of heritability for nine common cancers using data from
1220 genome-wide association studies in Chinese population. **140**, 329-336 (2017).
1221 20 Huang, S. F. *et al.* Familial aggregation of nasopharyngeal carcinoma in Taiwan. *Oral*
1222 *Oncol* **73**, 10-15, doi:10.1016/j.oraloncology.2017.07.020 (2017).
1223 21 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature*
1224 **461**, 747-753, doi:10.1038/nature08494 (2009).
1225 22 Genin, E. Missing heritability of complex diseases: case solved? *Human genetics* **139**,
1226 103-113, doi:10.1007/s00439-019-02034-4 (2020).
1227 23 Dai, W. *et al.* Whole-exome sequencing identifies MST1R as a genetic susceptibility
1228 gene in nasopharyngeal carcinoma. *Proceedings of the National Academy of Sciences*
1229 **113**, 3317-3322, doi:10.1073/pnas.1523436113 (2016).
1230 24 Yu, G. *et al.* Whole-Exome Sequencing of Nasopharyngeal Carcinoma Families
1231 Reveals Novel Variants Potentially Involved in Nasopharyngeal Carcinoma. *Scientific*
1232 *reports* **9**, doi:10.1038/s41598-019-46137-4 (2019).
1233 25 Wang, T. M. *et al.* Whole-Exome Sequencing Study of Familial Nasopharyngeal
1234 Carcinoma and Its Implication for Identifying High-Risk Individuals. *J Natl Cancer*
1235 *Inst* **114**, 1689-1697, doi:10.1093/jnci/djac177 (2022).
1236 26 Xiao, R. W. *et al.* Rare POLN mutations confer risk for familial nasopharyngeal
1237 carcinoma through weakened Epstein-Barr virus lytic replication. *EBioMedicine* **84**,
1238 104267, doi:10.1016/j.ebiom.2022.104267 (2022).
1239 27 Chan, K. A. *et al.* Analysis of plasma Epstein–Barr virus DNA to screen for
1240 nasopharyngeal cancer. **377**, 513-522 (2017).
1241 28 Li, T. *et al.* Anti–Epstein–Barr Virus BNLF2b for Mass Screening for
1242 Nasopharyngeal Cancer. **389**, 808-819 (2023).
1243 29 Winter, J. R. *et al.* Predictors of Epstein-Barr virus serostatus in young people in
1244 England. **19**, 1-9 (2019).
1245 30 Hatton, O. L., Harris-Arnold, A., Schaffert, S., Krams, S. M. & Martinez, O. M. J. I. r.
1246 The interplay between Epstein–Barr virus and B lymphocytes: implications for
1247 infection, immunity, and disease. **58**, 268-276 (2014).

31 Su, Z. Y., Siak, P. Y., Leong, C.-O. & Cheah, S.-C. J. F. i. m. The role of Epstein–Barr virus in nasopharyngeal carcinoma. **14**, 1116143 (2023).

32 Liu, Y. *et al.* Tumour heterogeneity and intercellular networks of nasopharyngeal carcinoma at single cell resolution. *Nature communications* **12**, 741, doi:10.1038/s41467-021-21043-4 (2021).

33 Chen, Y.-P. *et al.* Single-cell transcriptomics reveals regulators underlying immune cell diversity and immune subtypes associated with prognosis in nasopharyngeal carcinoma. **30**, 1024-1042 (2020).

34 Altshuler, D. M. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

35 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).

36 Trinh, C. T. H. *et al.* LMP1-EBV Gene Deletion Mutations and HLA Genotypes of Nasopharyngeal Cancer Patients in Vietnam. *Pathophysiology* **30**, 1-12, doi:10.3390/pathophysiology30010001 (2023).

37 Tabula Sapiens, C. *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896, doi:10.1126/science.abl4896 (2022).

38 Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. **583**, 699-710 (2020).

39 Tsao, S. W., Tsang, C. M. & Lo, K. W. Epstein-Barr virus infection and nasopharyngeal carcinoma. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **372**, doi:10.1098/rstb.2016.0270 (2017).

40 Yasmin-Karim, S., King, M. R., Messing, E. M. & Lee, Y.-F. J. O. E-selectin ligand-1 controls circulating prostate cancer cell rolling/adhesion and metastasis. **5**, 12097 (2014).

