nature portfolio

Corresponding author(s): Miao Xu, Jianjun Liu, Weiwei Zhai, Zhonghua Liu

Last updated by author(s): Mar 6, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Host genotype data were collected using the Illumina Asian Screening Array Chip or Illumina OmniZhongHua-8 Chip.
EBV whole-genome sequencing was performed on the Illumina NextSeq platform.
EBV genotyping was performed using the MassArray iPLEX platform (Agena Bioscience).
HLA typing data were generated using the Illumina NovaSeq 6000 platform.
Biolayer interferometry data were collected with Octet BLI Discovery v12.2.
Flow cytometry data were collected on a CytoFLEX LX flow cytometer (Beckman Coulter).
Luciferase activity was measured using a Spark 10M multimode microplate reader (Tecan).
ELISPOT data were collected using an AID ELISPOT Reader.

Data analysis

Host genotype quality control and PCA were performed using PLINK (v1.9). HLA region imputation was conducted using SNP2HLA (v1.0.3). HLA typing was performed using HLA-HD (v1.2.0). EBV raw reads were trimmed and filtered using fastp. The processed paired-end reads were aligned using the BWA (v0.7.17). Duplicated reads were removed by Picard (v2.18.14). VerifyBamID (v1.1.3) was employed to detect mixed EBV infections. Following the GATK best practice workflows (v4.1.8.1), EBV variants were identified after base quality score recalibration. The functional annotation of the EBV variants was performed using the SNPEff package (v5.0e). EBV variants were imputed with Beagle (v5.4). We constructed a viral genetic relatedness matrix (vGRM) from EBV variants using the GEMMA software (v0.98.5), and subsequently fit a logistic mixed model in GMMAT (v1.4.2). Permutations were implemented in BRASS, which accounts for sample clonality and relatedness. Bootstrap resampling was performed using R package boot (v1.3.30). Population attributable fraction was calculated using R package AF (v0.1.5). HLA-restricted epitopes were predicted by netMHCpan4.1. Molecular graphics were generated in PyMOL (v3.1.6.1). Biolayer interferometry data were analyzed with the Octet Analysis Studio (v12.2) software. Flow cytometry data were analyzed using CytExpert (Beckman Coulter) and FlowJo v10 (TreeStar). Multiple sequence alignment was performed using MAFFT (v7.490). EBV variant quality control and Fst were performed using VCFtools (v0.1.13). The maximum likelihood of the phylogenetic relationship was inferred using IQ-TREE 2 (v2.2.0). All

phylogenetic trees were visualized and annotated using R packages ape (v5.8.1), treeio (v1.30.0), ggtree (v3.14.0), and ggtreeExtra (v1.16.0). Chromosome painting analysis was performed using ChromoPainter, embedded within fineSTRUCTURE (v4.1.0). Consensus genomes for each EBV clade were generated using EMBOSS (v6.6.0.0) and analyzed for recombination events with RDP5. Extended haplotype homozygosity was estimated using the R package rehh (v3.2.2). Scripts for the stepwise host–virus interaction scans and the EBV phylogenetic and recombination analyses have now been deposited in Zenodo (https://doi.org/10.5281/zenodo.18258512).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The GRCh37/hg19 reference genome is publicly available (please refer to https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.13/). The accession numbers for EBV genome data retrieved from NCBI are listed in Supplementary Table 16. The paired EBV–host sequencing and genotype data have been deposited in the Genome Sequence Archive for Human at the National Genomics Data Center and are available at https://ngdc.cncb.ac.cn/gsa-human/browse/HRA017196 (accession no. HRA017196; BioProject no. PRJCA040311). The EBV sequencing data have been deposited in NCBI database under BioProject ID PRJNA1268153, and the associated SRA link is https://www.ncbi.nlm.nih.gov/sra?LinkName=bioproject_sra_all&from_uid=1268153. Base maps are available through Natural Earth at https://www.naturalearthdata.com/.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	The study datasets included both sexes.
Reporting on race, ethnicity, or other socially relevant groupings	Population structure indicates the NPC cases and healthy controls for host-EBV genome-wide interaction study were of Southern Chinese ancestry.
Population characteristics	The characteristics of NPC cases and healthy controls for host-EBV genetic interaction study was summarized in Supplementary Table 2.
Recruitment	Participants from Southern China were enrolled through two independent recruitments (Sample Sets 1 and 2). Subjects in Sample Set 1 were participants in a population-based case-control study conducted in NPC-endemic regions of Southern China (Guangdong and Guangxi Provinces) between 2010 and 2014. Treatment-naive patients histologically confirmed with NPC were identified through a rapid case ascertainment system involving a network of local physicians. Population-based controls were randomly selected from local population registries; individuals with no history of malignancy were enrolled and frequency-matched to cases by sex, 5-year age group, and residential area. In Sample Set 2, an independent NPC case-control study was recruited from the same Chinese population in Southern China between 2013 and 2022. The Singapore dataset is a completely independent NPC case-control study recruited in Singapore. An additional dataset of healthy individuals was recruited from multiple regions of China (Northern: Shandong, Shanxi, Inner Mongolia, Xinjiang, Hebei and Liaoning; Eastern: Fujian; Southwestern: Sichuan). For details of recruitment, see Methods. We cannot fully exclude selection-related biases inherent to case–control recruitment (e.g., hospital-ascertained cases and non-participation), which may affect generalizability; matching and covariate adjustment reduce the likelihood of major bias in relative-risk estimates.
Ethics oversight	The study was approved by the institutional ethics committee of the Sun Yat-sen University Cancer Center, Guangzhou, China.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Step 1 leveraged our larger dataset (discovery: 747 cases/1,251 controls; validation: 644 cases/880 controls; host genotyping and high-risk

