HGG
Advances

ARTICLE

Germline *HLA* heterozygosity is associated with decreased lung cancer risk

Taotao Tan,[1,2,40] Vikram R. Shaw,[1,40] Jinyoung Byun,[3,4] Hyun-Sung Lee,[5,6] Younghun Han,[3,4] Yafang Li,[3,4] Rayjean J. Hung,[7,8] David C. Christiani,[9,10] Xin-An Wang,[9] Mattias Johansson,[11] Xiangjun Xiao,[3] David Zaridze,[12] Stig Egil Bojesen,[13,14] Sanjay Shete,[15] Demetrios Albanes,[16] Melinda C. Aldrich,[17] Adonina Tardon,[18] Guillermo Fernandez-Tardon,[19]

(Author list continued on next page)

Summary

Heterozygosity at human leukocyte antigen (*HLA*) loci may improve lung cancer immunosurveillance by increasing recognition of the tumor by the immune system. Previous studies utilizing data from population-level biobanks, such as the United Kingdom Biobank and FinnGen, have identified an association between germline *HLA* class II (*HLA*-II) heterozygosity and reduced lung cancer risk in smokers. In the present study, we evaluate the association between *HLA* heterozygosity and lung cancer in a large case-control study (15,302 cases and 14,580 controls) with imputed *HLA* allele-type information, comparing differences in *HLA* heterozygosity between smokers and non-smokers, among lung cancer subtypes, and at 2- and 4-digit *HLA* allele resolution. We identify a strong protective association of *HLA*-II heterozygosity in smokers compared to non-smokers, particularly at the *HLA-DPB1* and *HLA-DPA1* loci, and provide subtype-specific resolution. Finally, analysis of the additive effects of *HLA* allele heterozygosity in smokers identified significant associations with several 4-digit *HLA* alleles, including *HLA-B**08:01, *HLA-A**01:01, *HLA-C**07:01, *HLA-DQA1**05:01, *HLA-DRB1**03:01, and *HLA-C**03:04. Our study provides additional evidence, with added histologic subtype information, that germline *HLA*-II heterozygosity is inversely associated with lung cancer risk.

Introduction

Lung cancer is the leading cause of global cancer-related mortality.[1,2] Tobacco smoke is the primary risk factor for lung cancer development, but other factors contribute to its risk, including environmental and occupational exposures (e.g., arsenic,[3] radon,[4] and asbestos[4]), chronic lung disease, and lung infections.[5] Histologically, lung cancer tumors are divided into two main categories, specifically small-cell lung carcinoma (SCLC) and non-small-cell

[1]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA;[2]Department of Integrative Physiology, Baylor College of Medicine, Houston, TX, USA;[3]Department of Internal Medicine, University of New Mexico, Albuquerque, NM, USA;[4]University of New Mexico Comprehensive Cancer Center, University of New Mexico, Albuquerque, NM, USA;[5]Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA;[6]David Sugarbaker Division of Thoracic Surgery, Michael E. DeBakey Department of Surgery, Baylor College of Medicine, Houston, TX, USA;[7]Prosserman Centre for Population Health Research, Lunenefeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada;[8]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada;[9]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA;[10]Pulmonary and Critical Care Division, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA;[11]Genomic Epidemiology Branch, International Agency for Research on Cancer, 69366 Lyon, France;[12]Clinical Epidemiology, N.N. Blokhin National Medical Research Centre of Oncology, Moscow, Russia;[13]Department of Clinical Biochemistry, Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen, Denmark;[14]Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen, Denmark;[15]Department of Biostatistics, University of Texas, M.D. Anderson Cancer Center, 7007 Bertner Ave., Houston, TX, USA;[16]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA;[17]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA;[18]Public Health Department, University of Oviedo, and Health Research Institute of Asturias, ISPA, Av. del Hospital Universitario, s/n, 33011 Oviedo, Asturias, Spain;[19]University of Oviedo and CIBERESP, Faculty of Medicine, Oviedo, Spain;[20]Population Sciences in the Pacific Program, University of Hawaii Cancer Center, Honolulu, HI, USA;[21]Clalit National Cancer Control Center at Carmel Medical Center and Technion Faculty of Medicine, Mikhal St. 7, Haifa 3436212, Israel;[22]University Medical Center Göttingen, Institute of Genetic Epidemiology, Humboldtallee 32, 37073 Göttingen, Germany;[23]Helmholtz-Munich Institute of Epidemiology, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany;[24]Department of Biosciences and Medical Biology, Center for Tumor Biology and Immunology, University of Salzburg, 5020 Salzburg, Austria;[25]Cancer Cluster Salzburg, Salzburg, Austria;[26]Division of Cancer Epigenomics, DKFZ – German Cancer Research Center, Heidelberg, Germany;[27]Roy Castle Lung Cancer Research Programme, The University of Liverpool, Liverpool, UK;[28]Department of Molecular and Clinical Cancer Medicine, The University of Liverpool, Liverpool, UK;[29]Department of Oncology, University of Sheffield, Sheffield, UK;[30]Radboud University Medical Center, Nijmegen, The Netherlands;[31]National Institute of Occupational Health, Oslo, Norway;[32]British Columbia Cancer Agency, Vancouver, BC, Canada;[33]Department of Radiation Sciences, Umeå University, Umeå, Sweden;[34]Department of Medical Biosciences, Umeå University, Umeå, Sweden;[35]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA;[36]Department of Epidemiology, Geisel School of Medicine, Hanover, NH, USA;[37]Department of Pharmaceutical Sciences, School of Pharmacy and Pharmaceutical Sciences, University at Buffalo (SUNY), Buffalo, NY, USA;[38]Markey Cancer Center, University of Kentucky, Lexington, KY, USA;[39]Section of Genetics, International Agency for Research on Cancer, World Health Organization, 25 Avenue Tony Garnier, CS 90627, 69366 Cedex 07 Lyon, France

(Affiliations continued on next page)

Loïc Le Marchand,[20] Gad Rennert,[21] Heike Bickeböller,[22] H.-Erich Wichmann,[23] Angela Risch,[24,25,26] John K. Field,[27,28] Michael Davies,[27,28] Penella Woll,[29] Lambertus A. Kiemeney,[30] Aage Haugen,[31] Shanbeh Zienolddiny,[31] Stephen Lam,[32] Mikael Johansson,[33] Kjell Grankvist,[34] Matthew B. Schabath,[35] Angeline Andrew,[36] Philip Lazarus,[37] Susanne M. Arnold,[38] Dakai Zhu,[1,3] Maria Teresa Landi,[16] James McKay,[39] Christopher Amos,[1,3,4] and Chao Cheng[1,5,41,*]

lung carcinoma (NSCLC), with the latter contributing to the majority of cases.[6] The main subtypes of NSCLC include adenocarcinoma (ADC), squamous cell carcinoma (SCC), and large-cell lung carcinoma (LCLC).[6] A smaller group of neuroendocrine tumors also exists, with carcinoid tumors being the primary subset.

