nature portfolio

Peer Review File

EBV strain interacts with host HLA to drive nasopharyngeal carcinoma risk

Corresponding Author: Professor Miao Xu

Parts of this Peer Review File have been redacted as indicated to maintain the confidentiality of unpublished data.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Chen and colleagues report on a genome-to-genome analysis of Epstein-Barr virus (EBV) and human genetics, specifically using an environmental/interaction effect framework to test for individual and joint effects on nasopharyngeal carcinoma (NPC). They recapitulate known strong effects of specific HLA types and high-risk EBV genotypes on NPC risk, but most interestingly find an interaction effect between a specific EBV variant and HLA type which increase risk further. The authors identify a 'dual-risk' subgroup of the population with this interaction, which is particularly prevalent in Southern China, and could explain the higher than expected burden of disease in this region. The authors propose and validate a mechanism of modified binding efficacy between the modified peptide and HLA variant. The large effect size, novel interaction, and potentially large population effect (i.e. expected cases due to this effect) make this a potentially important finding.

Elements that elevate this study beyond some other genome-to-genome efforts are: the large effect size of the tested effects on cancer, and the interaction effect specifically; replication efforts in other populations; a well-powered sample size; lab validation of binding effects.

Some parts of the paper remained unclear to me:

1) The description of odds-ratios is not always clear, and how the newly detected interaction effect reported here expands upon previous knowledge of increased risk from HLA and specific viral lineages. Specifically:
- A clear explanation of the interaction effect specifically in the SC population would help, perhaps added to figure 3. What is the rate per million in the population without any risk factor, what is the rate per millionwith: one of the HLA variants, with high risk EBV, and with the interaction effect? [or instead of per million as the denominator, the Southern China population size could be used, as the authors did to estimate 35k preventable cases]. My reading of these from the current text is that the dual-risk group has an increase in risk of around double (OR =~ 10 for HLA variant, then OR =~ 18 with the viral variant too). It might be a bit misleading to only quote an OR of 18 in the abstract, as this doesn't refer to the interaction effect alone. Additional context of OR for the HLA and high-risk EBV effects could be included. Clearly noting the low population prevalence of NPC upfront would help too (which is done, just later in the text). Explaining the interpretation of the PAF in the main text could also help.
- Fig 1d&c: It's not right to start the plot for these ORs at 0. ORs are not additive either. This plot should probably be logged. Overall these panels are quite unclear without the text, and I wonder if another presentation of interaction effects might be considered.
- I wanted to double-check that ascertainment (i.e. over-sampling of NPC cases in the collection) has been correctly reported on. I do see the authors mention in the methods that this is a rare disease, but could they comment on if/how this oversampling of NPC cases in the collection affects the interpretation of the reported OR in the interaction term specifically?

2) The interaction test does not seem to control for viral population structure sufficiently. The authors use 4 MDS components, but this is not going to be able to control for the level of clonality seen in figure 6 (I didn't see this specifically

tested/shown in a QQ plot or similar?).

While in figure 6 it looks like the interaction SNP does not completely correlate with high-risk EBV, it is definitely still associated with specific lineages. Typically viral GWAS uses a linear mixed model to control for the very strong population structure. Also, permutation tests tend to underestimate the significance threshold as samples are not iid due to their relatedness.

Can the authors convince us that the interaction effect is likely the reported SNP, and that this SNP is not just a marker for a specific lineage/lineages?

3) The two step process is not fully statistically justified. The authors dismiss a full genome-to-genome scan as underpowered with the sample size, and only check viral variants associated with HLA-A*11:01. The p-value they find is not very strong. Overall I do not think this is a disaster, because the strength of the effect appears large, it is found in two independent populations, this does look like a plausible GWAS peak (no strong population structure effects in figure 2), and there appears to be good functional evidence to support the result from the binding assay. But still, reporting the full genome-to-genome results against all human variants (as is typical for such studies) would be useful.

4) How, mechanistically, this interaction leads to cancer risk is not totally clear to me, especially as someone without a background in virology or oncology. Likely this just needs more explanation in the text. For example, by what mechanism is EBV thought to cause NPC? Is it because of increased mutagenesis? Or causing selection on somatic variants? Is there anything that can be said about the reduced binding efficacy to EBV enhancing this effect? (i.e. is an increased viral load known to increase cancer risk, and how/why?)

Minor comments:

- There are actually only ~750 samples with virus sequence, perhaps reporting this as well as the headline number of 3500 would be helpful to understand the study design.

- I am not an expert here, but the follow-up in vivo for binding efficiency seemed convincing. I did wonder whether screening interaction strength with AlphaFold3 model confidence (or a similar open source tool) would help further increase confidence in these findings?

- Where was statistical power reported? This is mentioned in the methods, but I didn't find it anywhere else?

- The effect on viral load shown in figure 4, while significant in change in mean, is a lot smaller than the population variation (shown by the violin range).

- L297: The word 'paradoxically' caused me to started wondering if the binding results were actually in the expected direction of effect. Looking at figure 4 I think they are, but I wonder if this suggests some more clarity around effect direction discussed in this part may be possible.

- I didn't see that previous genome-to-genome studies were referenced.

- In the text of the results, there are too many context-dependent or highly subjective terms such as 'largest/highest-quality' and 'intriguing/powerful'.

Referee #2

(Remarks to the Author)
Chen et al. have compiled and studied a large, highly-informative genetic dataset of host/EBV pairs to identify synergistic variants that impact risk of nasopharyngeal carcinoma (NPC). They observe a clear interaction between HLA-A*11:01, the locus associating most strongly (genome wide) with NPC, and a variant encoding a non-synonymous amino acid change in the viral tumor suppressor molecule EBNA3B. The genetic epidemiological data are quite convincing. They go on to map the epitope in which the variant resides (position 9 of the epitope) and show that it binds to A*11:01. They show that among individuals infected with high-risk EBV strains, the protective A*11:01 allele associates with lower viral load and the susceptible A*02:07 allele associates with higher viral load when combined with the 85841G. These data were not strong, but did go in the expected directions. Additional geographical and evolutionary findings were provided, complementing the genetic and functional data.
This paper nicely uncovers the major basis for A*11:01 protection against NPC (the data remain unknown for A*02:07 susceptibility, though somehow this also seems to involve 85841G). Overall, the approach the investigators took was excellent and the paper is very nicely written.
I have a number of questions/suggestions that I think should be addressed.

1. Does the VVILENVSR epitope bind to A*02:07? Was it predicted to bind A*02:07 by NetMHCpan 4.1? This could be tested in the BLI assay. I'm doubtful that it does, as the binding motifs for these two allotypes are completely different; A*02:07 strongly prefers leucine or valine at position 9, and A*11:01 strongly prefers lysine or arginine at this position. Why

would A*02:07 interact with the EBV SNP 85841? What could be the mechanism if it does not involve epitope binding to A*02:07?
2. Why is 85841A so protective against NPC among the high-risk strains? Maybe more to the point, why is 85841G bad, even in the presence of A*11:01? Is it because it is in LD with bad alleles at BALF2 and 7048? Is there anything known regarding the functional basis for the BALF2-CCT or 7048C association with high-risk EBV strains? Could the authors briefly address this in the Discussion or Introduction? It appears that the variants marking high risk EBV strains have a particularly stronger effect on risk of NPC (even independent of A*11:01).
3. It would be helpful if the authors could provide a table of pairwise linkage disequilibrium values for SNP 85841, BALF2, and 7048. This could be provided as supplementary material. Figure 3 seems to indicate that 85841G further refines the high-risk phenotype.
4. What is the distribution of the two immunodominant EBNA3B epitope (IVT and AVF) variants among high risk and lower risk EBV strains? The authors state that “Paradoxically, individuals carrying A*11:01 had lower NPC risk when infected with EBV strains carrying these epitope-loss mutations than non-carriers”, but was this in the high-risk EBV-infected group or in all subjects? Suppl Table 12 seems to indicate this was in the high-risk only grouping. This data needs more discussion. Why would the IVT and AVF A*11:01-binding alleles confer susceptibility to NPC, whereas the binding allele of VVI confers protection? The authors might want to concentrate on this, and provide one or more hypotheses. Is it possible to study CD8+ T cells from individuals with high-risk EBV who are A*11:01+ to identify differences in responses to these three epitopes (cytokine production, killing of infected cells, etc)? This would be especially interesting!
5. It would be helpful if the authors could please be very careful to note when individuals infected with high risk EBV strains (vs all strains) are being used. For example, were all 747 NPC cases and 1,251 controls (discovery), and all 644 cases and 880 controls in the validation infected with high-risk EBV strains (lines 117-122)? I realize that this information is sometimes given in suppl material (in this case, in Suppl Fig. 1), but it would help to have that clarified throughout the text as well. Line 211 is another example, and there are more.
6. If you analyze all subjects with non-high risk EBV (regardless of the variant at 85841), does A*11:01 show significant protection (it is not possible to tell from Fig. 3, but looks as if A*11:01 might be protective in the non-high risk groupings)? This could be done as a meta-analysis in order to provide power. If it does show significant protection in the non-high risk group, then one might conclude that A*11:01 renders some protection in addition to the interaction with 85841G.
7. Can the authors conclude anything regarding selection pressures on the host or the virus based on data described in the section on geographical co-enrichment of interacting host-viral factors? How much of the apparent viral spread could be due to (1) an ascertainment bias towards NPC cases in the samples that were sequenced, (2) variability in human demographics (war, famine) in South China over the past 4000 years that may have increased representation among carriers of particular EBV strains, or (3) an improved transmission rate of the C1 virus from individual to individual?
8. Do the authors think that 85841G arose and was maintained because it enhances fitness of the virus or is carried by the sweep of another advantageous SNP? It seems paradoxical that the C1 clade is considered to be selectively favored while at the same time associated with a reduced viral load due to the recognition of 85841G by HLA-A*11:01. Would the relationship of recombinant clades C1 (and X1) to other clades be altered if the ML tree (Fig 6) was constructed following exclusion of the recombinant N1-derived segment (about 30-40kb)?
9. Does SNP 85841 associate with the presence of IgA against VCA among NPC negative subjects?
10. Lines 172-174 and Lines 181-184: These two sentences seem to contradict one another. Is there an independent effect of A*02:07 after conditioning on A*11:01, or not?
11. Lines 219-224: The authors mention A*11:01 specifically in lines 219-220 with no mention of A*02:07, but the next sentence seems to suggest that conditioning on viral SNP 85841 also abolished the effect of A*02:07?
12. Line 317: Does “the HLA-interacting EBV strains” mean high-risk EBV strains with 85841G? It might be a bit confusing to change terminology, if indeed these mean the same thing. A similar issue applies to line 323.

Referee #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #4

(Remarks to the Author)
This manuscript by Chen et al. the authors use a paired human–EBV genomic dataset comprising >3,500 individuals and provide a stepwise genome-to-genome analysis to uncover host–viral genetic interactions underlying nasopharyngeal carcinoma (NPC) in a specific southern Chinese population, as the title implies. Using risk association with BALF-CCT, which is a methodology that has been used before by previous authors included many of the authors in this group, the authors identify HLA-A11:01 as a protective and HLA-A02:07 as deleterious host alleles that putatively interacts with EBV SNP 85841G in EBNA3B, leading to a “dual-risk” subgroup that apparently explains ~47% of NPC burden in Southern China. The validate this on smaller cohorts in China and Singapore. This is an interesting association study, but remains an association study at that, and seems to incorporate a number of previously known lines of evidence with the main novel discovery here of the EBV high risk SNP described here. There are a number of issues I wanted to raise in this study that needs to be addressed for the strength of this novelty to be upheld in this study.
1. The dataset used here likely comprises previous data that has been published members of this group before, and I have trouble understanding the resource value of this 'largest dataset eve', as claimed here. How much of this is data is new and how much overlaps with previous analyses.
2. There are a number of contradictions between this study and a recently published paper by the same authors in Cell

genomics a year prior, where several HLA and EBV interacting factors have also been identified. The authors make some reference to these but do not resolve the contradictions completely. Are you suggesting that the EBV factorsnidentified in your previous paper to be no longer significant, and therefore those claims should be withdrawn? similarly, are the HLA SNPs there still true? one was clarified but the other was largely ignored.
3.The HLA data is not particularly novel, and have been previously published. It is the associated with the EBNA 3B SNP that drives th novelty here and the potential co-evolution. Given that, I feel that there is limited data showing interaction between the peptide and A2-07, which is the deleterious /high risk SNP, and therefore in my mind, the more important one to show rather than the protective SNP. This asymmetry of data (which also cuts across other statistical analyses) with HLA A2-07, which to me is the more important 'positive' effect that needs stronger supporting evidence.
4. Even if the claim here is all about the protective nature of A11-01, and given that the main novelty here is the EBNA SNP, the authors need to show g immunological consequences to strengthen causal inference, including additional T cell assays (e.g., CTL recognition, cytokine release, killing assays).
5. Given the circular argument of finding host factors that are linked to to BALF-CCT and the re connecting that with the EBNA SNP, is there an a priori connection between these to begin with?
6. Given the lack of ancient EBV genomes, there should be some caveats into the evolutionary analyses of the EBV genome and co-evolutionary events as described here.
7. Finally, the translational claims are overreaching, I agree that screening could be an important strategy for high risk individuals, but to date, the concept of EBV vaccination has failed and so a moot point in the discussion.
8. There is a suggestion of no original code was included; please clarify if custom scripts for interaction scans/phylogenetics could be shared to improve reproducibility even if they are not novel per se. I was of the understanding that all codes used need to be made available.

Referee #5

(Remarks to the Author)
Manuscript Nr.: 2025-06-16134
Chen et al., “Human-EBV Interactions Drive Nasopharyngeal Carcinoma Endemicity in Southern China”

The authors demonstrate that a mutation in EBNA3B in Epstein Barr virus (EBV) strains that seem to have emerged from recombination of Northern and Southern Chinese viral strains synergize with HLA mediated risk for nasopharyngeal carcinoma (NPC). Namely, the variant encodes an EBNA3B peptide that binds to HLA-A*1101 and thereby presumably induces protective CD8+ T cell responses, while HLA-A*0207 confers risk. Without the high risk EBV strains the HLA locus seemed to confer only minimally increased risk for NPC. The authors identify an EBNA3B derived peptide (VVILENVGR) of the high risk strain that binds HLA-A*1101 better than its counterpart in the low risk viral strains (VVILENVSQ) and could increase protective T cell responses that lower NPC risk. 47% of endemic NPC cases seem to carry the combination of HLA-A*1101-HLA-A*0207+ and high risk EBV strain. Therefore, the authors argue that this combination is a major determinant of NPC development.

The reported interaction between NPC risk conferring EBV strain variants and HLA haplotype is interesting and novel. However, it should be put into context of other HLA-A11 restricted T cell responses, immune escape from HLA-A*0207 restricted CD8+ T cell responses and T cell infiltration, as well as EBV latency II infection in NPC.

Major comments:
1. Does the VVILENVSR peptide elicit significant CD8+ T cell responses in individuals with high risk EBV strains and HLA-A*1101 expression? Does the frequency of such cells correlate with lower EBV loads?
2. The authors argue that a recombination event generated the EBNA3B variant carrying virus. Does the EBNA3B variant influence T cell attraction that has previously been linked to EBNA3B dependent CXCL9 and CXCL10 production? Would a possibly decreased T cell attracting function explain the selection of this recombinant in Southern China?
3. Previously the EBNA3B aa416-424 epitope presented on HLA-A*1101 was described to be mutated in EBV strains of Papua New-Guinea as reported by the Massucci group. Is this epitope intact in the reassortments of Northern and Southern China EBV strains that increase NPC risk? How do IVTDFSVIK and VVILENVSR specific CD8+ T cell responses compare in healthy Southern Chinese individuals that carry HLA-A*1101.
4. NPC tumor cells usually do not express EBNA3B. How do the authors envision that the VVILENVSR specific CD8+ T cells protect from NPC risk?
5. The authors were not able to associate the HLA-A*0207 allele with sequence variation in the NPC risk EBV strains due to lower frequency of this MHC molecule in the studied population. However, it would still be helpful to know if any of the known HLA-A2 restricted epitopes and putative HLA-A*0207 restricted epitopes are mutated in the NPC risk associated EBV strains. This might explain why this allele is enriched in NPC patients that carry these viruses.

Minor comments:
1. The authors have primarily investigated HLA interactions with EBV strain variants. Therefore, they could high-light this in the title along the lines of “EBV strain variations interact with HLA haplotypes to increase NPC risk.”

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
This is a comprehensive revision that addresses my previous concerns, adding both statistical analyses and further lines of validation, ultimately strengthening the main conclusions.

I would still ask that the results from full GxG screen alluded to in the response to point three is added as supplementary material, or deposited in an appropriate resource if possible (e.g. zenodo or osf). Even if underpowered now, it may be useful for future validation or meta-analysis efforts. (This very minor comment may be a matter for the editor to decide on importance/necessity.)

Referee #2

(Remarks to the Author)
The authors have provided detailed responses to all of my comments, and have performed many experiments to address each comment definitively.
I congratulate the authors.

Referee #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #4

(Remarks to the Author)
Thank you for all your clarification and revisions in this version of the manuscript. Most of my major concerns have been addressed and I am especially heartened to see the immune assays used to demonstrate the actual T-cell effect of HLA A11-01 and the high risk peptide in a number of donor PBMCs. I also appreciate the attempt to analyse the different peptides from the high risk EBV lineage that may or may not associate with the A02-07 genotype- it is difficult to prove a negative result which may be the case here but nevertheless these negative results are important and interesting.

One final query, is there a speculation on whether the protective nature of A11-01 here shaped the evolutionary distribution of this HLA type across china and Southeast Asia? this is well outside the scope of this manuscript but more an interesting side-bar to the data presented here.

Referee #5

(Remarks to the Author)
Manuscript Nr.: 2025-06-16134A
Chen et al., “EBV strain variation interacts 1 with host HLA-A to determine
2 nasopharyngeal carcinoma risk”

The authors demonstrate that a mutation in EBNA3B in Epstein Barr virus (EBV) strains that seem to have emerged from recombination of Northern and Southern Chinese viral strains synergize with HLA mediated risk for nasopharyngeal carcinoma (NPC). Namely, the variant encodes an EBNA3B peptide that binds to HLA-A*1101 and thereby presumably induces protective CD8+ T cell responses, while HLA-A*0207 confers risk. Without the high risk EBV strains the HLA locus seemed to confer only minimally increased risk for NPC. The authors identify an EBNA3B derived peptide (VVILENVGR) of the high risk strain that binds HLA-A*1101 better than its counterpart in the low risk viral strains (VVILENVSQ) and could increase protective T cell responses that lower NPC risk. 47% of endemic NPC cases seem to carry the combination of HLA-A*1101-HLA-A*0207+ and high risk EBV strain. Therefore, the authors argue that this combination is a major determinant of NPC development.

In their revised manuscript the authors addressed all of my concerns but there was no direct correlation between VVILENVSQ specific HLA-A*1101 restricted CD8+ T cell responses and salivary viral loads. They further demonstrate that the high risk EBNA3B variant allows less efficient CXCL9 and CXCL10 production which could diminish T cell attraction into the tumor microenvironment, but the authors prefer to report these findings in another manuscript. They now also document that additional HLA-A*1101 restricted epitopes are lost in the high risk EBV strains in the Southern Chinese population. I also consider the argument that diminished EBNA3B specific CD8+ T cell responses would allow for an increased EBV reservoir to seed epithelial cells valid. However, would then a correlation between diminished HLA-A*1101 restricted CD8+ T cell responses and viral shedding or circulating viral loads not be even more important? Finally, the authors were not able to observe differences with HLA-A2 restricted EBV high and low risk variant derived peptide pools but low reactivity against an LMP1 derived HLA-A2 restricted peptide that is not present in Southern Chinese EBV variants. Therefore, they have toned down their conclusions with respect to the NPC risk increase by HLA-A*0207 even so this is the HLA allele most frequently found to be associated with NPC in genome wide association studies.

In summary, the authors have included a large amount of additional work into their revised manuscript. However, it seems like HLA-A*1101 restricted CD8+ T cell mediated immune control of EBV is compromised by Southern Chinese EBV variants beyond the described polymorphism and no prominent association of the VVILENVSQ specific HLA-A*1101

restricted CD8+ T cell response could be documented. Moreover, the authors prefer to include the interesting data on lower CXCL9 and CXCL10 production by the high risk EBNA3B variant into a different manuscript, even so this might more generally explain diminished CD8+ T cell mediated immune control of the high risk EBV variant beyond the mutation in one particular CD8+ T cell epitope.

Point-by-point responses to reviewers' comments

We would like to thank the editor and reviewers for your supportive comments and constructive suggestions, and we are grateful for the opportunity to improve this work. Following your recommendations, we have strengthened the statistical interaction analyses, added T cell functional assays, and revised the overall presentation of our study. Changes in the manuscript are highlighted in red, and our point-by-point responses are provided below.

Referees' comments:

Referee #1:

Chen and colleagues report on a genome-to-genome analysis of Epstein-Barr virus (EBV) and human genetics, specifically using an environmental/interaction effect framework to test for individual and joint effects on nasopharyngeal carcinoma (NPC). They recapitulate known strong effects of specific HLA types and high-risk EBV genotypes on NPC risk, but most interestingly find an interaction effect between a specific EBV variant and HLA type which increase risk further. The authors identify a 'dual-risk' subgroup of the population with this interaction, which is particularly prevalent in Southern China, and could explain the higher than expected burden of disease in this region. The authors propose and validate a mechanism of modified binding efficacy between the modified peptide and HLA variant. The large effect size, novel interaction, and potentially large population effect (i.e. expected cases due to this effect) make this a potentially important finding.

Elements that elevate this study beyond some other genome-to-genome efforts are: the large effect size of the tested effects on cancer, and the interaction effect specifically; replication efforts in other populations; a well-powered sample size; lab validation of binding effects.

Response: We sincerely appreciate your supportive summary and constructive comments regarding the novelty of our work.

Some parts of the paper remained unclear to me:

1) The description of odds-ratios is not always clear, and how the newly detected interaction effect reported here expands upon previous knowledge of increased risk from HLA and specific viral lineages. Specifically:

a) A clear explanation of the interaction effect specifically in the SC population would help, perhaps added to figure 3. What is the rate per million in the population without any risk factor, what is the rate per million with: one of the HLA variants, with high risk EBV, and with the interaction effect? [or instead of per million as the denominator, the Southern China population size could be used, as the authors did to estimate 35k preventable cases]. My reading of these from the current text is that the dual-risk group has an increase in risk of around double (OR =~ 10 for HLA variant, then OR =~ 18 with the viral variant too). It might be a bit misleading to only quote an OR of 18 in the abstract, as this doesn't refer to the interaction effect alone. Additional context of OR for the HLA and high-risk EBV effects could be included. Clearly noting the low population prevalence of NPC upfront would help too (which is done, just later in the text). Explaining the interpretation of the PAF in the main text could also help.

Response: We thank you for these constructive suggestions. Following your advice, we now decompose the joint effect for the dual-risk group on the relative risk (RR) scale (approximated by ORs given NPC's low incidence, also see Response to R1Q1.c), following the published interaction tutorial[1]. We have reported these components in the revised Abstract (**Lines 48-50**) and Results (**Lines 461-465**), with supporting results in revised **Figure 5** (**Response Figure 1**) and **Supplementary Table 15**. Specifically, the reported OR of 18.6 refers to the joint effect in the dual-risk subgroup and the corresponding total relative excess risk is ~17.6 (OR-1), of which ~9.2 is attributable to interaction (RERI), ~0.7 to susceptible HLA background alone, and ~7.6 to high-risk-85841G EBV infection alone. For clear presentation in the main figure, we have labeled the risk strata in the newly added **Figure 5c**: neither risk factor (baseline), HLA-risk-only, EBV-risk-only, and dual-risk, so that the OR for each stratum and the interaction effect (RERI) for the dual-risk subgroup can be read directly (**Response Figure 1**).

In light of your comment to provide an intuitive population impact for endemic Southern China, we have now reported absolute NPC incidence rates per 100,000 person-years across host and viral genetic backgrounds, based on regional incidence and the observed HLA allele and viral strain frequencies, using the corresponding OR estimates[2-4] (**Response Figure 1**, revised as **Figure 5c**; **Lines 470-475**). We also translate the population attributable fraction (PAF) into the estimated number of attributable cases under the counterfactual assumption[5,6] that exposure to specific high-risk EBV strains could be eliminated (**Lines 628-630**; **Response Figure 1**, revised as **Figure 5c**).

EBV subtype	HLA-A allele	Population frequency	OR (95% CI)	RERI (95% CI)	PAF (95% CI)	NPC incidence / 100,000 PY	Preventable cases	
Non-high-risk	Protective	23.2%	Baseline	–	15.6% (8.3%, 23.0%)	3.1	–	Neither
	Susceptible	31.0%	1.7 (1.0, 2.9)	–	6.6% (−2.6%, 15.9%)	5.3	–	HLA-risk
High-risk -85841A	Protective	3.3%	5.4 (2.5, 11.5)	–	3.6% (−4.2%, 11.3%)	16.5	~1.5k	
	Susceptible	6.0%	6.4 (3.4, 11.7)	0.3 (−5.2, 4.5)	7.2% (−0.4%, 14.9%)	19.6	~2.9k	EBV-risk
High-risk -85841G	Protective	16.2%	8.6 (5.4, 14.1)	–	20.0% (12.8%, 27.3%)	26.4	~8.2k	
	Susceptible	20.5%	18.6 (11.6, 29.8)	9.2 (4.9, 16.3)	46.9% (42.0%, 51.8%)	57.0	~19.1k	Dual-risk

Response Figure 1. Joint and interaction effects of HLA-A and EBV subtypes on NPC risk. Joint effect odds ratios (ORs), relative excess risks due to interaction (RERIs), and population burden metrics, including population attributable fraction (PAF), estimated NPC incidence, and preventable cases, are shown across strata defined by EBV strains and HLA-A background in the Southern China population. EBV infection was classified as non-high-risk, high-risk-85841A, or high-risk-85841G. HLA-A subgroups were defined as protective (A*11:01$^+$ and A*02:07$^-$) or susceptible (A*11:01$^-$ or A*02:07$^+$). The reference stratum was the protective HLA-A subgroup infected with non-high-risk EBV. ORs and RERIs were estimated using logistic regression models adjusted for age, sex, and the first four human MDS components, under the RERI framework, with 95% confidence intervals (CIs) calculated via bootstrap resampling. NPC incidence rates were estimated based on ORs for each stratum and overall regional incidence of approximately 20 per 100,000 person-years. Preventable cases denote estimated counterfactual counts under elimination of the corresponding EBV exposure. The dual-risk stratum is highlighted in red.

Furthermore, to make Figure 3 easier to interpret, following your suggestion we have added a concise explanation of the interaction effect into the main text (**Lines 258-262**) and figure legend (**Figure 3**; see revised legend) that "*the interaction effect was quantified using the Relative Excess Risk due to Interaction (RERI), a classical epidemiologic measure that captures the excess joint effect of carrying both risk factors beyond the sum of their individual effects, thereby reflecting the contribution of the host–virus interaction to the overall disease burden in the population*[1,7]". Details are provided in the Methods (**Methods file**: **Lines 301-325**).

Finally, we have followed your suggestion and further moved the description of the low NPC incidence in Southern China earlier in the Introduction (**Lines 67-69**) and added a brief interpretation and the reference for population attributable fraction (PAF) in the Results (**Lines 451-453**). We hope these revisions resolve the ambiguity noted by the reviewer and make the risk architecture and population relevance of the interaction clearer.

b) Fig 1d&c: It's not right to start the plot for these ORs at 0. ORs are not additive either. This plot should probably be logged. Overall these panels are quite unclear without the text, and I wonder if another presentation of interaction effects might be considered.

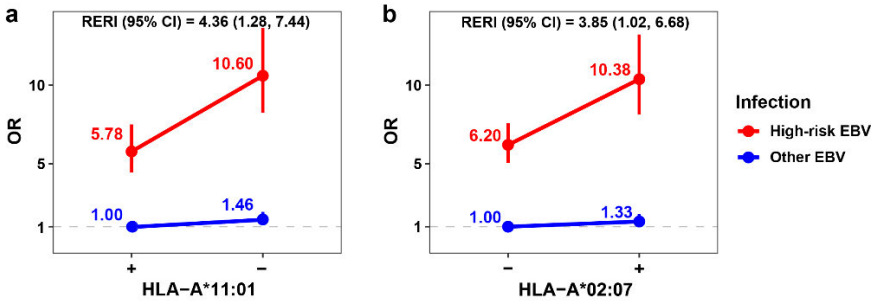
Response: Thank you for pointing out the gap regarding the presentation of interaction effects, and we fully agree that ORs are not additive. Following your advice to avoid potential misunderstanding of additive ORs and to better visualize interaction, we have revised this figure using a classical line interaction plot and made the following improvements:

i. replaced the previous bars with line plots, setting the null line at OR = 1.

ii. displayed point estimates with 95% CIs.

iii. reported the increase in risk due to interaction, formally quantified as the relative excess risk due to interaction (RERI), with 95% CIs in the revised **Figure 1d-e** (**Response Figure 2**).

In this format, the non-parallel slopes of the two lines convey the presence and magnitude of interaction.



Response Figure 2. Interaction between high-risk EBV and HLA-A alleles on NPC risk. Interaction line plots showing odds ratios (ORs) with 95% confidence intervals (CIs) for NPC risk stratified by infection with high-risk EBV strains (BALF2-CCT, in red) *versus* other non-high-risk strains (in blue) and by HLA-A*11:01 (**a**) or HLA-A*02:07 (**b**) status (present "+" or absent "−"). ORs were estimated by logistic regression, adjusted for age, sex, and the first four human MDS components in the meta-analysis of discovery and validation datasets. 95% CIs were obtained via bootstrap resampling. The dashed gray line represents the reference level (OR = 1), corresponding to HLA-A*11:01^{+} (a) or HLA-A*02:07^{-} (b) individuals infected with non-high-risk EBV.

c) I wanted to double-check that ascertainment (i.e. over-sampling of NPC cases in the collection) has been correctly reported on. I do see the authors mention in the methods that this is a rare disease, but could they comment on if/how this oversampling of NPC cases in the collection affects the interpretation of the reported OR in the interaction term specifically?