41 Yoshida, M. *et al.* E-selectin polymorphism associated with myocardial infarction causes enhanced leukocyte-endothelial interactions under flow conditions. **23**, 783-788 (2003).

42 Dornbos, P. *et al.* A combined polygenic score of 21,293 rare and 22 common variants improves diabetes diagnosis based on hemoglobin A1C levels. *Nature genetics* **54**, 1609-1614, doi:10.1038/s41588-022-01200-1 (2022).

43 Van Den Bogaert, A. *et al.* The DTNBP1 (dysbindin) gene contributes to schizophrenia, depending on family history of the disease. **73**, 1438-1443 (2003).

44 Coccia, M., Rossi, A., Riccio, A., Trotta, E. & Santoro, M. G. J. P. o. t. N. A. o. S. Human NF-κB repressing factor acts as a stress-regulated switch for ribosomal RNA processing and nucleolar homeostasis surveillance. **114**, 1045-1050 (2017).

45 Yi, M. *et al.* Rediscovery of NF-kappaB signaling in nasopharyngeal carcinoma: How genetic defects of NF-kappaB pathway interplay with EBV in driving oncogenesis? *J Cell Physiol* **233**, 5537-5549, doi:10.1002/jcp.26410 (2018).

46 Elia, A., Vyas, J., Laing, K. G. & Clemens, M. J. J. E. j. o. b. Ribosomal protein L22 inhibits regulation of cellular activities by the Epstein-Barr virus small RNA EBER-1. **271**, 1895-1905 (2004).

47 Shen, C.-L. *et al.* Ribosome protein L4 is essential for Epstein–Barr virus nuclear antigen 1 function. **113**, 2229-2234 (2016).

48 Klinge, S., Voigts-Hoffmann, F., Leibundgut, M., Arpagaus, S. & Ban, N. J. S. Crystal structure of the eukaryotic 60 S ribosomal subunit in complex with initiation factor 6. **334**, 941-948 (2011).

1296 49 Zhang, Z., Zhang, Y., Qiu, Y., Mo, W. & Yang, Z. J. B. Human/eukaryotic ribosomal
1297 protein L14 (RPL14/eL14) overexpression represses proliferation, migration, invasion
1298 and EMT process in nasopharyngeal carcinoma. **12**, 2175-2186 (2021).
1299 50 Huang, X.-P. *et al.* Alteration of RPL14 in squamous cell carcinomas and
1300 preneoplastic lesions of the esophagus. **366**, 161-168 (2006).
1301 51 Telen, M. J. J. H. O. C. Cellular adhesion and the endothelium: E-selectin, L-selectin,
1302 and pan-selectin inhibitors. **28**, 341-354 (2014).
1303 52 Köhler, S., Ullrich, S., Richter, U. & Schumacher, U. J. B. j. o. c. E-/P-selectins and
1304 colon carcinoma metastasis: first in vivo evidence for their crucial role in a clinically
1305 relevant model of spontaneous metastasis formation in the lung. **102**, 602-609 (2010).
1306 53 Zen, K. *et al.* CD44v4 is a major E-selectin ligand that mediates breast cancer cell
1307 transendothelial migration. **3**, e1826 (2008).
1308 54 Esposito, M. *et al.* Bone vascular niche E-selectin induces mesenchymal–epithelial
1309 transition and Wnt activation in cancer cells to promote bone metastasis. **21**, 627-639
1310 (2019).
1311 55 Zhang, B. *et al.* NOTCH4ΔL12_16 sensitizes lung adenocarcinomas to EGFR-TKIs
1312 through transcriptional down-regulation of HES1. **14**, 3183 (2023).
1313 56 Zhou, J. *et al.* Long noncoding RNA MALAT1 sponging miR-26a-5p to modulate
1314 Smad1 contributes to colorectal cancer progression by regulating autophagy.
1315 *Carcinogenesis* **42**, 1370-1379, doi:10.1093/carcin/bgab069 (2021).
1316 57 Sato, G. *et al.* Pan-cancer and cross-population genome-wide association studies
1317 dissect shared genetic backgrounds underlying carcinogenesis. **14**, 3671 (2023).
1318 58 Ma, Y. *et al.* Integrating single-cell sequencing data with GWAS summary statistics
1319 reveals CD16+ monocytes and memory CD8+ T cells involved in severe COVID-19.
1320 **14**, 16 (2022).
1321 59 Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia.
1322 **50**, 825-833 (2018).
1323 60 Witte, J. S., Visscher, P. M. & Wray, N. R. The contribution of genetic variants to
1324 disease depends on the ruler. *Nature reviews. Genetics*, doi:10.1038/nrg3786 (2014).
1325 61 Li, Q. & Cohen, J. I. J. C. c. m. r. Epstein-Barr virus and the human leukocyte antigen
1326 complex. **6**, 175-181 (2019).
1327 62 Su, C. *et al.* EBNA2 driven enhancer switching at the CIITA-DEXI locus suppresses
1328 HLA class II gene expression during EBV infection of B-lymphocytes. **17**, e1009834
1329 (2021).
1330 63 Li, Z. *et al.* Association of rare CYP39A1 variants with exfoliation syndrome
1331 involving the anterior chamber of the eye. **325**, 753-764 (2021).
1332 64 Zhang, L. *et al.* Genomic Analysis of Nasopharyngeal Carcinoma Reveals TME-
1333 Based SubtypesGenomic Subtypes of Nasopharyngeal Carcinoma. **15**, 1722-1732
1334 (2017).
1335 65 Jin, S. *et al.* Single-cell transcriptomic analysis defines the interplay between tumor
1336 cells, viral infection, and the microenvironment in nasopharyngeal carcinoma. **30**,
1337 950-965 (2020).
1338 66 Zhao, J. *et al.* Single cell RNA-seq reveals the landscape of tumor and infiltrating
1339 immune cells in nasopharyngeal carcinoma. **477**, 131-143 (2020).
1340 67 Gong, L. *et al.* Comprehensive single-cell sequencing reveals the stromal dynamics
1341 and tumor-specific characteristics in the microenvironment of nasopharyngeal
1342 carcinoma. **12**, 1-18 (2021).
1343 68 Aldana, R. & Freed, D. Data Processing and Germline Variant Calling with the
1344 Sentieon Pipeline. *Methods in molecular biology* **2493**, 1-19, doi:10.1007/978-1-
1345 0716-2293-3_1 (2022).

