RESEARCH ARTICLE SUMMARY

IMMUNOGENETICS

An immunogenetic basis for lung cancer risk

Chirag Krishna†, Anniina Tervi†, Miriam Saffern†, Eric A. Wilson†, Seong-Keun Yoo, Nina Mars, Vladimir Roudko, Byuri Angela Cho, Samuel Edward Jones, Natalie Vaninov, Myvizhi Esai Selvan, Zeynep H Gümüş, FinnGen, Tobias L. Lenz, Miriam Merad, Paolo Boffetta, Francisco Martínez-Jiménez, Hanna M. Ollila*‡, Robert M. Samstein*‡, Diego Chowell*‡

INTRODUCTION: Whether the host immune system naturally protects against cancer has long been the subject of intense debate. The cancer immunosurveillance theory ascribes a protective function to the adaptive immune system, whereby T cell–mediated recognition of neoantigens presented by the major histocompatibility complex suppresses early neoplasia. Studies in mice have provided support for the cancer immunosurveillance theory, yet evidence for a protective role of the immune system against cancer in humans has been relatively lacking. In lung cancer, genetic variation in the human leukocyte antigen (*HLA*) locus is linked to tumor evolution and treatment outcomes, but whether *HLA* polymorphisms reduce lung cancer risk—which would imply a role of the host immune system in preventing lung cancer—is currently unclear. Population-scale biobank analysis coupling host genetics with longitudinal clinical data enables a systematic investigation of how *HLA* polymorphism influences lung cancer risk together with smoking and other established risk factors.

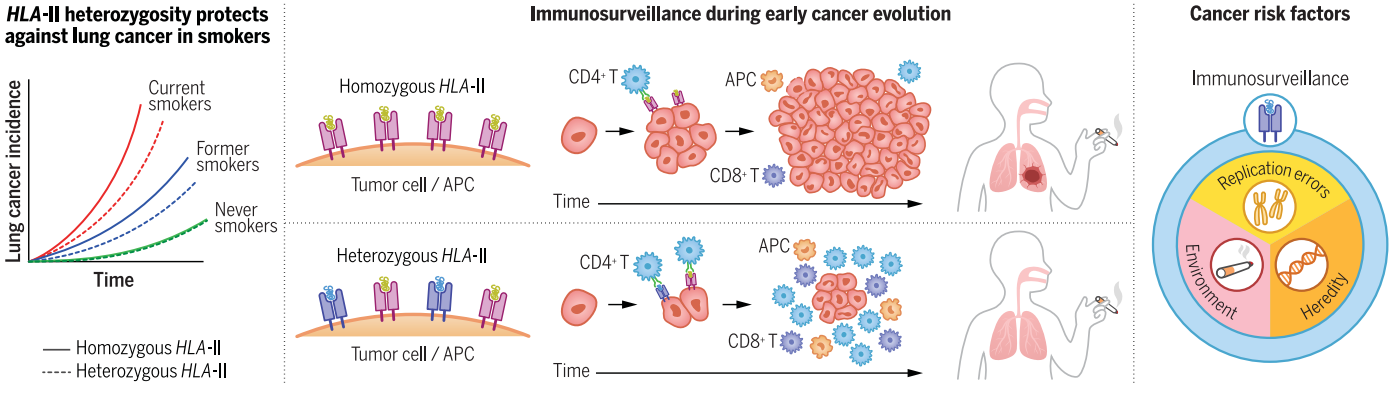
RATIONALE: Understanding the molecular determinants of cancer risk is critical for early cancer detection and strategies to limit cancer mortality. Tobacco smoking increases lung cancer risk and is associated with a heightened somatic mutation rate that drives neoplastic potential, but whether there are additional risk factors that further modify lung cancer susceptibility, even among smokers, is unclear. The *HLA* heterozygote advantage theory posits that an *HLA* genotype encoding two different allomorphs enables presentation of a more diverse antigenic peptide repertoire to the immune system—and subsequent superior immune control of infected or cancerous cells—than does an *HLA* genotype encoding two equivalent allomorphs. Thus, heterozygous *HLA* allomorphs may present more neoantigens arising from smoking-derived somatic mutations. In this study, we evaluated the effect of *HLA* heterozygosity on lung cancer risk, leveraging genetic and longitudinal clinical data from the UK Biobank and FinnGen together with multimodal genomic analyses of nonmalignant and lung tumor samples.