Sample size	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 power under this stepwise design, we performed simulations following the method in "Increasing the Power of Identifying Gene × Gene Interactions in Genome-wide Association Studies". 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; r2 = 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 × 10^-8) was 83% at the Step 1 discovery sample size (747 cases/1,251 controls). 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, the power to detect the A*11:01 × 85841G interaction at the EBV genome-wide significance threshold (Bonferroni-corrected for the EBV scan) was 91%. These results are consistent with prior human G × G power analyses and support that our stepwise design is statistically efficient under current sample-size realities.
Data exclusions	For data exclusion, we removed samples lacking demographic information or EBV genotype data, due to failed genotyping caused by low EBV titers in biosamples. Additionally, host genotyping, EBV whole-genome sequencing, and EBV genome sequences used for phylogenetic analysis were processed using standard quality control procedures, as detailed in Supplementary Figures 1, 3, 8, and 19.
Replication	In the host genome-wide scan for interaction with EBV, the interactions between HLA variants and the high-risk EBV subtype were replicated using an independent validation dataset from Southern China (Cases: n=644, Controls: n=880). Furthermore, the interactions between HLA-A alleles and the 85841G-carrying high-risk EBV strains were validated in two independent datasets: one from Southern China (Cases: n=388, Controls: n=543) and the other completely independent dataset from Singapore (Cases: n=223, Controls: n=204). Replication numbers for the other experiments are provided in the corresponding figure legends, and consistent results were obtained.
Randomization	This is a case-control study. Randomization is not applicable.
Blinding	All sequencing and data collection were performed by researchers who were blinded to disease status. And, the investigators involved in the data analysis and result interpretation were not involved in sample recruitments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Antibody / Source / Catalogue number / clone number / Dilution APC/Cyanine7 anti-human CD45 / BioLegend / 304014 / HI30 / 1:100 PE-Cy7 anti-human CD3 / BD Biosciences / 557749 / SP34-2 / 1:100 PerCP-Cy5.5 anti-human CD8 / BD Biosciences / 560662 / RPA-T8 / 1:100 Alexa Fluor 647 anti-human IFN-γ / BD Biosciences / 557729 / B27 / 1:100 Brilliant Violet 421 anti-human CD19 / BioLegend / 302233/ HIB19 / 1:100 FITC anti-human CD3 / BioLegend / 300305 / HIT3a / 1:100 anti-human CD28 / BD Biosciences / 555725 / CD28.2 / 1:1000 anti-human CD49d / BD Biosciences / 555502 / 9F10 / 1:500
Validation	All antibodies used for flow cytometry were validated by the manufacturers for the corresponding applications. See the manufacturers' websites for further information (https://www.biolegend.com/en-gb and https://www.bdbiosciences.com). Antibodies were also independently validated in-house through preliminary experiments using either cell lines or primary cells before use in formal experiments.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Non-human primate cell lines B95-8 (ATCC), CNE2-Akata (Zeng YX's lab), CNE2-M81 (Zeng YX's lab), and COS-7 cells expressing HLA-A*11:01, A*02:07 or A*02:01 (Li G's lab). B95-8, CNE2-Akata, and CNE2-M81 were used as EBV-producer cell lines. PBMCs used in this study were obtained from nine female and seventeen male donors.