In addition to the increased lung cancer risk conferred by exogenous exposures, studies have also identified germline genetic risk variants that contribute to lung cancer pathogenesis.[7–9] Variations in genes including *TP53*, *EGFR*, *CHRNA5*, *BRCA2*, and *TERT*, in addition to several others, have been implicated in lung cancer etiology.[8,9] One set of associations, between polymorphisms in human leukocyte antigen (*HLA*) class I (*HLA*-I) and class II (*HLA*-II) genes and lung cancer development, has highlighted a potential role of the immune system in lung cancer development.[8–10] The *HLA*-I and *HLA*-II genes are integral to the process of presenting endogenous and exogenous antigens, respectively, to T cells. Our previous fine-mapping study of the HLA region identified associations with *HLA-B*08:01* and *HLA-DQB1*06* in a European population and *HLA-DQB1*04:01* and *HLA-DRB1*07:01* in an Asian population.[11] The *HLA* genes have a diverse array of alleles, contributing to significant polymorphism that allows for the detection of a wide range of antigens.[12]

Increased *HLA* diversity may improve immunosurveillance of lung cancer, leading to earlier tumor recognition by the immune system.[12] Accordingly, the heterozygote advantage hypothesis[13] suggests that heterozygosity at *HLA* loci improves protection from disease due to improved tumor antigen presentation.[12] Data from recent studies have supported this hypothesis in lung cancer, with studies identifying somatic *HLA*-I loss as a mechanism of immune evasion.[14,15] One study identified that *HLA* loss of heterozygosity (LoH) occurs in 40% of NSCLC and is associated with a high subclonal neoantigen burden and enrichment in metastatic sites.[15] Another study identified a higher incidence of *HLA*-I LoH in lung cancer with brain metastasis, in addition to reduced CD8^{+} T cell infiltration into these tumors.[16] Taken together, these observations in the somatic setting suggest that germline *HLA* homozygosity may predispose an individual to NSCLC development, while heterozygosity may be protective. A recent study addressed this question using data from the United Kingdom Biobank (UKB) and FinnGen (FG), finding that *HLA*-II heterozygosity was associated with reduced lung cancer risk in current and former, but not never, smokers.[12] The authors suggested that the differential results for smokers and non-smokers may result from smoking increasing the number of mutations and smoking-derived antigens, which increase the importance of the immune response to early neoplastic disease in this cohort of patients.[12]

In the current study, we sought to replicate the previously mentioned result in a larger sample to determine the association between heterozygosity at various *HLA* loci with lung cancer risk, stratified by smoking status. We evaluate this question in a large study of European ancestry ($N = 29{,}882$ with $n = 15{,}302$ cases), compare differences in *HLA* heterozygosity between smokers and non-smokers, highlight histologic subtype-specific effects, and provide granular information about *HLA* allelic subtypes to the 2- and 4-digit levels.

Material and methods

Sample collection and genotypes

The study sample, along with component studies, quality control, and processing steps, has been previously described[11] and will be described briefly here. Participants came from the OncoArray study, a collection of 30 case-control studies that form part of the Integrative Analysis of Lung Cancer Risk and Etiology (INTEGRAL) and the Lung Cancer Cohort Consortium (LC3).[11] Informed consent for all participants and institutional review board approval from each institution was obtained.[11] In the studied cohort, we analyzed 29,882 samples, including 15,302 cases, of European ancestry (Tables 1 and S1). Genotyping was performed using the OncoArray genotyping platform, a custom Illumina array focused on cancer and with additional coverage of the *HLA* region.[11] Sample and variant quality control metrics have been previously described, and in brief included the removal of samples with low genotyping success rates (<95%), mismatched genetically inferred and reported sex, or excess identity by descent sharing relative to other samples.[11]

HLA imputation and validation

Utilizing the high coverage of the *HLA* region by the OncoArray SNP, genotyping data from 25 to 35 Mb at chromosome 6 was used to impute classical 2- and 4-digit *HLA* alleles.[11] The 2-digit *HLA* allele (e.g., *HLA-DRB1*03*) provides information about the broad serologic family, while the 4-digit *HLA* allele (e.g., *HLA-DRB1*03:01*) specifies the specific protein sequence of the peptide-binding region amino acids. Reference data for *HLA* imputation were collected by the Type 1 Diabetes Genetics Consortium (T1DGC) and includes 5,225 individuals of European origin with

[40] These authors contributed equally

[41] Lead contact

*Correspondence: chao.cheng@bcm.edu

https://doi.org/10.1016/j.xhgg.2026.100567.

Table 1. Sample demographic and histologic information for controls and lung cancer cases

	Total cohort ($N = 29{,}882$)	Controls ($n = 14{,}580$)	Cases ($n = 15{,}302$)
Sex (male, %)	17,818 (59.6)	8,649 (59.3)	9,169 (59.9)
Age (median, IQR)	63 (56–70)	61 (55–68)	64 (57–71)
Smoking status (yes, %)	24,379 (81.6)	10,524 (72.2)	13,855 (90.5)
Lung cancer	–		15,302 (100)
Adenocarcinoma (%)			6,887 (45.0)
Squamous cell carcinoma (%)			4,075 (26.6)
Small cell lung cancer (%)			1,739 (11.4)
Non-small cell carcinoma (%)			991 (6.5)
Mixed (%)			915 (6.0)
Large cell lung cancer (%)			524 (3.4)
Carcinoids (%)			171 (1.1)

genotyping data for 8,534 SNPs and 424 classical *HLA*-I (*HLA-A*, *HLA-B*, and *HLA-C*) and *HLA*-II (*HLA-DRB1*, *HLA-DQB1*, *HLA-DQA1*, *HLA-DPB1*, and *HLA-DPA1*) genes.[11] Imputation was performed strictly within the HLA region (chr6:27970031–33965553) using the Michigan Imputation Server 4-digit HLA reference panel; no intensity data was used for HLA typing.[11] Using this approach, each participant had specific alleles at the HLA loci, which were used to define loss-of-HLA-allele (LoA), gain-of-HLA-allele (GoA), homozygosity, and heterozygosity (defined in the next section), utilizing a framework similar to that of the LoH in human leukocyte antigen (LOHHLA) tool.[15] Validation of the *HLA*-imputed data was performed as previously described.[11]

Statistical analysis

We defined *HLA* heterozygosity, GoA, and LoA as follows. *HLA* heterozygosity is defined as an individual who carries two different 4-digit HLA alleles (e.g., an individual who carries *HLA_A*01:01* and *HLA_A*01:02*). *HLA* GoA is defined as an individual who carries more than two copies of an *HLA* allele (e.g., an individual who carries *HLA_A*01:01*, *HLA_A*01:01*, and *HLA_A*01:02*). *HLA* LoA is defined as an individual who carries fewer than two copies of an *HLA* allele (e.g., an individual who carries only *HLA_A*01:01*). Several multivariable models were tested, varying the definitions of the *HLA* variable as described below. Each model was adjusted for principal components (PCs) 1–5, sex, age, and smoking status, and nominal significance was considered at $p < 0.05$.