Response: We thank you for raising this important point. As NPC is a rare disease, a case-control design with intentional oversampling of cases is essential to achieve sufficient statistical power. As you have pointed out, this design does not bias the estimated effect sizes in logistic regression, because odds ratios (ORs) are invariant to case-control sampling[8]. Under the rare disease assumption, ORs provide good approximations of relative risk (RR), and this approximation also extends to interaction effects quantified by the relative excess risk due to interaction (RERI), as demonstrated in the published tutorial paper (included in **Supplementary Text**: **Lines 48-70**)[1].

Briefly, we quantified the increase in relative risk due to the interaction effect using RERI, a classical epidemiologic measure that captures the excess joint effect in the dual-risk group (RR_{11}) beyond the sum of the individual effects of EBV (RR_{01}) and HLA (RR_{10}), thereby reflecting the contribution of the host-virus interaction to the overall population disease burden (**Response Figure 3**)[1,7]. In the context of NPC's low incidence rate (age-standardized incidence of ~20 per 100,000 person-years)[3,4], ORs closely approximate relative risks (RRs)[8], allowing this decomposition of the joint effect as follows:

$$RERI = RR_{11} - RR_{10} - RR_{01} + 1 \approx OR_{11} - OR_{10} - OR_{01} + 1$$

a

The incidence of NPC based on high-risk HLA and EBV exposure.

High-risk HLA	High-risk EBV	Incidence
0	0	p_{00}
1	0	p_{10}
0	1	p_{01}
1	1	p_{11}

Excess risk due to interaction:
$p_{11} - p_{10} - p_{01} + p_{00}$

b

Relative excess risk due to HLA × EBV in NPC.

High-risk HLA	High-risk EBV	Relative risk
0	0	RR_{00} = 1 (reference)
1	0	$RR_{10} = p_{10} / p_{00}$
0	1	$RR_{01} = p_{01} / p_{00}$
1	1	$RR_{11} = p_{11} / p_{00}$

Relative excess risk due to interaction:
$RERI = RR_{11} - RR_{10} - RR_{01} + 1$

c

Calculation of relative risk (RR) based on odds ratio (OR) and disease incidence.

OR	p_0 (Baseline incidence)	
	10 / 100,000 PY	20 / 100,000 PY
3	RR = 3.00	RR = 3.00
5	RR = 5.00	RR = 5.00
10	RR = 9.99	RR = 9.98
18	RR = 17.97	RR = 17.94

$RR = \frac{p_1}{p_0}$; $OR = \frac{p_1 / (1-p_1)}{p_0 / (1-p_0)}$
For the rare disease, $OR \approx RR$;
Then, $RERI \approx OR_{11} - OR_{10} - OR_{01} + 1$

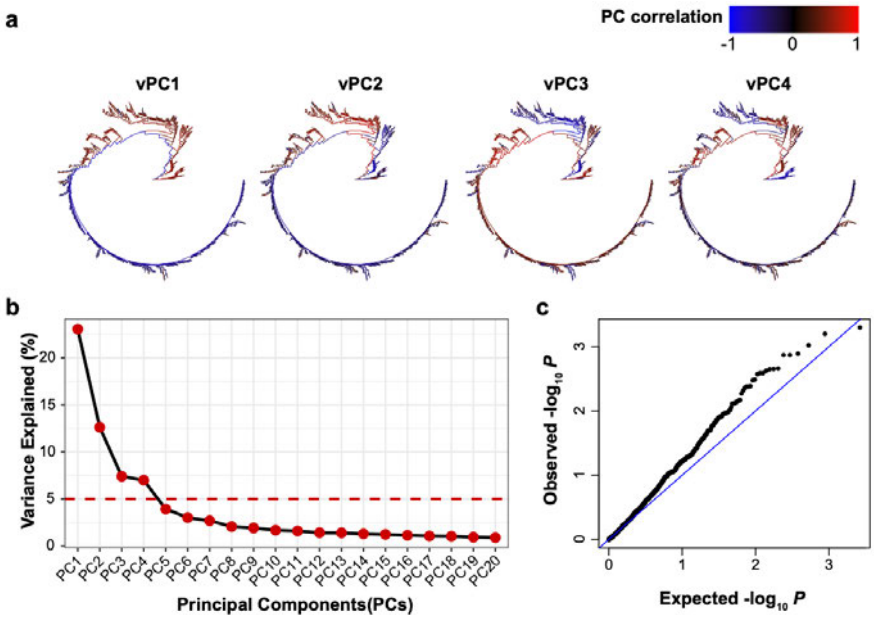
Response Figure 3. Conceptual framework for assessing relative excess risk due to interaction (RERI). a, NPC incidence across four exposure strata defined by the joint presence or absence of high-risk HLA and high-risk EBV. The reference group consists of individuals unexposed to both factors. **b**, Relative excess risk due to HLA×EBV interaction in NPC. **c**, Calculation of relative risk based on odds ratio and disease incidence.

Thus, the reported ORs and the interaction effect (RERI) remain valid and interpretable under a case-control design and the rare disease assumption. To further complement relative risk estimates, we have

also followed your advice and added absolute risk estimates for each subgroup to the Results (**Figure 5c**; **Lines 470-475**). The corresponding methods and a schematic clarification have been incorporated into the Methods (**Methods file**: **Lines 367-375**) and supplementary materials (**Supplementary Text**: **Lines 48-70**).

2a) The interaction test does not seem to control for viral population structure sufficiently. The authors use 4 MDS components, but this is not going to be able to control for the level of clonality seen in figure 6 (I didn't see this specifically tested/shown in a QQ plot or similar?).

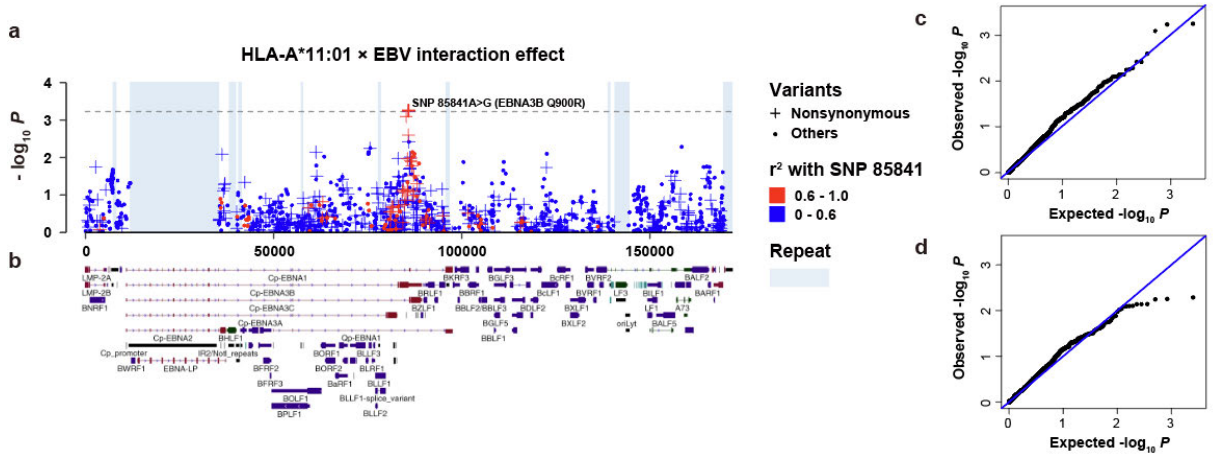
Response: We appreciate this constructive comment and fully agree that controlling for viral population structure is important. In our original analysis, we controlled viral population structure using four viral principal components (vPCs) and included age, sex, and the top host MDS components as covariates to control for host population structure. The top four vPCs captured the major clade structure of EBV in our data, with the clonal clade primarily represented by vPC1 and the major structure within non-clonal clades captured by vPC2 (**Response Figure 4a**). Beyond these top four vPCs, subsequent components each explain <5% of the variance and mainly reflect finer-scale viral structures (**Response Figure 4b**), which will be further captured by viral genetic relatedness matrix (vGRM), as suggested by the reviewer (see below). The original QQ plot showed moderate inflation (**Response Figure 4c**), which can be expected for compact pathogen genomes with extensive genome-wide linkage disequilibrium[9,10].



Response Figure 4. Principal components of EBV genetic variation and their relationship to EBV lineages, and quantile-quantile (QQ) plot for the EBV genome-wide interaction analysis. a, Leading principal components of EBV genetic variation correspond to major lineages in the EBV genealogy. For all EBV genomes analyzed, branches of the phylogeny are colored by their correlation with each of the first four principal components (PC1-PC4). Positive and negative correlations are shown in red and blue, respectively, with brighter colors indicating stronger correlations. **b,** Variance explained by the top 20 EBV PCs. The red dashed line marks the 5% variance-explained threshold. **c,** QQ plot for the EBV genome-wide interaction analysis in original Fig. 2.

Following your advice and consistent with the general mixed-model principle used in previous pathogen-GWAS (e.g., Earle et al[10]), we have further strengthened control of viral population structure in the EBV genome-wide interaction fine-mapping analysis by using a generalized linear mixed model (GLMM) for the binary outcome (NPC case *versus* control), with viral relatedness modeled as a random effect. Specifically, we constructed a vGRM from EBV variants after LD-based pruning using the GEMMA software[11], following recommended practices for highly linked genomes[12-14]. We then fit the logistic mixed model in the GMMAT software, using this vGRM as a random effect and including age, sex, the top four viral PCs, and the top four host MDS components as fixed effects[15]. With this improved framework, the top four vPCs captured major clade structure (**Response Figure 4a-b**, revised as **Supplementary Figure 11a-b**), whereas the vGRM random effect captured fine-scale structure and relatedness that the four major vPCs did not fully account for[14-16]. Together, this model adjusted for broad clade structure (vPCs) and residual viral relatedness (vGRM), as well as host population structure (**Methods file**: **223-283**;

Supplementary Text: **Lines 226-243**).



Response Figure 5. EBV genome-wide scan for the interaction with HLA-A*11:01 using GLMM. a, Manhattan plot showing *P*-values for EBV variant × HLA-A*11:01 interaction. The interaction analysis included 324 NPC cases (281 individuals infected with high-risk EBV strains (BALF2-CCT or 7048C) and 43 individuals infected with other non-high-risk strains) and 410 controls (173 individuals infected with high-risk EBV strains (BALF2-CCT or 7048C) and 237 individuals infected with other non-high-risk strains), using the generalized linear mixed model adjusted for age, sex, host population structure (first four MDS components), and viral population structure (first four EBV PCs), and with viral genetic relatedness matrix (vGRM) modeled as random effect. The top interacting EBV variant was SNP 85841A>G ($P_{interaction}$ = 5.57 × 10^{-4}), which encodes an amino-acid substitution at EBNA3B residue 900. The $-\log_{10}$-transformed *P*-values (y-axis) for interaction effects are plotted by EBV genomic position (x-axis). The dashed gray line indicates the EBV genome-wide significance threshold (P = 5.92 × 10^{-4}). Nonsynonymous variants are marked with "+", and synonymous or non-coding variants are shown as circles. Variant colors denote the corresponding linkage disequilibrium (r^2) with the leading SNP 85841. **b,** Schematic of EBV genes across the genome. **c,** QQ plot for the genome-wide interaction analysis in a. **d,** QQ plot after excluding the top variant 85841 and its ± 1kb flanking region.

Using this upgraded approach, the QQ plot showed that this new model provided better control of viral population structure with modest inflation, as expected due to the linkage across the EBV genome (**Response Figure 5c**, revised as **Supplementary 11c**). Particularly, when we further excluded the top variant (85841) and the variants within its ± 1kb flanking region, the QQ plot became deflated (**Response Figure 5d**, revised as **Supplementary 11c**), confirming that residual inflation in the overall association results was minimal. Importantly, the HLA-A*11:01 × EBNA3B 85841G interaction remained to be the top signal, with the new *P* value ($P_{interaction}$ = 5.57×10^{-4}) surpassing the threshold of viral genome-wide significance ($P_{threshold}$ = 5.92×10^{-4}) and largely unchanged effect size (**Response Figure 5a-b**, revised as **Figure 2**).

We really appreciate the constructive suggestion from the reviewer. These new analyses further improved the robustness of the results for HLA-A*11:01 × EBV 85841G interaction in NPC and significantly strengthened the statistical foundation of our conclusions. We have updated the Results (**Lines 210-224**), Methods (**Methods file**: **Lines 223-283**), and relevant figures (**Figure 2** and **Supplementary Figure 11**) accordingly to incorporate these new results and methodological improvements.

2b) While in figure 6 it looks like the interaction SNP does not completely correlate with high-risk EBV, it is definitely still associated with specific lineages. Typically viral GWAS uses a linear mixed model to control for the very strong population structure. Also, permutation tests tend to underestimate the significance threshold as samples are not iid due to their relatedness.

Response: We appreciate your constructive comment regarding the need to control for strong viral population structure in permutation when iid does not hold. We fully agree that naive permutations (e.g., directly permuting genotypes or case-control labels) can potentially inflate type 1 error when samples are not iid. To obtain a more valid empirical distribution under the null, we used the method presented in "BRASS: Permutation methods for binary traits in genetic association studies with structured samples"[17]. BRASS first fit a GLMM that included host covariates (age, sex, host MDS components), vPCs and vGRM to control the viral population structure and sample relatedness, excluding the viral SNP, host allele and their interaction (the null model). We computed the residuals and decorrelated them to obtain approximately exchangeable residuals. We then permuted these decorrelated residuals 10,000 times, and

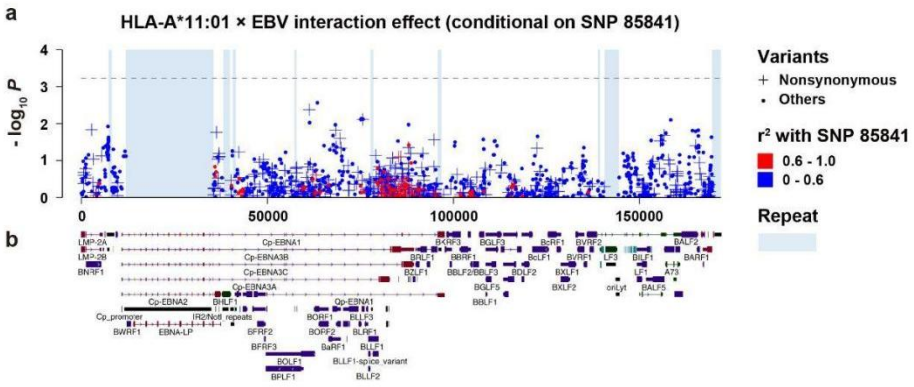
re-fit the interaction model on each replicate. This approach preserves the host-viral relatedness structure and accounts for clonality and relatedness to the viral genome while breaking only the association of interest. Using the minP across the EBV genome in each permutation, we set the empirical genome-wide threshold at $\alpha = 0.05$, yielding an empirical genome-wide threshold of 5.92×10^{-4} for the EBV interaction scan. Using this new permutation procedure for structured samples and the new threshold of genome-wide significance, the HLA-A*11:01×EBNA3B 85841G interaction remains the top signal with viral genome-wide significance. We have updated the new permutation in the revised Results (**Lines 224-227**) and Methods (**Methods file**: **Lines 273-283**) sections.

2c) Can the authors convince us that the interaction effect is likely the reported SNP, and that this SNP is not just a marker for a specific lineage/lineages?

Response: We thank you for raising this important question. We agree that SNP 85841G is associated with specific lineages (X1 and C1), as shown in Figure 6. Given the high linkage and near-clonality of viral genomes, distinguishing causal variants from closely associated lineage proxies is inherently challenging. While 85841 was statistically pinpointed as the top variant with the strongest evidence for the interaction with HLA-A*11:01, we have further applied three orthogonal approaches to support it as the most plausible causal driver of this interaction.

i. Solid statistical evidence with GLMM and genome-wide significance, as well as independent replication.

We re-analyzed EBV interaction fine-mapping using a GLMM with a viral GRM as a random effect, alongside viral PCs and host covariates (age, sex, top host MDS components) to control viral population structure and relatedness. The corresponding QQ plots showed well-controlled calibration with only modest genome-wide inflation (**Response Figure 5c**, revised as **Supplementary 11c**). In this new analysis, SNP 85841 consistently retained the top rank across the EBV genome (**Response Figure 5a-b,** revised as **Figure 2**), and conditioning on SNP 85841 abolished the evidence of interaction across the whole viral genome (**Response Figure 6**, revised as **Supplementary Figure 12**). We did not observe residual inflation after excluding SNP 85841 and its ±1 kb flanking region (**Response Figure 5d**, revised as **Supplementary 11d**). Together, these results prioritize 85841 as the lead SNP and the most plausible driver of the HLA-A*11:01 × EBV interaction (**Figure 2**, **Lines 203-240**). Importantly, this interaction was replicated in the independent datasets from Southern China and Singapore with consistent direction and similarly large effect sizes across different datasets (**Figure 3**, **Lines 257-277**).

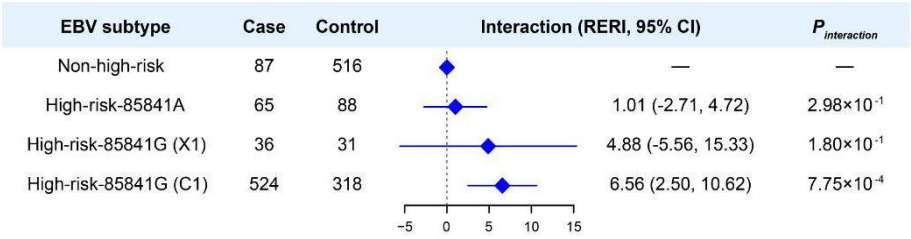


Response Figure 6. GLMM and conditioning of HLA-A*11:01 × EBV interactions identify SNP 85841 as the lead variant. a, Manhattan plot showing *P*-values for EBV variant × HLA-A*11:01 interaction. The interaction analysis was performed using the generalized linear mixed model, conditioned on EBV SNP 85841 and adjusted for age, sex, host population structure (first four MDS components), as well as viral population structure (first four EBV PCs), and with viral genetic relatedness matrix (vGRM) modeled as random effect. The $-\log_{10}$-transformed *P*-values (y-axis) for interaction effects are plotted by EBV genomic position (x-axis). The dashed gray line indicates the EBV genome-wide significance threshold ($P = 5.92 \times 10^{-4}$). Nonsynonymous variants are denoted by "+", and synonymous and non-coding variants are denoted by circles. Variant colors denote the corresponding linkage disequilibrium (r^2) with SNP 85841. **b,** Schematic of EBV genes.

ii. Independent origin and consistent interaction.

As lineages X1 and C1 both carry this 85841G variant, we performed chromosome painting and found that

both X1 and C1 lineages harbor partially overlapping, yet distinct, N1-like segments containing 85841G, consistent with at least two independent recombination events introducing the same variant (**Figure 6**) into the southern S1-like strains. Concordantly, both C1 and X1 showed consistent trend of interaction effects with HLA-A*11:01 on NPC risk (RERI = 6.56 for C1 and 4.88 for X1). Only C1 interaction reached statistical significance, as X1 was much rarer and underpowered (**Response Figure 7** and **Supplementary Figure 22**). This recurrent presence of 85841G in independently recombinant lineages, together with consistent interaction effects across lineages, argues against the idea that 85841G is merely a lineage marker, but rather a causal variant (**Lines 535-540**).



Response Figure 7. Interactions between HLA-A*11:01 and EBV subtypes in NPC risk. The interaction effects (relative excess risk due to interaction, RERI), their 95% CIs, and interaction *P*-values were estimated using logistic regression under the RERI framework in the Southern China dataset used in Figures 3a and 3b. EBV subtypes were grouped into four categories based on the presence of high-risk EBV markers (BALF2-CCT or 7048C), the genotype of SNP 85841 (A or G), and phylogenetic analysis, including non-high-risk EBV, 85841A-carrying high-risk EBV (S1), and subgroups of 85841G-carrying high-risk EBV (C1 and X1).

As pointed out in Earle et al[10], homoplasy, such as recurrent mutations or recombination events that break linkage, can enhance detection power in pathogen GWASs. This is the case for 85841G, where the consistent interaction effects from independent recombination events (e.g., in X1 and C1) have allowed us to detect the interaction signal (**Lines 540-544**).

iii. Functional evidence.

Recognizing that statistical evidence can support causal inference but cannot establish the molecular mechanism, we performed additional functional experiments to test the molecular interaction between the 85841G-encoded peptide and HLA-A*11:01 and its impact on T cell responses (**Lines 313-362**). T cell functional assays and binding assays demonstrate that the 85841G-encoded peptide VVILENVSR (VVI) is a bona fide HLA-A*11:01-restricted epitope (**Figure 4a-l**). It bound A*11:01 (**Figure 4a-b** and **Supplementary Figure 15a**) and elicited A*11:01-restricted T cell responses that specifically lysed VVI-presenting A*11:01^{+} cells (**Figure 4c-l**). By contrast, the 85841A-encoded low-risk peptide VVILENVGQ did not bind A*11:01 and failed to stimulate T cell responses (**Figure 4a-l**). Among individuals infected with 85841G-positive strains, A*11:01 carriers also showed lower saliva EBV load and lower NPC risk than non-carriers (**Figures 4n-o**), consistent with stronger T cell responses and tighter immune control in A*11:01 carriers (**Lines 364-371**). This reduced mucosal EBV burden would plausibly lessen opportunities for reseeding the nasopharyngeal epithelium, an early step toward NPC initiation (For working model, see **Response Figure 10**).

Lastly, we also evaluated the adjacent SNP 85837 G>A, which was one of the two signals that surpassed the significance threshold in the A*11:01×EBV interaction fine-mapping and is carried on haplotypes tightly linked to the top SNP 85841. The residue serine encoded by 85837A corresponds to position 8 of the candidate 9-mer peptide and is not predicted to occupy a key HLA-A*11:01 anchor position. Using a peptide-exchangeable HLA-A*11:01 monomer, we performed enzyme-linked immunosorbent assays (ELISA) with positional substitution peptides and found that the C-terminal arginine at position 9 encoded by 85841G, rather than serine at position 8 encoded by 85837A, was critical for HLA-A*11:01 binding, consistent with the known A*11:01 binding motif and structural predictions (**Lines 322-325**; **Supplementary Figure 15a**; AlphaFold3 model shown in the Response to R1 Minor Comment 2, **Response Figure 12**). This signal likely emerged due to high LD with the 85841G variant. Together, these findings support the interpretation that 85837 is a correlated, non-driving marker, whereas 85841G is the most plausible functional variant underlying the A*11:01 × EBV interaction.

In summary, these three lines of evidence including the logistic mixed model (**Figure 2**) and independent replication (**Figure 3**), independent recombination with consistent interaction across lineages (X1 and C1 *versus* S1; **Figure 6**), functional assays, and saliva viral titers (**Figure 4**) support the interpretation that

85841G is the most plausible driver of the interaction with HLA-A*11:01. We have updated the Results, Figures, and Methods sections accordingly to incorporate these new results.

In line with the reviewer's broader point and prior pathogen-GWAS literature[10,18,19], we have discussed that the trade-off between discovery power and controlling for population structure makes fine-mapping in near-clonal genomes particularly challenging and highlighted the importance of functional validation for causal interpretation (**Lines 652-655**) in the revised manuscript.

3) The two step process is not fully statistically justified. The authors dismiss a full genome-to-genome scan as underpowered with the sample size, and only check viral variants associated with HLA-A*11:01. The p-value they find is not very strong. Overall I do not think this is a disaster, because the strength of the effect appears large, it is found in two independent populations, this does look like a plausible GWAS peak (no strong population structure effects in figure 2), and there appears to be good functional evidence to support the result from the binding assay. But still, reporting the full genome-to-genome results against all human variants (as is typical for such studies) would be useful.

Response: Thank you for raising this important point and for summarizing the key evidence supporting the host–EBV interaction. We agree that a full genome-to-genome (G×G) scan is the long-term goal. However, with the current paired host–EBV whole-genome dataset (324 cases/410 controls), it is severely underpowered because of the multiple-testing burden. We therefore justify the stepwise design on biological and statistical grounds and add power simulations (and an exploratory full G×G scan) to support this design.

i. Rationale for the stepwise G×G design

Prior studies indicate that NPC risk is shaped by both host HLA variation and infection with high-risk EBV lineages, yet the causal determinants for host−EBV interactions have not been resolved on either side because prior analyses relied on lineage-level viral markers and limited paired genomes[12,20-28]. To overcome this resolution gap, we adopted a stepwise genome-to-genome analysis, which sequentially addresses three causal questions:

a) Whether additional host loci (beyond the HLA) interact with high-risk EBV to influence NPC risk?

b) Which HLA allele drives the interaction effect with high-risk EBV infection?

c) Which specific viral variant within the high-risk lineage drives this interaction?

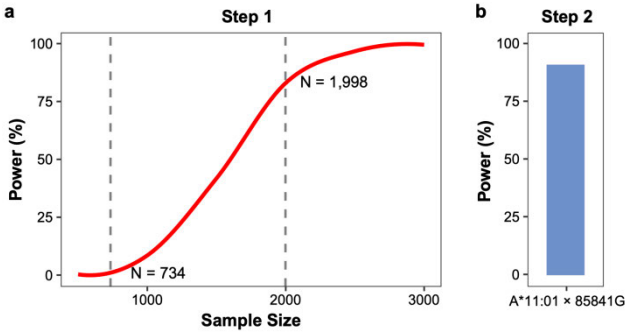
To answer the first and second questions, we performed the genome-wide scan of the host genome and further fine mapping analysis of the HLA region through the HLA imputation (Step 1) to identify loci whose effects on NPC depend on high-risk EBV infection. Then, to address the last question, we performed another genome-wide scan of the viral genome (Step 2) to fine-map the variants within the high-risk EBV lineage that drive the interaction with HLA-A*11:01, the major host interacting variant identified by the Step 1 analysis. We have added this rationale for stepwise design in the revised Introduction (**Lines 80-83**) and Results (**Lines 95-102)**.

ii. Statistical justification and Power simulations

Because EBV infection is near-universal, interaction mapping should be performed under a case–control design with NPC as the outcome[29]. In this setting, exhaustive host–virus pairwise testing is severely underpowered due to the multiple-testing burden[30,31]. Prior G×G studies have therefore shown that prioritizing loci with strong disease relevance (or marginal effects) substantially improves power compared with exhaustive G×G scans[32-35].

For our dataset, this consideration is crucial because paired host–virus whole-genome data and, especially, EBV whole genomes from population controls remain scarce in the field. Despite substantial efforts to collect this paired host–EBV genome dataset (324 cases/410 controls), which is large in the context of currently available EBV resources, it is still far below what would be required for a fully powered G×G interaction scan. Under these practical constraints, Step 1 leveraged our larger genotyping dataset (discovery: 747 cases/1,251 controls; validation: 644 cases/880 controls; host genotyping and high-risk EBV infection status) to identify host variants whose effects on NPC depend on high-risk EBV infection, and Step 2 restricted EBV fine-mapping in the paired host–viral genome dataset to the statistically supported HLA-A allele from Step 1.

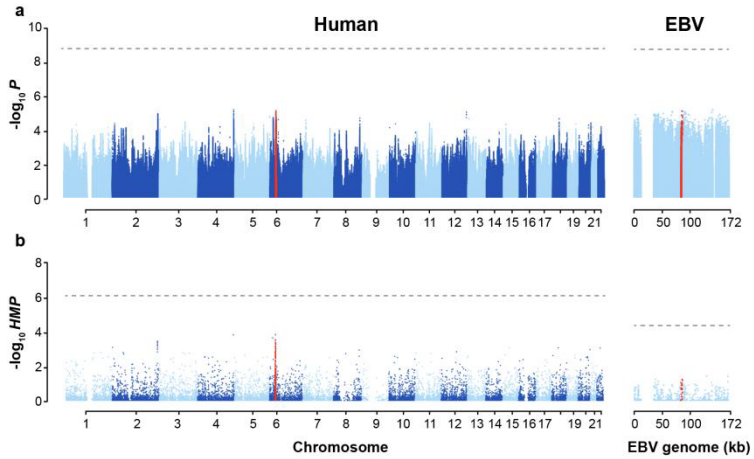
To quantify the power gain under this stepwise design, we performed simulations following the method in "Increasing the Power of Identifying Gene × Gene Interactions in Genome-wide Association Studies"[35]. In Step 1, we resampled the discovery dataset 1,000 times and simulated NPC case-control status using the estimated main and interaction effects of the top interaction signal rs417162 (tagging HLA-A*11:01; r^2 = 0.73), high-risk EBV (BALF2-CCT), and their interaction. The power to rediscover the top interaction at the genome-wide significance threshold ($P < 5 \times 10^{-8}$) was <20% at the paired-WGS sample size (324 cases/410 controls), but increased to 83% at the Step 1 discovery sample size (747 cases/1,251 controls; **Response Figure 8a**). In Step 2, we randomly resampled individuals and their EBV genomes from the paired host-EBV genome dataset and simulated NPC case–control status using the estimated effects of HLA-A*11:01, EBV SNP 85841, and their interaction. Across 1,000 simulations, power to detect the A*11:01 × 85841G interaction at the EBV genome-wide threshold (Bonferroni-corrected for the EBV scan) was 91% (**Response Figure 8b**). These results are consistent with prior human G×G power analyses and support that our stepwise design is statistically efficient and logistically feasible under current sample size[32-35]. We have added these power analyses to the Methods (**Methods file**: **Lines 285-298**) and summarized them in the revised manuscript (Results: **Lines 102-103** and Discussion: **Lines 655-658)** and supplementary materials (**Supplementary Text**: **Lines 72-113)**.