Authentication	EBER staining and DNA sequencing were used to confirm the presence of EBV genomes in B95-8, CNE2-Akata, and CNE2-M81 cells and to verify the EBV strain identity. The absence of endogenous human HLA expression in parental COS-7 was confirmed by flow cytometry.
Mycoplasma contamination	Cell cultures tested negative for micoplasma contamination.
Commonly misidentified lines (See ICLAC register)	EBV-infected CNE2 cell lines were used only as virus-producer cells to generate infectious EBV virions (Akata or M81). They were not used in any experiments addressing NPC or tumor cell biology, and no conclusions in this study rely on this cell line background.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Sample preparation is described in detail in the Methods. In T cell activation assay, surface staining was performed for 1 h at 4℃ using Fixable Viability Dye eFluor 506 (eBioscience), APC/Cyanine7 anti-human CD45 (HI30, BioLegend), PE-Cy7 anti-human CD3 (SP34-2, BD Biosciences), and PerCP-Cy5.5 anti-human CD8 (RPA-T8, BD Biosciences). Cells were then washed, fixed, and permeabilized for 1-5 h, followed by intracellular staining with Alexa Fluor 647 anti-human IFN-γ (BD Biosciences). T cell-mediated cytotoxicity against LCL targets was assessed by flow cytometric analysis of target cell death. After co-incubation of LCLs and peptide-expanded T cells, cells were harvested and washed twice with PBS and then incubated with BV421 anti-human CD19 (Biolegend) antibody for 30 min at 4℃. Then, cells were washed with binding buffer (Biolegend) three times and stained with APC Annexin V Apoptosis Detection Kit (Biolegend) together with SYTOX AADvanced (Invitrogen) according to the manufacturers' instructions. For tetramer staining, PBMCs from NPC patients or healthy controls were stained for 1 h at 4℃ with a PE-conjugated HLA-A*11:01 tetramer loaded with VVILENVSR (MBL), AVFDRKSDAK (BetterGen), or IVTDFSVIK (BetterGen), and with an APC-conjugated HLA-A*11:01 tetramer loaded with ATAAAAAAK (MBL) as a negative control. Cells were subsequently surface-stained with Fixable Viability Dye eFluor 506 (eBioscience), FITC anti-human CD3 (HIT3a, BioLegend), and PerCP-Cy5.5 anti-human CD8 (RPA-T8, BD Biosciences).
Instrument	Flow cytometry data were collected on a CytoFLEX LX flow cytometer (Beckman Coulter).
Software	Flow cytometry data were analyzed using CytExpert (Beckman Coulter) and FlowJo v10 (TreeStar).
Cell population abundance	Subset abundance was determined by gating.
Gating strategy	The gating strategy for T cell assays is described in detail in Supplementary Figure 25. To assess the frequency and activation of EBV antigen-expanded CD8+ T cell, cells were stained with the Live/Dead dye and antibodies against CD45, CD3, CD8 and for intracellular IFN-γ production. To assess T cell-mediated cytotoxicity, cocultures were stained with an anti-human CD19 antibody to gate LCLs, and viable target cells were defined as Annexin V-negative and SYTOX AADvanced-negative. PBMCs from NPC patients or healthy controls were stained with a PE-conjugated HLA-A*11:01 tetramer loaded with EBV-derived peptides (VVILENVSR, AVFDRKSDAK, or IVTDFSVIK), an APC-conjugated HLA-A*11:01 tetramer loaded with ATAAAAAAK as a negative control, Live/Dead dye, and antibodies against CD3 and CD8.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.