69 Kendig, K. I. *et al.* Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. **10** (2019).
70 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. **43**, 491 (2011).
71 Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575, doi:Doi 10.1086/519795 (2007).
72 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76-82, doi:10.1016/j.ajhg.2010.11.011 (2011).
73 Ka, S. *et al.* HLAscan: genotyping of the HLA region using next-generation sequencing data. **18**, 1-11 (2017).
74 Robinson, J. *et al.* Ipd-imgt/hla database. **48**, D948-D955 (2020).
75 Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. **8**, e64683 (2013).
76 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. J. G. b. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **10**, R25 (2009).
77 Anders, S., Pyl, P. T. & Huber, W. J. B. HTSeq—a Python framework to work with high-throughput sequencing data. **31**, 166-169 (2015).
78 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, doi:10.1093/nar/gkq603 (2010).
79 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
80 Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761-763, doi:10.1093/bioinformatics/btu703 (2015).
81 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research* **47**, D886-D894, doi:10.1093/nar/gky1016 (2019).
82 Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature genetics* **47**, 276-283, doi:10.1038/ng.3196 (2015).
83 Li, Q. & Wang, K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *American journal of human genetics* **100**, 267-280, doi:10.1016/j.ajhg.2017.01.004 (2017).
84 Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. **89**, 82-93 (2011).
85 Li, B. & Leal, S. M. J. T. A. J. o. H. G. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. **83**, 311-321 (2008).
86 Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. J. T. A. J. o. H. G. Sequence kernel association tests for the combined effect of rare and common variants. **92**, 841-853 (2013).
87 Viechtbauer, W. J. J. o. s. s. Conducting meta-analyses in R with the metafor package. **36**, 1-48 (2010).
88 Becker, B. J. J. T. h. o. r. s. Combining significance levels. 215-230 (1994).
89 M, D. metap: meta-analysis of significance values. *R package version 1.8* (2022).