Response Figure 8. Power simulations for the two-step interaction design. a, Power simulation for Step 1, illustrating the power to detect interaction effects as a function of total case–control sample size. The vertical dashed gray lines indicate the sample sizes of the paired host–virus genome dataset (n = 734, <20% power) and the discovery interaction scan dataset (n = 1,998, >80% power). **b,** Power simulation for Step 2, showing the power to detect the HLA-A*11:01 × EBV 85841G interaction.

iii. Exploratory genome-to-genome (G×G) interaction analysis

In response to your request, we performed an exploratory full G×G interaction scan in the paired host–EBV whole-genome dataset (324 cases/410 controls), testing >10^7 host–EBV variant pairs. Given this high multiple-testing burden at the current paired-sample size, the scan was not expected to achieve a stringent family-wise error threshold for either marginal or interaction effects (see power simulations; **Response Figure 9a**). Nevertheless, signals in the HLA region and at the EBNA3B locus containing 85841G were enriched among the top signals above background, yet remained largely sub-threshold. We additionally applied harmonic mean *P*-value (HMP) aggregation, which stabilized ranking under multiple testing[36] but did not change the overall conclusion (**Response Figure 9b**). Under these constraints, the full scan cannot reliably distinguish true signals or tag signals driven by LD and noise. Accordingly, we present the G×G scan as an exploratory analysis in the **Supplementary Text Lines 115-139**, but it is not conclusive for fine-mapping additional interaction loci with the current sample size.



Response Figure 9. Joint human–EBV genome-wide interaction scan analysis of NPC in 324 cases and 410 controls. a, Human and EBV variants were subjected to quality control filters, including a call rate >90% and MAF >10%. For each tested human–EBV variant pair, we additionally required a minimum minor-allele count of 30 for each human SNP within each stratum defined by case–control status and EBV genotype to ensure stable inference in logistic regression and score tests. Interaction tests were performed under the genome-wide interaction framework using logistic regression models adjusted for age, sex, and the top four human MDS components. Results are plotted by genomic position for 79,726 human SNPs and 1,114 EBV SNPs. **b,** Significance of each human SNP was evaluated using the harmonic mean *P*-value (HMP) over all pairwise interaction tests with EBV SNPs and vice versa. Red points denote SNPs in the HLA region (chr6: 28–34 Mb) in the human genome panel (left), and the EBV SNP 85841 together with viral SNPs within ± 2 kb in the EBV genome panel (right). Bonferroni-corrected significance thresholds at α = 0.05 are shown as dashed gray lines.

iv. Summary

In summary, our two-step design was chosen because it directly addresses the biologically motivated questions, maximizes power given our current datasets, and follows recommended principles for G×G interaction mapping[32-35]. The A*11:01 × 85841G interaction is supported by (i) the logistic mixed-model analyses and empirical BRASS thresholds, (ii) replication across independent populations, and (iii) functional assays that confirm the predicted epitope and its HLA-A*11:01-restricted T cell response. We have now clarified the design rationale and power simulations in the Introduction (**Lines 80-83**), Results (**Lines 95-103**), Methods (**Methods file**: **Lines 153-298**), and supplementary materials (**Supplementary Text**: **Lines 72-113**) of the revised manuscript.

Looking forward, we share your view that a comprehensive G×G interaction analysis is an important goal worth pursuing in the future. Many pathogen G×G studies apply case-only designs, but this is not applicable for EBV due to its near-universal infection[29]. Within a case-control framework, a fully powered G×G interaction scan would require paired host-viral genome datasets that are orders of magnitude larger than those currently available[30,31], particularly for EBV whole-genome sequences from population controls. We therefore view our two-step design as a statistically efficient and biologically grounded approach under current sample-size constraints. As larger paired datasets accumulate, future studies should enable more comprehensive G×G maps and reveal additional host–EBV interaction signals (**Lines 658-659**).

4) How, mechanistically, this interaction leads to cancer risk is not totally clear to me, especially as someone without a background in virology or oncology. Likely this just needs more explanation in the text. For example, by what mechanism is EBV thought to cause NPC? Is it because of increased mutagenesis? Or causing selection on somatic variants? Is there anything that can be said about the reduced binding efficacy to EBV enhancing this effect? (i.e. is an increased viral load known to increase cancer risk, and how/why?)

Response: We appreciate your suggestion to clarify how the HLA-A*11:01 × EBNA3B 85841G interaction could translate into NPC risk. In response, we have added new functional T cell experiments showing that 85841G creates a bona fide HLA-A*11:01-restricted T cell epitope (VVILENVSR) that elicits EBV-specific T cell responses and expanded the Results (**Lines 313-362**; **Figure 4**) and Discussion (**Lines 571-581** and **607-618**) to place the host–virus interaction in the established framework of EBV-driven NPC, together with a schematic model (**Response Figure 10**, revised as **Supplementary Figure 23**).

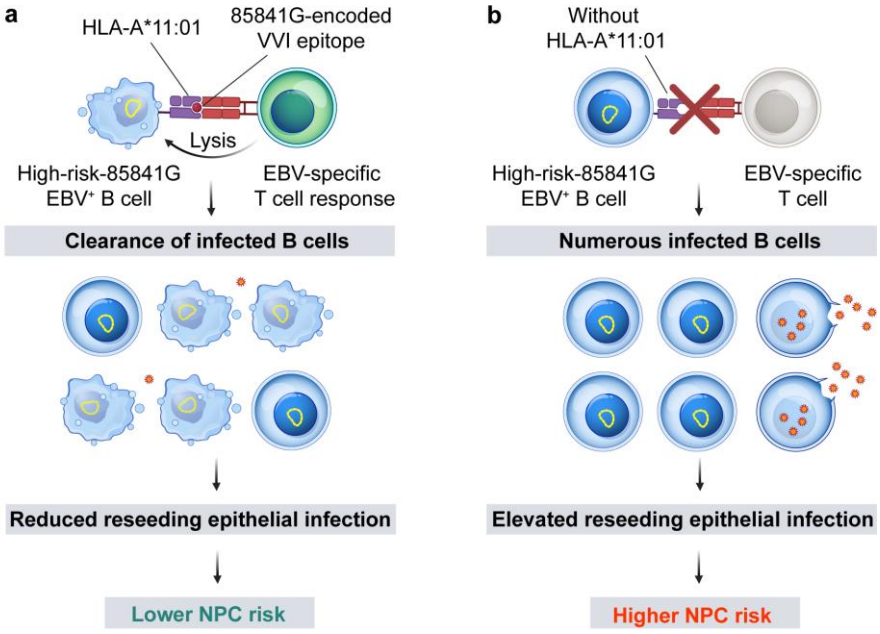
i. How EBV is thought to drive NPC.

NPC arises from the clonal expansion of EBV-infected basal epithelial cells in the nasopharynx. EBV entry into nasopharyngeal epithelial cells is an early, pre-tumor event[37-39]. After primary infection, EBV establishes lifelong latency in B cells and reactivates intermittently within oropharyngeal/nasopharyngeal lymphoid tissues, reseeding the mucosal epithelium and shedding through saliva[40]. In most healthy carriers, EBV reactivation is effectively restrained and is not directly oncogenic[41]. However, if a pre-neoplastic epithelial lesion already exists, EBV can be retained in basal epithelial cells and contribute to tumor initiation and progression. It promotes NPC tumorigenesis primarily through oncogenic viral proteins and non-coding RNAs that drive tumor cell transformation and proliferation, immune evasion, and epigenetic reprogramming[42,43]. In line with this model, increased mucosal EBV activity precedes NPC diagnosis by years, as reflected by elevated EBV-specific IgA responses[44-47], and circulating EBV DNA is elevated in patients with NPC and is used for early detection[48], indicating a pre-tumor window of heightened viral activity[49]. In this setting, more frequent EBV reactivation and weaker immune control are expected to increase epithelial exposure and facilitate the establishment and maintenance of infection within susceptible epithelial compartments.

ii. Where the interaction fits.

Our revised Results provide a mechanistic link from genetic interaction to immune control of EBV. The 85841G substitution creates an HLA-A*11:01-restricted epitope, VVILENVSR (VVI), that elicits $CD8^+$ T cell activation and cytotoxicity against EBV^+ B cells transformed by the high-risk-85841G strain M81 (**Results**: **Lines 313-362**; **Figure 4a-l**). Among healthy carriers infected with high-risk-85841G EBV, HLA-A*11:01 is associated with lower salivary EBV DNA levels, consistent with tighter viral control and lower NPC risk in this subgroup (**Results**: **Lines 364-371**; **Figure 4n-o**). Taken together, these data support a model that: in $A*11:01^+$ individuals infected with high-risk-85841G EBV, VVI-specific A*11:01-restricted $CD8^+$ T cells help restrain the EBV-infected B cell reservoir and limit reactivation, thereby reducing epithelial reseeding and the probability of NPC initiation. In individuals lacking HLA-A*11:01, recognition of the VVI epitope is absent, and high-risk-85841G EBV infection is associated with weaker immune control and higher NPC risk (**Response Figure 10**, revised as **Supplementary Figure 23**).

We have incorporated this mechanism overview (**Supplementary Figure 23**), together with supporting literature, into the revised Discussion (**Lines 571-581** and **607-618**) and added the new T cell functional results (**Lines 313-362**; **Figure 4**).



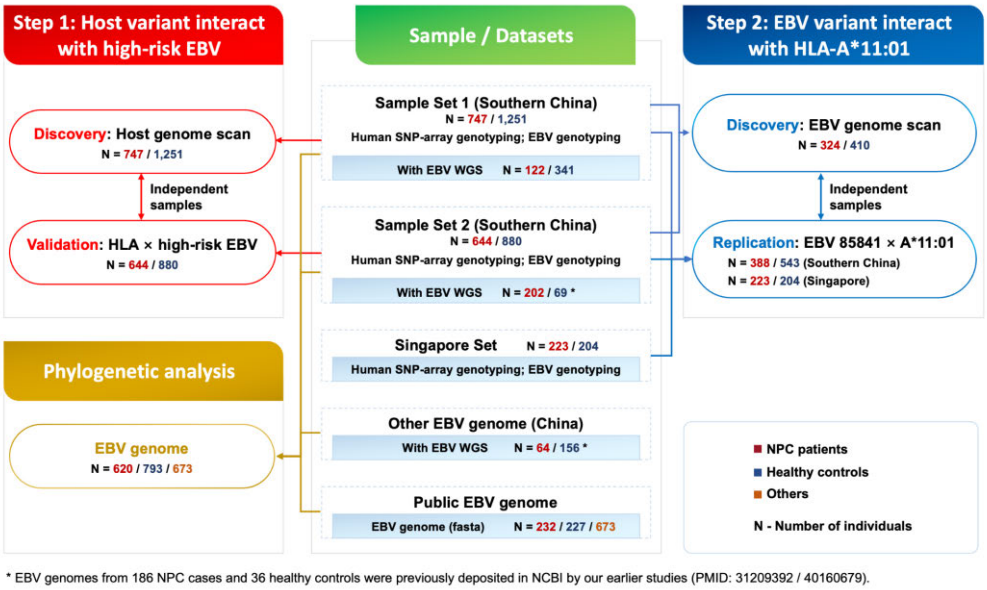
Response Figure 10. Schematic model of EBV immune control and NPC risk associated with HLA-A*11:01. a, In individuals carrying HLA-A*11:01 and infected with 85841G-carrying EBV strains, the EBNA3B 85841G substitution generates an epitope VVILENVSR (VVI) that is efficiently presented by HLA-A*11:01 to EBV-specific $CD8^+$ T cells. These HLA-A*11:01-restricted VVI-specific T cells are capable of lysing EBV^+ B cells infected with high-risk-85841G strains, potentially reducing the size of the infected B cell reservoir and epithelial reseeding, thus lowering NPC risk. **b,**

In individuals lacking HLA-A*11:01, presentation of the 85841G-encoded epitope to EBV-specific $CD8^+$ T cells is suboptimal. EBV^+ B cells infected by high-risk-85841G strains are less efficiently cleared, leading to a larger pool of infected and reactivating B cells, elevated epithelial reseeding, and consequently higher NPC risk.

Minor comments:

- There are actually only ~750 samples with virus sequence, perhaps reporting this as well as the headline number of 3500 would be helpful to understand the study design.

Response: We appreciate your suggestion on clarifying the samples and datasets used in each analysis. Our study uses larger case–control datasets for host genome-wide interaction scan with the high-risk EBV status (Step 1: 747 cases/1,251 controls in discovery; 644 cases/880 controls in validation), and a subset with paired host–EBV genome data for EBV interaction fine-mapping (Step 2: 324 cases and 410 controls). In light of your comment, we have added a schematic summarizing the workflow and the sample size used at each step (**Response Figure 11,** revised as **Supplementary Figure 1**), and we now clearly report the sample sizes in the main text (**Lines 97-102**) and corresponding figure legends (**Figures 1** and **2**).

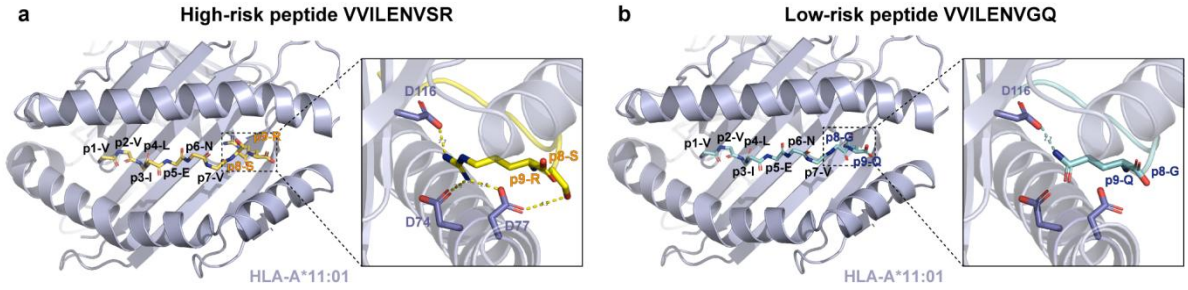


Response Figure 11. Overview of study design and data analysis workflows. In Step 1, a host genome-wide scan in a Southern China discovery dataset was used to identify host variants interacting with high-risk EBV, and the interactions were then evaluated in an independent validation dataset. In Step 2, an EBV genome-wide scan in Southern China case-control samples was performed to identify viral variants interacting with HLA-A*11:01, and the lead variant (EBV SNP 85841) was replicated in independent Southern China and Singapore datasets. EBV genomes from China and public datasets were combined for phylogenetic analysis. Numbers shown in the corresponding colors indicate the sample sizes of NPC patients (red), healthy controls (blue), and other individuals (orange), respectively.

- I am not an expert here, but the follow-up in vivo for binding efficiency seemed convincing. I did wonder whether screening interaction strength with AlphaFold3 model confidence (or a similar open source tool) would help further increase confidence in these findings?

Response: We appreciate this helpful suggestion. In the revision, we have added structural modeling of the peptide-HLA-A*11:01 complex as supportive evidence alongside the biochemical binding (ELISA/BLI) and functional T cell assays. Using an available high-resolution HLA-A*11:01 structure as a template (PDB code: 6JOZ), we modeled the peptide-HLA complexes for VVILENVSR (85841G) and VVILENVGQ (85841A) using AlphaFold3. The predicted structures (**Response Figure 12**, revised as **Supplementary Figure 16**) are concordant with our experimental results: peptide VVILENVSR adopts an A*11:01-compatible configuration with a well-supported C-terminal anchoring via the 85841G-encoded residue Arg at position 9 (R9), whereas VVILENVGQ does not adopt a comparable C-terminal anchoring configuration with the 85841A-encoded Gln (Q9). We present these models as supporting evidence,

consistent with the experiments showing that 85841G-encoded R9 is critical for A*11:01 binding (**Supplementary Figure 15a**). We have included the predicted structures in **Supplementary Figure 16** and summarized these results (**Lines 322-325**) in the revised manuscript.



Response Figure 12. Predicted structures of HLA-A*11:01 in complex with high-risk or low-risk EBNA3B peptides. a, Predicted structure of HLA-A*11:01 (gray) bound to the 85841G-encoded peptide VVILENVSR (yellow). The peptide adopts a stable HLA-A*11:01-compatible configuration with positions p1–p9 indicated. The inset highlights the C-terminal region, where the arginine at p9 (R9) forms a well-supported interaction network with A*11:01 residues D74, D77 and D116 in the peptide-binding groove. **b,** Predicted structure of HLA-A*11:01 bound to the 85841A-encoded peptide VVILENVGQ (cyan). In this model, the glutamine at p9 (Q9) fails to adopt a stable anchoring geometry as R9 and shows markedly reduced contact with the MHC groove, consistent with a less favorable peptide-MHC interface.

- Where was statistical power reported? This is mentioned in the methods, but I didn't find it anywhere else?

Response: We thank the reviewer for this comment. We have now added power summary and corresponding figures for detecting host–virus interactions in the Result (**Lines 102-103** and **293-296**), Methods (**Methods file**: **Lines 285-298**) and supplementary materials (**Supplementary Text**: **Lines 72-113**).

Briefly, we performed empirical simulations by resampling individuals in each analysis, using the main and interaction effect sizes estimated in our datasets to generate NPC case–control status across 1,000 replicates, and quantified power as the proportion of replicates in which the prespecified interaction test reached the significance threshold used in the corresponding scan. Simulations were parameterized using rs417162 (top interaction signal) and BALF2-CCT for Step 1, and EBV SNP 85841 and HLA-A*11:01 for Step 2, together with their estimated interaction effects. In Step 1 (host genome-wide interaction scan; discovery dataset n = 1,998), the estimated power was 83%. In Step 2 (EBV interaction fine-mapping; paired host–viral genome dataset n = 734), the estimated power to detect the EBV SNP 85841 × HLA-A*11:01 interaction was 91%.

- The effect on viral load shown in figure 4, while significant in change in mean, is a lot smaller than the population variation (shown by the violin range).

Response: We appreciate this comment regarding the large variation in salivary viral load. Because EBV undergoes intermittent reactivation and episodically sheds into saliva, salivary EBV DNA fluctuates within individuals and shows wide inter-host variability, likely reflecting the random process such as the timing of reactivation, shedding, and sampling[50]. Accordingly, we do not expect host genotype alone to explain the majority of this variance in healthy carriers. Despite this variability, we observed modest but significant mean differences in viral load across HLA-A genotypes among carriers of the HLA-interacting high-risk-85841G strain, but not among carriers of non-interacting strains, supporting an HLA- and EBV strain-dependent immune control. We have added this biological background and relevant references in the revised Results (**Lines 364-368**).

- L297: The word 'paradoxically' caused me to started wondering if the binding results were actually in the expected direction of effect. Looking at figure 4 I think they are, but I wonder if this suggests some more clarity around effect direction discussed in this part may be possible.

Response: We thank you for pointing out this un-intended confusion. Yes, our binding and functional results are in the expected direction. We originally used “paradoxically” to describe that high-risk-85841G strains carry the previously reported epitope-loss variants in two immunodominant A*11:01-restricted

EBNA3B epitopes (AVF/IVT), which diminish A*11:01-restricted EBV-specific T cell responses, suggesting the presence of a different causal variant that could explain the protective interaction effect of A*11:01. We found that the 85841G-encoded peptide VVILENVSR (VVI), but not the 85841A-encoded peptide, binds HLA-A*11:01 and elicits A*11:01-restricted EBV-specific $CD8^+$ T cell responses, consistent with tighter immune control and lower NPC risk in A*11:01 carriers. By contrast, in individuals lacking A*11:01, this VVI-specific recognition is absent and high-risk-85841G EBV infection is associated with a higher NPC risk, consistent with the direction of the genetic interaction. To avoid potential confusion, we have removed the word "paradoxically" and rephrased the text to state the direction of the binding, immune, and risk effects (**Lines 308-311**).

- I didn't see that previous genome-to-genome studies were referenced.

Response: We thank you for this helpful comment. We have now added references to prior genome-to-genome studies in EBV, HIV, HCV, HBV, M. tuberculosis, pneumococcal meningitis, and Malaria[9,19,36,51-55] into the Introduction to provide context for our approach (**Lines 83-85**).

- In the text of the results, there are too many context-dependent or highly subjective terms such as 'largest/highest-quality' and 'intriguing/powerful'.

Response: We appreciate this helpful suggestion. We have carefully reviewed the entire manuscript and removed or toned down subjective wording, including 'largest', 'highest-quality', 'intriguing', 'powerful', and similar terms, replacing them with more objective, data-supported descriptions (e.g., specifying sample size, sequencing depth, or statistical significance) to ensure a more neutral and precise presentation.

Referee #2:

Chen et al. have compiled and studied a large, highly-informative genetic dataset of host/EBV pairs to identify synergistic variants that impact risk of nasopharyngeal carcinoma (NPC). They observe a clear interaction between HLA-A*11:01, the locus associating most strongly (genome wide) with NPC, and a variant encoding a non-synonymous amino acid change in the viral tumor suppressor molecule EBNA3B. The genetic epidemiological data are quite convincing. They go on to map the epitope in which the variant resides (position 9 of the epitope) and show that it binds to A*11:01. They show that among individuals infected with high-risk EBV strains, the protective A*11:01 allele associates with lower viral load and the susceptible A*02:07 allele associates with higher viral load when combined with the 85841G. These data were not strong, but did go in the expected directions. Additional geographical and evolutionary findings were provided, complementing the genetic and functional data.

This paper nicely uncovers the major basis for A*11:01 protection against NPC (the data remain unknown for A*02:07 susceptibility, though somehow this also seems to involve 85841G). Overall, the approach the investigators took was excellent and the paper is very nicely written.

Response: We sincerely appreciate your thoughtful summary and constructive comments.

I have a number of questions/suggestions that I think should be addressed.

1. Does the VVILENVSR epitope bind to A*02:07? Was it predicted to bind A*02:07 by NetMHCpan 4.1? This could be tested in the BLI assay. I'm doubtful that it does, as the binding motifs for these two allotypes are completely different; A*02:07 strongly prefers leucine or valine at position 9, and A*11:01 strongly prefers lysine or arginine at this position. Why would A*02:07 interact with the EBV SNP 85841? What could be the mechanism if it does not involve epitope binding to A*02:07?

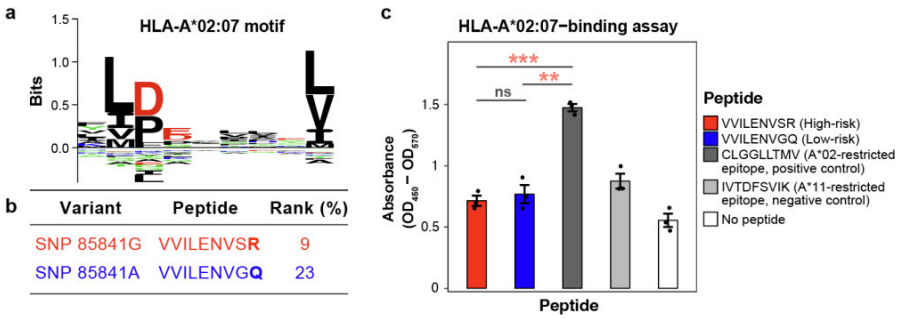
Response: We appreciate your thoughtful question on potential mechanism underlying the putative interaction between A*02:07 and SNP 85841. In our current study, the evidence for an independent A*02:07 interaction was only suggestive and did not reach genome-wide significance. In the discovery host genome-wide scan for interaction with the high-risk EBV BALF2-CCT, A*11:01 was the only interaction signal with robust evidence, and conditioning on A*11:01 abolished all remaining interaction

signals in the genome. In the meta-analysis of discovery and validation datasets, the A*02:07 × high-risk EBV interaction did not remain genome-wide significant ($P < 5 \times 10^{-8}$) after conditioning on A*11:01 (before conditioning, $P = 1.72 \times 10^{-8}$; after conditioning, $P = 4.86 \times 10^{-4}$; **Supplementary Table 5**). Likewise, an exploratory test focusing on the A*02:07 × EBV SNP 85841G interaction (rather than a viral genome-wide scan, given limited power) yielded only weak evidence ($P = 1.12 \times 10^{-2}$), much weaker than the A*11:01 × 85841G interaction ($P = 5.57 \times 10^{-4}$). Taken together, the A*02:07 × EBV interaction signal in NPC is weaker than our primary A*11:01 interaction finding and remains to be validated in larger datasets at both the host and viral genome level, particularly given the lower frequency of A*02:07 than A*11:01 in Southern China. In light of these results, we have now toned down the statement for A*02:07 × EBV interaction in the main text (**Lines 189-195** and **293-299**).

Despite the limited statistical support, we followed your suggestion and performed binding assays. We found no evidence that the 85841G-encoded high-risk peptide VVILENVSR or the 85841A-encoded low-risk peptide VVILENVGQ binds to A*02:07, consistent with the distinct A*02:07 and A*11:01 binding motifs (**Lines 319-322**). Given the marginal genetic evidence and the lack of detectable binding, we did not pursue deeper functional experiments or detailed speculation about the potential mechanism of A*02:07 × 85841G interaction at this stage. The experimental details and results are described below.

i. Experimental data showing no binding of HLA-A*02:07 to 85841-encoded peptides

To test whether A*02:07 can bind the 85841G or 85841A-encoded peptide, we performed both *in silico* prediction and experimental binding assays. NetMHCpan 4.1 predicted that neither the high-risk 85841G-encoded peptide (VVILENVSR) nor the low-risk 85841A-encoded peptide (VVILENVGQ) binds A*02:07, and neither conforms to the canonical A*02:07 binding motif (**Response Figure 13a-b**). A sliding-window screen of 9 to 11-mer peptides encompassing the 85841-encoded residue also showed no predicted binders of A*02:07 (all ranks >5%). Consistent with these predictions, peptide-HLA binding assays using ELISA confirmed that A*02:07 did not bind either peptide (**Response Figure 13c**, revised as **Supplementary Figure 15b**). These results argue strongly against a direct peptide-level mechanism for A*02:07 at SNP 85841, in contrast to the well-supported A*11:01-restricted VVILENVSR epitope (**Lines 319-322**).

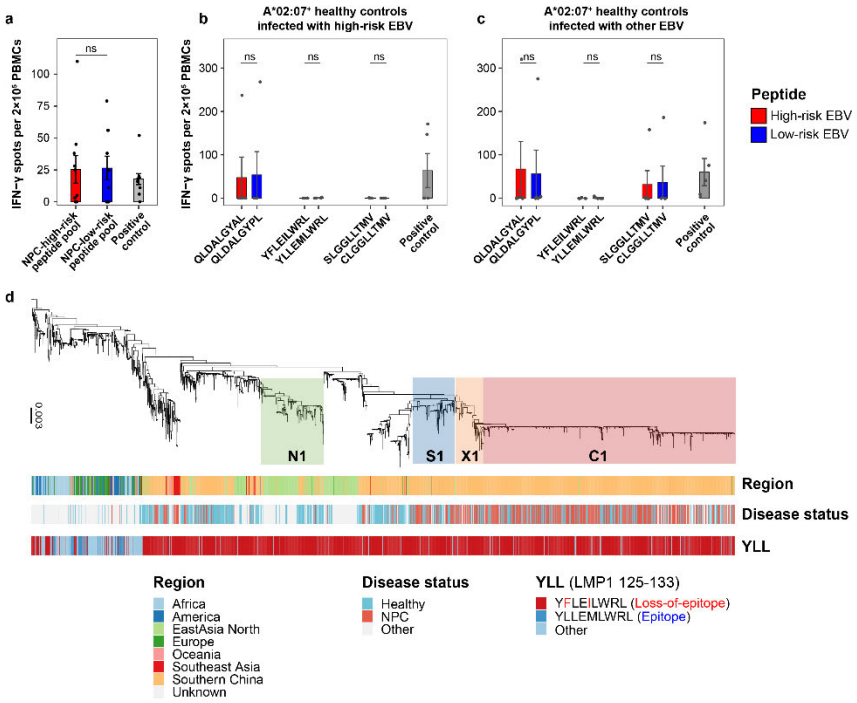


Response Figure 13. HLA-A*02:07 binding motif and lack of binding to the 85841G or 85841A-encoded peptide. **a**, Sequence logo of HLA-A*02:07-bound peptides. **b**, NetMHCpan 4.1 predicted ranks of binding affinities for peptides encoded by 85841G and 85841A variants. **c**, ELISA-based peptide-HLA-A*02:07 binding assay for VVILENVSR and VVILENVGQ. Peptide CLGGLLTMV, an HLA-A*02:07-restricted epitope, was used as a positive control. IVTDFSVIK, an unrelated peptide with negligible HLA-A*02:07 binding, and the no-peptide condition were used as negative controls. In the absence of peptide binding, HLA class I complexes are unstable and undergo degradation, resulting in reduced ELISA signals with lower OD values. Absorbance was calculated as $OD_{450}–OD_{570}$. Data are shown as mean ± s.e.m. (n = 3). Statistical significance was evaluated using two-tailed *t*-tests. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ns, not significant.

ii. Exploratory analyses of potential A*02:07-restricted epitopes

Motivated by your question and the established strong genetic association of A*02:07 with NPC risk, we systematically examined the reported HLA-A*02-restricted EBV epitopes[56-66] (**Supplementary Tables 20** and **21**) as well as predicted candidate epitopes (binding affinity ranks < 2%, predicted by NetMHCpan 4.1) that carry variants in strong linkage disequilibrium ($r^2 > 0.5$) with the high-risk viral variants (e.g., BALF2-CCT, EBER2-7048C) or EBNA3B-85841G (**Supplementary Table 21**). These peptide pairs from high-risk and low-risk strains were tested for T cell reactivity by ELISPOT assays using PBMCs from A*02:07$^+$ individuals, with previously reported A*02-restricted epitopes that are conserved (i.e., not carrying common variants) as positive controls. The T cell responses against the NPC-high-risk and NPC-low-risk peptide pools showed no significant differences (**Response Figure 14a**). Among all peptide pairs

tested, two pairs elicited detectable T cell responses in A*02$^+$ individuals: the known CLG epitope and its SLG high-risk counterpart, and a predicted peptide pair (low-risk peptide QLDALGYPL and high-risk peptide QLDALGYAL). However, neither pair showed differential reactivity between high-risk and low-risk versions (**Response Figure 14b-c**). All other predicted A*02:07-restricted epitopes were non-reactive in these ELISPOT assays. Together, the lack of differential reactivity across the tested high-risk/low-risk peptide pairs argues against a single epitope gain/loss mechanism that differentiates high-risk from low-risk EBV strains as the basis for the putative A*02:07 interaction. We also revisited the known LMP1 A*02-escape variant YFLEILWRL (immunogenic wild-type: YLLEMLWRL)[57,67], which is nearly fixed across Asian EBV strains (**Response Figure 14d**). As expected, this escape variant showed no T cell reactivity in ELISPOT assays using PBMCs from A*02:07$^+$ healthy donors (**Response Figure 14b-c**), suggesting that loss of this YLL epitope may weaken A*02:07-restricted T cell recognition of all Asian EBV strains.