RESULTS: In both the UK Biobank and FinnGen, we found that heterozygosity at the HLA class II (*HLA*-II) loci was associated with reduced risk of lung cancer over more than a decade of follow-up. *HLA*-II heterozygosity was associated with reduced risk of lung cancer in both current and former but not never-smokers, suggesting that smoking-derived antigens may augment the immune response to early neoplastic disease. *HLA*-II homozygosity conferred substantial lifetime risk of disease (e.g., in the UK Biobank, 13.9% for current smokers homozygous at *HLA-DRB1*) and was independent of known clinical and genetic risk factors, including a genome-wide polygenic risk score. Heterozygosity of amino acid sites within the *HLA*-II peptide binding groove was also associated with reduced risk of lung cancer, whereas analysis of single-cell RNA-sequencing data from nonmalignant and tumor lung samples showed that lung macrophages and epithelial cells express *HLA*-II and are affected by smoking. Analysis of tumor genomes from the The Cancer Genome Atlas (TCGA) cohort, the Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort, and Hartwig Medical Foundation cohort revealed widespread loss of heterozygosity (LOH) of the *HLA*-II loci in lung cancer, with rates of LOH equaling those of *HLA*-I. An analysis of neoantigen repertoires between lung cancer tumors with and without *HLA*-II LOH showed that *HLA*-II LOH favors the loss of alleles with larger neopeptide repertoires, underscoring the importance of the *HLA*-II loci and the $CD4^+$ T cell response in lung cancer.

CONCLUSION: The association of *HLA*-II heterozygosity with reduced risk of lung cancer implies that genetic variation in immunosurveillance is a feature of cancer susceptibility, together with environmental exposures, hereditary risk, and DNA replication errors. Our findings broaden understanding of the role of the host immune system in cancer risk and may motivate the incorporation of immunogenetics into lung cancer screening programs. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: diego.chowell@mssm.edu (D.C.); robert.samstein@mountsinai.org (R.M.S.); hanna.m.ollila@helsinki.fi (H.M.O)
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
Cite this article as C. Krishna *et al.*, *Science* **383**, eadi3808 (2024). DOI: 10.1126/science.adi3808

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adi3808



The immunogenetics of lung cancer risk. Genetic epidemiological analyses in the UK Biobank and FinnGen coupled with multimodal genomics reveal that *HLA*-II heterozygosity is associated with reduced lung cancer risk in smokers. The data suggest that heterozygous *HLA* allomorphs promote immune control of early neoplasia through presentation of diverse smoking-related antigens. Thus, genetic variation in immunosurveillance is a key cancer risk factor. APC, antigen-presenting cell.

RESEARCH ARTICLE

IMMUNOGENETICS

An immunogenetic basis for lung cancer risk

Chirag Krishna[1]†, Anniina Tervi[2]†, Miriam Saffern[3,4]†, Eric A. Wilson[3,4,5]†, Seong-Keun Yoo[3,4,5], Nina Mars[2], Vladimir Roudko[3,4], Byuri Angela Cho[3,4,5], Samuel Edward Jones[2], Natalie Vaninov[3,4], Myvizhi Esai Selvan[6], Zeynep H Gümüş[6,7], FinnGen§, Tobias L. Lenz[8], Miriam Merad[3,9,4,10,11], Paolo Boffetta[12,13], Francisco Martínez-Jiménez[14,15], Hanna M. Ollila[1,2,16,17]*‡, Robert M. Samstein[3,4,7,18]*‡, Diego Chowell[3,9,4,5]*‡

Cancer risk is influenced by inherited mutations, DNA replication errors, and environmental factors. However, the influence of genetic variation in immunosurveillance on cancer risk is not well understood. Leveraging population-level data from the UK Biobank and FinnGen, we show that heterozygosity at the *human leukocyte antigen (HLA)*-II loci is associated with reduced lung cancer risk in smokers. Fine-mapping implicated amino acid heterozygosity in the *HLA*-II peptide binding groove in reduced lung cancer risk, and single-cell analyses showed that smoking drives enrichment of proinflammatory lung macrophages and *HLA*-II+ epithelial cells. In lung cancer, widespread loss of *HLA*-II heterozygosity (LOH) favored loss of alleles with larger neopeptide repertoires. Thus, our findings nominate genetic variation in immunosurveillance as a critical risk factor for lung cancer.