90 Preston, R. C. *et al.* E-selectin ligand complexes adopt an extended high-affinity conformation. *J Mol Cell Biol* **8**, 62-72, doi:10.1093/jmcb/mjv046 (2016).
91 Qian, J. *et al.* A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. **30**, 745-762 (2020).
92 Zhang, M. *et al.* Dissecting transcriptional heterogeneity in primary gastric adenocarcinoma by single cell RNA sequencing. **70**, 464-475 (2021).
93 McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. J. C. s. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. **8**, 329-337. e324 (2019).
94 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. **184**, 3573-3587. e3529 (2021).
95 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. **16**, 1289-1296 (2019).
96 Wolock, S. L., Lopez, R. & Klein, A. M. J. C. s. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. **8**, 281-291. e289 (2019).
97 Young, M. D. & Behjati, S. J. G. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. **9**, giaa151 (2020).
98 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529, doi:10.1016/j.cell.2021.04.048 (2021).
99 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. **344**, 1396-1401 (2014).
100 Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics* **47**, 1114-1120, doi:10.1038/ng.3390 (2015).
101 Bei, J.-X. *et al.* A GWAS Meta-analysis and Replication Study Identifies a Novel Locus within CLPTM1L/TERT Associated with Nasopharyngeal Carcinoma in Individuals of Chinese AncestryNPC GWAS Follow-up Identifies CLPTM1L/TERT Locus. **25**, 188-192 (2016).
102 Delaneau, O., Marchini, J. & Zagury, J.-F. J. N. m. A linear complexity phasing method for thousands of genomes. **9**, 179-181 (2012).
103 Howie, B. N., Donnelly, P. & Marchini, J. J. P. g. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. **5**, e1000529 (2009).
104 Luo, C. L. *et al.* RBFOX2/GOLIM4 Splicing Axis Activates Vesicular Transport Pathway to Promote Nasopharyngeal Carcinogenesis. *Adv Sci (Weinh)*, e2004852, doi:10.1002/advs.202004852 (2021).
105 Xu, X. C. *et al.* RNA-binding motif protein RBM47 promotes tumorigenesis in nasopharyngeal carcinoma through multiple pathways. *J Genet Genomics* **48**, 595-605, doi:10.1016/j.jgg.2021.05.006 (2021).

1444 **Figure and table legends**

1445 **Figure 1. Study overview.** A two-stage association study design was applied to investigate the
1446 genetic factors associated with NPC. In the discovery stage, a total of 5,022 samples, including
1447 the GD-SYSUCC cohort from Guangdong in China and the SG cohort from Singapore, were
1448 genotyped with WES and analyzed to identify four independent variants (two of which are
1449 novel), four novel genes, and 59 pathways associated with NPC. The associations were
1450 subsequently validated in the replication stage, which included 9,047 samples from the GD-ZS
1451 cohort in Guangdong and the HK cohort in Hong Kong, China. Bioinformatic analyses and
1452 functional experiments were conducted to explore the biological functions and the clinical
1453 implications of the novel loci.
1454

1455 **Figure 2. Significant pathways and sentinel genes associated with NPC. A**. Significant
1456 pathways associated with NPC. ALLNPC denotes all NPC samples in the discovery stage,
1457 whereas FHNPC represents NPC cases with a familial NPC history. Color bars from white to
1458 purple indicate minus log-transformed P values. **B**. Sentinel genes for NPC-associated
1459 pathways. Genes indicated atop are highlighted in orange if they are part of a specific pathway.
1460 The listed genes are those belonging to at least two NPC-associated pathways and exhibiting a
1461 gene-based association P value below 0.01. **C**. Locations of the rare variants associated with
1462 NPC at significance level of $P < 1\times10^{-4}$ within the genic regions of SELE. Grey arrows denote
1463 non-coding variants, blue arrows represent synonymous coding variants, and the red arrow
1464 indicates non-synonymous coding variant predicted as "deleterious". The rs5361 minor allele
1465 introduces a missense mutation at position 149, resulting in an amino acid substitution from S
1466 to R. The index SNV rs3917410 showed the most significant P value in the association tests in
1467 the discovery stage.
1468

1469 **Figure 3. Expression patterns of novel and known NPC-associated genes across diverse**
1470 **cell types in NPC tumor tissues.** Single-cell transcriptomic analyses of 223,593 cells derived
1471 from NPC tumor tissues in 56 patients. **A**. UMAP plot of 223,593 single cells grouped into
1472 seven major cell clusters. **B**. Violin plot illustrating normalized expression of NPC-associated
1473 genes across these major cell clusters. All epithelial cells captured in NPC tumor were
1474 malignant (see Methods). **C**. The expression of the marker gene (*VWF*) for endothelial cells
1475 and alongside the novel NPC-associated gene *SELE*; top panel: initial UMAP plot, bottom