Response Figure 14. T cell responses to A*02:07-restricted EBV epitopes and potential epitopes, and distribution of an LMP1 A*02 escape variant. a, No difference in T cell responses to peptide pools derived from NPC-high-risk *versus* NPC-low-risk EBV strains. PBMCs from A*02:07$^+$ individuals (n = 10) were used to assess peptide-pool–specific T cell responses by IFN-γ ELISPOT. Previously reported HLA-A*02-restricted epitopes that are conserved were included as positive controls. **b-c**, No differential T cell responses to A*02:07-restricted EBV epitope pairs between high-risk and low-risk versions. PBMCs from A*02:07$^+$ individuals infected with high-risk EBV (b; n = 5) or non-high-risk EBV (c; n = 5) were used to assess peptide-specific T cell responses. Polyclonal T cells were expanded against A*02:07-restricted peptide pairs and then restimulated with each peptide (2 μM) for 18 h. T cell activation was assessed by IFN-γ ELISPOT. Data are shown as mean ± s.e.m.. Spot counts were normalized by subtracting the negative control (DMSO), and values lower than negative control were set to zero. Statistical significance was evaluated using two-tailed *t*-tests, with $P > 0.05$ considered non-significant (ns). **d**. Distribution of the HLA-A*02-restricted LMP1 epitope and escape variants across EBV lineages. Maximum-likelihood phylogeny of 1,852 Type 1 EBV genome sequences, annotated by host geographic region, disease status, and the previously reported LMP1 A*02-restricted epitope (YLLEMLWRL) and escape variant (YFLEILWRL). In the YLL epitope track, red indicates epitope-loss variants, and blue indicates epitope or non-escape variants. Four major lineages are highlighted: the clonally expanded 85841G-carrying high-risk lineage (C1, red), the non-clonal 85841G-carrying high-risk lineage (X1, orange), the Southern China 85841A-carrying high-risk lineage (S1, blue), and the Northern China 85841G-carrying non-high-risk lineage (N1, green).

iii. A possible explanation for the suggestive A*02:07 × high-risk EBV interaction

HLA class I molecules are central to antiviral immunity, shaping CD8$^+$ T cell recognition of infected cells. Thus, unlike EBV risk variants where "high-risk" naturally implies increased pathogenic potential, it is not immediately intuitive that a host HLA allele would show a risk-increasing direction of effect. It is worth noting that A*02:07 has consistently emerged as one of the strongest NPC risk-associated HLA alleles[23,24,26,27,68] and is often the only significant risk allele identified in human GWAS of NPC, whereas other top HLA signals (minor alleles) tend to be protective[26]. This pattern suggests that A*02:07 may be less effective in presenting EBV antigens compared with other HLA-A alleles. When individuals carrying

A*02:07 are infected with highly lytic, high-risk 85841G strains, as seen in the M81 EBV strain[69], the effect of this limited immune recognition may be further exacerbated, leading to increased susceptibility to EBV-driven oncogenesis. Accordingly, the A*02:07-associated risk increase, observed as a suggestive A*02:07 × high-risk EBV interaction on NPC risk, could reflect a spillover of the dominant high-risk EBV effect. However, given that the interaction signal is only suggestive and our epitope analyses for A*02:07 are exploratory, the EBV × A*02:07 interaction and its mechanism remain to be elucidated in future studies. Thus, we view this narrative as a plausible hypothesis rather than a definitive conclusion.

In response, we have toned down our statements regarding EBV × A*02:07 interaction in the revised **Results** (**Lines 189-195** and **293-299**) and have added the results of A*02:07 binding and epitope assays, together with a discussion of this hypothesis, into supplementary materials (**Supplementary Text**: **Lines 293-396**).

2. Why is 85841A so protective against NPC among the high-risk strains? Maybe more to the point, why is 85841G bad, even in the presence of A*11:01? Is it because it is in LD with bad alleles at BALF2 and 7048? Is there anything known regarding the functional basis for the BALF2-CCT or 7048C association with high-risk EBV strains? Could the authors briefly address this in the Discussion or Introduction? It appears that the variants marking high risk EBV strains have a particularly stronger effect on risk of NPC (even independent of A*11:01).

Response: We greatly appreciate your insightful questions and fully agree that high-risk-85841G EBV strains have strong risk effect on NPC risk, independent of HLA-A alleles and even in the presence of A*11:01. This enhanced carcinogenicity likely involves at least two processes: (i) sufficient EBV activity and epithelial reseeding despite immune surveillance at a pre-tumor stage, and (ii) viral properties that support persistence and oncogenic cellular programs in infected epithelial cells.

Consistent with this view, the representative high-risk-85841G strain, the NPC-derived M81 EBV, has been reported to display increased epithelial tropism and elevated lytic replication in both B cells and epithelial cells, compatible with enhanced reactivation and epithelial reseeding[69]. Sequence variation in other loci enriched in high-risk lineages (for example, Zta and EBER2) has also been implicated in viral replication, cell growth, and immune modulation[69-71]. Even though EBNA3B variant effects are not yet defined, the established functions of EBNA3B suggest that its sequence variation may contribute to strain-specific oncogenic and T cell-attraction properties[72,73]. By contrast, HLA-restricted EBV-specific T cell immunity is expected to primarily affect immune control. Although 85841G generates a novel A*11:01-restricted epitope, these high-risk strains simultaneously carry epitope-loss mutations in the two immunodominant A*11:01 epitopes AVF and IVT, thereby reducing overall A*11:01-mediated T cell surveillance[74-77]. These linked variants within the high-risk-85841G genetic background may act in concert to shape the strain-level phenotypes, but their individual contributions remain to be disentangled.

In light of your comment, we have incorporated these possible mechanisms underlying the enhanced carcinogenic potential of high-risk EBV strains into Discussion (**Lines 587-599**).

3. It would be helpful if the authors could provide a table of pairwise linkage disequilibrium values for SNP 85841, BALF2, and 7048. This could be provided as supplementary material. Figure 3 seems to indicate that 85841G further refines the high-risk phenotype.

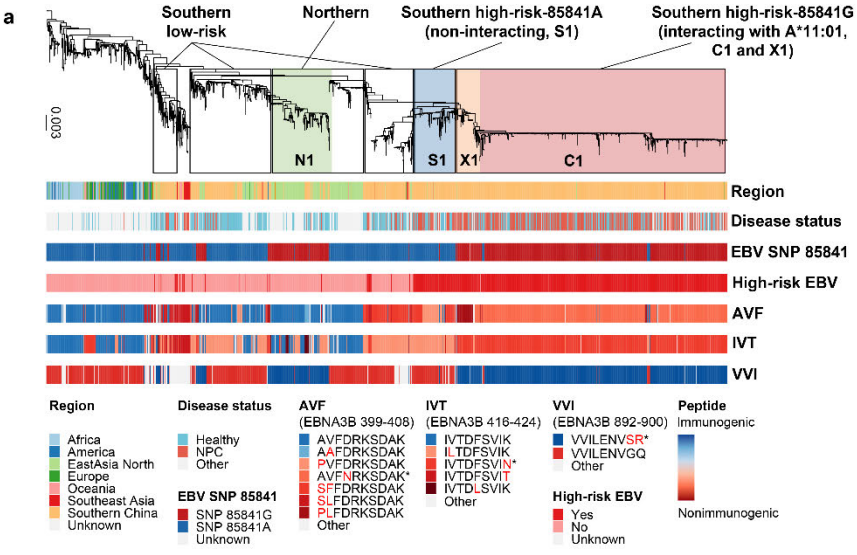
Response: Thank you for this helpful suggestion. We have now added **Supplementary Figure 2** showing the moderate linkage disequilibrium (LD) between SNP 85841 and the high-risk EBV markers BALF2-CCT and 7048C, and the high LD between BALF2-CCT and 7048C (**Response Figure 15**). Consistent with your understanding, within the high-risk viral background defined by BALF2-CCT or 7048C, 85841G further refines the viral haplotype that confers the strongest NPC risk and shows the most significant interaction with HLA-A*11:01 (**Figure 3**, nested lineage structure shown in **Supplementary Figure 2**).

Variant 1	Variant 2	r^2	D'
EBNA3B-85841G	BALF2-CCT	0.48	0.69
EBNA3B-85841G	EBER2-7048C	0.56	0.77
BALF2-CCT	EBER2-7048C	0.70	0.86

Response Figure 15. Pairwise linkage disequilibrium values (r^2 and D') for EBNA3B-85841G, BALF2-CCT, and EBER2-7048C in EBV genomes from Southern China.

4. What is the distribution of the two immunodominant EBNA3B epitope (IVT and AVF) variants among high risk and lower risk EBV strains? The authors state that "Paradoxically, individuals carrying A*11:01 had lower NPC risk when infected with EBV strains carrying these epitope-loss mutations than non-carriers", but was this in the high-risk EBV-infected group or in all subjects? Suppl Table 12 seems to indicate this was in the high-risk only grouping. This data needs more discussion. Why would the IVT and AVF A*11:01-binding alleles confer susceptibility to NPC, whereas the binding allele of VVI confers protection? The authors might want to concentrate on this, and provide one or more hypotheses. Is it possible to study CD8+ T cells from individuals with high-risk EBV who are A*11:01+ to identify differences in responses to these three epitopes (cytokine production, killing of infected cells, etc)? This would be especially interesting!

Response: Thank you for raising this important question regarding the protective effect of A*11:01 and T cell responses to EBNA3B epitopes. In our dataset, IVT and AVF epitopes are almost completely lost in all high-risk strains (intact IVT frequency: 0.6%; intact AVF frequency: 2%), including 85841G-carrying lineages (C1 and X1) as well as the 85841A-carrying high-risk lineage S1, whereas low-risk Southern China strains partially preserve the AVF epitope (intact AVF frequency: 35%) and nearly completely lack IVT epitope (intact IVT frequency: 3%; **Response Figure 16a-b**, revised as **Supplementary Figure 14**; included in **Results**: **Lines 304-309 and Discussion**: **Lines 602-607**). This lineage-specific pattern is consistent with previous studies evaluating AVF/IVT-specific T cell responses in Chinese donors[74-76]. As the reviewer inferred, the strong protective effect of A*11:01 was observed primarily among individuals infected with high-risk 85841G-positive EBV, while a suggestive weaker protective effect was detected in those carrying low-risk strains in our meta-analysis combining the discovery and validation datasets (included in **Results**: **Lines 181-182**; **Response Figure 16c**, revised as **Supplementary Figure 7a**). We speculate that the preservation of AVF in some low-risk strains may contribute to the protection observed for A*11:01 in this group (**Lines 602-604**).



b

Subpopulation carrying different EBV strains	Frequency of AVF epitope	Frequency of IVT epitope	Frequency of VVI epitope
Southern low-risk	34.9%	2.9%	19.4%
Southern high-risk (S1, C1 and X1)	2.3%	0.6%	82.0%
Southern high-risk-85841G (C1 and X1)	1.2%	0.1%	97.0%
Northern	87.9%	64.7%	47.8%

c

Subpopulation carrying different EBV strains	HLA-A*11:01 OR (95% CI)	P
Low-risk EBV	0.68 (0.54, 0.85)	7.18E-04
High-risk EBV	0.58 (0.49, 0.69)	1.05E-09

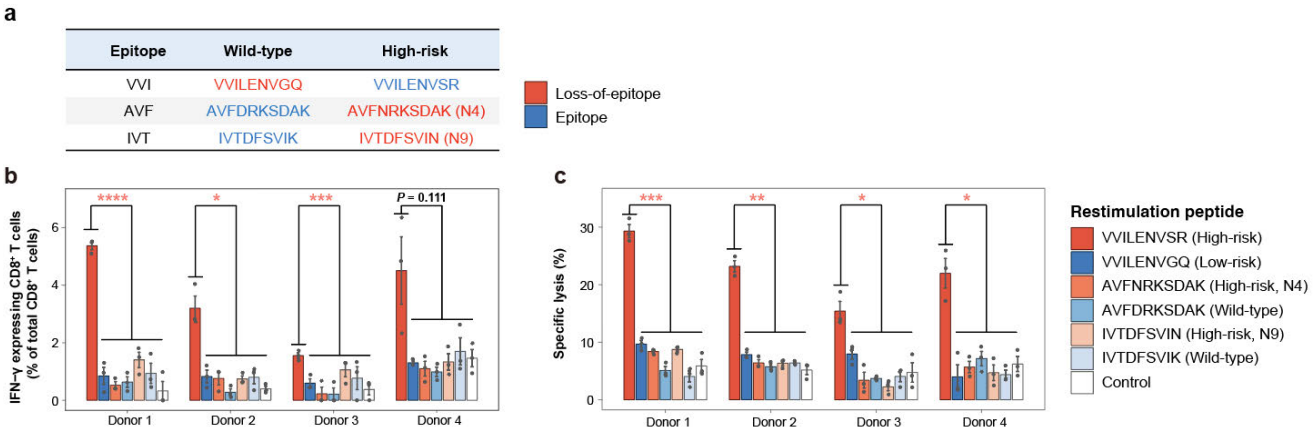
Response Figure 16. Distribution of EBNA3B epitopes AVF, IVT, and VVI. a, Maximum-likelihood phylogeny of 1,852 Type 1 EBV genome sequences, annotated by host geographic region, disease status, EBV SNP 85841 genotype, high-risk EBV subtype (defined by BALF2-CCT or 7048C) and the A*11:01-restricted EBNA3B epitopes AVF, IVT and VVI. Four major lineages are highlighted: the clonally expanded 85841G-carrying high-risk lineage (C1, red), the non-clonal 85841G-carrying high-risk lineage (X1, orange), the Southern China 85841A-carrying high-risk lineage (S1, blue) and the Northern China 85841G-carrying non-high-risk lineage (N1, green). In the AVF, IVT and VVI tracks, colors denote peptide immunogenicity defined by $CD8^+$ T cell responses (blue, immunogenic; red, non-immunogenic). Amino-acid changes relative to the reference epitope sequence are highlighted in red, and asterisks (*) indicate epitope variants

present in the clonally expanded 85841G-carrying high-risk lineage C1. **b**, Frequencies of the A*11:01-restricted EBNA3B epitopes AVF, IVT and VVI across subpopulations carrying different EBV strains. **c**, Associations of HLA-A*11:01 with NPC stratified by EBV subtypes in Southern China. Associations were estimated using logistic regression models adjusted for age, sex, and the first four human MDS components, in the meta-analysis of the discovery and validation datasets.

Following your suggestion, we performed $CD8^+$ T cell assays to address why the 85841G-encoded peptide VVILENVSR confers strong protection in the presence of A*11:01.

i. 85841G-encoded peptide elicits specific A*11:01-restricted T cell responses.

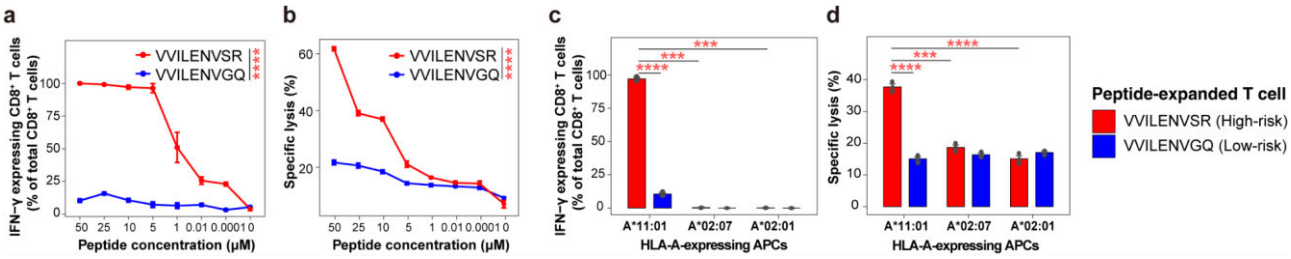
Using PBMCs from A*11:01$^+$ healthy individuals infected with high-risk-85841G EBV strains, we expanded polyclonal T cells with three EBNA3B peptide pairs (representing three epitopes and their epitope-loss variants): AVF *versus* AVF/N4, IVT *versus* IVT/N9, and VVILENVSR *versus* VVILENVGQ (**Response Figure 17a**). These T cells were then cocultured with COS-7 cells engineered to express HLA-A*11:01 and β_2M. The COS-7 cell line, a fibroblast-like cell line derived from African green monkey and lacking endogenous human HLA expression, is commonly used as the antigen-presenting cell (APC) after HLA-β_2M transfection[78-81]. IFN-γ intracellular cytokine staining (ICS) showed that the 85841G-encoded peptide VVILENVSR elicited markedly stronger $CD8^+$ T cell activation than either AVF or IVT, and their corresponding epitope-loss peptides (AVF/N4 and IVT/N9), whereas the 85841A-encoded peptide VVILENVGQ did not activate peptide-specific $CD8^+$ T cell responses (**Response Figure 17b**). Consistently, cytotoxicity assays showed specific killing of A*11:01-expressing target cells pulsed with VVILENVSR, but not with any of the other five peptides (**Response Figure 17c**). In line with prior reports, given AVF and IVT epitope-loss variants in high-risk-85841G EBV, we observed little AVF- or IVT-specific responses among A*11:01 carriers infected with this viral strain. These data indicate that in A*11:01$^+$ individuals infected with high-risk-85841G EBV strains lacking AVF and IVT, the 85841G-encoded VVILENVSR peptide is capable of triggering specific $CD8^+$ T cell responses (**Figure 4c-d**; **Lines 330-341**).



Response Figure 17. 85841G-encoded peptide VVILENVSR elicits A*11:01-restricted T cell responses. a, EBNA3B peptide pairs used in $CD8^+$ T cell experiments. **b-c,** Specific T cell responses against three EBNA3B peptide pairs were evaluated by $CD8^+$ T cell activation (b) and cytotoxicity (c) after polyclonal T cell coculture with peptide-pulsed HLA-A*11:01-expressing COS-7 cells in four A*11:01$^+$ healthy individuals infected with high-risk-85841G EBV. Polyclonal T cells were expanded from PBMCs using the three peptide pairs and then co-cultured with luciferase-expressing HLA-A*11:01$^+$ COS-7 target cells pulsed with each peptide (10 μM) as indicated. In (b-c), data are presented as mean ± s.e.m.. (n=3). In (b), $CD8^+$ T cell activation was assessed by IFN-γ intracellular cytokine staining following co-culture. In (c), specific cell lysis was measured by the luciferase activity of the remaining viable target cells after 24h of co-culture. Statistical analysis was performed using two-tailed *t*-tests. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

ii. Peptide titration confirms VVILENVSR is a bona fide A*11-restricted epitope.

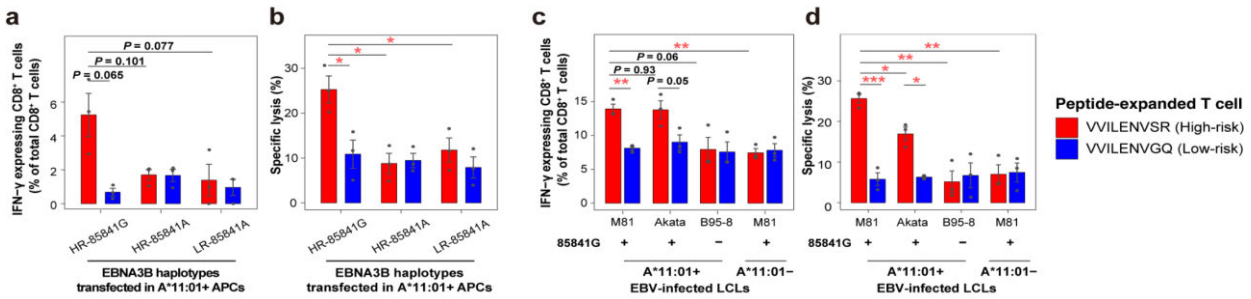
We performed peptide titration and observed dose-dependent T cell activation and cytotoxicity for VVILENVSR, but not VVILENVGQ, when presented by A*11:01-expressing COS-7 APCs (**Response Figure 18a-b**). This T cell response was HLA-A*11:01-restricted, as APCs expressing HLA-A*02:07 or A*02:01 did not support robust VVILENVSR-specific activation or killing (**Response Figure 18c-d**). These results confirm that VVILENVSR is a bona fide A*11:01-restricted epitope to elicit specific T cell responses (**Figure 4e-h**; **Lines 342-345**).



Response Figure 18. 85841G-encoded peptide VVILENVSR is a bona fide A*11:01-restricted epitope. a-b, Peptide titration showing dose-dependent CD8$^+$ T cell activation (a) and cytotoxicity (b). Primary T cells expanded with either VVILENVSR (85841G) or VVILENVGQ (85841A) from PBMCs of A*11:01$^+$ individuals infected with high-risk-85841G EBV were co-cultured with luciferase and HLA-A*11:01 co-expressing COS-7 target cells pulsed with the corresponding peptides at different concentrations. **c-d,** HLA restriction assays using COS-7 cells expressing HLA-A*11:01, A*02:07 or A*02:01, pulsed with VVILENVSR (85841G) or VVILENVGQ (85841A). CD8$^+$ T cell activation (c) and cytotoxicity (d) were evaluated by co-culture of peptide-specific T cells with COS-7 target cells co-expressing luciferase and different HLA-A molecules and pulsed with either VVILENVSR or VVILENVGQ (10 μM) as indicated. Peptide-specific T cells were expanded with either VVILENVSR (85841G) or VVILENVGQ (85841A) from PBMCs of A*11:01$^+$ individuals infected with high-risk-85841G EBV. In (a, c), T cell activation was assessed by IFN-γ intracellular cytokine staining following co-culture. In (b, d), specific cell lysis was measured by the luciferase activity of the remaining viable target cells after 24h of co-culture. Data are presented as mean ± s.e.m. (n=3). Statistical analysis was performed using two-way repeated-measures ANOVA (a-b) and two-tailed *t*-tests (c-d). *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

iii. VVI-specific T cells target M81 (85841G$^+$) LCLs

Furthermore, endogenous antigen-presentation experiments using A*11:01$^+$ APCs expressing EBNA3B haplotypes indicate that only the high-risk-85841G EBNA3B-encoded peptide, but not the peptide from either high-risk-85841A or the low-risk EBNA3B haplotype, can be processed and presented to trigger VVILENVSR-specific CD8$^+$ T cell cytotoxicity (**Response Figure 19a-b**, revised as **Figure 4i-j**; **Lines 346-349**). Consistent with these findings, VVILENVSR-specific CD8$^+$ T cells showed stronger IFN-γ-producing responses to A*11:01$^+$ lymphoblastoid cell lines (LCLs) transformed by the NPC-derived M81 EBV (high-risk-85841G$^+$) as well as Akata EBV (N1 lineage), which also carries the 85841G-encoded VVILENVSR epitope (**Response Figure 19c**, revised as **Figure 4k**). In contrast, VVILENVGQ-expanded T cells did not show specific responses. These VVILENVSR-specific T cells also selectively lysed A*11:01$^+$ LCLs transformed by M81 or Akata, with substantially reduced killing of A*11:01$^+$ LCLs transformed by the infectious mononucleosis-derived B95-8 strain (85841A$^+$) or A*11:01$^-$ LCLs (**Response Figure 19d**, revised as **Figure 4l**; **Lines 350-356**).

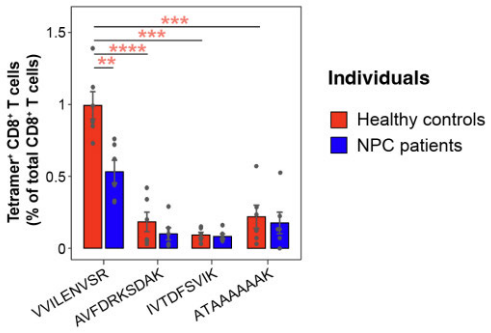


Response Figure 19. Peptide-specific T cell responses to endogenously expressed EBNA3B haplotypes. a-b, Specific T cell responses to endogenously expressed EBNA3B haplotypes were evaluated by CD8$^+$ T cell activation (a) and T cell cytotoxicity (b) assays. Luciferase and HLA-A*11:01 co-expressing COS-7 target cells were transfected with EBNA3B haplotypes from different EBV lineages, representing high-risk (HR)-85841G, high-risk (HR)-85841A, and low-risk (LR)-85841A EBV from Southern China. These COS-7 target cells were co-cultured with primary T cells expanded with either VVILENVSR (85841G) or VVILENVGQ (85841A) from PBMCs of A*11:01$^+$ individuals infected with high-risk-85841G EBV as indicated. **c,d,** Specific T cell responses to EBV-transformed lymphoblastoid cell lines (LCLs) established from A*11:01$^+$ healthy donors (n=3) were evaluated by CD8$^+$ T cell activation (c) and cytotoxicity (d) assays. LCLs were transformed by B95-8 (low-risk-85841A), Akata (low-risk-85841G, Northern China N1 lineage), or M81 (Southern China high-risk-85841G) EBV. Target LCLs were co-cultured with primary T cells expanded from PBMCs of heterologous A*11:01$^+$ healthy donors infected with high-risk-85841G EBV with either VVILENVSR (85841G) or VVILENVGQ (85841A), as indicated. M81-transformed LCLs established from A*11:01$^-$ healthy donors (n=3) were included as controls. LCLs were labeled by anti-CD19 staining and cell death was assessed by Annexin V and SYTOX AADvanced staining to differentiate viable, dying and dead cells. LCLs co-cultured without peptide-specific T cells served as the baseline control. In (a,c), T cell activation was assessed by IFN-γ intracellular cytokine staining following co-culture. In (b), specific cell lysis was measured by the luciferase activity of the remaining viable target cells after 48h of co-culture. In (d), specific cell lysis was measured by flow cytometry and quantified as the percentage of dying or dead LCLs after 24 h of coculture. Data are presented as mean ± s.e.m. (n=3). Statistical analysis was performed using two-

tailed *t*-tests. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

iv. Such T cell responses are substantially diminished in NPC patients.

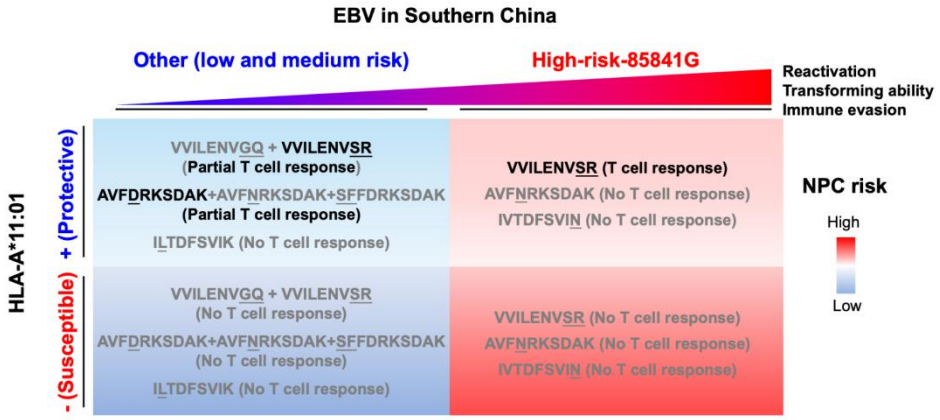
Moreover, among individuals infected with high-risk-85841G EBV, the VVILENVSR-specific A*11:01-tetramer$^+$ CD8$^+$ T cells were markedly reduced in NPC cases compared with healthy EBV carriers (**Response Figure 20**, revised as **Figure 4m**; **Lines 357-359**), consistent with prior reports of generally reduced EBV-specific T cell immunity in NPC[82-85]. This reduction coincided with the heightened EBV-specific IgA levels in NPC cases, suggesting a disease-associated shift toward increased viral antigen exposure and diminished virus-specific cellular immunity.



Response Figure 20. Frequency of HLA-A*11:01 tetramer$^+$ EBV-specific CD8$^+$ T cells in NPC patients and healthy controls. Tetramer staining of PBMCs from NPC patients (n=6) and healthy controls (n=6) infected with high-risk-85841G EBV. PBMCs were stained with a PE-conjugated HLA-A*11:01 tetramer loaded with EBV-derived peptides (VVILENVSR, AVFDRKSDAK, or IVTDFSVIK), and an APC-conjugated HLA-A*11:01 tetramer loaded with ATAAAAAAK as a negative control. The frequency of tetramer$^+$ CD8$^+$ T cells is shown as mean ± s.e.m.. Statistical analysis was performed using two-tailed *t*-tests. **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

Together, these results establish VVILENVSR as a functional HLA-A*11:01-restricted epitope generated by 85841G and show that it elicits specific CD8$^+$ T cell responses against high-risk-85841G EBV-transformed B lymphocytes despite the loss of the immunodominant AVF and IVT epitopes (**Response Figure 20**, revised as **Figure 4c-m**). IVT and AVF are established immunodominant A*11:01-restricted EBNA3B epitopes[74,75], and their loss in high-risk EBV strains diminishes the strong A*11:01-mediated immune surveillance. The 85841G mutation creates a new A*11:01-restricted epitope (VVILENVSR), which partially restores immune recognition and confers protection specifically in A*11:01 carriers infected with high-risk EBV (**Response Figure 21**; **Lines 360-362**).

We have incorporated these new analyses and experiments into the revised manuscript (**Results**: **Lines 329-362**; **Methods file**: **Lines 416-522**) and updated the corresponding figures (**Figure 4**). We have also revised the original wording ("Paradoxically…") to clarify that A*11:01 likely confers protection in carriers infected with high-risk-85841G EBV through alternative epitopes given the loss of IVT and AVF epitopes to avoid misunderstanding (**Lines 308-311**).



Response Figure 21. Schematic summary of EBNA3B epitope immunorecognition patterns underlying the A*11:01 × EBV interaction effect on NPC risk.