Lung cancer is currently the leading cause of worldwide cancer mortality (*1*–*3*). Although diagnosis rates for advanced-stage disease continue to decline, rates for early-stage disease have increased (*2*), highlighting the need for research clarifying the factors underpinning lung cancer risk.

Smoking causes lung cancer through DNA damage and other mechanisms and accounts for more than 80% of lung cancer deaths (*4*). The role of smoking in lung cancer risk and mortality was initially defined in seminal work by Richard Doll over 70 years ago (*5*) and validated in countless studies since then, including recent meta-analyses highlighting a severe dose-response relationship between the number of packs smoked and mortality of lung cancer and other diseases (*6*). Genetic studies have implicated germline genetic variation in lung cancer risk, including mutations in *TP53*, epidermal growth factor receptor (*EGFR*), and others (*1*, *7*). Together, these studies have established lung cancer as a multifactorial disease with diverse genetic and environmental triggers (*8*). However, our understanding of the full spectrum of lung cancer risk factors and how they interact remains incomplete; for example, genome-wide association studies (GWAS) explain only a tiny proportion of the genetic variability in lung cancer risk (*9*). Indeed, there exists wide variability in lung cancer risk even among smokers (*9*, *10*).

The importance of the immune system in conferring protection against pathogens is well established (*11*). However, there is a long-standing debate regarding whether the immune system also protects against cancer. The cancer immunosurveillance hypothesis, initially developed by Ehrlich, Thomas, and Burnet (*12*–*16*), posits that lymphocytes constantly survey tissues for neoplastic cells presenting mutation-derived neoantigens, an activity that could trigger an effective immune response that eliminates developing cancers. The cancer immunoediting theory suggests that the immune system plays dual protective and promoting roles in neoplastic transformation (*17*). Moreover, large cohort studies have noted an increased risk of diverse infection-related and unrelated cancer among solid organ transplant recipients (*18*). A plausible interpretation of these seminal studies is that abrogation or differences in the strength of immune surveillance may lead to variations in cancer risk (*19*–*21*).

Lung cancer is an exemplary disease for the study of immunosurveillance in cancer because the healthy lung is among the most heavily T cell–infiltrated tissues (*22*). Additionally, metastatic lung cancers demonstrate encouraging responses to immune checkpoint blockade (ICB) agents targeting T cells via the PD-L1/PD-1 and CTLA-4 axes (*23*–*25*), highlighting an important role for neoantigen-driven cytotoxic activity in the disease. Furthermore, key studies investigating the basis for ICB response in lung cancer and other tumor types (*26*) have shown that the elevated mutation rate caused by smoking (*23*) promotes increased visibility of neoantigens to cytotoxic T cells (*27*). Thus, these previous studies suggest that there may exist interactions between smoking and the immune system in the development of lung cancer. Yet, such interactions—and the role of the immune system in cancer risk in general—are not well understood.

One clue as to how the immune system is involved in lung cancer risk has arisen from GWAS, which have implicated individual single-nucleotide polymorphisms (SNPs) and alleles of the human leukocyte antigen (*HLA*) class I (*HLA*-I) and II (*HLA*-II) genes in lung cancer susceptibility (*28*–*30*). The *HLA* genes are highly polymorphic (*31*) and encode the major histocompatibility complex (MHC) molecules, which serve as critical gatekeepers of the adaptive immune response through the presentation of self- and foreign antigens for recognition by T cells. Previous work has highlighted the somatic loss of *HLA*-I as a mechanism of immune evasion in lung cancer (*32*, *33*). Furthermore, *HLA*-I and *HLA*-II genotypes influence the oncogenic-driver landscape (*34*, *35*). However, whether and how *HLA* polymorphism interacts with smoking and other risk factors in driving lung cancer risk over time has not currently been addressed.