Estimating effects of HLA gene heterozygosity—e.g., homozygous, heterozygous, GoA, and LoA—on lung cancer

To study how *HLA* heterozygosity contributes to lung cancer risk, we first stratified the dataset by smoking status for both cases and controls (24,379 smokers and 5,503 non-smokers). Among smokers, there were 13,855 cases and 10,524 controls. Among non-smokers, there were 1,447 cases and 4,056 controls. For each cohort, we then fitted a logistic regression to predict a binary disease status for each *HLA* gene factor variable, which consists of homozygous, heterozygous, GoA, or LoA. In each regression model, we controlled for the top five PCs, age, and sex. Treating *HLA* homozygosity as the baseline, we calculated the effects of heterozygosity, GoA, and LoA, and computed their p value and standard error with the Wald test. The odds ratio (OR $= \exp(\beta)$) was visualized together with the 95% confidence interval (CI), which was also on the OR scale. Fixed-effect meta-analysis of ORs and 95% CIs was performed using the "metagen" function from the meta R package. Log(OR) and standard errors were pooled for each HLA locus, and summary effect estimates with heterogeneity statistics (I^2) were extracted. To test whether HLA heterozygosity demonstrated different effects among smokers and non-smokers, we conducted interaction analysis using the model logit[Disease] $=$ HLA-gene $+$ Smoke $+$ HLA-gene $\times$ Smoke $+$ 5 PCs $+$ age $+$ sex.

Estimating effects of HLA gene heterozygosity—e.g., homozygous, heterozygous, GoA, and LoA—on lung cancer subtypes

To study the association between *HLA* heterozygosity and each lung cancer subtype, we performed identical analyses stratified by lung cancer subtype. We constructed strata by combining all individuals with a specific type of lung cancer (e.g., adenocarcinoma) and compared them to all individuals without any form of lung cancer. We then performed logistic regression for each disease and smoking stratum, conditioning on the top five PCs, age, and sex. Analysis was not repeated for GoA and LoA owing to very small sample sizes.

Estimation of additive effects of 2-digit/4-digit HLA alleles on lung cancer subtypes

To estimate the additive effect of 2-digit and 4-digit *HLA* alleles, we also performed logistic regression for smokers and non-smokers. Importantly, besides the top five PCs, age, and sex, we also included the heterozygosity status of the indexing *HLA* gene as a covariate. For example, when regressing disease status on *HLA-A*01:01* allele numbers, we also included *HLA-A* status, such as homozygous, heterozygous, GoA, or LoA, as a covariate. We ran logistic regressions for 147 2-digit *HLA* alleles and 423 4-digit *HLA* alleles and visualized the p value using Manhattan plots. The Bonferroni correction was applied according to the number of alleles and plotted as a gray dashed line in the Manhattan plot (0.00034 threshold for 2-digit *HLA* alleles and 0.00019 threshold for 4-digit *HLA* alleles).

Results

Patients with lung cancer demonstrate less germline *HLA* heterozygosity

The demographic and histologic information for the 15,302 cases with lung cancer and 14,580 controls is

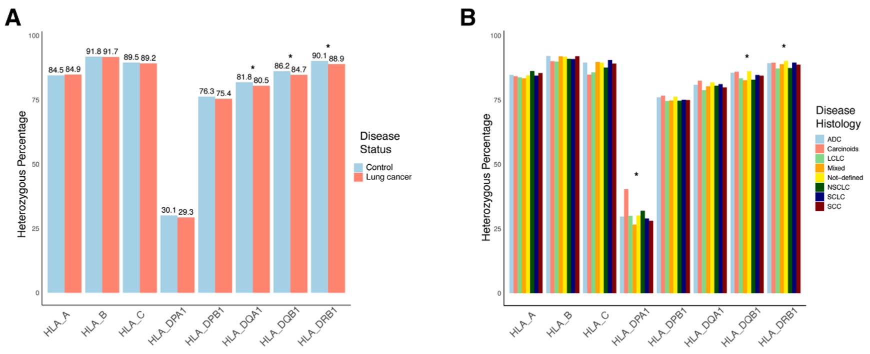

Figure 1. Demographic and histologic information stratified by *HLA* allele

Heterozygote percentage by (A) disease status and (B) disease histology. Significance established using the chi-squared test. *$p < 0.05$ for indicated *HLA* allele.

presented in Tables 1 and S1. The predominant histologic subtype of lung cancer was ADC (45.0%), followed by SCC (26.6%) and SCLC (11.4%). In univariable analyses, no differences in *HLA* heterozygosity percentages were observed between cases and controls for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DPA1*, or *HLA-DPB1* (Figure 1). Significant differences, however, were observed for *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB1*, with less *HLA* heterozygosity observed in the lung cancer cases. Upon analysis of the histologic subtypes, significant differences in *HLA* heterozygosity were observed for *HLA-DPA1*, *HLA-DQB1*, and *HLA-DRB1*, with ADC generally demonstrating increased heterozygosity compared to SCC.

Germline *HLA* heterozygosity is associated with overall lung cancer risk in multivariable analysis

Several models testing different underlying *HLA* variable constructs were implemented to analyze the association between germline *HLA* heterozygosity and overall lung cancer risk (Table 2). A protective effect of *HLA* heterozygosity was observed upon building an *HLA* variable counting the total number of heterozygotes per sample for all *HLA* loci (OR, 0.98; 95% CI, 0.96–0.99; $p = 2.72 \times 10^{-3}$), in addition to another *HLA* variable counting only samples with heterozygosity at all *HLA* loci (OR, 0.92; 95% CI, 0.86–0.98; $p = 8.86 \times 10^{-3}$). The effect appeared to be specific to major histocompatibility complex (*MHC*) class II loci, as significance was lost upon analysis building an *HLA* variable with just *HLA-A*, *HLA-B*, and *HLA-C* (OR, 1.00; 95% CI, 0.96–1.03; $p = 0.93$) but was retained when the *HLA* variable was based on *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB1* (OR, 0.96; 95% CI, 0.94–0.98; $p = 2.33 \times 10^{-4}$). Furthermore, we performed a fixed-effect meta-analysis investigating the association between *HLA* heterozygosity with lung cancer in the overall cohort from our study with the data previously published by Krishna et al.[12] from the UKB and FG (Table S2). We demonstrated consistent effects across the studies, with no significant associations among *HLA-I* genes but significant associations with low heterogeneity in *HLA-DPA1* (OR, 0.94; 95% CI, 0.90–0.99; $p = 0.01$; $I^2 = 0$) and *HLA-DPB1* (OR, 0.93; 95% CI, 0.89–0.97; $p = 4.48 \times 10^{-4}$; $I^2 = 0.08$).

HLA heterozygosity has heterogeneous effects among smokers and non-smokers

We next tested whether *HLA* heterozygosity (including three *HLA-I* and five *HLA-II* genes) was significantly associated with lung cancer, stratified by smoking status ($n = 24{,}379$ smokers and $n = 5{,}503$ non-smokers). We found that among smokers, *HLA-DPA1*, *HLA-DPB1*, and *HLA-DRB1* showed significant associations with disease development, whereas among non-smokers, only *HLA-DQB1* heterozygosity was associated with reduced lung cancer risk (Figure 2). We also investigated the association between LoA and GoA with lung cancer risk (Figures S1A and S1B), although the sample size was very small. We additionally estimated the effects of heterozygosity by removing individuals with either GoA and LoA HLA genes (Figure S1C) and obtained similar results. To formally assess whether the HLA genes have different effects among smokers and non-smokers, we performed interaction analyses for eight HLA genes (Table S3). We found that *HLA-DPA1* and *HLA-DPB1* heterozygosity exhibited significantly different effects among smokers and non-smokers ($p = 0.0063$ for *HLA-DPA1* and $p = 0.0089$ for *HLA-DPB1*).