5. It would be helpful if the authors could please be very careful to note when individuals infected with high risk EBV strains (vs all strains) are being used. For example, were all 747 NPC cases and 1,251 controls (discovery), and all 644 cases and 880 controls in the validation infected with high-risk EBV strains (lines 117-122)? I realize that this information is sometimes given in suppl material (in this case, in Suppl Fig. 1), but it would help to have that clarified throughout the text as well. Line 211 is another example, and there are more.

Response: Thank you for pointing this out. We confirm that there was no strain-based selection: all NPC cases and controls in both the discovery (747 cases/1,251 controls) and validation (644 cases/880 controls) datasets were recruited without knowing EBV strain status. EBV genotyping and WGS were done after recruitment. To address your concern, we have revised the main text (**Lines 119-120** and **206-209**) and Methods (**Methods file**: **Lines 20-21**) to state this explicitly and revised the figure legends (**Figures 1** and **2**) showing the numbers of high-risk *versus* non-high-risk EBV carriers in each analysis.

6. If you analyze all subjects with non-high risk EBV (regardless of the variant at 85841), does A*11:01 show significant protection (it is not possible to tell from Fig. 3, but looks as if A*11:01 might be protective in the non-high risk groupings)? This could be done as a meta-analysis in order to provide power. If it does show significant protection in the non-high risk group, then one might conclude that A*11:01 renders some protection in addition to the interaction with 85841G.

Response: Thank you for this insightful question. Following your suggestion, we performed a meta-analysis to test whether HLA-A*11:01 shows a protective effect among individuals infected with non-high-risk EBV strains, irrespective of the variant at EBNA3B 85841. We observed a protective trend of A*11:01 in this non-high-risk EBV subgroup, although this association ($P = 7.18\times10^{-4}$) was much weaker than the A*11:01 association observed within the high-risk EBV subgroup ($P = 1.05\times10^{-9}$) and did not reach the conventional genome-wide significance ($P < 5 \times 10^{-8}$; **Response Figure 16c**, revised as **Supplementary Figure 7a**). This result suggests that A*11:01 provides some protection in the context of non-high-risk EBV infection (**Lines 181-182**). Together with prior studies[74-77] and our observation that the AVF epitope is preserved in 34.9% of non-high-risk EBV strains in Southern China (**Response Figure 16a-b**, revised as **Supplementary Figure 14**), this protection could plausibly be mediated by A*11:01-restricted recognition of the AVF epitope among A*11:01$^+$ individuals carrying the non-high-risk EBV. We have added these results (**Supplementary Figures 7a** and **14**) and revised the Discussion (**Lines 602-604**) accordingly.

7. Can the authors conclude anything regarding selection pressures on the host or the virus based on data described in the section on geographical co-enrichment of interacting host-viral factors? How much of the apparent viral spread could be due to (1) an ascertainment bias towards NPC cases in the samples that were sequenced, (2) variability in human demographics (war, famine) in South China over the past 4000 years that may have increased representation among carriers of particular EBV strains, or (3) an improved transmission rate of the C1 virus from individual to individual?

Response: Thank you very much for these thoughtful comments. We address each point below.

i. Potential ascertainment bias

We agree that oversampling NPC cases could bias evolutionary inferences. To address this concern, we re-conducted the evolutionary analyses using EBV genomes from healthy controls only in Southern China (excluding all NPC cases; **Supplementary Text**: **Lines 441-456**). The phylogenetic features supporting the clonal expansion of C1, e.g., short tip lengths and high sequence identity, remain robust under this control-only analysis. In addition, our selection analyses of the C1 lineage were also conducted with Southern China EBV genomes from healthy individuals only (**Supplementary Text**: **Lines 421-439**). Consistent with these phylogenetic findings, in our data the C1 lineage is markedly enriched among healthy individuals from Southern China (31% on average) but is rare among healthy individuals from Northern China (< 1%), supporting a regional expansion rather than an artifact of oversampling cases (**Supplementary Table 14**; **Supplementary Text**: **Lines 421-456**).

ii. Relationship between C1 evolution and human demographic history

This is an excellent question. At present, information is very limited because ancient EBV genomes are

unavailable. EBV resides predominantly in B cells and oropharyngeal epithelial tissues and is unlikely to be well preserved in most archaeological samples. After consulting with Prof. Qiaomei Fu and surveying available ancient human datasets, we did not find usable EBV genomes, and the only reported ancient EBV fragment[86] has too few high-quality reads for any evolutionary analysis.

Nevertheless, historical and genetic data suggest multiple North-South migration waves over the past 3000 to 5000 years[87], which could have shaped EBV lineage structure through recombination, founder effects and clonal drift. The earlier work by Profs Masucci, Rickinson and colleagues showed that HLA-restricted CTL pressure can drive the recurrent selection for loss of immunodominant A*11:01-restricted EBNA3B epitopes in multiple high-A*11:01 populations[74-77], providing a precedent for host-virus coadaptation.

iii. Causes of the increased frequency of the C1 lineage

The precise selective forces remain uncertain. EBV evolution is generally shaped by traits enhancing its persistence in hosts and transmission between hosts, rather than oncogenicity. Sequence features of the C1 recombinant block that harbor the loss of two immunodominant AVF/IVT A*11:01-epitopes and other functional changes (e.g., in EBNA3B and Zta) affecting viral replication, persistence, and immune modulation[69,72] could have been advantageous in Southern Chinese populations. Demographic expansion and drift may also have contributed. Our current data cannot pinpoint the exact target of the selective force in the EBV genome. We have added these discussions in the revised Discussion (**Lines 632-645**).

8. Do the authors think that 85841G arose and was maintained because it enhances fitness of the virus or is carried by the sweep of another advantageous SNP? It seems paradoxical that the C1 clade is considered to be selectively favored while at the same time associated with a reduced viral load due to the recognition of 85841G by HLA-A*11:01. Would the relationship of recombinant clades C1 (and X1) to other clades be altered if the ML tree (Fig 6) was constructed following exclusion of the recombinant N1-derived segment (about 30-40kb)?

Response: Thank you for these insightful questions. We have addressed each point in detail below:

i. The targets of natural selection

At present, the selective target remains unresolved, but the A*11:01 × 85841G interaction that protects against NPC is unlikely to have been directly selected. NPC occurs late in life, is rare at the population level, and this interaction enhances T cell recognition and viral control. Thus, it is probably a byproduct rather than a direct driver of selection.

We agree that your suggestion of hitchhiking is highly plausible. The 85841G variant lies within the segment that also carries the IVT/AVF epitope-loss mutations[74,75] (**Response Figure 16,** revised as **Supplementary Figure 14)** and other functional variants, e.g., in EBNA3B and Zta, potentially influencing viral persistence, reactivation and immune modulation[69,72]. In this context, selection acting on any of these features could have increased the frequency of the entire block, including 85841G, through linkage. Although we think the 85841G × A*11:01 interaction is unlikely to be the direct target of selection, our data cannot formally distinguish direct selection from hitchhiking. We have clarified this point and added a brief discussion of potential selection targets in the revised manuscript (**Lines 634-643**).

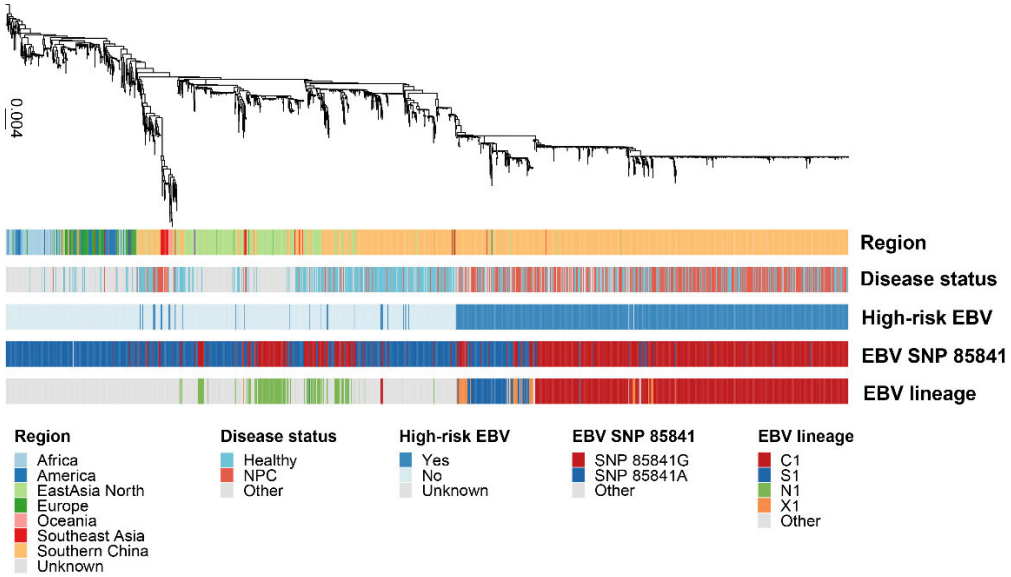
ii. Reconciling C1 expansion with HLA-A*11:01-dependent viral-load effects

This is an interesting question. The selective forces driving the population expansion of the C1 lineage remain unclear, and salivary viral load may not directly reflect population-level viral fitness. In our data, epitope-loss variants affecting two immunodominant A*11:01-restricted EBNA3B epitopes (AVF and IVT) are nearly fixed across the C1 lineage, consistent with these being the primary A*11:01-related immune escape events[74-77]. Against this background, the 85841G mutation introduces the VVI epitope, partially restoring A*11:01-restricted immune control and associating with slightly lower viral load and reduced NPC risk specifically in A*11:01 carriers when compared with non-carriers. This pattern suggests AVF/IVT loss and VVI gain may fine-tune A*11:01-restricted immune pressure, without necessarily causing a detrimental effect on viral transmission or population-level fitness. Furthermore, because these epitopes are HLA-A*11:01-restricted, any epitope-associated effects on immune control, including the mild viral-load differences we observe, are expected to be A*11:01-dependent and therefore would not operate directly in non-carriers. Altogether, C1 expansion is a population-level outcome that reflects the net fitness effects across host backgrounds and genome-wide viral variation, while VVI is more plausibly

hitchhiking with other linked variants than driving expansion on its own. We have included a discussion of selection and potential hitchhiking in the revised Discussion (**Lines 634-643**).

iii. The evolutionary history of the C1/X1 after excluding the N1-derived recombinant regions

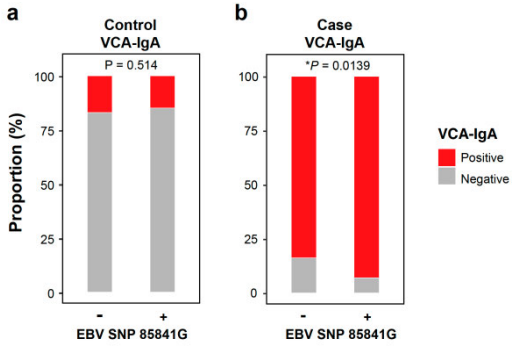
Following the reviewer's suggestion, we reconstructed the phylogeny after excluding the N1-derived recombinant region (~54-130 kb). The C1 and X1 lineages still cluster with Southern China NPC high-risk S1 strains, consistent with the original recombination analysis (**Response Figure 22**). We have added this tree and clarified this point in supplementary materials (**Supplementary Text**: **Lines 484-496**) of the revised manuscript.



Response Figure 22. Maximum-likelihood tree of EBV genomes after exclusion of the N1-derived recombinant region (~54-130 kb), annotated by host geographic origin, disease status, high-risk EBV subtype (defined by BALF2-CCT or 7048C), EBV SNP 85841 genotype, and EBV lineages. The EBV lineages are the same as those defined in Figure 6: the clonally expanded 85841G-carrying high-risk lineage (C1, red), the 85841A-carrying high-risk lineage predominantly from Southern China (S1, blue), the 85841G-carrying non-high-risk lineage predominantly from Northern China (N1, green), and the non-clonal 85841G-carrying high-risk lineage (X1, orange).

9. Does SNP 85841 associate with the presence of IgA against VCA among NPC negative subjects?

Response: Thank you for this important question. We examined the association between SNP 85841G and EBV VCA-IgA seropositivity among healthy individuals (n = 792). The majority of individuals (~84%) were VCA-IgA negative, and we did not observe an association with 85841G (**Response Figure 23a)**. This is consistent with tight immune control and often asymptomatic EBV reactivation in most healthy carriers. As an exploratory analysis, we also tested this association in NPC patients (n = 483) and observed a different pattern. As expected, the majority of NPC cases (~91%) were positive for VCA-IgA. Under this disease state, carriers of 85841G^{+} strains showed significantly higher VCA-IgA positivity (P = 0.01, **Response Figure 23b**), likely reflecting the substantially heightened EBV reactivation and antigen exposure in NPC that makes strain-associated differences in lytic activity more detectable than in healthy carriers.



Response Figure 23. Distribution of EBV-specific IgA serostatus by EBV SNP 85841. Stacked bar plots showing the distribution of VCA-IgA serostatus (positive or negative) among healthy controls (**a**) and NPC cases (**b**), stratified by infection with 85841G-positive EBV strain. Differences between the two groups were determined using χ^2 tests. *, $P < 0.05$.

10. Lines 172-174 and Lines 181-184: These two sentences seem to contradict one another. Is there an independent effect of A*02:07 after conditioning on A*11:01, or not?

Response: Thank you for pointing out this issue. In brief, we do not find a robust, genome-wide significant, independent interaction between A*02:07 and high-risk EBV after conditioning on A*11:01. Specifically, in the discovery dataset, conditioning on A*11:01 or the top associated SNP (an A*11:01 tagging SNP) abolished all remaining interaction signals across the genome (**Supplementary Figure 12**). In the meta-analysis of the discovery and validation datasets, the A*02:07 × high-risk EBV interaction did not remain genome-wide significant ($P < 5 \times 10^{-8}$) after conditioning on A*11:01 (before conditioning, $P = 1.72 \times 10^{-8}$; after conditioning, $P = 4.86 \times 10^{-4}$; **Supplementary Table 5**). Therefore, the A*02:07 interaction signal remains suggestive and will require validation in larger future studies. We have clarified these results in the revised text (**Lines 189-195** and **293-299**) to avoid misunderstanding.

11. Lines 219-224: The authors mention A*11:01 specifically in lines 219-220 with no mention of A*02:07, but the next sentence seems to suggest that conditioning on viral SNP 85841 also abolished the effect of A*02:07?

Response: Thank you for pointing out this wording issue. These sentences were intended to describe the effect of A*11:01, not A*02:07 or other HLA-A alleles. We have corrected the wording in the revised manuscript (**Lines 227-228** and **233-234**).

12. Line 317: Does "the HLA-interacting EBV strains" mean high-risk EBV strains with 85841G? It might be a bit confusing to change terminology, if indeed these mean the same thing. A similar issue applies to line 323.

Response: Thank you for pointing out this ambiguity. In the original text, "HLA-interacting EBV strains" was intended to refer to the high-risk EBV strains carrying the 85841G variant. To maintain consistency and avoid confusion, we have revised the text to use the term "high-risk-85841G" EBV strains throughout the manuscript.

Referee #3:

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #4:

This manuscript by Chen et al. the authors use a paired human–EBV genomic dataset comprising >3,500 individuals and provide a stepwise genome-to-genome analysis to uncover host–viral genetic interactions underlying nasopharyngeal carcinoma (NPC) in a specific southern Chinese population, as the title implies. Using risk association with BALF-CCT, which is a methodology that has been used before by previous authors included many of the authors in this group, the authors identify HLA-A11:01 as a protective and HLA-A02:07 as deleterious host alleles that putatively interacts with EBV SNP 85841G in EBNA3B, leading to a "dual-risk" subgroup that apparently explains ~47% of NPC burden in Southern China. The validate this on smaller cohorts in China and Singapore. This is an interesting association study, but remains an association study at that, and seems to incorporate a number of previously known lines of evidence with the main novel discovery here of the EBV high risk SNP described here. There are a number of issues I wanted to raise in this study that needs to be addressed for the strength of this novelty to be upheld in this study.

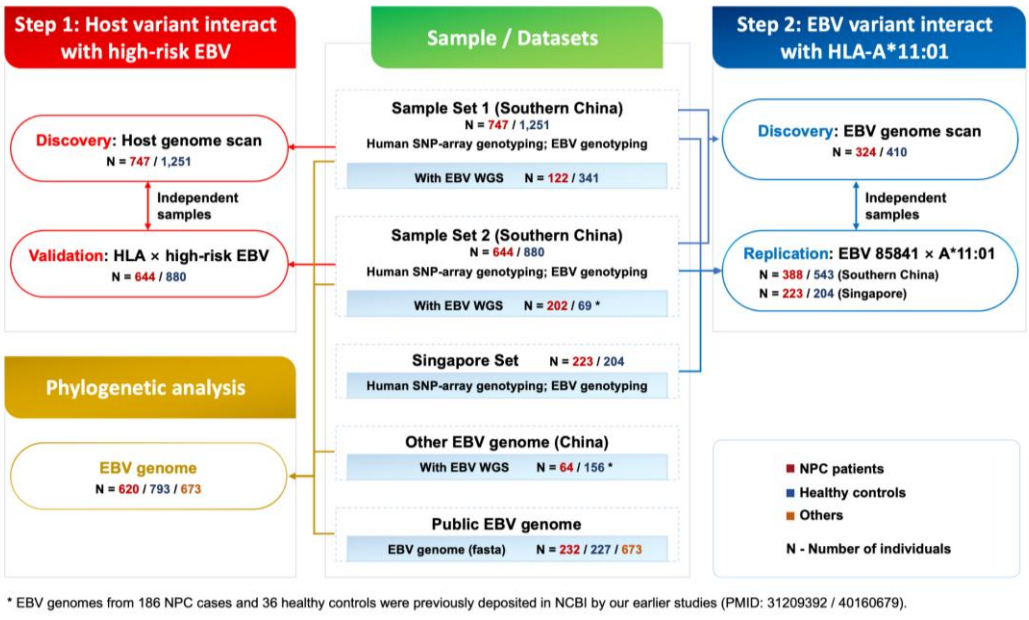
Response: We are grateful for your constructive summary and for the positive comments on our dataset, stepwise genome-to-genome framework, and cross-cohort validation. We understand your underlying

concern: given the substantial recent progress in NPC genetics (including our prior work), it may not have been sufficiently clear in the original submission what is truly new here and whether the conclusions go beyond association. We have taken substantial efforts in the revised work, in particular, we have (i) performed additional functional experiments to establish mechanistic basis for the genetic interaction between HLA-A*11:01 and EBNA3B 85841G, and (ii) revised the text to make the conceptual advances more explicit. We address each point in detail below.

1. The dataset used here likely comprises previous data that has been published members of this group before, and I have trouble understanding the resource value of this 'largest dataset eve', as claimed here. How much of this is data is new and how much overlaps with previous analyses.

Response: Thank you for pointing out this gap in the original submission. As this study integrates newly generated data with publicly available EBV genomes, we agree that it would be helpful to clarify the relationship between new data with our prior datasets. We have now added a schematic chart (**Response Figure 24** below; revised as **Supplementary Figure 1**) summarizing the dataset composition across analyses and indicating which components are newly generated, publicly sourced, or overlapping with our prior publications, and we have clarified this point in the Methods (**Methods file**: **51-59**).

The majority of the paired data used for the host–EBV interaction analyses are new. For host genome-wide genotyping, all data used in the genome-wide interaction scan (n = 3,522) were generated for this study and have not been reported in our prior publications. For EBV whole-genome sequences, this study in total includes 732 newly sequenced EBV genomes, together with 1,354 EBV genomes from NCBI databases. The EBV genome-wide interaction fine-mapping was conducted in the paired host–EBV genome dataset (n = 734), within which 154 out of 734 EBV genomes were previously deposited in NCBI in our earlier work[12,88]. Thus, most host and EBV data contributing to interaction analyses are new to this study. We hope these revisions (**Supplementary Figure 1** and **Methods file**: **Lines 51-59**) help to clarify the dataset usage in this study.



Response Figure 24. Overview of study design and data analysis workflows. In Step 1, a host genome-wide scan in a Southern China discovery dataset was used to identify host variants interacting with high-risk EBV, and the interactions were then evaluated in an independent validation dataset. In Step 2, an EBV genome-wide scan in Southern China case-control samples was performed to identify viral variants interacting with HLA-A*11:01, and the lead variant (EBV SNP 85841) was replicated in independent Southern China and Singapore datasets. EBV genomes from China and public datasets were combined for phylogenetic analysis. Numbers shown in the corresponding colors indicate the sample sizes of NPC patients (red), healthy controls (blue), and other individuals (orange), respectively.

2. There are a number of contradictions between this study and a recently published paper by the same authors in Cell genomics a year prior, where several HLA and EBV interacting factors have also been

identified. The authors make some reference to these but do not resolve the contradictions completely. Are you suggesting that the EBV factors identified in your previous paper to be no longer significant, and therefore those claims should be withdrawn? similarly, are the HLA SNPs there still true? one was clarified but the other was largely ignored.

Response: Thank you for asking this clarification question concerning two points that were not conveyed clearly enough in the original manuscript: what is new relative to our prior Cell Genomics study, and whether the conclusions are consistent across the two papers. Our Cell Genomics work employed a candidate variant-based design to study marker-level host–EBV interaction. In the present study, we move beyond marker-level inference toward a comprehensive genome-wide interrogation of interactions with expanded host and EBV datasets. This enables us to pinpoint the primary interaction signal in NPC risk to a specific host–virus pair (HLA-A*11:01 × EBNA3B 85841G), and we provide mechanistic evidence supporting this interaction, thereby refining the earlier marker-level signal to causal determinants of the interaction on both the host and viral sides. Importantly, the genetic interaction conclusions across the two studies are consistent in the direction of effect. The key advance of the current work lies not in revisiting the established HLA main effects, but in improving resolution and providing a mechanistic basis for the host–EBV interaction. We now summarize the key differences and connections in **Response Figure 25**.

Category	Cell Genomics	Current manuscript	New contributions in this study
Study design / scope	**2 host SNPs + 1 viral marker**	Stepwise **genome-to-genome interaction scan**, enabling causal resolution	**Host genome-wide** and **viral genome-wide** fine-mapping
Causal host variant for interaction	**Unknown**	**HLA-A*11:01**	Resolves **causal interacting HLA-A allele**
Causal viral variant for interaction	**Unknown**	**EBNA3B 85841G** (functional variant nested within the 163364T clade)	Identifies the **causal interacting EBV variant**
Host–virus interaction effect size	**rs2860580 × EBV163364** RERI = 4.08	**HLA-A*11:01 × High-risk-85841G** RERI = 7.27	Demonstrates **stronger causal** interaction
Functional mechanism	**NO**	**YES** 85841G encodes an **A*11-restricted epitope** eliciting the T cell response	Provides **mechanistic evidence** for interaction
Evolutionary context	**NO**	**YES** Interacting EBV lineages originated from **independent recombinations**	Provides **evolutionary** insight of interacting EBV lineages
Public health significance	**NO**	**YES** Defined **dual-risk** subgroup explained **~47% NPC** in Southern China	Quantifies the **public-health impact** of the interaction

Response Figure 25. Key advances of our current study compared to previous Cell Genomics study.

i. From a candidate-variant design to a stepwise genome-to-genome study design; differences in "significance" reflect different testing burden across study designs

In our prior Cell Genomics work, we used a candidate-variant design and reported the interactions between a high-risk EBV marker (BALF2-CCT marker 163364T) and two top susceptibility markers within HLA region (rs2860580 and rs2894207) on NPC risk. Because that analysis tested two host markers, the significance threshold was set to be 0.025 after Bonferroni correction for two tests. That study provided initial, marker-level evidence suggesting a host–virus interaction, but by design it could not determine (i) whether loci outside the HLA also interact with high-risk EBV to influence NPC risk, or (ii) which specific host allele(s) and viral variant(s) within the HLA and the high-risk EBV background drive the interaction effect.

The present study was designed to address these questions. Using expanded paired host and EBV datasets and a stepwise genome-to-genome framework, we first evaluated host × high-risk EBV interaction at genome-wide scale and confirmed that the HLA region is the primary host contributor to interaction with

high-risk EBV strains. We then fine-mapped the interaction signal to a specific host–virus pair, HLA-A*11:01 and EBNA3B 85841G, and added functional evidence to support their causal interaction. Under this genome-wide framework, the corresponding significance thresholds reflect the larger multiple testing burden (5×10^{-8} for the host genome and 5.92×10^{-4} for the EBV genome).

Therefore, there is no contradiction between the two studies. The marker-level statistical interactions reported previously remain valid within their original candidate-variant design. Differences in "significance" mainly reflect the different multiple-testing burdens and thresholds across study designs (candidate-variant *versus* genome-wide), rather than a change in the direction or validity of the signal.

ii. Consistent in the direction of interaction effects

Across the two studies, the interaction signals are consistent in direction. One main advance of the present study is improved resolution, mapping earlier marker-level inference to a refined, causal host–virus pair. Below we clarify how the earlier markers relate to the refined signal.

EBV side. The previously reported high-risk EBV BALF2-CCT marker (163364T) remains strongly associated with NPC and serves as a valid marker of the high-risk clade. In the present study, we further refined the A*11:01 interaction to EBNA3B 85841G, with 85841G-carrying strains representing the subset within this high-risk clade that drives the host–virus interaction (**Figures 3 and 6**).

Host side. Following your question, we re-examined rs2860580 and rs2894207 in the current expanded dataset. rs2860580 shows a strong interaction with high-risk EBV ($P = 6.8 \times 10^{-10}$), and fine-mapping indicates that this signal is explained by its LD with HLA-A*11:01 ($r^2 = 0.72$): conditioning on A*11:01 abolishes the rs2860580 × high-risk EBV interaction (**Response Figure 26**), consistent with rs2860580 tagging the functional allele A*11:01. For rs2894207, the interaction evidence remains directionally consistent, but is weaker than the primary A*11:01 interaction finding and does not meet significance threshold in the present genome-wide framework (before conditioning on A*11:01, $P = 9.1 \times 10^{-7}$; after conditioning, $P = 2.3 \times 10^{-4}$), suggesting that additional HLA × EBV interactions will require larger samples to resolve at genome-wide significance.

Host SNP	Interaction effect (95% CI)	$P_{interaction}$	$P_{interaction}$ conditional on HLA-A*11:01
rs2860580	4.26 (1.93, 6.59)	6.82E-10	1.20E-01
rs2894207	3.72 (1.17, 6.26)	9.11E-07	2.30E-04

Response Figure 26. Interactions between host SNPs and EBV subtypes in NPC risk. The relative excess risks due to interaction (Interaction effects; RERIs), 95% CIs, and $P_{interaction}$-values for NPC risk were estimated by logistic regression under the RERI framework in the host genome-wide discovery dataset. All analyses were adjusted for age, sex, and the first four human MDS components.

In summary, our prior Cell Genomics findings remain valid at the marker level within the original candidate-variant framework, and the present study refines them by showing that (i) the high-risk EBV BALF2-CCT marker and the host HLA markers rs2860580 or rs2894207 largely tag the same underlying host–virus interactions, and (ii) the rs2860580 × high-risk EBV interaction is best captured by HLA-A*11:01 × EBNA3B 85841G, with a potential interaction signal for rs2894207 that will require larger datasets for confirmation at genome-wide significance. Accordingly, we have discussed that, as larger paired host–EBV datasets become available, comprehensive genome-to-genome scans may uncover additional host–EBV interactions beyond those captured here in the revised manuscript (**Discussion**: **Lines 658-659**). We have also revised the Introduction (**Lines 71-78** and **80-83**) and Discussion (**Lines 557-559**) to more clearly position the present study relative to prior host-only, EBV-only, and marker-level interaction work, and summarized the connections and distinctions between the two studies in supplementary materials (**Supplementary Text**: **Lines 498-570**).

3.The HLA data is not particularly novel, and have been previously published. It is the associated with the EBNA 3B SNP that drives th novelty here and the potential co-evolution. Given that, I feel that there is limited data showing interaction between the peptide and A2-07, which is the deleterious /high risk SNP, and therefore in my mind, the more important one to show rather than the protective SNP. This asymmetry of data (which also cuts across other statistical analyses) with HLA A2-07, which to me is the more important 'positive' effect that needs stronger supporting evidence.

Response: We appreciate your thoughtful comment and agree with your assessment that the EBNA3B 85841G interaction signal is the key novelty in this study, particularly for moving beyond known main effects toward a resolved host–virus interaction. In light of your suggestions, we have therefore made substantial efforts to strengthen the mechanistic evidence for the interaction between 85841G and its host partner HLA-A*11:01, as this is the most strongly supported interaction in our dataset (see Response to R4Q4 below). We also agree that HLA-A*02:07 is a major NPC risk allele that has been consistently validated across studies. However, in our genome-to-genome analyses, evidence for an A*02:07 × high-risk EBV interaction is weaker than that for A*11:01 and remains suggestive, and we therefore interpret A*02:07 × EBV genetic interaction findings cautiously. Following your comment, we performed additional SNP85841-encoded peptide–A*02:07 binding assays and exploratory T cell assays, and we retained A*02:07 for risk stratification as a validated host risk factor. Details are provided below.