The heterozygote advantage hypothesis is a foundational principle of the evolution of the *HLA* system and of *HLA*-mediated protection against disease. According to this hypothesis (*36*), individuals heterozygous at *HLA* are afforded greater protection against disease because they present more antigens for T cell recognition through their two different *HLA* allomorphs than do homozygous individuals and consequently clear infected or neoplastic cells more efficiently. Although evidence for heterozygote advantage theory has been demonstrated most clearly in the context of clinical outcomes among individuals who already have the disease [i.e., in delaying progression to AIDS among individuals with HIV (*37*, *38*), clearance of hepatitis B (*39*), or response to ICB in metastatic cancer (*40*–*46*)], whether there

[1]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [2]Institute for Molecular Medicine, Finland (FIMM), HiLIFE, University of Helsinki, Helsinki 00290, Finland. [3]The Marc and Jennifer Lipschultz Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [4]Department of Immunology and Immunotherapy, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [5]Icahn Genomics Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [6]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [7]Center for Thoracic Oncology, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [8]Research Unit for Evolutionary Immunogenomics, Department of Biology, Universität Hamburg, 20146 Hamburg, Germany. [9]Department of Oncological Sciences, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [10]Division of Hematology and Medical Oncology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [11]Human Immune Monitoring Center, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [12]Department of Medical and Surgical Sciences, Alma Mater Studiorum University of Bologna, 40138 Bologna, Italy. [13]Stony Brook Cancer Center, Stony Brook University, New York, NY 11794, USA. [14]Vall d'Hebron Institute of Oncology, Barcelona 08035, Spain. [15]Hartwig Medical Foundation, Amsterdam 1098 XH, the Netherlands. [16]Center for Genomic Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA. [17]Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA 02114, USA. [18]Department of Radiation Oncology, Mount Sinai Hospital, New York, NY 10029, USA.

***Corresponding author. Email: diego.chowell@mssm.edu (D.C.); robert.samstein@mountsinai.org (R.M.S.); hanna.m.ollila@helsinki.fi (H.M.O)**

†These authors contributed equally to this work.

‡These authors contributed equally to this work.

§FinnGen consortium members are listed in the supplementary materials.

exists a protective effect of *HLA* heterozygosity against the development of lung cancer (or other cancer types) is currently unknown. Such an effect, together with established risk factors such as smoking and age, may underscore *HLA* heterozygosity and the immune system in general as critical factors in lung cancer risk.

In this study, we hypothesized that heterozygosity at the *HLA* genes is associated with reduced lung cancer risk over time, on the basis of the assumption that two different *HLA* allomorphs will present more neoplastic antigens than will a single *HLA* allomorph (*47*), thus increasing the likelihood of a cytotoxic reaction against mutated neoplastic cells. To test this hypothesis, we leveraged clinical, genetic, environmental, and longitudinal data from two large-scale population cohorts: the UK Biobank (N = 391,182) and FinnGen (N = 183,163) (fig. S1). We then employed multiple approaches, including fine-mapping and structural analyses of the peptide binding groove and genomic profiling of the healthy lung through single-cell RNA sequencing (scRNA-seq), to clarify the mechanisms underlying *HLA*-mediated protection against lung cancer. Lastly, we investigated somatic loss of heterozygosity (LOH) of the *HLA*-I and *HLA*-II loci in lung cancer tumors from The Cancer Genome Atlas (TCGA) cohort, the Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort, and the Hartwig Medical Foundation cohort.

Immunogenetic and demographic characterization of individuals in the UK Biobank and FinnGen

We sought to examine the effects of *HLA* heterozygosity on lung cancer risk at the population level, defined here as the odds ratio (OR) or hazard ratio (HR) corresponding to diagnosis or death due to lung cancer as a function of *HLA* zygosity and other clinical variables. Thus, we first assembled individual-level genetic, clinical, environmental, and longitudinal clinical data from the UK Biobank and FinnGen (*48*, *49*) (table 1). The UK Biobank and FinnGen are distinctive in size and scope, with rich longitudinal phenotypic and health-related information available through linkage to medical records for each participant followed over time. In addition, the UK Biobank and FinnGen consist of roughly 500,000 and 350,000 genotyped individuals from the UK and Finland, respectively, including the imputation of genotypes at the classical *HLA*-I (*HLA-A*, *HLA-B*, *HLA-C*) and *HLA*-II (*HLA-DRB1*, *HLA-DQB1*, *HLA-DQA1*, *HLA-DPB1*, *HLA-DPA1*) genes. FinnGen in particular has employed a Finnish-specific reference panel for *HLA* imputation (*50*). Thus, the UK Biobank and FinnGen are well suited to address whether *HLA* heterozygosity affects cancer risk.