HLA heterozygosity across different lung cancer subtypes

Next, we performed a stratified analysis to investigate the effect of *HLA* heterozygosity on lung cancer subtypes. We stratified the cohort by cancer subtype and conducted logistic regression separately for smokers and non-smokers (Figure 3). We found that all *HLA*-II genes exhibited protective effects for SCC among smokers but not for non-smokers. In contrast, heterozygosity of *HLA-DPA1* showed risk associations for carcinoids and NSCLC among non-smokers. Interaction analysis indicated *HLA-DPA1* has differential effects between smokers and non-smokers for ADC ($p = 0.0356$) and carcinoids ($p = 0.0471$) (Table S4). No significant differential effects between smokers and non-smokers were observed for SCC, possibly due to the small number of SCC cases among non-smokers.

Additive effects of *HLA* allele heterozygosity

Inspired by genome-wide association studies (GWASs), we further investigated additive effects of heterozygosity at *HLA* alleles on lung cancer risk. We typed 147 2-digit *HLA* alleles and 423 4-digit *HLA* alleles. We then performed stratified analysis for smokers and non-smokers (Figure 4 and Tables S5–S8). We performed logistic

Table 2. Multivariable logistic regression models varying *HLA* variable definition

	Model description	OR (95% CI)	*p* value
Model 1	*HLA* variable counts the total number of heterozygotes per sample for all *HLA* loci	0.98 (0.96, 0.99)	2.72×10^{-3}
Model 2	*HLA* variable counts only samples with heterozygosity at all HLA loci	0.92 (0.86, 0.98)	8.86×10^{-3}
Model 3	*HLA* variable counts the total number of heterozygotes per sample for *HLA-A*, *HLA-B*, and *HLA-C*	1.00 (0.96, 1.03)	0.93
Model 4	*HLA* variable counts the total number of heterozygotes per sample for *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB1*	0.96 (0.94, 0.98)	2.33×10^{-4}
Model 5	*HLA* variable counts the total number of losses of an *HLA* allele per sample	0.94 (0.83, 1.06)	0.32
Model 6	*HLA* variable counts the total number of gains of an *HLA* allele per sample	0.84 (0.53, 1.34)	0.46
Model 7	*HLA* variable counts the total number of LoA for *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB1*	1.03 (0.87, 1.23)	0.70
Model 8	*HLA* variable counts the total number of GoA for *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB1*	1.02 (0.45, 2.32)	0.95

Each model is adjusted for PCs 1–5, sex, age, and smoking status.

regression for each *HLA* allele while conditioning on covariates, including the top five PCs, age, sex, and heterozygosity, of the corresponding *HLA* gene. For instance, to estimate the additive effects of a 2-digit allele *HLA-A*01*, we performed logistic regression with logit[Disease] = HLA-A*01 + HLA-A + PC + age + sex. Importantly, we condition on *HLA-A* (a factor variable including "Homozygous," "Heterozygous," "GoA," and "LoA") to remove gene-level effects, which produces more accurate additive effect estimates for HLA alleles.

We did not detect strong additive effects for *HLA* alleles among non-smokers, whereas several *HLA* alleles showed significant associations among smokers. For instance, *HLA-B:08* (4-digit: *HLA-B*08:01*) and *HLA-A*01* (4-digit: *HLA-A*01:01*) showed significant effects ($p = 5.42 \times 10^{-9}$ and $p = 9.99 \times 10^{-8}$, respectively). Compared to our previous study,[11] we replicated the top HLA alleles, including *HLA-A*01:01*, *HLA-B*08:01*, *HLA-C*07:01*, *HLA-DQA1*05:01*, and *HLA-DRB1*03:01*, and identified an association at *HLA-C*03:04*. We also found a significant association for *HLA-DQB1*06*, consistent with our prior work.[11] To identify independently associated HLA alleles, we performed conditional analyses by including all significant HLA alleles in a joint logistic regression (Table S9). *HLA-A*01:01* and *HLA-C*03:04* remained significant in the conditional analysis ($p = 1.25 \times 10^{-2}$ and $p = 9.91 \times 10^{-4}$, respectively). We repeated the analysis stratified by histology in SCC and ADC (Figures 5 and S2, respectively), which demonstrated no significant associations for ADC but significant 4-digit associations for the following loci in SCC: *HLA-A*01:01*, *HLA-B*07:02*, *HLA-B*08:01*, *HLA-C*03:04*, *HLA-C*07:01*, *HLA-C*07:02*, *HLA-DQA1*05:01*, and *HLA-DRB1*03:01*. Epitopes associated with each of these 4-digit HLA types were queried in the Immune Epitope Database and Tools (IEDB; https://www.iedb.org/) database,[17] and the top 50 results for each are displayed in Table S10.

Discussion

The present study characterizes the association between germline *HLA* region heterozygosity and lung cancer risk in a large consortium, identifying that patients with lung cancer have reduced germline *HLA* heterozygosity for class II loci. Furthermore, we identified differential effects when we stratified by both smoking status and histologic lung cancer subtype. Specifically, we identified a stronger protective effect of *HLA*-II heterozygosity in smokers compared to non-smokers, particularly at the *HLA-DPB1* and *HLA-DPA1* loci. Finally, analysis of the additive effects of *HLA* allele heterozygosity in smokers identified significant associations with several 4-digit *HLA* alleles, including *HLA-B*08:01*, *HLA-A*01:01*, *HLA-C*07:01*, *HLA-DQA1*05:01*, *HLA-DRB1*03:01*, and *HLA-C*03:04*.

The heterozygote advantage hypothesis suggests that heterozygosity at *HLA* loci may improve protection from disease due to improved tumor antigen presentation.[12,13] The results from the present study are largely consistent

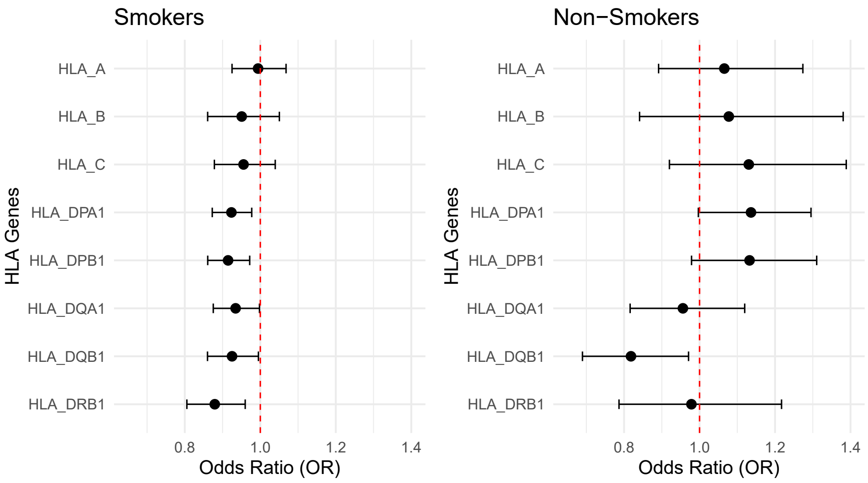

Figure 2. Effects of germline *HLA* heterozygosity on lung cancer risk among smokers and non-smokers
We ran logistic regression for each HLA gene using logit[Disease] = HLA-gene + 5 PCs + age + sex. “Disease” is a binary variable to indicate if an individual has any form of lung cancer. “HLA-gene” is a factor variable that includes “Homozygosity,” “Heterozygosity,” “GoA,” and “LoA.” In this analysis, we calculated the OR for “Heterozygosity” by treating “Homozygosity” as the baseline. Results for “GoA” are shown in Figure S1A; results for “LoA” are shown in Figure S1B. 95% confidence intervals are shown in both panels.