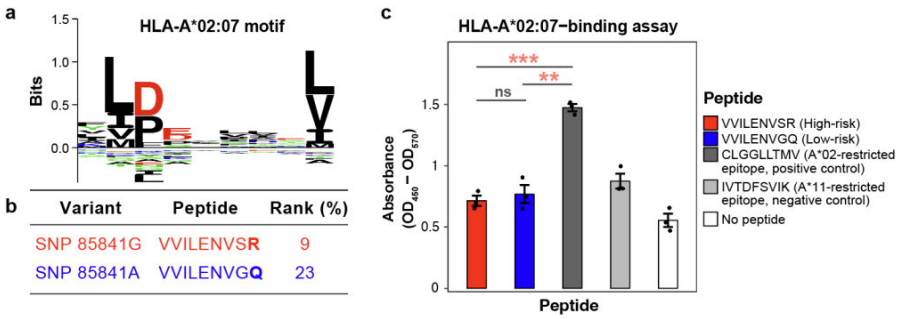
i. Strength of evidence differs between A*11:01 and A*02:07

In our genome-to-genome analyses, the A*02:07 × high-risk EBV interaction signal is weaker and remains suggestive, in contrast to the robust and reproducible A*11:01 interaction. Specifically, in the discovery host genome-wide scan for interaction with the high-risk EBV subtype BALF2-CCT, the A*02:07 × high-risk EBV interaction signal ($P = 3.61 \times 10^{-6}$) was much weaker than that for A*11:01 ($P = 6.55 \times 10^{-10}$), and A*11:01 was the only interaction signal with robust evidence. Conditioning on A*11:01 abolished all remaining interaction signals in the genome (**Supplementary Figure 12**). In meta-analysis of discovery and validation datasets, the A*02:07 × high-risk EBV interaction did not remain genome-wide significant ($P < 5 \times 10^{-8}$) after conditioning on A*11:01 (before conditioning, $P = 1.72 \times 10^{-8}$; after conditioning, $P = 4.86 \times 10^{-4}$; **Supplementary Table 5**). Therefore, while A*02:07 is a major host risk allele, its interaction with high-risk EBV in NPC requires further validation. Accordingly, we focused this study on the A*11:01 interaction. Larger sample sizes will be required to evaluate the A*02:07 × EBV interaction in future work, given the lower frequency of A*02:07 than A*11:01 in Southern China (**Lines 189-195** and **293-299**).

ii. Experimental data showing no binding of HLA-A*02:07 to 85841-encoded peptides

Despite the limited statistical support, we followed your suggestion to explore whether A*02:07 directly interacts with the 85841G or 85841A-encoded peptide. We found no evidence that the 85841G-encoded high-risk peptide VVILENVSR or the 85841A-encoded low-risk peptide VVILENVGQ binds to A*02:07, consistent with the distinct A*02:07 and A*11:01 binding motifs. The experimental details and results are described below.

To test whether A*02:07 can bind the 85841G or 85841A-encoded peptide, we performed both in silico prediction and experimental binding assays. NetMHCpan 4.1 predicted that neither the high-risk 85841G-encoded peptide (VVILENVSR) nor the low-risk 85841A-encoded peptide (VVILENVGQ) binds A*02:07, and neither conforms to the canonical A*02:07 binding motif (**Response Figure 27a-b**). A sliding-window screen of 9 to 11-mer peptides encompassing the 85841-encoded residue also showed no predicted binders of A*02:07 (all predicted ranks > 5%). Consistent with these predictions, peptide-HLA binding assays using ELISA confirmed that A*02:07 did not bind either peptide (**Response Figure 27c**, revised as **Supplementary Figure 15b**). These results provide strong evidence against a direct peptide-level mechanism for A*02:07 at SNP 85841, in contrast to the well-supported A*11:01-restricted VVILENVSR epitope (**Lines 319-322**).



Response Figure 27. HLA-A*02:07 binding motif and lack of binding to the 85841G or 85841A-encoded peptide. **a**, Sequence logo of HLA-A*02:07-bound peptides. **b**, NetMHCpan 4.1 predicted ranks of binding affinities for peptides encoded by 85841G and 85841A variants. **c**, ELISA-based peptide-HLA-A*02:07 binding assay for VVILENVSR and VVILENVGQ. Peptide CLGGLLTMV, an HLA-A*02:07-restricted epitope, was used as a positive control. IVTDFSVIK, an unrelated peptide with negligible HLA-A*02:07 binding, and the no-peptide condition were used as negative controls. In the absence of peptide binding, HLA class I complexes are unstable and undergo degradation, resulting in reduced

ELISA signals with lower OD values. Absorbance was calculated as OD_{450}–OD_{570}. Data are shown as mean ± s.e.m. (n = 3). Statistical significance was evaluated using two-tailed *t*-tests. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ns, not significant.

iii. A possible explanation for the suggestive A*02:07 × high-risk EBV interaction

It is worth noting that A*02:07 has consistently emerged as one of the strongest NPC risk-associated HLA alleles[23,24,26,27,68] and is often the only significant risk allele identified in human GWAS of NPC, whereas other top HLA signals (minor alleles) tend to be protective[26]. HLA class I molecules are central to antiviral immunity, shaping $CD8^+$ T cell recognition of infected cells. Thus, unlike EBV risk variants where "high-risk" naturally implies increased pathogenic potential, it is not immediately intuitive that a host HLA allele would show a risk-increasing direction of effect. This consistent risk association suggests that A*02:07 may be less effective in presenting EBV antigens compared with other HLA-A alleles. When individuals carrying A*02:07 are infected with highly lytic, high-risk 85841G strains, as seen in the M81 EBV strain[69], the effect of this limited immune recognition may be further exacerbated, leading to increased susceptibility to EBV-driven oncogenesis. Accordingly, the A*02:07-associated risk increase, observed as a suggestive A*02:07 × high-risk EBV interaction on NPC risk, could reflect a spillover of the dominant high-risk EBV effect. However, given that the interaction signal is only suggestive, the EBV × A*02:07 interaction and its mechanism remain to be elucidated in future studies. Thus, we view this narrative as a plausible hypothesis rather than a definitive conclusion.

Given the suggestive genetic evidence and the lack of detectable binding, we feel it would be premature to conduct deeper functional studies or to speculate in detail about the potential mechanism of an A*02:07 × EBV interaction at this stage. Accordingly, we have toned down the statements regarding A*02:07 × EBV interaction throughout the manuscript (**Lines 189-195** and **293-299**) and have added the results of peptide-A*02:07 binding assays into revised Results (**Lines 321-322**) and supplementary materials (**Supplementary Figure 15b**; **Supplementary Text**: **Lines 293-334** and **382-396**).

iv. Population-level burden and risk stratification

For population-level risk stratification, we retained HLA-A*02:07 as the major host risk factor that has been confirmed in prior studies and in our data, while noting that its interaction with high-risk EBV remains suggestive (**Lines 449-451**). We included A*02:07 together with A*11:01 to define the host HLA-A background for population burden estimation. Population burden was assessed using the population attributable fraction (PAF) that denotes the proportion of NPC cases in the population attributable to a given stratum, and we defined the dual-risk subgroup (stratum) as individuals jointly defined by high-risk-85841G EBV infection and a susceptible HLA-A background ($A*11:01^-$ or $A*02:07^+$). We have added a new **Supplementary Table 18** comparing two host-background definitions. Using A*11:01 alone already yields a high dual-risk PAF (~40%), and additionally incorporating A*02:07 increases the dual-risk PAF to ~47% (an increase of ~7 percentage points; **Lines 618-623**). This analysis is presented for risk stratification, but does not constitute validation evidence for an A*02:07 × EBV interaction (**Lines 498-500**).

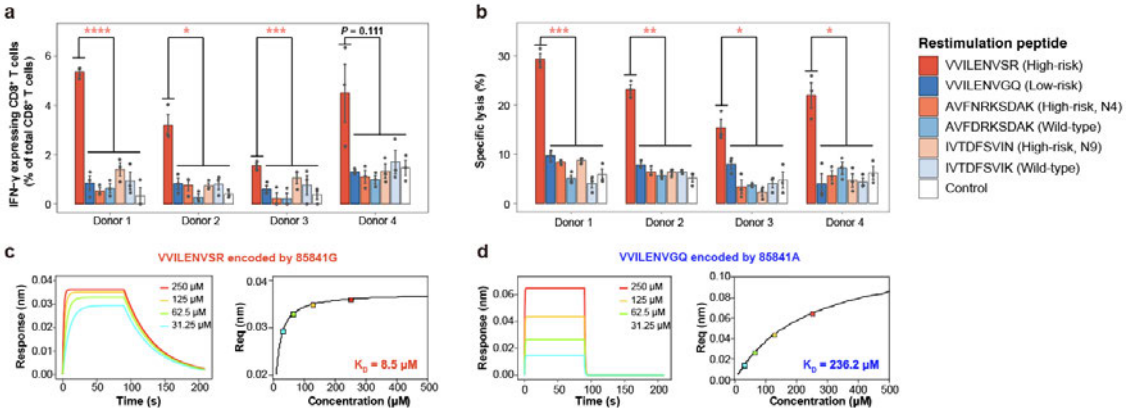
4. Even if the claim here is all about the protective nature of A11-01, and given that the main novelty here is the EBNA SNP, the authors need to show g immunological consequences to strengthen causal inference, including additional T cell assays (e.g., CTL recognition, cytokine release, killing assays).

Response: We greatly appreciate this constructive suggestion and agree that functional validation is essential to establish the immunological consequences of the A*11:01 × EBNA3B-85841G interaction. Following your advice, we conducted a series of functional T cell experiments including cytokine release, killing assays, peptide titration, endogenous presentation, and recognition of $85841G^+$ M81 EBV-transformed LCLs, in addition to peptide-HLA binding assays (**Lines 313-356** and **Figure 4a-l**). Together, these new data demonstrate that the 85841G-encoded peptide VVILENVSR is a bona fide A*11:01-restricted epitope, eliciting $CD8^+$ T cell responses against LCLs transformed by high-risk-85841G EBV strain M81. Because NPC is a virus-driven cancer in which EBV is the key etiologic factor and an invariable risk factor, any protective effect that modifies EBV-related risk is particularly important. HLA-A*11:01 is among the strongest protective alleles for NPC[23,24,26,27,68], and its interaction with high-risk EBV accounts for ~85% of the overall protective association of A*11:01 with NPC (**Lines 185-187**). These functional results provide a molecular basis for this interaction. The experimental details and results are described below.

i. 85841G-encoded peptide elicits specific A*11:01-restricted T cell responses.

Using PBMCs from A*11:01$^+$ healthy individuals infected with high-risk-85841G EBV strains, we expanded polyclonal T cells with three EBNA3B peptide pairs. One pair comprises the 85841G-encoded high-risk peptide VVILENVSR (VVI) and the 85841A-encoded low-risk peptide VVILENVGQ. The other two pairs correspond to the immunodominant HLA-A*11:01-restricted EBNA3B epitopes AVFDRKSDAK and IVTDFSVIK and their epitope-loss variants (AVF/N4 and IVT/N9), which have been consistently reported in prior studies of NPC-derived EBV strains from Southern China[74-76] (**Response Figure 28a-b**). In line with these reports, loss of AVF and IVT is near-universal in high-risk lineages in our data (**Supplementary Figure 14**, **Lines 304-311**). We therefore included these two EBNA3B peptide pairs as control groups.

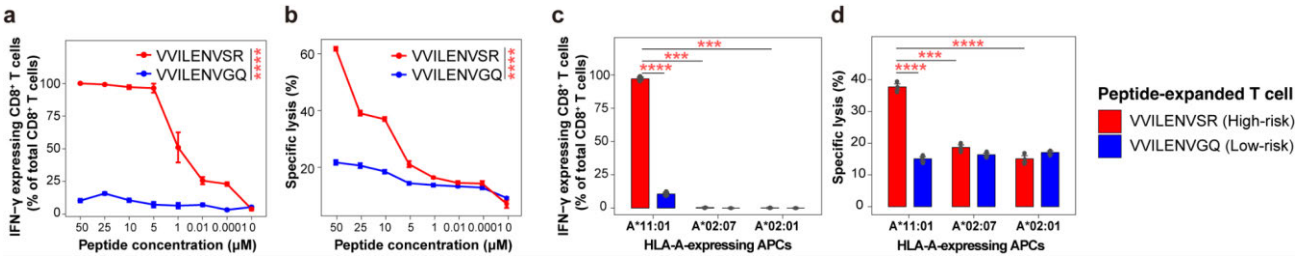
These *ex vivo* expanded T cells were then cocultured with COS-7 cells engineered to express HLA-A*11:01 and β2M. The COS-7 cell line, a fibroblast-like cell line derived from African green monkey and lacking endogenous human HLA expression, is commonly used as the antigen-presenting cell (APC) after HLA-β2M transfection[78-81]. IFN-γ intracellular cytokine staining (ICS) showed that the 85841G-encoded peptide VVILENVSR elicited markedly stronger CD8$^+$ T cell activation than either AVF or IVT, and their corresponding epitope-loss peptides (AVF/N4 and IVT/N9), whereas the 85841A-encoded peptide VVILENVGQ did not activate peptide-specific CD8$^+$ T cell responses (**Response Figure 28a**). Consistently, cytotoxicity assays showed specific killing of A*11:01-expressing target cells pulsed with VVILENVSR, but not with any of the other five peptides (**Response Figure 28b**). This lack of AVF- or IVT-specific responses among A*11:01 carriers infected with high-risk-85841G EBV is consistent with prior reports showing that these epitope-loss variants abolished AVF/IVT-specific T cell responses. Together with the BLI binding assay results that only VVILENVSR bound A*11:01 (**Response Figure 28c-d**), these data indicate that in A*11:01$^+$ individuals infected with high-risk-85841G EBV strains, despite loss of the A*11:01-restricted AVF and IVT epitopes, the 85841G-encoded peptide VVILENVSR is capable of triggering specific CD8$^+$ T cell responses (**Figure 4b-d**; **Lines 325-327** and **330-341**).



Response Figure 28. 85841G-encoded peptide VVILENVSR elicits A*11:01-restricted T cell responses. a-b, Specific T cell responses against three EBNA3B peptide pairs were evaluated by CD8$^+$ T cell activation (a) and cytotoxicity (b) after polyclonal T cell coculture with peptide-pulsed HLA-A*11:01-expressing COS-7 cells in four A*11:01$^+$ healthy individuals infected with high-risk-85841G EBV. Polyclonal T cells were expanded from PBMCs using the three peptide pairs and then co-cultured with luciferase-expressing HLA-A*11:01$^+$ COS-7 target cells pulsed with each peptide (10 μM) as indicated. In (a-b), data are presented as mean ± s.e.m. (n=3). In (a), CD8$^+$ T cell activation was assessed by IFN-γ intracellular cytokine staining following co-culture. In (b), specific cell lysis was measured by the luciferase activity of the remaining viable target cells after 24h of co-culture. Statistical analysis was performed using two-tailed *t*-tests. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$. **c-d,** Binding affinities of the peptides VVILENVSR encoded by 85841G and VVILENVGQ encoded by 85841A to HLA-A*11:01 were measured using biolayer interferometry (BLI). Biotinylated HLA-A*11:01 was immobilized on SSA biosensors, and peptide binding was measured across a dilution series. BLI fitting curves are shown for each peptide dilution. BLI steady-state analyses and K_D values are presented.

ii. Peptide titration confirms VVILENVSR is a bona fide A*11-restricted epitope.

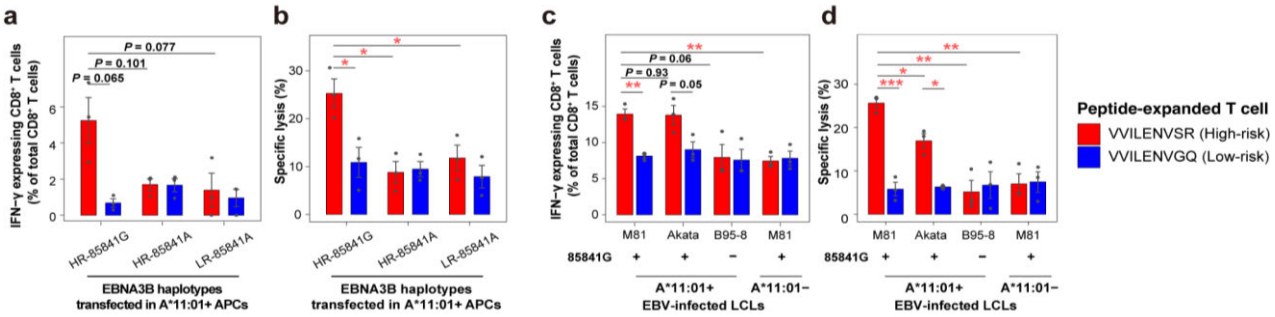
We performed peptide titration and observed dose-dependent T cell activation and cytotoxicity for VVILENVSR, but not VVILENVGQ, when presented by A*11:01-expressing COS-7 APCs (**Response Figure 29a-b**). This T cell response was HLA-A*11:01-restricted, as APCs expressing HLA-A*02:07 or A*02:01 did not support robust VVILENVSR-specific activation or killing (**Response Figure 29c-d**). These results confirm that VVILENVSR is a bona fide A*11:01-restricted epitope to elicit specific T cell responses (**Figure 4e-h**; **Lines 342-345**).



Response Figure 29. 85841G-encoded peptide VVILENVSR is a bona fide A*11:01-restricted epitope. a-b, Peptide titration showing dose-dependent CD8$^+$ T cell activation (a) and cytotoxicity (b). Primary T cells expanded with either VVILENVSR (85841G) or VVILENVGQ (85841A) were co-cultured with luciferase and HLA-A*11:01 co-expressing COS-7 target cells pulsed with the corresponding peptides at different concentrations. **c-d,** HLA restriction assays using COS-7 cells expressing HLA-A*11:01, A*02:07 or A*02:01, pulsed with VVILENVSR (85841G) or VVILENVGQ (85841A). CD8$^+$ T cell activation (c) and cytotoxicity (d) were evaluated by co-culture of peptide-specific T cells with COS-7 target cells co-expressing luciferase and different HLA-A molecules and pulsed with either VVILENVSR or VVILENVGQ (10 μM) as indicated. In (a-d), primary T cells were expanded from PBMCs of HLA-A*11:01$^+$ healthy donors infected with high-risk-85841G EBV with either VVILENVSR (85841G) or VVILENVGQ (85841A), as indicated. In (a, c), T cell activation was assessed by IFN-γ intracellular cytokine staining following co-culture. In (b, d), specific cell lysis was measured by the luciferase activity of the remaining viable target cells after 24h of co-culture. Data are presented as mean ± s.e.m. (n=3). Statistical analysis was performed using two-way repeated-measures ANOVA (a-b) and two-tailed *t*-tests (c-d). *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

iii. VVI-specific T cells target M81 (85841G$^+$) LCLs

Furthermore, endogenous antigen-presentation experiments using A*11:01$^+$ APCs expressing EBNA3B haplotypes indicate that only the high-risk-85841G EBNA3B-encoded peptide, but not the peptide from either high-risk-85841A or the low-risk EBNA3B haplotype, can be processed and presented to trigger VVILENVSR-specific CD8$^+$ T cell cytotoxicity (**Response Figure 30a-b**, revised as **Figure 4i-j**; **Lines 346-349**). Consistent with these findings, VVILENVSR-specific CD8$^+$ T cells showed stronger IFN-γ-producing responses to A*11:01$^+$ lymphoblastoid cell lines (LCLs) transformed by the NPC-derived M81 EBV (high-risk-85841G$^+$) as well as Akata EBV (N1 lineage), which also carries the 85841G-encoded VVILENVSR epitope (**Response Figure 30c**, revised as **Figure 4k**; **Lines 350-352**). In contrast, VVILENVGQ-expanded T cells did not show specific responses. These VVILENVSR-specific T cells also selectively lysed A*11:01$^+$ LCLs transformed by M81 or Akata, with substantially reduced killing of A*11:01$^+$ LCLs transformed by the infectious mononucleosis-derived B95-8 strain (85841A$^+$) or A*11:01$^-$ LCLs (**Response Figure 30d**, revised as **Figure 4l**; **Lines 353-356**).



Response Figure 30. Peptide-specific T cell responses to endogenously expressed EBNA3B haplotypes. a-b, Specific T cell responses to endogenously expressed EBNA3B haplotypes were evaluated by CD8$^+$ T cell activation (a) and T cell cytotoxicity (b) assays. Luciferase and HLA-A*11:01 co-expressing COS-7 target cells were transfected with EBNA3B haplotypes from different EBV lineages, representing high-risk (HR)-85841G, high-risk (HR)-85841A, and low-risk (LR)-85841A EBV from Southern China. These COS-7 target cells were co-cultured with primary T cells expanded with either VVILENVSR (85841G) or VVILENVGQ (85841A) from PBMCs of A*11:01$^+$ individuals infected with high-risk-85841G EBV, as indicated. c,d, Specific T cell responses to EBV-transformed lymphoblastoid cell lines (LCLs) established from A*11:01$^+$ healthy donors (n=3) were evaluated by CD8$^+$ T cell activation (c) and cytotoxicity (d) assays. LCLs were transformed by B95-8 (low-risk-85841A), Akata (low-risk-85841G, Northern China N1 lineage), or M81 (Southern China high-risk-85841G) EBV. Target LCLs were co-cultured with primary T cells expanded from PBMCs of heterologous A*11:01$^+$ healthy donors infected with high-risk-85841G EBV with either VVILENVSR (85841G) or VVILENVGQ (85841A), as indicated. M81-transformed LCLs established from A*11:01$^-$ healthy donors (n=3) were included as controls. LCLs were labeled by anti-CD19 staining and cell death was assessed by Annexin V and SYTOX AADvanced staining to differentiate viable, dying and dead cells. LCLs co-cultured without peptide-specific T cells served as the baseline control. In (a,c), T cell activation was assessed by IFN-γ intracellular cytokine staining following co-culture. In (b), specific cell lysis was measured by the luciferase activity of the remaining viable target cells after 48h of

co-culture. In (d), specific cell lysis was measured by flow cytometry and quantified as the percentage of dying or dead LCLs after 24 h of coculture. Data are presented as mean ± s.e.m. (n=3). Statistical analysis was performed using two-tailed *t*-tests. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

Taken together, these results establish VVILENVSR as a functional HLA-A*11:01-restricted epitope generated by 85841G and show that it elicits specific $CD8^+$ T cell responses against high-risk-85841G EBV-transformed B lymphocytes, providing mechanistic support for the A*11:01 × 85841G interaction. We have incorporated these new analyses and experiments into the revised manuscript (**Results**: **Lines 329-356** and **Methods file**: **Lines 416-500**) and updated the corresponding figures (**Figure 4c-l**).

5. Given the circular argument of finding host factors that are linked to to BALF-CCT and the re connecting that with the EBNA SNP, is there an a priori connection between these to begin with?

Response: Thank you for raising this concern. We understand now how our stepwise design, first establishing the interaction effect and subsequently fine-mapping the causal host and viral variants, might give you the impression of a circular analysis. We would like to take this opportunity to clarify our analytical procedure. There is no a priori connection between the BALF2-CCT high-risk lineage marker and the EBNA3B 85841G variant, and the 85841G-encoded peptide was not known to be an A*11:01 epitope before this study. Establishing the genetic and molecular interaction between A*11:01 and 85841G is the novel finding of our study.

In Step 1, we performed a host genome-wide scan and HLA imputation to identify the allele interacting with high-risk EBV infection to determine NPC risk. No viral SNPs other than the high-risk lineage markers (BALF2-CCT) were used in this step. To fine-map the viral causal variant driving this interaction with HLA-A*11:01, in Step 2 we then performed viral genome-wide interaction mapping, this time testing individual EBV SNPs across the viral genome. In this step, BALF2-CCT was not used. Thus, the discovery of 85841G did not rely on any prior assumption about its relationship to BALF2-CCT nor did we re-test the same marker that defined the original high-risk lineage.

In this framework, each step uses the results of the previous step only to refine the search space, rather than to re-test the same hypothesis, thereby preserving statistical independence while gaining power. Our approach thus follows a sequential, stepwise interaction-mapping strategy[32-35], rather than a circular analysis. In light of your question, we have now made the rationale for this stepwise design explicit in the revised manuscript (**Lines 80-83** and **95-102**) to avoid confusion.

6. Given the lack of ancient EBV genomes, there should be some caveats into the evolutionary analyses of the EBV genome and co-evolutionary events as described here.

Response: We thank you for this insightful comment. We agree that, in the absence of ancient EBV genomes, molecular dating and co-evolution events should be interpreted with caution. Nevertheless, our conclusions regarding the mosaic origin of C1/X1, the recombination events introducing 85841G, and the clonal expansion of C1 in Southern China are supported by multiple clock-independent analyses (phylogeny, recombination detection, chromosome painting and population-genetic statistics) and therefore do not rely on ancient EBV DNA. In light of your comment, we have added explicit caveats in the Results (**Lines 527-529**) that the recombination date and the inferred selection coefficient cannot be directly anchored and are model-dependent, and we have revised the Discussion (**Lines 643-645**) to note that reconstruction of host–EBV coevolution history will benefit from paired ancient human–EBV genomes.

7. Finally, the translational claims are overreaching, I agree that screening could be an important strategy for high risk individuals, but to date, the concept of EBV vaccination has failed and so a moot point in the discussion.

Response: We appreciate your comment and agree that EBV vaccination has not yet achieved clinical success. Following your advice, we have removed the discussion regarding EBV vaccination.

8. There is a suggestion of no original code was included; please clarify if custom scripts for interaction scans/phylogenetics could be shared to improve reproducibility even if they are not novel per se. I was of the understanding that all codes used need to be made available.

Response: Thank you for this helpful suggestion. Scripts for the stepwise host–virus interaction scans and the EBV phylogenetic and recombination analyses have now been deposited in Zenodo (reviewer token: https://zenodo.org/records/18258512?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6ImUyOThkNDdiLTQwNTItNGNmZS05OGNhLTgwZTllYzEyMmVmYyIsImRhdGEiOnt9LCJyYW5kb20iOiJkNGVlZmQwZGYwZjUzYzQwMDgzMTk0ODRmYmIwZTYyYSJ9.9sY87ZesCCDeSRqT3K_KYThwpSs5yBhkbqdmKMPYyiQiyuSdQzpfkG1bsFnICkg0EBlIlPnifJrlz0lUaDzDZQ) and will be made publicly available upon publication. These analytic pipelines and scripts were developed and written by our team using existing R and Python packages. We have updated the Code Availability statement accordingly (**Lines 709-716**).

Referee #5:

Manuscript Nr.: 2025-06-16134

Chen et al., "Human-EBV Interactions Drive Nasopharyngeal Carcinoma Endemicity in Southern China"

The authors demonstrate that a mutation in EBNA3B in Epstein Barr virus (EBV) strains that seem to have emerged from recombination of Northern and Southern Chinese viral strains synergize with HLA mediated risk for nasopharyngeal carcinoma (NPC). Namely, the variant encodes an EBNA3B peptide that binds to HLA-A*1101 and thereby presumably induces protective CD8+ T cell responses, while HLA-A*0207 confers risk. Without the high risk EBV strains the HLA locus seemed to confer only minimally increased risk for NPC. The authors identify an EBNA3B derived peptide (VVILENVGR) of the high risk strain that binds HLA-A*1101 better than its counterpart in the low risk viral strains (VVILENVSQ) and could increase protective T cell responses that lower NPC risk. 47% of endemic NPC cases seem to carry the combination of HLA-A*1101-HLA-A*0207+ and high risk EBV strain. Therefore, the authors argue that this combination is a major determinant of NPC development.

The reported interaction between NPC risk conferring EBV strain variants and HLA haplotype is interesting and novel. However, it should be put into context of other HLA-A11 restricted T cell responses, immune escape from HLA-A*0207 restricted CD8+ T cell responses and T cell infiltration, as well as EBV latency II infection in NPC.

Response: We greatly appreciate your summary and recognition regarding the novelty of our findings.

Major comments:

1. Does the VVILENVSR peptide elicit significant CD8+ T cell responses in individuals with high risk EBV strains and HLA-A*1101 expression? Does the frequency of such cells correlate with lower EBV loads?

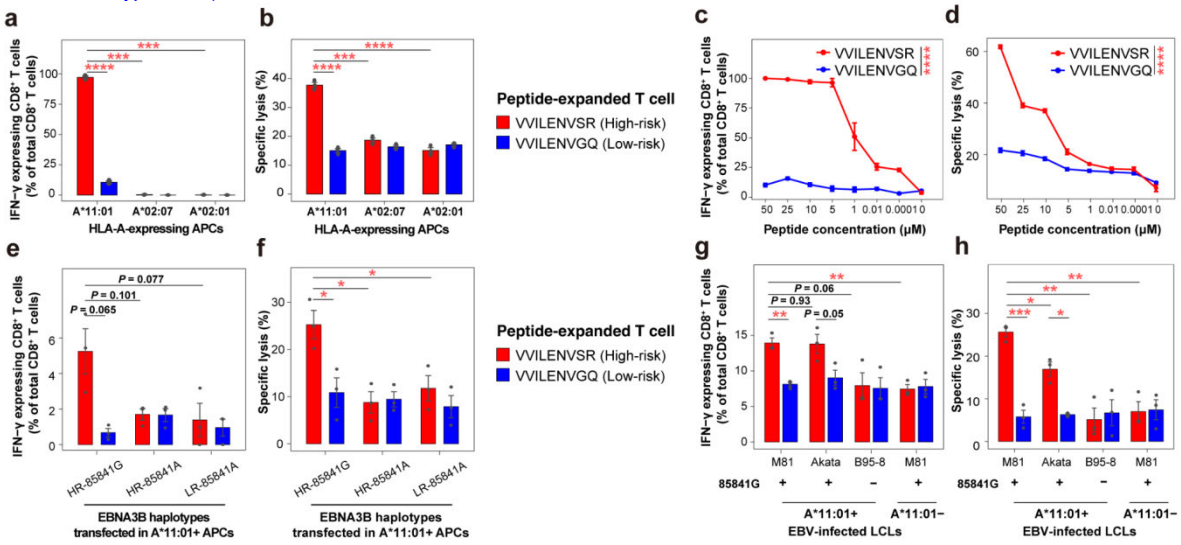
Response: We thank you for raising this important question. In response, we performed T cell functional assays, and we summarize the new supporting data below.

i. VVILENVSR is a bona fide HLA-A*11:01-restricted epitope and elicits specific $CD8^+$ T cell responses in A*11:01^+ healthy carriers infected with high-risk-85841G strains.