After performing quality control of *HLA* genotypes as recommended by the UK Biobank and filtering out any individuals with a cancer diagnosis before the start of the UK Biobank study (*48*) (fig. S2; Materials and methods), a cohort of 391,182 individuals was identified for further analysis. The primary clinical endpoint of interest in our study was a first diagnosis or death due to lung cancer [defined by ICD-10 codes (ICD, International Statistical Classification of Diseases and Related Health Problems)] over a roughly 14-year follow-up period, with participants recruited between March 2007 and October 2011. We documented 2468 individuals in the UK Biobank fitting these criteria of lung cancer case, with the remaining individuals designated as healthy controls (N = 384,928) after excluding individuals with missing data (fig. S2). Consistent with previous reports, the most common histological subtypes of lung cancer were adenocarcinoma (N = 700), squamous cell carcinoma (N = 338), and small cell carcinoma (N = 192); with the remaining patients representing other or missing histologies. Of the lung cancer cases, 86.8% recorded as current or former smokers and the remainder as never-smokers. Gender was roughly evenly split between males and females in both cases and controls. To replicate our findings from the UK Biobank, we assembled lung cancer case and control data from FinnGen (*51*–*53*) by using the same criteria applied to the UK Biobank. We identified 3480 lung cancer cases in FinnGen (N = 183,163 total individuals after filtering; fig S3). Although the distribution of lung cancer subtypes in FinnGen was similar to that in the UK Biobank, a key difference was the proportion of smokers in each category among lung cancer cases (for instance, 70.8% current smokers in FinnGen as compared with 41.8% current smokers in the UK Biobank). The percentage of male lung cancer patients (78.3%) far exceeded the percentage of female lung cancer patients (21.7%) in FinnGen, whereas the distribution was more balanced in the UK Biobank.

Although several prior studies have used the imputed *HLA* genotypes provided by the UK Biobank for bespoke analyses (*54*), we undertook several additional quality checks to validate the quality of imputed *HLA* genotyping in the UK Biobank. We first compared the allele frequency of 2-field (i.e., four-digit) alleles in the UK Biobank to population-level allele frequencies from the Allele Frequency Net Database (AFND) (*55*) (Fig. 1A); the frequencies were highly correlated ($P < 0.0001$; Spearman's $\rho = 0.91$), suggesting that allele genotyping in the UK Biobank is representative of the allele genotypes in the wider UK population. We observed similar results comparing allele frequencies in FinnGen with Finnish population allele-frequency data ($P < 0.0001$; Spearman's $\rho = 0.85$) (Fig. 1B). We then directly compared allele frequencies in the UK Biobank with those in FinnGen (Fig. 1C); allele frequencies were generally correlated ($P < 0.0001$; Spearman's $\rho = 0.66$) except for a few *HLA*-I and *HLA*-II alleles that approached allele frequencies of up to 10% in the individual cohorts. The strong correlations between allele frequencies in the UK Biobank or FinnGen with the general

Table 1. Clinical and demographic characteristics of lung cancer cases and controls in UK Biobank and FinnGen.

Characteristic	UK Biobank Full Cohort	UK Biobank Lung Cancer	FinnGen Full Cohort	FinnGen Lung Cancer
Total individuals	391,182	2468	183,163	3480
Healthy controls	384,928		179,233	
Lung cancer subtype				
Adenocarcinoma (%)		700 (28.3%)		801 (23.0%)
Squamous (%)		338 (13.7%)		679 (19.5%)
Small cell (%)		192 (7.8%)		336 (9.7%)
Other/unknown (%)		1238 (50.2%)		1664 (47.8%)
Smoking status				
Current (%)	40,674 (10.5%)	1107 (44.9%)	50,179 (27.4%)	2464 (70.8%)
Former (%)	135,410 (35.0%)	1035 (41.9%)	43,493 (23.7%)	666 (19.1%)
Never (%)	211,312 (54.5%)	326 (13.2%)	89,041 (48.6%)	350 (10.1%)
Sex				
Male	177,259 (54.2%)	1270 (48.5%)	89,052 (48.6%)	2724 (78.3%)
Female	210,137 (45.8%)	1297 (51.4%)	93,661 (51.1%)	756 (21.7%)
Age (I.Q.R.)	58 (50-63)	67 (63-71)	63 (49-74)	75 (70-80)