with this hypothesis and previous work by Krishna et al.,[12] complementing their study with a larger sample size and additional histologic information. Krishna et al. analyzed participants from the UKB and FG, finding that germline *HLA*-II heterozygosity was associated with a decreased risk of lung cancer.[12] Interestingly, while *HLA*-I heterozygosity has been previously associated with other metastatic cancers,[18–22] we replicate Krishna et al.’s[12] finding that *HLA*-II heterozygosity is primarily associated with lung cancer, despite identifying a small number of *HLA*-I associations that increase risk in our additive analysis. Furthermore, Krishna et al. demonstrated that *HLA*-II homozygosity conferred a higher lifetime risk of lung cancer, with 13.9% risk for current smokers who were also homozygous

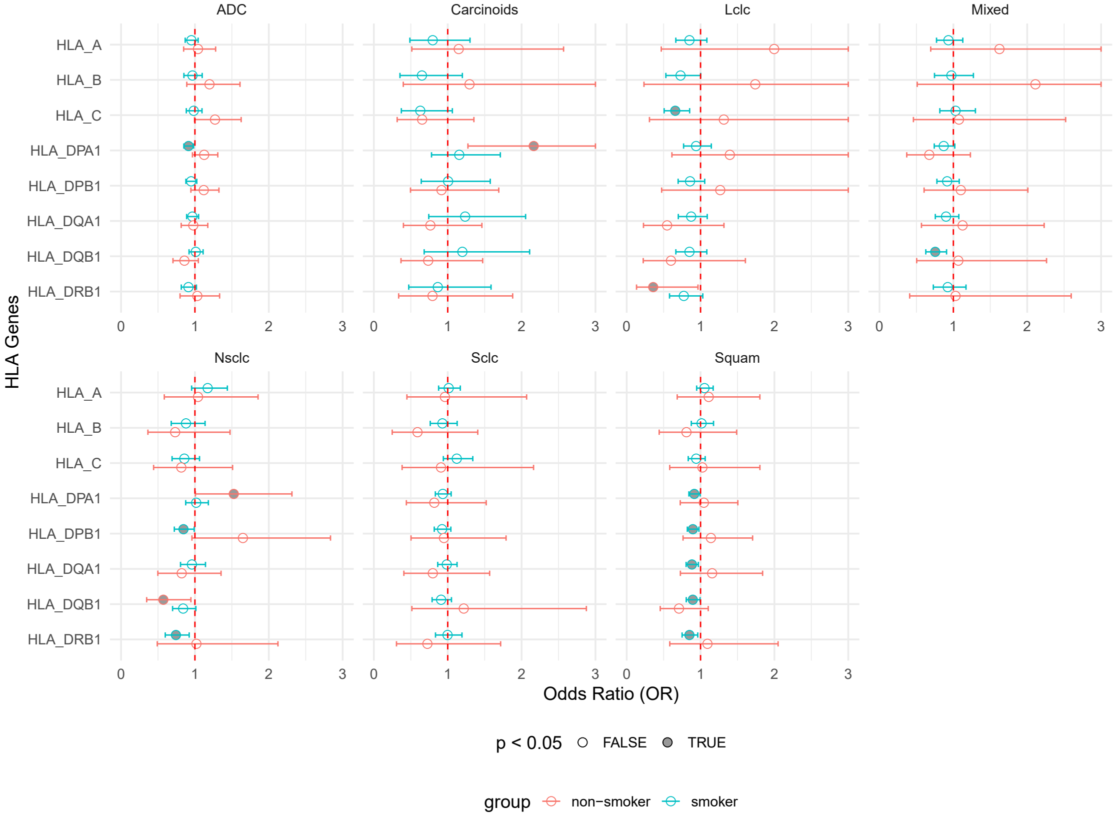

Figure 3. Effects of *HLA* heterozygosity on seven subtypes of lung cancer, stratified by smoking status
Confidence intervals are truncated at OR = 3 for visualization. Significant associations are denoted by filled circles. ADC, adenocarcinoma; Lclc, large cell lung carcinoma; Mixed, multiple types; Nsclc, non-small cell lung cancer; Sclc, small cell lung cancer; Squam, squamous cell carcinoma. 95% confidence intervals are shown in all panels.