1476 panel: re-normalized UMAP emphasizing cells highlighted by the red circles in the initial
1477 UMAP plot.
1478

1479 **Figure 4. rs2276868 regulates the expression of RPL14**. **A**. Single-cell transcriptome
1480 analysis showed the mRNA expression of RPL14 in 15,623 malignant epithelial cells from 35
1481 NPC samples with different rs2276868 genotypes. **B**. Relative luciferase activity changes in
1482 293T cells transfected with plasmids containing rs2276868 -[C], -[T], or control vectors. **C**.
1483 Pearson correlation analysis indicates the relationship between *RPL14* and *NKRF* expression
1484 (measured as transcript per million, TPM) in transcriptome data of NPC patients (n=87). **D**.
1485 RT-qPCR illustrates the mRNA expression of RPL14 in NPC cells transfected with NKRF
1486 siRNAs or control siRNA. **E-F**. Relative luciferase activity in 293T cells co-transfected with
1487 the rs2276868 -[C] or -[T] plasmids and NKRF siRNA (F) or NKRF overexpression vectors
1488 (G). Corresponding statistics are presented at the right. **G**. Schematic diagram indicates primer
1489 pairs used for PCR amplification of RPL14 fragments. **H-I**. ChIP assay in S26 cells transfected
1490 with Flag-NKRF and control vectors. ChIP PCR (H) and qPCR (I) showed RPL14 expression
1491 level in cells. P1-3 denote primer pairs targeting genomic regions shown in G. **J**. Pearson
1492 correlation analysis indicates the relationship between NKRF and RPL14 expression in single-
1493 cell transcriptome data are different for NPC patients with different genotypes. rs2276868 -
1494 [CC/CT] patients have a stronger correlation than rs2276868-[TT] patients. *$p < 0.05$, **$p <$
1495 $0.01$, ***$p < 0.001$.
1496

1497 **Figure 5. *RPL14* inhibits EBV infection and lytic cycle activation in NPC cells. A.** The top
1498 20 pathways significantly associated with RPL14 expression in malignant epithelial cells from
1499 NPC tumor. **B.** Correlation analysis between EBV-activity scores and *RPL14* expression within
1500 malignant epithelial cells. **C.** Single-cell transcriptome analysis showing *LMP1* and *LMP2*
1501 expression in NPC samples with either rs2276868 -[C] or -[T] genotypes. **D.** Western blot
1502 examination of RPL14 and LMP1 protein levels in NPC cell lines S26 and HK-1 cells infected
1503 with EBV at different timepoints. **E**. Western blot assessment of the knockdown efficiency of
1504 RPL14 siRNAs or control siRNA in S26 and HK-1 cells. **F.** Flow cytometry quantification of
1505 GFP intensity for the EBV infection efficiency in the NPC cells described in E. **G**. Western
1506 blot assay showing RPL14 protein expression in S26 and HK-1 cells infected with lentivirus
1507 stably expressing RPL14. ACTIN served as a loading control. **H**. Flow cytometry assessment
1508 of EBV infection efficiency in the cells described in G. **I-J**. Changes of gene expression in

1509 CNE2-EBV cells subjected to EBV lytic induction with TPA and NAB. RT-qPCR (I) and
1510 western blotting (J) were performed to detect the mRNA and protein expression, respectively,
1511 of RPL14 and EBV genes at different timepoints of EBV lytic induction. **K-L**. RT-qPCR
1512 analysis of EBV lytic gene expression in CNE2-EBV and C666-1 cells transfected with RPL14
1513 siRNAs or control siRNA (K), or infected with lentivirus stably expressing RPL14 vector or
1514 control vector (L). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ***$p < 0.0001$.
1515