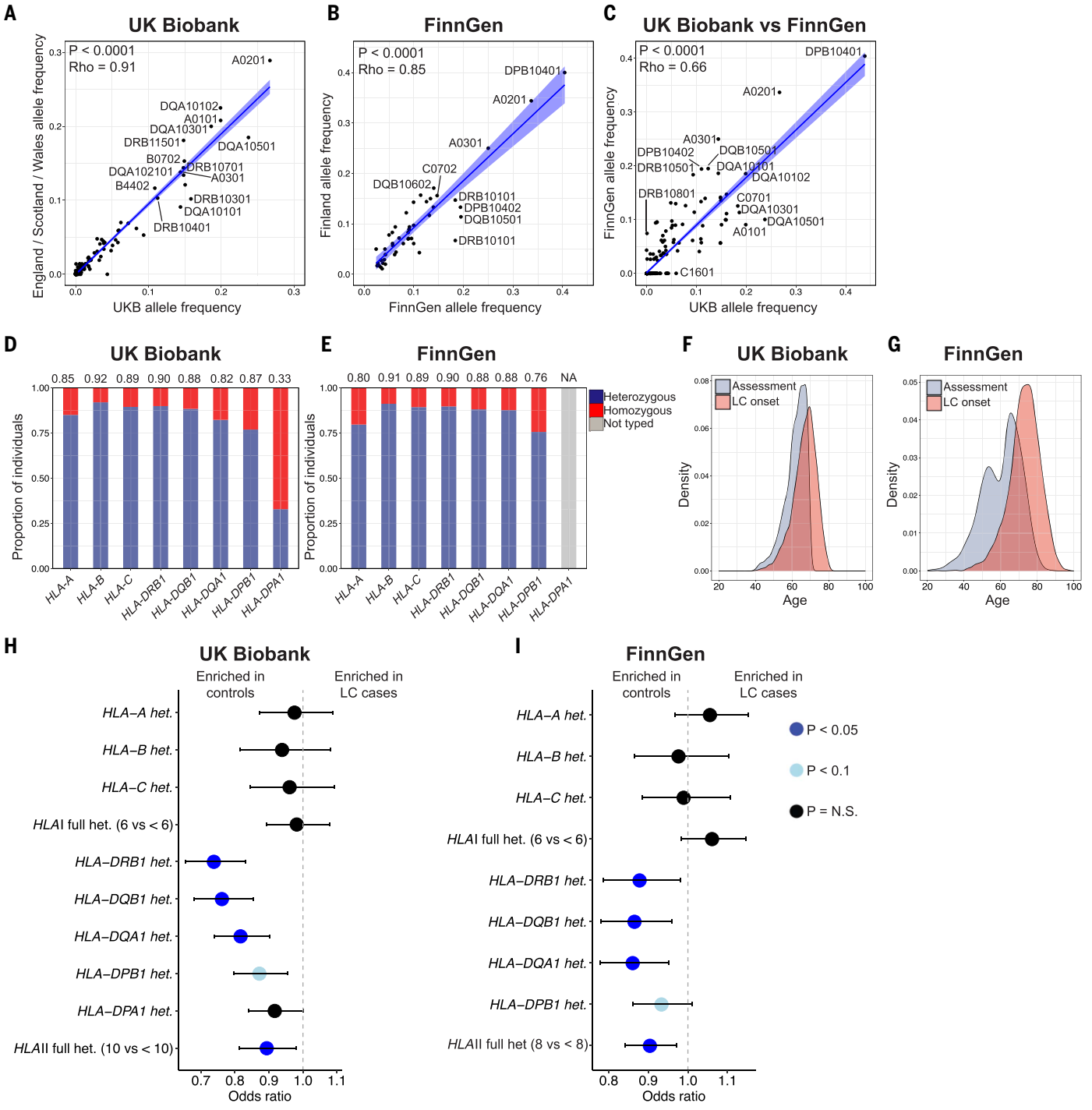


Fig. 1. *HLA* genotype and associations with lung cancer risk in UK Biobank and FinnGen. (**A**) Correlation of *HLA* allele frequencies in the UK Biobank with mean allele frequencies across England, Scotland, and Wales was obtained from the Allele Frequency Net Database (AFND). *P* value computed using Spearman correlation. (**B**) Correlation of *HLA* allele frequencies in FinnGen with allele frequencies from Finland obtained from AFND. *P* value calculated with Spearman correlation. (**C**) Correlation of *HLA* allele frequencies in UK Biobank with allele frequencies in FinnGen. *P* value calculated with Spearman correlation. (**D**) Rates of heterozygosity at four-digit allele resolution in UK Biobank. (**E**) Rates of heterozygosity at four-digit allele resolution in FinnGen. *HLA-DPA1* genotypes were not imputed in FinnGen and are thus left gray. (**F**) Distribution of age at onset among lung cancer cases compared with age at first assessment in UK Biobank. (**G**) Distribution of age at onset among lung cancer cases compared with age at first assessment in FinnGen. (**H**) Multivariable logistic regression analyses testing heterozygosity at the indicated locus together with all clinical and demographic covariates for associations with lung cancer case and control status in UK Biobank. Forest plots depict odds ratio from logistic regression and 95% CI. (**I**) Multivariable logistic regression analyses testing heterozygosity at the indicated locus and all clinical and demographic covariates for associations with lung cancer case and control status in FinnGen. Forest plots depict OR from logistic regression and 95% CI.

population—and with each other—were also observed when stratifying the correlation analyses by locus (fig. S4 and tables S1 to S3). Earlier literature demonstrates that *HLA* allele frequencies differ across geographic locations and across ethnic groups (*56*). Furthermore, previous comparisons of genetic ancestry and population structure in large cohorts such as gnomAD and FinnGen have shown that the Finnish population is genetically isolated from the rest of Europe. Moreover, because of relatively recent bottlenecks, the Finnish population is enriched in alleles that have not yet been selected out (*49*, *57*). However, our data suggest that in general, *HLA* allele frequencies are correlated between the two studies. Consistent with earlier studies (*38*, *40*, *41*), we defined heterozygosity at each of the eight *HLA*-I and *HLA*-II loci as different alleles at 2-field resolution, since the 2-field allele codes represent variation at the amino acid sequence level of the *HLA* molecules (*58*). Consistent with our analyses showing broad concordance of allele frequencies between the two cohorts, we found that the rates of *HLA* allele heterozygosity between UK Biobank and FinnGen were also highly comparable (Fig. 1, D and E).