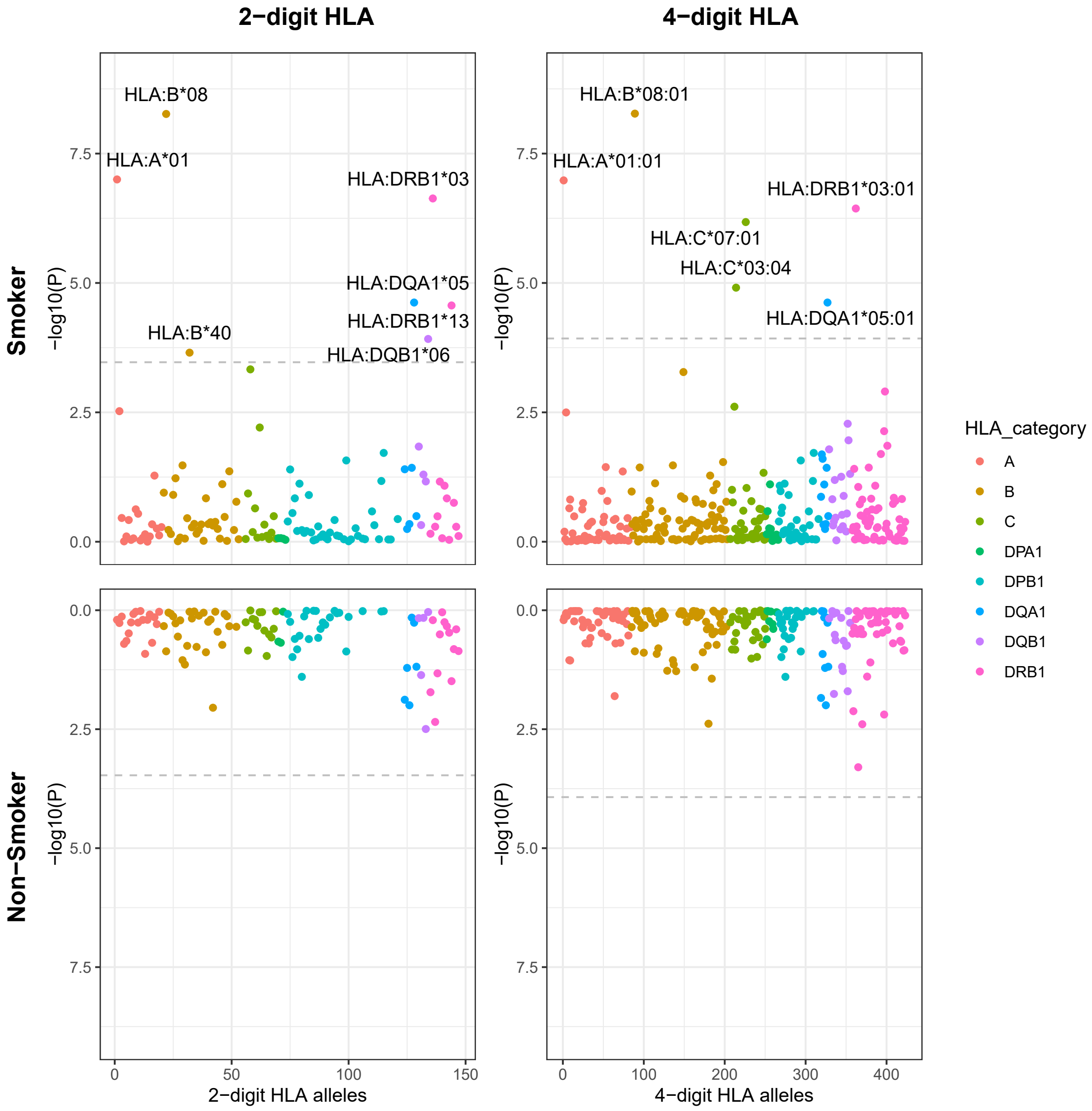

Figure 4. Logistic regression analysis for each 2-digit/4-digit HLA allele with aggregated disease status: logit[Disease] = HLA-allele + HLA-gene + PC + age + sex

"HLA-allele" is the dosage of any 2-digit/4-digit HLA alleles (e.g., HLA:A*01 dosage, which consists of 0, 1, 2). "HLA-gene" is a factor variable for the corresponding HLA gene (e.g., HLA-A, which consists of "Homozygous," "Heterozygous," "GoA," and "LoA"). The OR and p values are reported for all 2-digit/4-digit HLA alleles. The $-\log_{10}(p \text{ value})$ is visualized for each HLA allele (top row: smokers; bottom row: non-smokers; left column: 2-digit HLA alleles; right column: 4-digit HLA alleles).

at *HLA-DRB1*.[12] Molecular studies have provided additional evidence: (1) *HLA*-II peptide-binding groove amino acid heterozygosity is associated with reduced lung cancer risk, (2) single-cell RNA-sequencing data demonstrated alterations in *HLA*-II expressing lung macrophages and epithelial cells in smokers, and (3) neoantigen repertoire analysis demonstrated a loss of alleles with larger neopeptide repertoires in *HLA*-II LoH samples.[12] *HLA* heterozygosity research has also demonstrated that a careful balance between cancer and autoimmunity exists, with several studies demonstrating a complex association between *HLA* heterozygosity and immune-mediated disease risk as *HLA* heterozygosity may confer an increased risk of immune-mediated diseases.[23–25]

Our study also aimed to characterize effects from specific *HLA*-II loci on risks for lung cancer. *HLA-DRB1* heterozygosity was replicated as an important locus for associated with increased (e.g., homozygous) or decreased (e.g., heterozygous) disease risk in our study, with the effect appearing for smokers and in the SCC and NSCLC subtypes. Our histology-specific results suggested that the unique protective effect of *HLA*-II heterozygosity exists among SCC patients with a history of smoking but not ADC patients with a history of smoking, consistent with previous work.[11] We also identified several *HLA* loci at 4-digit resolution associated with increased risk in the additive analysis, including *HLA-B*08:01*, *HLA-A*01:01*, *HLA-C*07:01*, *HLA-DQA1*05:01*, *HLA-DRB1*03:01*, and *HLA-C*03:04*. In our prior classical

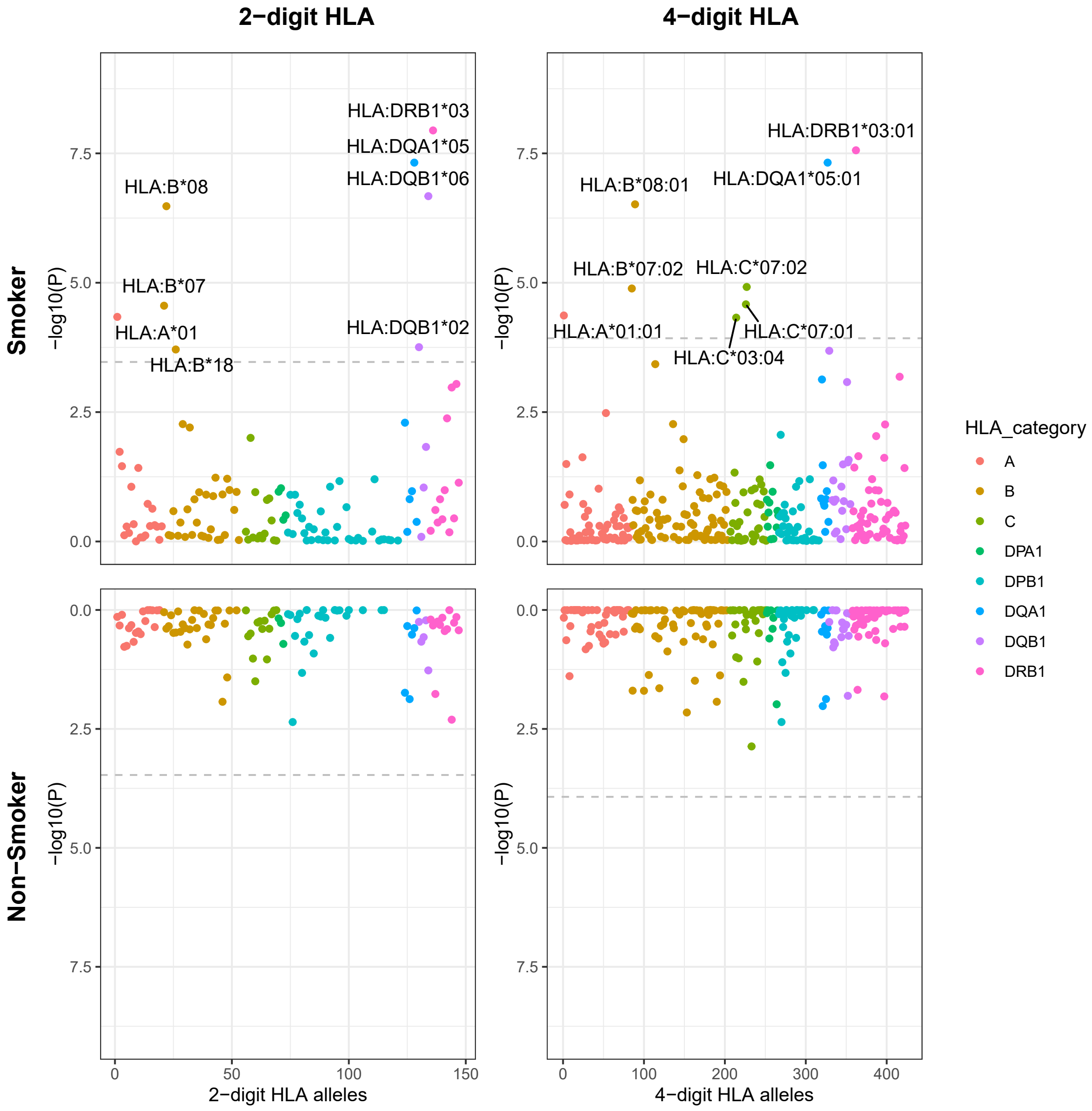

Figure 5. Logistic regression analysis for each 2-digit/4-digit HLA allele in SCC: logit[Disease] = HLA-allele + HLA-gene + PC + age + sex