Using PBMCs from A*11:01^+ individuals infected with high-risk-85841G EBV strains, we expanded polyclonal T cells with either 85841G-encoded peptide VVILENVSR (high-risk) or 85841A-encoded peptide VVILENVGQ (low-risk). These T cells were then cocultured with COS-7 cells engineered to express human HLA-A alleles and β2M. The COS-7 cell line, a fibroblast-like cell line derived from African green monkey and lacking endogenous human HLA expression, is commonly used as the antigen-presenting cell (APC) after HLA-β2M transfection[78-81]. IFN-γ intracellular cytokine staining (ICS) showed that peptide VVILENVSR elicited substantially stronger $CD8^+$ T cell activation than VVILENVGQ when presented by HLA-A*11:01-expressing COS-7 APCs (**Response Figure 31a**). Consistently, cytotoxicity assays showed specific killing of A*11:01-expressing target cells pulsed with VVILENVSR, but not with VVILENVGQ (**Response Figure 31b**). This T cell response was HLA-A*11:01-restricted, as COS-7 cells expressing HLA-A*02:07 or A*02:01 did not support VVILENVSR-specific activation or killing

(**Response Figure 31a-b**, revised as **Figure 4g-h**). Peptide titration further showed dose-dependent T cell activation and cytotoxicity for VVILENVSR, but not VVILENVGQ, when presented by A*11:01-expressing APCs (**Response Figure 31c-d**, revised as **Figure 4e-f**). Furthermore, endogenous antigen-presentation experiments using A*11:01$^+$ APCs expressing EBNA3B haplotypes indicate that only the high-risk-85841G EBNA3B-encoded peptide, but not the peptide from either high-risk-85841A or the low-risk EBNA3B haplotype, can be processed and presented to trigger VVILENVSR-specific CD8$^+$ T cell cytotoxicity (**Response Figure 31e-f**, revised as **Figure 4i-j**).

Consistent with these findings, VVILENVSR-specific CD8$^+$ T cells showed stronger IFN-γ-producing responses to A*11:01$^+$ lymphoblastoid cell lines (LCLs) transformed by the NPC-derived M81 EBV (high-risk-85841G$^+$) as well as Akata EBV (N1 lineage), which also carries the 85841G-encoded VVILENVSR epitope (**Response Figure 31g**, revised as **Figure 4k**). In contrast, VVILENVGQ-expanded T cells did not show specific responses. These VVILENVSR-specific T cells also selectively lysed A*11:01$^+$ LCLs transformed by M81 or Akata, with substantially reduced killing of A*11:01$^+$ LCLs transformed by the infectious mononucleosis-derived B95-8 strain (85841A$^+$) or A*11:01$^-$ LCLs (**Response Figure 33h**, revised as **Figure 4l**).

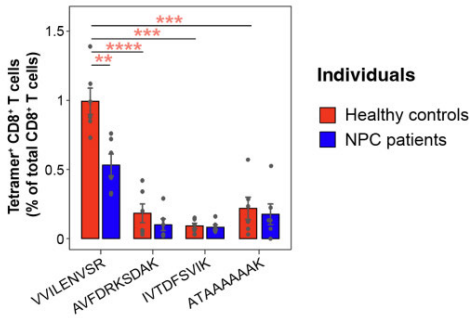


Response Figure 31. 85841G-encoded peptide VVILENVSR elicits A*11:01-restricted T cell responses. a-b, HLA restriction assays using COS-7 cells expressing HLA-A*11:01, A*02:07 or A*02:01, pulsed with VVILENVSR (85841G) or VVILENVGQ (85841A). CD8$^+$ T cell activation (a) and cytotoxicity (b) were evaluated by co-culture of peptide-specific T cells with COS-7 target cells co-expressing luciferase and different HLA-A molecules and pulsed with either VVILENVSR or VVILENVGQ (10 μM) as indicated. **c-d**, Peptide titration showing dose-dependent CD8$^+$ T cell activation (c) and cytotoxicity (d). Primary T cells expanded with either VVILENVSR (85841G) or VVILENVGQ (85841A) were co-cultured with luciferase and HLA-A*11:01 co-expressing COS-7 target cells pulsed with the corresponding peptides at different concentrations. **e-f**, Specific T cell responses to endogenously expressed EBNA3B haplotypes were evaluated by CD8$^+$ T cell activation (e) and T cell cytotoxicity (f) assays. Luciferase and HLA-A*11:01 co-expressing COS-7 target cells were transfected with EBNA3B haplotypes from different EBV lineages, representing high-risk (HR)-85841G, high-risk (HR)-85841A, and low-risk (LR-85841A) EBV from Southern China. These COS-7 target cells were co-cultured with primary T cells expanded with either VVILENVSR (85841G) or VVILENVGQ (85841A) as indicated. **g-h,** Specific T cell responses to EBV-transformed lymphoblastoid cell lines (LCLs) established from A*11:01$^+$ healthy donors (n=3) were evaluated by CD8$^+$ T cell activation (g) and cytotoxicity (h) assays. LCLs were transformed by B95-8 (low-risk-85841A), Akata (low-risk-85841G, Northern China N1 lineage), or M81 (Southern China high-risk-85841G) EBV. Target LCLs were co-cultured with primary T cells expanded from PBMCs of heterologous A*11:01$^+$ healthy donors with either VVILENVSR (85841G) or VVILENVGQ (85841A), as indicated. M81-transformed LCLs established from A*11:01$^-$ healthy donors (n=3) were included as controls. In (h), specific cell lysis was measured by flow cytometry and quantified as the percentage of dying or dead LCLs after 24 h of coculture. LCLs were labeled by anti-CD19 staining and cell death was assessed by Annexin V and AAD staining to differentiate viable, dying and dead cells. LCLs co-cultured without peptide-specific T cells served as the baseline control. In (a-h), primary T cells were expanded from PBMCs of HLA-A*11:01$^+$ healthy donors infected with high-risk-85841G EBV with either VVILENVSR (85841G) or VVILENVGQ (85841A), as indicated. In (a, c, e, g), T cell activation was assessed by IFN-γ intracellular cytokine staining following co-culture. In (b, d, f), specific cell lysis was measured by the luciferase activity of the remaining viable target cells after 24h (b, d) or 48h (f) of co-culture. In (a-h), data are presented as mean ± s.e.m. (n=3). Statistical analysis was performed using two-tailed *t*-tests in (a-b, e-h) and two-way repeated-measures ANOVA in (c-d). *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

ii. Such T cell responses are substantially diminished in NPC patients.

Among individuals infected with high-risk-85841G EBV, VVILENVSR-specific A*11:01-tetramer$^+$ CD8$^+$

T cells were markedly reduced in NPC cases compared with healthy EBV carriers (**Response Figure 32**, revised as **Figure 4m**), consistent with prior reports of generally reduced EBV-specific T cell immunity in NPC[82-85]. This reduction coincided with the heightened EBV-specific IgA levels in NPC cases, suggesting a disease-associated shift toward increased viral antigen exposure and diminished virus-specific cellular immunity.



Response Figure 32. Frequency of HLA-A*11:01 tetramer⁺ EBV-specific CD8⁺ T cells in NPC patients and healthy controls. Tetramer staining of PBMCs from NPC patients (n=6) and healthy controls (n=6) infected with high-risk-85841G EBV. PBMCs were stained with a PE-conjugated HLA-A*11:01 tetramer loaded with EBV-derived peptides (VVILENVSR, AVFDRKSDAK, or IVTDFSVIK), and an APC-conjugated HLA-A*11:01 tetramer loaded with ATAAAAAAK as a negative control. The frequency of tetramer⁺ CD8⁺ T cells is shown as mean ± s.e.m.. Statistical analysis was performed using two-tailed *t*-tests. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

We attempted to test whether the frequency of VVI-specific T cells correlates with lower EBV loads, but in healthy carriers salivary EBV load shows large variability due to episodic shedding, which obscures correlations with immune cell frequencies. Particularly, because tetramer assays were feasible only in a small PBMC subset, we did not detect an association between VVILENVSR-specific tetramer⁺ CD8⁺ T cell frequency and EBV load. In our larger saliva dataset, when infected with high-risk-85841G EBV strains (n = 322), HLA-A*11:01 carriers showed a significant modest reduction in salivary EBV load compared with non-carriers (**Figure 4n**), supporting A*11:01-restricted immune control.

Overall, these results provide functional evidence that VVILENVSR-specific HLA-A*11:01-restricted T cell responses are engaged against high-risk-85841G EBV-transformed B cells but become markedly attenuated in patients with NPC, likely reflecting disease-associated immune dysfunction. We have incorporated these new analyses and experiments into the revised manuscript (**Results**: **Lines 342-359**; **Methods file**: **Lines 416-522**) and updated the corresponding figures (**Figure 4e-m**).

2. The authors argue that a recombination event generated the EBNA3B variant carrying virus. Does the EBNA3B variant influence T cell attraction that has previously been linked to EBNA3B dependent CXCL9 and CXCL10 production? Would a possibly decreased T cell attracting function explain the selection of this recombinant in Southern China?

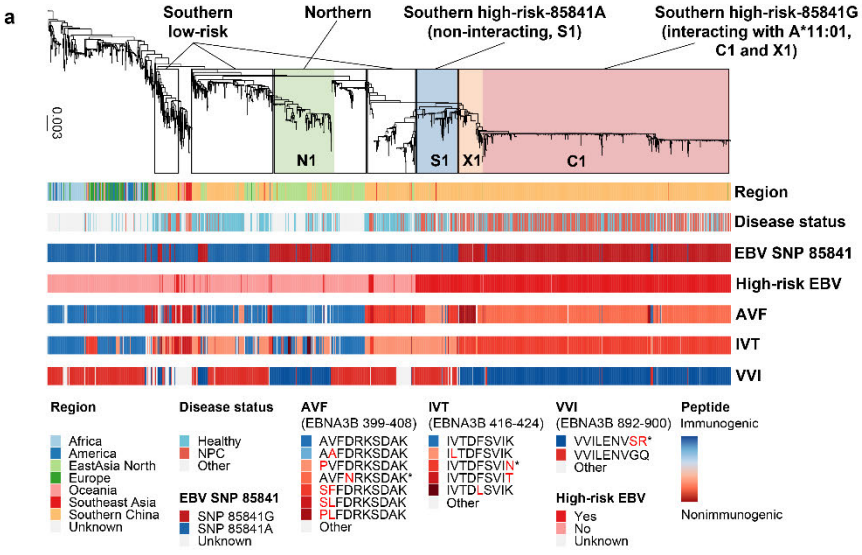
Response: We thank you for raising this important potential mechanism. EBNA3B has been reported to function as a tumor suppressor and to promote CXCL9 and CXCL10 expression, thereby facilitating T cell recruitment[72,73]. [REDACTED]. In the revised manuscript, we have cited the relevant EBNA3B literature and expanded the Discussion to note that, in addition to the HLA-dependent effects emphasized here, EBNA3B haplotype variation, implicated in viral replication, cell growth, and immune modulation, could contribute to strain-specific oncogenic potential

(**Lines 589-596**) and potentially to the selective advantage of the high-risk lineage in Southern China (**Lines 639-643**).

[REDACTED]

3. Previously the EBNA3B aa416-424 epitope presented on HLA-A*1101 was described to be mutated in EBV strains of Papua New-Guinea as reported by the Massucci group. Is this epitope intact in the reassortments of Northern and Southern China EBV strains that increase NPC risk? How do IVTDFSVIK and VVILENVSR specific CD8+ T cell responses compare in healthy Southern Chinese individuals that carry HLA-A*1101.

Response: Thank you for this insightful question. The previously described A*11:01-restricted EBNA3B epitope IVTDFSVIK (IVT, aa 416-424) is almost absent in EBV strains from Southern China (intact IVT frequency: ~1%; **Response Figure 34**). The nearby A*11:01-restricted EBNA3B epitope AVFDRKSDAK (AVF, aa 399-408) is lost in a substantial proportion of Southern China low-risk strains (intact AVF frequency: 35%) but is also rare in high-risk EBV strains (intact AVF frequency: 2% in S1, C1, and X1 lineages) from Southern China, where A*11:01 is the most common HLA allele (**Response Figure 34**). This pattern mirrors the epitope-escape observations reported in Papua New Guinea and has also been observed in Chinese donors by the Masucci and Rickinson groups[74-76]. In contrast, both epitopes remain largely intact in EBV strains from Northern China (intact IVT frequency: 65%; intact AVF frequency: 88%; **Response Figure 34**), where the frequency of HLA-A*11:01 (~26%) is substantially lower than in Southern China (~51%)[89]. These data have been incorporated into the main text (**Lines 304-308**, **602-604** and **637-638**) and presented in **Supplementary Figure 14**.

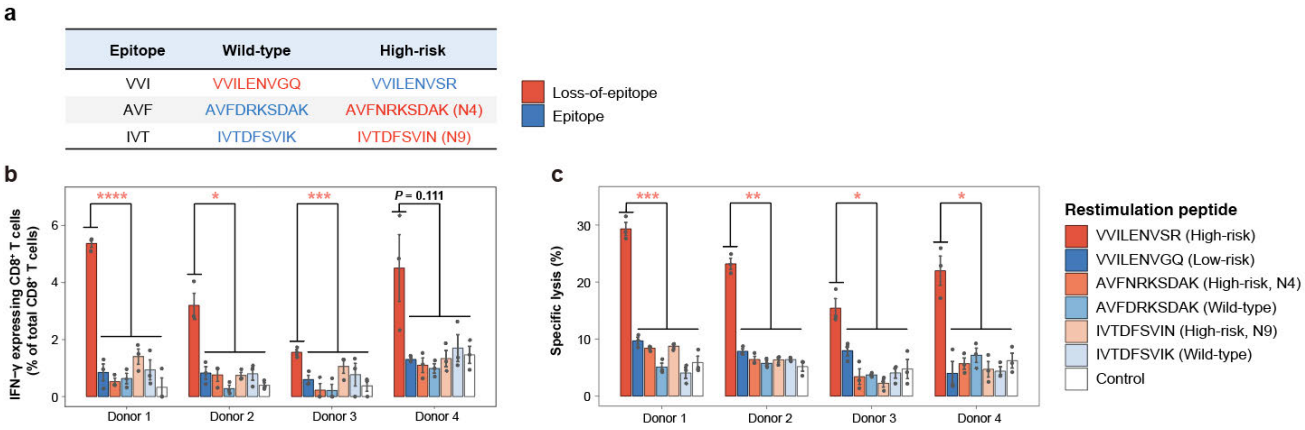


b

Subpopulation carrying different EBV strains	Frequency of AVF epitope	Frequency of IVT epitope	Frequency of VVI epitope
Southern low-risk	34.9%	2.9%	19.4%
Southern high-risk (S1, C1 and X1)	2.3%	0.6%	82.0%
Southern high-risk-85841G (C1 and X1)	1.2%	0.1%	97.0%
Northern	87.9%	64.7%	47.8%

Response Figure 34. Distribution of EBNA3B epitopes AVF, IVT, and VVI. a, Maximum-likelihood phylogeny of 1,852 Type 1 EBV genome sequences, annotated by host geographic region, disease status, EBV SNP 85841 genotype, high-risk EBV subtype (defined by BALF2-CCT or 7048C) and the A*11:01-restricted EBNA3B epitopes AVF, IVT and VVI. Four major lineages are highlighted: the clonally expanded 85841G-carrying high-risk lineage (C1, red), the non-clonal 85841G-carrying high-risk lineage (X1, orange), the Southern China 85841A-carrying high-risk lineage (S1, blue) and the Northern China 85841G-carrying non-high-risk lineage (N1, green). In the AVF, IVT and VVI tracks, colours denote peptide immunogenicity defined by $CD8^+$ T cell responses (blue, immunogenic; red, non-immunogenic). Amino-acid changes relative to the reference epitope sequence are highlighted in red, and asterisks (*) indicate epitope variants present in the clonally expanded 85841G-carrying high-risk lineage C1. **b**, Frequencies of the A*11:01-restricted EBNA3B epitopes AVF, IVT and VVI across subpopulations carrying different EBV strains.

Following your suggestion, we tested A*11:01-restricted $CD8^+$ T cell responses to the IVT, AVF, and the newly identified VVILENVSR (VVI) epitopes in A*11:01$^+$ healthy Southern Chinese individuals naturally infected with high-risk EBV strains. We expanded polyclonal T cells from PBMCs with three EBNA3B peptide pairs (representing three epitopes and their epitope-loss variants): AVF *versus* AVF/N4, IVT *versus* IVT/N9, and VVILENVSR *versus* VVILENVGQ (**Response Figure 35a**, revised as **Supplementary Figure 17**). These T cells were then cocultured with COS-7 APCs engineered to express HLA-A*11:01 and β_2M. IFN-γ intracellular cytokine staining (ICS) showed that the 85841G-encoded peptide VVILENVSR elicited markedly stronger $CD8^+$ T cell activation than either AVF or IVT, and their corresponding epitope-loss peptides (AVF/N4 and IVT/N9), whereas the 85841A-encoded peptide VVILENVGQ did not activate peptide-specific $CD8^+$ T cell responses (**Response Figure 35b**). Consistently, cytotoxicity assays showed specific killing of A*11:01-expressing target cells pulsed with VVILENVSR, but not with any of the other five peptides (**Response Figure 35c**). In line with prior reports, given AVF and IVT epitope-loss variants in high-risk-85841G EBV, we observed little AVF- or IVT-specific responses among A*11:01 carriers infected with this viral strain. These data indicate that in A*11:01$^+$ individuals infected with high-risk-85841G strains lacking AVF and IVT epitopes, VVILENVSR is capable of triggering HLA-A*11:01-restricted $CD8^+$ T cell responses, providing mechanistic support for the A*11:01 × 85841G interaction (**Lines 360-362**). We have incorporated these new analyses and experiments into the revised manuscript (**Results**: **Lines 330-341**) and added the corresponding figures (revised as **Figure 4c-d**).



Response Figure 35. 85841G-encoded peptide VVILENVSR elicits A*11:01-restricted T cell responses. a, EBNA3B peptide pairs used in $CD8^+$ T cell experiments. **b-c,** Specific T cell responses against three EBNA3B peptide pairs were evaluated by $CD8^+$ T cell activation (b) and cytotoxicity (c) after polyclonal T cell coculture with peptide-pulsed HLA-A*11:01-expressing COS-7 cells in four A*11:01$^+$ healthy individuals infected with high-risk-85841G EBV. Polyclonal T cells were expanded from PBMCs using the three peptide pairs and then co-cultured with luciferase-expressing HLA-A*11:01$^+$ COS-7 target cells pulsed with each peptide (10 μM) as indicated. In (b-c), data are presented as mean ± s.e.m. (n=3). In (b),$CD8^+$ T cell activation was assessed by IFN-γ intracellular cytokine staining following co-culture. In (c), specific cell lysis was measured by the luciferase activity of the remaining viable target cells after 24h of co-culture. Statistical analysis was performed using two-tailed *t*-tests. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

4. NPC tumor cells usually do not express EBNA3B. How do the authors envision that the VVILENVSR specific CD8+ T cells protect from NPC risk?

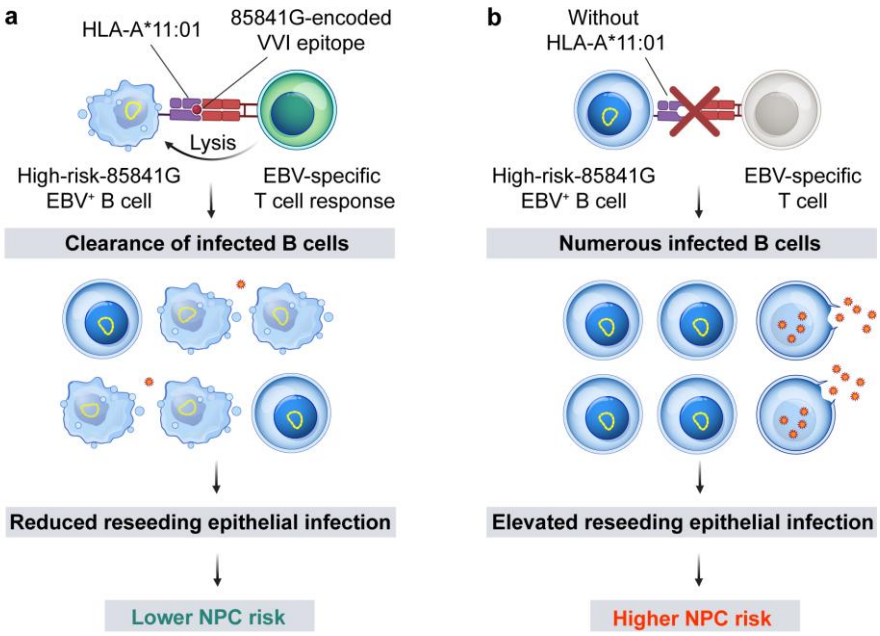
Response: We greatly appreciate this insightful question. Indeed, EBNA3B is not thought to be expressed in NPC tumor cells in Latency II. EBNA3B is co-expressed with other EBNAs and LMPs during Latency III, a transcriptional program that supports continuous growth of EBV-infected LCLs[90]. In healthy carriers, infected B cells in Latency III are mainly observed in tonsillar lymphoid tissue, a site of active viral

replication and new rounds of B cell infection[40,91]. Hence, this EBNA3B-expressing B cell compartment constitutes an EBV$^+$ B cell reservoir that may serve as a source for lytic reactivation and epithelial reseeding.

Consistent with this model, increased mucosal EBV activity precedes NPC diagnosis by years, as reflected by elevated EBV-specific IgA responses[44-47], and circulating EBV DNA is elevated in patients with NPC and is used for early detection[48], indicating a pre-tumor window of heightened viral activity[49]. Moreover, the NPC tumor-derived 85841G-carrying high-risk EBV strain, M81, has been shown to exhibit epithelial tropism and elevated lytic replication in both B cells and epithelial cells[69]. In this setting, more frequent EBV reactivation and weaker immune control are expected to increase epithelial exposure and facilitate the establishment and maintenance of infection within susceptible epithelial compartments.

In high-risk lineages, the immunodominant A*11:01-restricted EBNA3B epitopes (IVT and AVF) are lost. Instead, the 85841G substitution creates a new A*11:01-restricted epitope (VVILENVSR, VVI) that elicits CD8$^+$ T-cell activation and cytotoxicity against EBV$^+$ B cells transformed by the 85841G-positive strain M81 (**Results**: **Lines 313-356**; **Figure 4a-l**). Among healthy carriers infected with high-risk-85841G EBV, HLA-A*11:01 is associated with lower salivary EBV DNA levels, consistent with tighter viral control and lower NPC risk in this subgroup (**Results**: **Lines 368-371**; **Figure 4n**). Our study supports the model (**Response Figure 36**; revised as **Supplementary Figure 23**) that in A*11:01$^+$ individuals infected with high-risk-85841G EBV, VVILENVSR-specific A*11:01-restricted CD8$^+$ T cells help restrain the EBV-infected B cell reservoir and limit reactivation, thereby reducing epithelial reseeding and the probability of NPC initiation. In individuals lacking HLA-A*11:01, recognition of the VVI epitope is absent, and high-risk-85841G EBV infection is associated with weaker immune control and higher NPC risk.

We have incorporated this mechanism overview (**Supplementary Figure 23**), together with supporting literature, into the revised Discussion (**Lines 571-581** and **607-618**) and added the new T cell functional results (**Lines 313-362**; **Figure 4**).



Response Figure 36. Schematic model of EBV immune control and NPC risk associated with HLA-A*11:01. a, In individuals carrying HLA-A*11:01 and infected with 85841G-carrying EBV strains, the EBNA3B 85841G substitution generates an epitope VVILENVSR (VVI) that is efficiently presented by HLA-A*11:01 to EBV-specific CD8$^+$ T cells. These HLA-A*11:01-restricted VVI-specific T cells are capable of lysing EBV$^+$ B cells infected with high-risk-85841G strains, potentially reducing the size of the infected B cell reservoir and epithelial reseeding, thus lowering NPC risk. **b,** In individuals lacking HLA-A*11:01, presentation of the 85841G-encoded epitope to EBV-specific CD8$^+$ T cells is suboptimal. EBV$^+$ B cells infected by high-risk-85841G strains are less efficiently cleared, leading to a larger pool of infected and reactivating B cells, elevated epithelial reseeding, and consequently higher NPC risk.

5. The authors were not able to associate the HLA-A*0207 allele with sequence variation in the NPC risk EBV strains due to lower frequency of this MHC molecule in the studied population. However, it would still be helpful to know if any of the known HLA-A2 restricted epitopes and putative HLA-A*0207

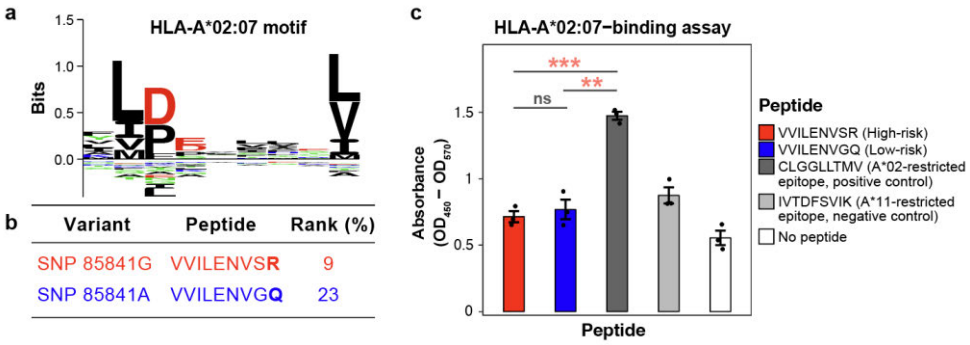
restricted epitopes are mutated in the NPC risk associated EBV strains. This might explain why this allele is enriched in NPC patients that carry these viruses.

Response: We appreciate your thoughtful question. In our study, we only observed a suggestive HLA-A*02:07 × high-risk EBV interaction that did not reach genome-wide significance. Specifically, in the discovery host genome-wide scan for interaction with the high-risk EBV subtype BALF2-CCT, the A*02:07 × high-risk EBV interaction signal ($P = 3.61 \times 10^{-6}$) was much weaker than that for A*11:01 ($P = 6.55 \times 10^{-10}$), and A*11:01 was the only interaction signal with robust evidence. Conditioning on A*11:01 abolished all remaining interaction signals in the genome (**Supplementary Figure 12**). Thus, the A*02:07 × EBV interaction in NPC remains to be validated. Since A*02:07 is less frequent in Southern China than A*11:01, larger sample sizes are needed in future work to more comprehensively investigate the A*02:07 × EBV interaction. We have now toned down the statement for A*02:07 × EBV interaction in the revised manuscript (**Lines 189-195 and 293-299**).

i. Experimental data showing no binding of HLA-A*02:07 to 85841-encoded peptides

Following your question, we performed binding assays. We found no evidence that the 85841G-encoded high-risk peptide VVILENVSR or the 85841A-encoded low-risk peptide VVILENVGQ binds to A*02:07, consistent with the distinct A*02:07 and A*11:01 binding motifs. The experimental details and results are described below.

First, we performed both in *silico* prediction and experimental binding assays to test whether A*02:07 can bind the 85841G or 85841A-encoded peptide. NetMHCpan 4.1 predicted that neither the high-risk 85841G-encoded peptide (VVILENVSR) nor the low-risk 85841A-encoded peptide (VVILENVGQ) binds A*02:07, and neither conforms to the canonical A*02:07 binding motif (**Response Figure 37a-b**). A sliding-window screen of 9 to 11-mer peptides encompassing the 85841-encoded residue also showed no predicted binders of A*02:07 (all ranks >5%). Consistent with these predictions, peptide-HLA binding assays using ELISA confirmed that A*02:07 did not bind either peptide (**Response Figure 37c**, revised as **Supplementary Figure 15b**). These results argue strongly against a direct peptide-level mechanism for A*02:07 at SNP 85841, in contrast to the well-supported A*11:01-restricted VVILENVSR epitope (**Lines 319-322**).

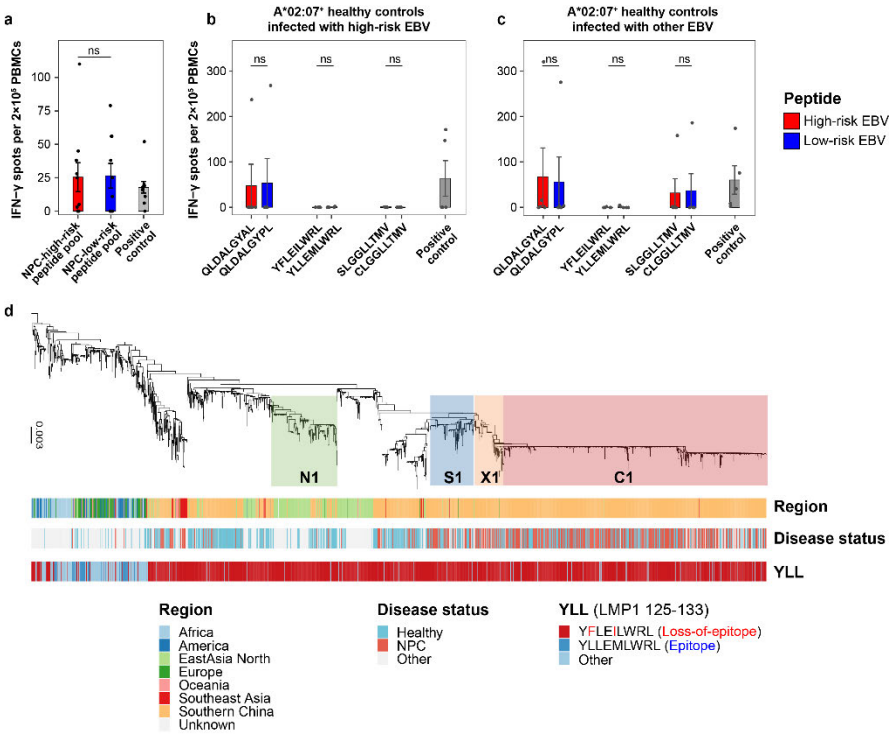


Response Figure 37. HLA-A*02:07 binding motif and lack of binding to the 85841G or 85841A-encoded peptide. **a**, Sequence logo of HLA-A*02:07-bound peptides. **b**, NetMHCpan 4.1 predicted ranks of binding affinities for peptides encoded by 85841G and 85841A variants. **c**, ELISA-based peptide-HLA-A*02:07 binding assay for VVILENVSR and VVILENVGQ. Peptide CLGGLLTMV, an HLA-A*02:07-restricted epitope, was used as a positive control. IVTDFSVIK, an unrelated peptide with negligible HLA-A*02:07 binding, and the no-peptide condition were used as negative controls. In the absence of peptide binding, HLA class I complexes are unstable and undergo degradation, resulting in reduced ELISA signals with lower OD values. Absorbance was calculated as $OD_{450}–OD_{570}$. Data are shown as mean ± s.e.m. (n = 3). Statistical significance was evaluated using two-tailed *t*-tests. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ns, not significant.

ii. Exploratory analyses of potential A*02:07-restricted epitopes

Furthermore, motivated by your question and the established strong genetic association of A*02:07 with NPC risk, we systematically examined the reported HLA-A*02-restricted EBV epitopes[56-66] as well as predicted candidate epitopes (binding affinity ranks < 2%, predicted by NetMHCpan 4.1) that carry variants in strong linkage disequilibrium ($r^2 > 0.5$) with the high-risk viral variants (e.g. BALF2-CCT, EBER2-7048C) or EBNA3B-85841G (**Supplementary Table 21**). These peptide pairs from high-risk and low-risk strains were tested for T cell reactivity by ELISPOT assays using PBMCs from A*02:07$^+$ individuals, with previously reported A*02-restricted epitopes that are conserved (i.e., not carrying

common variants) as positive controls. The T cell responses against the NPC-high-risk and NPC-low-risk peptide pools showed no significant differences (**Response Figure 38a**). Among all peptide pairs tested, two pairs elicited detectable $CD8^+$ T cell responses in $A*02^+$ individuals: the known CLG epitope and its SLG high-risk counterpart, and a predicted peptide pair (low-risk peptide QLDALGYPL and high-risk peptide QLDALGYAL). However, neither pair showed differential reactivity between high-risk and low-risk versions (**Response Figure 38b-c**). All other predicted A*02:07-restricted epitopes were non-reactive in these ELISPOT assays (**Response Figure 38b-c**). Together, the lack of differential reactivity across the tested high-risk/low-risk peptide pairs argues against a single epitope gain/loss mechanism that differentiates high-risk from low-risk EBV strains as the basis for the putative A*02:07 interaction. We also revisited the known LMP1 A*02-escape variant YFLEILWRL (immunogenic wild-type: YLLEMLWRL)[57,67], which is nearly fixed across Asian EBV strains (**Response Figure 38d**). As expected, this escape variant showed no T cell reactivity in ELISPOT assays using PBMCs from $A*02:07^+$ healthy donors (**Response Figure 38b-c**), suggesting that loss of this YLL epitope may weaken A*02:07-restricted T cell recognition of all Asian EBV strains.