1516 **Figure 6. Tumor suppressive function of *RPL14* in NPC. A**. Western blot analysis showing
1517 RPL14 protein levels in NPC cells infected with RPL14 overexpressing lentivirus. ACTIN
1518 serves as the loading control. **B**. Cell growth curves of the cells described in A. **C**. Colony
1519 formation assay for the cells described in A. Corresponding statistical analysis is shown to the
1520 right. **D**. Transwell migration assay evaluating the migration ability of the cells described in A.
1521 The statistical analysis is presented on the right. **E-H**. Tumor growth evaluation in a nude mice
1522 model with subcutaneous injection of NPC cells described in A. Tumor volumes were
1523 measured every three days. Visual presentation of tumor post-sacrifice (F) and weight (G) were
1524 presented. IF detection of Ki-67 expression (H) in the tumors described in F, and the
1525 corresponding statistical analysis is shown on the right. **I**. Transcriptomic analysis showcasing
1526 mRNA levels of RPL14 in NPC (n=87) versus control samples (n=10). **J**. Kaplan-Meier
1527 survival analysis linking RPL14 expression to survival time in NPC patients using bulk RNA-
1528 seq data (n=88). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

1529

1530 **Figure 7. *SELE*-S149R mutation in endothelial cells promotes the tumorigenesis and**
1531 **metastasis of NPC cells. A.** Structure prediction of E-selection paired with glycomimetic
1532 antagonist ligand 2-acetamido-2-deoxy-beta-D-glucopyranose (PDB code 4C16). The spatial
1533 proximity between Ser149 and the ligand is highlighted. Both the Ser149 side chain and the
1534 ligand are shown as sticks, with the distances between the OG atom of Ser149 and the C8 or
1535 O7 atoms of the ligand specified. **B.** Western blot examination of SELE protein level in
1536 HUVEC cells infected with lentivirus overexpressing SELE-WT, S149R mutant or control
1537 vectors. **C.** Transwell migration assay evaluating the migration ability of the cells described in
1538 A, with statistical analysis presented at the bottom. Scale bar, 100 μm. **D.** Tube formation
1539 assay with cells described in A. The statistical analysis is presented on the right. Scale bar, 100
1540 μm. **E.** Transwell migration assay assessing the migration ability of the NPC cell lines (S26
1541 and HK-1) co-cultured with HUVEC cells from A (S26/HK-1: HUVEC = 10:1), accompanied

1542 by statistical analysis presented at the bottom. Scale bar, 200 μm. **F-H.** Tumor growth
1543 evaluation in xenograft model with subcutaneous injection of S26 cells described in E. Tumor
1544 volumes. (F) are measured and visual presentation of tumor post-sacrifice (G) and weight (H)
1545 are presented. **I-K.** Tumor lymphatic metastasis of S26 cells described in E. Lymph nodes are
1546 visualized (I) and measured (J), with HE staining conducted to assess metastasis in these lymph
1547 nodes (K). Scale bar, 100 μm. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
1548

1549 **Figure 8. Contribution of common and rare variants to NPC susceptibility. A**. Fractional
1550 representation of NPC heritability attributable to non-HLA WES-SNVs, categorized by minor
1551 allele frequency (MAF) and linkage disequilibrium (LD). For each variant, MAF was
1552 calculated using the discovery samples, and LD score represented the aggregated $R^2$ with
1553 adjacent variants spanning a 200kb window. **B**. Density plots illustrating the PRS incorporating
1554 the novel loci and previously known GWAS loci (refined composite PRS, rcPRS) for cases
1555 and controls in both the discovery and replication cohorts. **C**. Receiver operating characteristic
1556 curves comparing the rcPRS and previously reported GWAS PRS (gPRS) for NPC across
1557 different cohorts. **D**. Relative odds ratio comparing the rcPRS or gPRS bins and the 5% lowest
1558 quantile group in different cohorts. **E**. Correlation of rcPRS with disease status in individuals
1559 categorized by their EBV VCA-IgA status.

1560

1561 **Table 1. Association of single variants with NPC susceptibility.** Notes: CHR, chromosome;
1562 A1, effective allele/amino acid; OR (95%CI), odds ratio estimate and the 95% confidence
1563 interval; $I^2$, the proportion of the variance in observed effect is due to variance in true effects
1564 rather than sampling error; Meta, meta-analyzed using the random effects model.

1565

1566 **Table 2. Associations of individual genes with NPC susceptibility.** Notes: ALLSNV, all
1567 SNVs; SKAT (C+R), SNP-set (Sequence) Kernel Association Test for the combined effect of
1568 common (C) and rare (R) variants; Burden, burden test; Meta, meta-analyzed using the
1569 summation of logits method; Adjusted P, the P value adjusted using the FDR method.
1570