"HLA-allele" is the dosage of any 2-digit/4-digit HLA alleles (e.g., HLA-A*01 dosage, which consists of 0, 1, 2). "HLA-gene" is a factor variable for the corresponding HLA gene (e.g., HLA-A, which consists of "Homozygous," "Heterozygous," "GoA," and "LoA"). The OR and p values are reported for all 2-digit/4-digit HLA alleles. The $-\log_{10}(p \text{ value})$ is visualized for each HLA allele (top row: smokers; bottom row: non-smokers; left column: 2-digit HLA alleles; right column: 4-digit HLA alleles).

GWAS, which tested the association between specific alleles and disease status, *HLA-B*08:01*, *HLA-A*01:01*, *HLA-C*07:01*, *HLA-DQA1*05:01*, and *HLA-DRB1*03:01* were identified.[11] *HLA-A*01:01*, *HLA-B*08:01*, *HLA-C*07:01*, and *HLA-DQA1*05:01* make up four of the five alleles (*HLA-DQB1*02:01*, the fifth, was not identified in our additive analysis) of the AH8.1 haplotype, which is a well-known and described Caucasian haplotype associated with immune-mediated diseases.[26–28] Our study provides evidence that reduced heterozygosity at alleles in this haplotype are associated with an increased germline risk of developing lung cancer and is consistent with our previous GWAS analysis,[11] highlighting a significant association between these alleles and lung cancer.

While the association between lung cancer and smoking is well described,[29] our study and previous work[12] identified a unique protective effect of heterozygosity at *HLA*-II in smokers. Krishna et al.[12] also demonstrated that the effect was specific to current and former but not never-smokers, which we also found in our larger study. Krishna et al. hypothesize that in smokers, *HLA*-II heterozygosity may improve recognition of smoking-related antigens in developing tumors, which could be presented by alveolar macrophages or dendritic cells for CD4^{+} T cell recognition and development of an antitumor response.[12]

The present study has some limitations. First, the non-smoking analytical group had a much smaller sample size, limiting our ability to resolve effects of HLA heterozygosity on lung cancer risk in this group. For most of the *HLA*-II region, heterozygosity showed little impact on risk for non-smokers, but *HLA-DQB1* heterozygosity was protective. Second, some cancer subtypes, such as mixed, LCLC, and carcinoids, demonstrated small sample sizes, also potentially preventing us from detecting significant effects. Finally, our study was focused on individuals of European ancestry, limiting the generalizability to other ancestral populations.

Taken together, our study validates a previous study of data from the UKB and FG, providing additional evidence that germline *HLA*-II heterozygosity is protective against developing lung cancer. Furthermore, our study highlights differences in protection across various *HLA*-II alleles at 2- and 4-digit resolution, in smokers and non-smokers, and across multiple histologic subtypes. Future research, as suggested by Krishna et al.,[12] may evaluate whether smokers who are homozygous at *HLA*-II may benefit from low-dose computed tomographic screening at an earlier age. Additional work may assess the relationship between immunotherapy response and *HLA* heterozygosity, as previous work in melanoma has demonstrated that a somatic loss of *HLA*-I heterozygosity was associated with poorer immune checkpoint blockade responses.[22] Our results suggest that the evaluation of *HLA*-II heterozygosity as a prognostic biomarker may be evaluated in conjunction with other clinical variables in future prospective clinical trials.

Data and code availability

Genotype data for the lung cancer OncoArray study have been deposited at the database of Genotypes and Phenotypes (dbGaP) under accession phs001273.v1.p1. Model statements utilized for analyses are included throughout the article. Additional figure code is available upon reasonable request to the lead contact.

Web resources

Immune Epitope Database and Tools database (IEDB), https://www.iedb.org/

Michigan Imputation Server Reference Panels, https://genepi.github.io/michigan-imputationserver/reference-panels/

Acknowledgments

This study is supported by National Institutes of Health (NIH) National Cancer Institute (NCI) grant no. U01CA209414 to D.C.C. and U19CA203654 to C.A. This study is also supported by the Cancer Prevention Research Institute of Texas (CPRIT) (RR180061 and RP240380 to C.C.) and the NCI of the NIH (1R01CA269764 and 3U54CA272691-02S1 to C.C.). Computation is supported by an NIH S10 shared instrument grant S10-OD032185. C.C. is a CPRIT Scholar in Cancer Research.

Declaration of interests

The authors declare no competing interests.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2026.100567.

Received: October 7, 2025
Accepted: January 7, 2026

References

1. Siegel, R.L., Miller, K.D., Fuchs, H.E., and Jemal, A. (2022). Cancer statistics, 2022. CA Cancer J. Clin. *72*, 7–33. https://doi.org/10.3322/caac.21708.
2. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J. Clin. *71*, 209–249. https://doi.org/10.3322/caac.21660.
3. Wei, S., Zhang, H., and Tao, S. (2019). A review of arsenic exposure and lung cancer. Toxicol. Res. *8*, 319–327. https://doi.org/10.1039/c8tx00298c.
4. Hubaux, R., Becker-Santos, D.D., Enfield, K.S.S., Lam, S., Lam, W.L., and Martinez, V.D. (2012). Arsenic, asbestos and radon: emerging players in lung tumorigenesis. Environ. Health *11*, 89. https://doi.org/10.1186/1476-069X-11-89.
5. Bade, B.C., and Dela Cruz, C.S. (2020). Lung Cancer 2020: Epidemiology, Etiology, and Prevention. Clin. Chest Med. *41*, 1–24. https://doi.org/10.1016/j.ccm.2019.10.001.
6. Rodriguez-Canales, J., Parra-Cuentas, E., and Wistuba, I.I. (2016). Diagnosis and Molecular Classification of Lung Cancer. Cancer Treat Res. *170*, 25–46. https://doi.org/10.1007/978-3-319-40389-2_2.
7. Byun, J., Han, Y., Li, Y., Xia, J., Long, E., Choi, J., Xiao, X., Zhu, M., Zhou, W., Sun, R., et al. (2022). Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. Nat. Genet. *54*, 1167–1177. https://doi.org/10.1038/s41588-022-01115-x.
8. McKay, J.D., Hung, R.J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D.C., Caporaso, N.E., Johansson, M., Xiao, X., Li, Y., et al. (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat. Genet. *49*, 1126–1132. https://doi.org/10.1038/ng.3892.
9. Long, E., Patel, H., Byun, J., Amos, C.I., and Choi, J. (2022). Functional studies of lung cancer GWAS beyond association. Hum. Mol. Genet. *31*, R22–R36. https://doi.org/10.1093/hmg/ddac140.
10. Bossé, Y., and Amos, C.I. (2018). A Decade of GWAS Results in Lung Cancer. Cancer Epidemiol. Biomarkers Prev. *27*, 363–379. https://doi.org/10.1158/1055-9965.EPI-16-0794.
11. Ferreiro-Iglesias, A., Lesseur, C., McKay, J., Hung, R.J., Han, Y., Zong, X., Christiani, D., Johansson, M., Xiao, X., Li, Y., et al. (2018). Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. Nat. Commun. *9*, 3927. https://doi.org/10.1038/s41467-018-05890-2.
12. Krishna, C., Tervi, A., Saffern, M., Wilson, E.A., Yoo, S.K., Mars, N., Roudko, V., Cho, B.A., Jones, S.E., Vaninov, N., et al. (2024). An immunogenetic basis for lung cancer risk. Science *383*, eadi3808. https://doi.org/10.1126/science.adi3808.
13. Penn, D.J., Damjanovich, K., and Potts, W.K. (2002). MHC heterozygosity confers a selective advantage against