Response Figure 38. T cell responses to A*02:07-restricted EBV epitopes and potential epitopes, and distribution of an LMP1 A*02 escape variant. a, No difference in T cell responses to peptide pools derived from NPC-high-risk versus NPC-low-risk EBV strains. PBMCs from $A*02:07^+$ individuals (n = 10) were used to assess peptide-pool–stimulated T cell responses by IFN-γ ELISPOT. Previously reported HLA-A*02-restricted epitopes that are conserved were included as positive controls. **b-c**, No differential T cell responses to A*02:07-restricted EBV epitope pairs between high-risk and low-risk versions. PBMCs from $A*02:07^+$ individuals infected with high-risk EBV (b; n = 5) or non-high-risk EBV (c; n = 5) were used to assess peptide-specific T cell responses. Polyclonal T cells were expanded against A*02:07-restricted peptide pairs and then restimulated with each peptide (2 μM) for 18 h. T cell activation was assessed by IFN-γ ELISPOT. Data are shown as mean ± s.e.m.. Spot counts were normalized by subtracting the negative control (DMSO), and values lower than negative control were set to zero. Statistical significance was evaluated using two-tailed *t*-tests, with $P > 0.05$ considered non-significant (ns). **d**. Distribution of the HLA-A*02-restricted LMP1 epitope and escape variants across EBV lineages. Maximum-likelihood phylogeny of 1,852 type 1 EBV genome sequences, annotated by host geographic region, disease status, and the previously reported LMP1 A*02-restricted epitope (YLLEMLWRL) and escape variant (YFLEILWRL). In the YLL epitope track, red indicates epitope-loss variants, and blue indicates epitope or non-escape variants. Four major lineages are highlighted: the clonally expanded 85841G-carrying high-risk lineage (C1, red), the non-clonal 85841G-carrying high-risk lineage (X1, orange), the Southern China 85841A-carrying high-risk lineage (S1, blue), and the Northern China 85841G-carrying non-high-risk lineage (N1, green).

iii. A possible explanation for the suggestive A*02:07 × high-risk EBV interaction

HLA class I molecules are central to antiviral immunity, shaping $CD8^+$ T cell recognition of infected cells. Thus, unlike EBV risk variants where "high-risk" naturally implies increased pathogenic potential, it is not immediately intuitive that a host HLA allele would show a risk-increasing direction of effect. It is worth noting that A*02:07 has consistently emerged as one of the strongest NPC risk-associated HLA

alleles[23,24,26,27,68] and is often the only significant risk allele identified in human GWAS of NPC, whereas other top HLA signals (minor alleles) tend to be protective[26]. This pattern suggests that A*02:07 may be less effective in presenting EBV antigens compared with other HLA-A alleles. When individuals carrying A*02:07 are infected with highly lytic, high-risk 85841G strains, as seen in the M81 EBV strain[69], the effect of this limited immune recognition may be further exacerbated, leading to increased susceptibility to EBV-driven oncogenesis. Accordingly, the A*02:07-associated risk increase, observed as a suggestive A*02:07 × high-risk EBV interaction on NPC risk, could reflect a spillover of the dominant high-risk EBV effect. However, given that the interaction signal is only suggestive and our epitope analyses for A*02:07 are exploratory, the A*02:07 × EBV interaction and its mechanism remain to be elucidated in future studies. Thus, we view this narrative as a plausible hypothesis rather than a definitive conclusion. Since this mechanism reflects composite differences in EBV control, A*02:07 and A*11:01-transgenic humanized mouse models, such as those used in recent EBV pathogenesis study[92], provide a more appropriate system to directly compare their EBV control across HLA-A backgrounds and to test this hypothesis.

In response, we have toned down our statements regarding A*02:07 × EBV interaction (**Lines 189-195** and **293-299)** in the revised manuscript and have added the results of A*02:07 binding and epitope assays, together with a discussion of this hypothesis, into Results (**Lines 321-322**) and supplementary materials (**Supplementary Figure 15b**; **Supplementary Text**: **Lines 293-396**). We also note that future investigation will require larger studies and advanced EBV infection models capable of directly dissecting EBV control across HLA background *in vivo* to fully resolve pathogenic HLA-EBV interactions (**Lines 658-661**).

Minor comments:

1. The authors have primarily investigated HLA interactions with EBV strain variants. Therefore, they could high-light this in the title along the lines of "EBV strain variations interact with HLA haplotypes to increase NPC risk."

Response: We greatly appreciate your thoughtful suggestion. We agree that our study focuses on the interaction between host HLA-A alleles and EBV strain variation in shaping NPC risk. In light of your comment, we have revised the title to: "EBV strain variation interacts with host HLA-A to determine nasopharyngeal carcinoma risk".

References

1 VanderWeele, T. J. & Knol, M. J. A Tutorial on Interaction. *Epidemiologic Methods* **3** (2014).
2 Zhang, J. & Yu, K. F. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* **280**, 1690-1691 (1998).
3 Zhang, L. F. *et al.* Incidence trend of nasopharyngeal carcinoma from 1987 to 2011 in Sihui County, Guangdong Province, South China: an age-period-cohort analysis. *Chin J Cancer* **34**, 350-357 (2015).
4 Chen, W. J. *et al.* Impact of an Epstein-Barr Virus Serology-Based Screening Program on Nasopharyngeal Carcinoma Mortality: A Cluster-Randomized Controlled Trial. *J Clin Oncol* **43**, 22-31 (2025).
5 Mansournia, M. A. & Altman, D. G. Population attributable fraction. *BMJ* **360**, k757 (2018).
6 Miettinen, O. S. Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am J Epidemiol* **99**, 325-332 (1974).
7 Usui, Y. *et al.* Helicobacter pylori, Homologous-Recombination Genes, and Gastric Cancer. *N Engl J Med* **388**, 1181-1190 (2023).
8 Prentice, R. L. & Pyke, R. Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403-411 (1979).
9 Lees, J. A. *et al.* Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat Commun* **10**, 2176 (2019).
10 Earle, S. G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* **1**, 16041 (2016).
11 Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821-824 (2012).

12 Xu, M. *et al.* Genome sequencing analysis identifies Epstein-Barr virus subtypes associated with high risk of nasopharyngeal carcinoma. *Nat Genet* **51**, 1131-1136 (2019).
13 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
14 Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet* **51**, 1749-1755 (2019).
15 Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet* **98**, 653-666 (2016).
16 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).
17 Mbatchou, J., Abney, M. & McPeek, M. S. BRASS: Permutation methods for binary traits in genetic association studies with structured samples. *PLoS Genet* **19**, e1011020 (2023).
18 Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nat Rev Genet* **18**, 41-50 (2017).
19 Bartha, I. *et al.* A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife* **2**, e01123 (2013).
20 Hui, K. F. *et al.* High risk Epstein-Barr virus variants characterized by distinct polymorphisms in the EBER locus are strongly associated with nasopharyngeal carcinoma. *Int J Cancer* **144**, 3031-3042 (2019).
21 Bei, J. X. *et al.* A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. *Nat Genet* **42**, 599-603 (2010).
22 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 Ancestry. *Cancer Epidemiol Biomarkers Prev* **25**, 188-192 (2016).
23 Tang, M. *et al.* The principal genetic determinants for nasopharyngeal carcinoma in China involve the HLA class I antigen recognition groove. *PLoS Genet* **8**, e1003103 (2012).
24 Tang, M. *et al.* Haplotype-dependent HLA susceptibility to nasopharyngeal carcinoma in a Southern Chinese population. *Genes Immun* **11**, 334-342 (2010).
25 Tse, K. P. *et al.* Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the HLA region at chromosome 6p21.3. *Am J Hum Genet* **85**, 194-203 (2009).
26 Ning, L. *et al.* Nasopharyngeal carcinoma MHC region deep sequencing identifies HLA and novel non-HLA TRIM31 and TRIM39 loci. *Commun Biol* **3**, 759 (2020).
27 Hildesheim, A. *et al.* Association of HLA class I and II alleles and extended haplotypes with nasopharyngeal carcinoma in Taiwan. *J Natl Cancer Inst* **94**, 1780-1789 (2002).
28 Xu, M. *et al.* Host genetic variants, Epstein-Barr virus subtypes, and the risk of nasopharyngeal carcinoma: Assessment of interaction and mediation. *Cell Genom* **4**, 100474 (2024).
29 Herrera-Luis, E., Benke, K., Volk, H., Ladd-Acosta, C. & Wojcik, G. L. Gene-environment interactions in human health. *Nat Rev Genet* **25**, 768-784 (2024).
30 WJ., G. & JM., M. QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies. (2006).
31 VanderWeele, T. J. Sample Size and Power Calculations for Additive Interactions. *Epidemiol Methods* **1**, 159-188 (2012).
32 Kawaguchi, E. S., Kim, A. E., Lewinger, J. P. & Gauderman, W. J. Improved two-step testing of genome-wide gene-environment interactions. *Genet Epidemiol* **47**, 152-166 (2023).
33 Wang, Z., Sul, J. H., Snir, S., Lozano, J. A. & Eskin, E. Gene-Gene Interactions Detection Using a Two-stage Model. *J Comput Biol* **22**, 563-576 (2015).
34 Lin, D. Y. Evaluating statistical significance in two-stage genomewide association studies. *Am J Hum Genet* **78**, 505-509 (2006).
35 Kooperberg, C. & Leblanc, M. Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet Epidemiol* **32**, 255-263 (2008).
36 Wilson, D. J. The harmonic mean p-value for combining dependent tests. *Proc Natl Acad Sci U S A* **116**, 1195-1200 (2019).
37 Sixbey, J. W., Nedrud, J. G., Raab-Traub, N., Hanes, R. A. & Pagano, J. S. Epstein-Barr virus replication in oropharyngeal epithelial cells. *N Engl J Med* **310**, 1225-1230 (1984).
38 Wong, K. C. W. *et al.* Nasopharyngeal carcinoma: an evolving paradigm. *Nat Rev Clin Oncol* **18**,

679-695 (2021).
39 Pathmanathan, R., Prasad, U., Sadler, R., Flynn, K. & Raab-Traub, N. Clonal proliferations of cells infected with Epstein-Barr virus in preinvasive lesions related to nasopharyngeal carcinoma. *N Engl J Med* **333**, 693-698 (1995).
40 Babcock, G. J., Decker, L. L., Volk, M. & Thorley-Lawson, D. A. EBV persistence in memory B cells in vivo. *Immunity* **9**, 395-404 (1998).
41 Taylor, G. S., Long, H. M., Brooks, J. M., Rickinson, A. B. & Hislop, A. D. The immunology of Epstein-Barr virus-induced disease. *Annu Rev Immunol* **33**, 787-821 (2015).
42 Bruce, J. P., Yip, K., Bratman, S. V., Ito, E. & Liu, F. F. Nasopharyngeal Cancer: Molecular Landscape. *J Clin Oncol* **33**, 3346-3355 (2015).
43 Tsao, S. W., Tsang, C. M., To, K. F. & Lo, K. W. The role of Epstein-Barr virus in epithelial malignancies. *J Pathol* **235**, 323-333 (2015).
44 Ji, M. F. *et al.* Sustained elevation of Epstein-Barr virus antibody levels preceding clinical onset of nasopharyngeal carcinoma. *Br J Cancer* **96**, 623-630 (2007).
45 Zeng, Y. *et al.* Prospective studies on nasopharyngeal carcinoma in Epstein-Barr virus IgA/VCA antibody-positive persons in Wuzhou City, China. *Int J Cancer* **36**, 545-547 (1985).
46 Warner, B. E. *et al.* The Epstein-Barr Virus Nuclear Antigen 1 Variant Associated with Nasopharyngeal Carcinoma Defines the Sequence Criteria for Serologic Risk Prediction. *Clin Cancer Res* **30**, 5207-5217 (2024).
47 Paudel, S. *et al.* Serologic Profiling Using an Epstein-Barr Virus Mammalian Expression Library Identifies EBNA1 IgA as a Prediagnostic Marker for Nasopharyngeal Carcinoma. *Clin Cancer Res* **28**, 5221-5230 (2022).
48 Chan, K. C. A. *et al.* Analysis of Plasma Epstein-Barr Virus DNA to Screen for Nasopharyngeal Cancer. *N Engl J Med* **377**, 513-522 (2017).
49 Mautner, J. & Middeldorp, J. M. Epstein-Barr Virus (EBV)-Specific Humoral Immune Responses in Health and Disease. *Curr Top Microbiol Immunol* (2025).
50 Gewurz, B. L., Longnecker, R. M. & Cohen, J. I. Epstein-Barr Virus (Chapter 11). *Fields Virology*, 324-388 (2021).
51 Rüeger, S. *et al.* The influence of human genetic variation on Epstein-Barr virus sequence diversity. *Sci Rep* **11**, 4586 (2021).
52 Ansari, M. A. *et al.* Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nat Genet* **49**, 666-673 (2017).
53 Phelan, J. *et al.* Genome-wide host-pathogen analyses reveal genetic interaction points in tuberculosis disease. *Nat Commun* **14**, 549 (2023).
54 Xu, Z. M. *et al.* Joint host-pathogen genomic analysis identifies hepatitis B virus mutations associated with human NTCP and HLA class I variation. *Am J Hum Genet* **111**, 1018-1034 (2024).
55 Band, G. *et al.* Malaria protection due to sickle haemoglobin depends on parasite genotype. *Nature* **602**, 106-111 (2022).
56 Lee, S. P. *et al.* HLA A2.1-restricted cytotoxic T cells recognizing a range of Epstein-Barr virus isolates through a defined epitope in latent membrane protein LMP2. *J Virol* **67**, 7428-7435 (1993).
57 Khanna, R., Burrows, S. R., Nicholls, J. & Poulsen, L. M. Identification of cytotoxic T cell epitopes within Epstein-Barr virus (EBV) oncogene latent membrane protein 1 (LMP1): evidence for HLA A2 supertype-restricted immune recognition of EBV-infected cells by LMP1-specific cytotoxic T lymphocytes. *Eur J Immunol* **28**, 451-458 (1998).
58 Saulquin, X. *et al.* A global appraisal of immunodominant CD8 T cell responses to Epstein-Barr virus and cytomegalovirus by bulk screening. *Eur J Immunol* **30**, 2531-2539 (2000).
59 Meij, P. *et al.* Identification and prevalence of CD8(+) T-cell responses directed against Epstein-Barr virus-encoded latent membrane protein 1 and latent membrane protein 2. *Int J Cancer* **99**, 93-99 (2002).
60 Straathof, K. C. *et al.* Characterization of latent membrane protein 2 specificity in CTL lines from patients with EBV-positive nasopharyngeal carcinoma and lymphoma. *J Immunol* **175**, 4137-4147 (2005).
61 Lee, S. P., Tierney, R. J., Thomas, W. A., Brooks, J. M. & Rickinson, A. B. Conserved CTL epitopes within EBV latent membrane protein 2: a potential target for CTL-based tumor therapy. *J Immunol* **158**, 3325-3334 (1997).

62 Duraiswamy, J. *et al.* Ex vivo analysis of T-cell responses to Epstein-Barr virus-encoded oncogene latent membrane protein 1 reveals highly conserved epitope sequences in virus isolates from diverse geographic regions. *J Virol* **77**, 7401-7410 (2003).
63 Kerr, B. M. *et al.* Identification of type B-specific and cross-reactive cytotoxic T-lymphocyte responses to Epstein-Barr virus. *J Virol* **70**, 8858-8864 (1996).
64 Burrows, S. R. *et al.* Five new cytotoxic T cell epitopes identified within Epstein-Barr virus nuclear antigen 3. *J Gen Virol* **75 (Pt 9)**, 2489-2493 (1994).
65 Pudney, V. A., Leese, A. M., Rickinson, A. B. & Hislop, A. D. CD8+ immunodominance among Epstein-Barr virus lytic cycle antigens directly reflects the efficiency of antigen presentation in lytically infected cells. *J Exp Med* **201**, 349-360 (2005).
66 Hislop, A. D., Annels, N. E., Gudgeon, N. H., Leese, A. M. & Rickinson, A. B. Epitope-specific evolution of human CD8(+) T cell responses from primary to persistent phases of Epstein-Barr virus infection. *J Exp Med* **195**, 893-905 (2002).
67 Lin, H. J., Cherng, J. M., Hung, M. S., Sayion, Y. & Lin, J. C. Functional assays of HLA A2-restricted epitope variant of latent membrane protein 1 (LMP-1) of Epstein-Barr virus in nasopharyngeal carcinoma of Southern China and Taiwan. *J Biomed Sci* **12**, 925-936 (2005).
68 Hildesheim, A. & Wang, C. P. Genetic predisposition factors and nasopharyngeal carcinoma risk: a review of epidemiological association studies, 2000-2011: Rosetta Stone for NPC: genetics, viral infection, and other environmental factors. *Semin Cancer Biol* **22**, 107-116 (2012).
69 Tsai, M. H. *et al.* Spontaneous lytic replication and epitheliotropism define an Epstein-Barr virus strain found in carcinomas. *Cell Rep* **5**, 458-470 (2013).
70 Li, Z. *et al.* Epstein-Barr virus ncRNA from a nasopharyngeal carcinoma induces an inflammatory response that promotes virus production. *Nat Microbiol* **4**, 2475-2486 (2019).
71 Li, Z. *et al.* The Epstein-Barr virus noncoding RNA EBER2 transactivates the UCHL1 deubiquitinase to accelerate cell growth. *Proc Natl Acad Sci U S A* **118** (2021).
72 White, R. E. *et al.* EBNA3B-deficient EBV promotes B cell lymphomagenesis in humanized mice and is found in human tumors. *J Clin Invest* **122**, 1487-1502 (2012).
73 White, R. E. *et al.* Extensive co-operation between the Epstein-Barr virus EBNA3 proteins in the manipulation of host gene expression and epigenetic chromatin modification. *PLoS One* **5**, e13979 (2010).
74 Midgley, R. S. *et al.* HLA-A11-restricted epitope polymorphism among Epstein-Barr virus strains in the highly HLA-A11-positive Chinese population: incidence and immunogenicity of variant epitope sequences. *J Virol* **77**, 11507-11516 (2003).
75 de Campos-Lima, P. O. *et al.* T cell responses and virus evolution: loss of HLA A11-restricted CTL epitopes in Epstein-Barr virus isolates from highly A11-positive populations by selective mutation of anchor residues. *J Exp Med* **179**, 1297-1305 (1994).
76 Midgley, R. S., Bell, A. I., McGeoch, D. J. & Rickinson, A. B. Latent gene sequencing reveals familial relationships among Chinese Epstein-Barr virus strains and evidence for positive selection of A11 epitope changes. *J Virol* **77**, 11517-11530 (2003).
77 de Campos-Lima, P. O. *et al.* HLA-A11 epitope loss isolates of Epstein-Barr virus from a highly A11+ population. *Science* **260**, 98-100 (1993).
78 Zhang, S. *et al.* Immunosequencing identifies signatures of T cell responses for early detection of nasopharyngeal carcinoma. *Cancer Cell* **43**, 1423-1441.e1410 (2025).
79 Malekzadeh, P. *et al.* Neoantigen screening identifies broad TP53 mutant immunogenicity in patients with epithelial cancers. *J Clin Invest* **129**, 1109-1114 (2019).
80 Lu, Y. C. *et al.* Efficient identification of mutated cancer antigens recognized by T cells associated with durable tumor regressions. *Clin Cancer Res* **20**, 3401-3410 (2014).
81 Caushi, J. X. *et al.* Transcriptional programs of neoantigen-specific TIL in anti-PD-1-treated lung cancers. *Nature* **596**, 126-132 (2021).
82 Moss, D. J. *et al.* Epstein-Barr virus specific T-cell response in nasopharyngeal carcinoma patients. *Int J Cancer* **32**, 301-305 (1983).
83 Lee, S. P. *et al.* CTL control of EBV in nasopharyngeal carcinoma (NPC): EBV-specific CTL responses in the blood and tumors of NPC patients and the antigen-processing function of the tumor cells. *J Immunol* **165**, 573-582 (2000).
84 Fogg, M. H., Wirth, L. J., Posner, M. & Wang, F. Decreased EBNA-1-specific CD8+ T cells in patients with Epstein-Barr virus-associated nasopharyngeal carcinoma. *Proc Natl Acad Sci U S A*

106, 3318-3323 (2009).
85 Li, J. *et al.* Functional inactivation of EBV-specific T-lymphocytes in nasopharyngeal carcinoma: implications for tumor immunotherapy. *PLoS One* **2**, e1122 (2007).
86 Jensen, T. Z. T. *et al.* A 5700 year-old human genome and oral microbiome from chewed birch pitch. *Nat Commun* **10**, 5520 (2019).
87 Yang, M. A. *et al.* Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* **369**, 282-288 (2020).
88 Zhang, X. *et al.* Out-of-Africa migration and clonal expansion of a recombinant Epstein-Barr virus drives frequent nasopharyngeal carcinoma in southern China. *Natl Sci Rev* **12**, nwae438 (2025).
89 Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res* **48**, D783-d788 (2020).
90 Münz, C. Latency and lytic replication in Epstein-Barr virus-associated oncogenesis. *Nat Rev Microbiol* **17**, 691-700 (2019).
91 Babcock, G. J., Hochberg, D. & Thorley-Lawson, A. D. The expression pattern of Epstein-Barr virus latent genes in vivo is dependent upon the differentiation stage of the infected B cell. *Immunity* **13**, 497-506 (2000).
92 Läderach, F. *et al.* EBV induces CNS homing of B cells attracting inflammatory T cells. *Nature* **646**, 171-179 (2025).

Part II. Responses to referees' comments

Referee #1:

This is a comprehensive revision that addresses my previous concerns, adding both statistical analyses and further lines of validation, ultimately strengthening the main conclusions.

I would still ask that the results from full GxG screen alluded to in the response to point three is added as supplementary material, or deposited in an appropriate resource if possible (e.g. zenodo or osf). Even if underpowered now, it may be useful for future validation or meta-analysis efforts.

(This very minor comment may be a matter for the editor to decide on importance/necessity.)

Response: We thank the referee for this suggestion. We have now included the results from the full G×G screen as supplementary material (**Supplementary Note 2** and **Supplementary Fig. 14**).

Referee #2:

The authors have provided detailed responses to all of my comments, and have performed many experiments to address each comment definitively.

I congratulate the authors.

Response: We thank the referee for the positive evaluation.

Referee #4:

Thank you for all your clarification and revisions in this version of the manuscript. Most of my major concerns have been addressed and I am especially heartened to see the immune assays used to demonstrate the actual T-cell effect of HLA A11-01 and the high risk peptide in a number of donor PBMCs. I also appreciate the attempt to analyse the different peptides from the high risk EBV lineage that may or may not associate with the A02-07 genotype- it is difficult to prove a negative result which may be the case here but nevertheless these negative results are important and interesting.

One final query, is there a speculation on whether the protective nature of A11-01 here shaped the evolutionary distribution of this HLA type across china and Southeast Asia? this is well outside the scope of this manuscript but more an interesting side-bar to the data presented here.

Response: We thank the referee for the supportive comments and for highlighting the value of the new immune assays. Regarding whether the protective nature of HLA-A*11:01 shaped its geographic distribution across China and Southeast Asia, we agree that this is an interesting question. While pathogen-driven selection and demographic history can both influence HLA allele frequencies, we do not have direct evidence to distinguish these forces in this context. We therefore refrain from drawing a conclusion and note this for future investigation (Lines 461-464).

Referee #5:

Manuscript Nr.: 2025-06-16134A

Chen et al., "EBV strain variation interacts with host HLA-A to determine nasopharyngeal carcinoma risk"

The authors demonstrate that a mutation in EBNA3B in Epstein Barr virus (EBV) strains that seem to have emerged from recombination of Northern and Southern Chinese viral strains synergize with HLA mediated risk for nasopharyngeal carcinoma (NPC). Namely, the variant encodes an EBNA3B peptide that binds to HLA-A*1101 and thereby presumably induces protective CD8+ T cell responses, while HLA-A*0207 confers risk. Without the high risk EBV strains the HLA locus seemed to confer only minimally increased risk for NPC. The authors identify an EBNA3B derived peptide (VVILENVGR) of the high risk strain that binds HLA-A*1101 better than its counterpart in the low risk viral strains (VVILENVSQ) and could increase protective T cell responses that lower NPC risk. 47% of endemic NPC cases seem to carry the combination of HLA-A*1101-HLA-A*0207+ and high risk EBV strain. Therefore, the authors

argue that this combination is a major determinant of NPC development.
In their revised manuscript the authors addressed all of my concerns but there was no direct correlation between VVILENVSQ specific HLA-A*1101 restricted CD8+ T cell responses and salivary viral loads. They further demonstrate that the high risk EBNA3B variant allows less efficient CXCL9 and CXCL10 production which could diminish T cell attraction into the tumor microenvironment, but the authors prefer to report these findings in another manuscript. They now also document that additional HLA-A*1101 restricted epitopes are lost in the high risk EBV strains in the Southern Chinese population. I also consider the argument that diminished EBNA3B specific CD8+ T cell responses would allow for an increased EBV reservoir to seed epithelial cells valid. However, would then a correlation between diminished HLA-A*1101 restricted CD8+ T cell responses and viral shedding or circulating viral loads not be even more important? Finally, the authors were not able to observe differences with HLA-A2 restricted EBV high and low risk variant derived peptide pools but low reactivity against an LMP1 derived HLA-A2 restricted peptide that is not present in Southern Chinese EBV variants. Therefore, they have toned down their conclusions with respect to the NPC risk increase by HLA-A*0207 even so this is the HLA allele most frequently found to be associated with NPC in genome wide association studies.
In summary, the authors have included a large amount of additional work into their revised manuscript. However, it seems like HLA-A*1101 restricted CD8+ T cell mediated immune control of EBV is compromised by Southern Chinese EBV variants beyond the described polymorphism and no prominent association of the VVILENVSQ specific HLA-A*1101 restricted CD8+ T cell response could be documented. Moreover, the authors prefer to include the interesting data on lower CXCL9 and CXCL10 production by the high risk EBNA3B variant into a different manuscript, even so this might more generally explain diminished CD8+ T cell mediated immune control of the high risk EBV variant beyond the mutation in one particular CD8+ T cell epitope.
Response: We thank the referee for the thoughtful and supportive comments. The central finding of this study is that the effect of high-risk EBV on NPC risk differs by HLA-A*11:01 status, and the manuscript therefore focuses on establishing the genetic and molecular basis for this interaction. We revised the Discussion to state explicitly that the mechanistic axis supported by our assays is unlikely to be the only driver of the observed host–EBV associations in NPC risk, and that additional viral genetic changes within the high-risk lineage, including EBNA3B variation, may also contribute to a broader, subtype-intrinsic oncogenic or immunomodulatory effect (Lines 466-476).

With respect to correlations with salivary shedding or circulating EBV DNA in healthy individuals, such *in vivo* relationships are inherently difficult to capture because EBV reactivation and shedding are intermittent and highly dynamic, and would likely require large-scale biobank or longitudinal data. Indeed, recent biobank studies have supported a role of MHC loci as major determinants of *in vivo* EBV activity in healthy populations[1,2]. In the current study, our *ex vivo* assays demonstrate HLA-A*11:01-restricted T cell killing of 85841G-carrying M81 EBV-transformed B cells, supporting HLA-A-mediated T cell immune control of high-risk EBV and

modulation of NPC risk.

We sincerely thank the editor and all the referees for their time, careful assessment, and constructive suggestions, which have substantially strengthened the manuscript.

References

1 Nyeo, S. S. *et al.* Population-scale sequencing resolves determinants of persistent EBV DNA. *Nature* **650**, 664-672, doi:10.1038/s41586-025-10020-2 (2026).

2 Schmidt, A. *et al.* Host control of persistent Epstein-Barr virus infection. *Nature*, doi:10.1038/s41586-026-10274-4 (2026).