multiple-strain infections. Proc. Natl. Acad. Sci. USA *99*, 11260–11264. https://doi.org/10.1073/pnas.162006499.

14. Martínez-Jiménez, F., Priestley, P., Shale, C., Baber, J., Rozemuller, E., and Cuppen, E. (2023). Genetic immune escape landscape in primary and metastatic cancer. Nat. Genet. *55*, 820–831. https://doi.org/10.1038/s41588-023-01367-1.
15. McGranahan, N., Rosenthal, R., Hiley, C.T., Rowan, A.J., Watkins, T.B.K., Wilson, G.A., Birkbak, N.J., Veeriah, S., Van Loo, P., Herrero, J., et al. (2017). Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. Cell *171*, 1259–1271.e11. https://doi.org/10.1016/j.cell.2017.10.001.
16. Han, J., Lin, Y., and Li, W. (2024). Effect of heterozygosity loss of HLA-I on immune evasion and promotion of non-small cell lung cancer brain metastasis and immunotherapy resistance. J. Clin. Oncol. *42*, 209. https://doi.org/10.1200/JCO.2024.42.23_suppl.209.
17. Vita, R., Blazeska, N., Marrama, D., IEDB Curation Team Members, Duesing, S., Bennett, J., Greenbaum, J., De Almeida Mendes, M., Mahita, J., Wheeler, D.K., et al. (2025). The Immune Epitope Database (IEDB): 2024 update. Nucleic Acids Res. *53*, D436–D443. https://doi.org/10.1093/nar/gkae1092.
18. Montesion, M., Murugesan, K., Jin, D.X., Sharaf, R., Sanchez, N., Guria, A., Minker, M., Li, G., Fisher, V., Sokol, E.S., et al. (2021). Somatic HLA Class I Loss Is a Widespread Mechanism of Immune Evasion Which Refines the Use of Tumor Mutational Burden as a Biomarker of Checkpoint Inhibitor Response. Cancer Discov. *11*, 282–292. https://doi.org/10.1158/2159-8290.CD-20-0672.
19. Takahashi, S., Narita, S., Fujiyama, N., Hatakeyama, S., Kobayashi, T., Kato, R., Naito, S., Sakatani, T., Kashima, S., Koizumi, A., et al. (2022). Impact of germline HLA genotypes on clinical outcomes in patients with urothelial cancer treated with pembrolizumab. Cancer Sci. *113*, 4059–4069. https://doi.org/10.1111/cas.15488.
20. Goodman, A.M., Castro, A., Pyke, R.M., Okamura, R., Kato, S., Riviere, P., Frampton, G., Sokol, E., Zhang, X., Ball, E.D., et al. (2020). MHC-I genotype and tumor mutational burden predict response to immunotherapy. Genome Med. *12*, 45. https://doi.org/10.1186/s13073-020-00743-4.
21. Chowell, D., Krishna, C., Pierini, F., Makarov, V., Rizvi, N.A., Kuo, F., Morris, L.G.T., Riaz, N., Lenz, T.L., and Chan, T.A. (2019). Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. Nat. Med. *25*, 1715–1720. https://doi.org/10.1038/s41591-019-0639-4.
22. Chowell, D., Morris, L.G.T., Grigg, C.M., Weber, J.K., Samstein, R.M., Makarov, V., Kuo, F., Kendall, S.M., Requena, D., Riaz, N., et al. (2018). Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. Science *359*, 582–587. https://doi.org/10.1126/science.aao4572.
23. Terao, C., Okada, Y., Ikari, K., Kochi, Y., Suzuki, A., Ohmura, K., Matsuo, K., Taniguchi, A., Kubo, M., Raychaudhuri, S., et al. (2017). Genetic landscape of interactive effects of HLA-DRB1 alleles on susceptibility to ACPA(+) rheumatoid arthritis and ACPA levels in Japanese population. J. Med. Genet. *54*, 853–858. https://doi.org/10.1136/jmedgenet-2017-104779.
24. Chen, G., Zhu, C., Chinoy, H., Amos, C.I., Morris, A.P., Lamb, J.A.; International Myositis Assessment & Clinical Studies Group (IMACS); and Myositis Genetics Scientific Interest Group (MYOGEN) (2025). HLA loci heterozygosity modulates genetic risk in idiopathic inflammatory myopathies. Ann. Rheum. Dis. *84*, 1696–1705. https://doi.org/10.1016/j.ard.2025.07.002.
25. Lenz, T.L., Deutsch, A.J., Han, B., Hu, X., Okada, Y., Eyre, S., Knapp, M., Zhernakova, A., Huizinga, T.W.J., Abecasis, G., et al. (2015). Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat. Genet. *47*, 1085–1090. https://doi.org/10.1038/ng.3379.
26. Price, P., Witt, C., Allcock, R., Sayer, D., Garlepp, M., Kok, C.C., French, M., Mallal, S., and Christiansen, F. (1999). The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. Immunol. Rev. *167*, 257–274. https://doi.org/10.1111/j.1600-065x.1999.tb01398.x.
27. Candore, G., Lio, D., Colonna Romano, G., and Caruso, C. (2002). Pathogenesis of autoimmune diseases associated with 8.1 ancestral haplotype: effect of multiple gene interactions. Autoimmun. Rev. *1*, 29–35. https://doi.org/10.1016/s1568-9972(01)00004-0.
28. Miller, F.W., Chen, W., O'Hanlon, T.P., Cooper, R.G., Vencovsky, J., Rider, L.G., Danko, K., Wedderburn, L.R., Lundberg, I.E., Pachman, L.M., et al. (2015). Genome-wide association study identifies HLA 8.1 ancestral haplotype alleles as major genetic risk factors for myositis phenotypes. Genes Immun. *16*, 470–480. https://doi.org/10.1038/gene.2015.28.
29. Nasim, F., Sabath, B.F., and Eapen, G.A. (2019). Lung cancer. Med. Clin. North Am. *103*, 463–473. https://doi.org/10.1016/j.mcna.2018.12.006.