To provide additional confidence in the robustness of our association results using imputed *HLA* genotypes, we assessed the concordance of *HLA* genotypes called from whole-exome sequencing data with imputed genotypes, both assessed in the same individuals from the UK Biobank. We used two independent, well-validated tools for genotyping of *HLA*-I and *HLA*-II—HLA*LA (*59*) and HLA-HD (*60*)—both of which have strong performance relative to other methods (*61*). Using these methods, we genotyped *HLA*-I and *HLA*-II from 43,000 individuals from the UK Biobank on the basis of whole-exome sequencing data available from blood. The proportion of individuals heterozygous at individual *HLA*-II loci across the three methods (imputation, exome typed with HLA*LA, and exome typed with HLA-HD) was comparable (fig. S5A) except for heterozygosity at *HLA-DQA1* assessed with HLA*LA (proportion of individuals heterozygous = 0.52). The frequencies of individual alleles were also concordant between imputed genotypes and exome genotypes, regardless of whether the exome genotypes were called with HLA*LA or HLA-HD (fig. S5B). We then calculated the concordance between zygosity (either heterozygous or homozygous) defined with the imputed *HLA* genotypes provided by the UK Biobank and zygosity defined with genotypes obtained from exome data that used HLA*LA or HLA-HD. The concordance rates were 95% or higher for all loci except for *HLA-DQA1* (69%) (fig. S5C), suggesting that any significant association results for *HLA-DQA1* should be interpreted with caution and replicated in an independent cohort.

HLA-II heterozygosity is associated with reduced lung cancer risk

Having validated the high quality of *HLA* genotyping in the UK Biobank, we next asked whether *HLA* heterozygosity is associated with reduced lung cancer risk in the UK Biobank by performing a multivariable logistic regression analysis. We controlled for clinical and demographic covariates that are known to influence lung cancer risk and outcomes in the UK Biobank (*62*). We reasoned that a multivariable model accounting for all covariates would be especially critical given the drastic difference in age between lung cancer cases in the UK Biobank (median 67) (table 1 and Fig. 1F) and those in FinnGen (median 75) (table 1 and Fig. 1G). Specifically, we fit an independent multivariable logistic regression for each *HLA* locus testing heterozygosity at the locus as a predictor together with clinical and demographic covariates, including smoking status (Materials and methods) (Fig. 1H). The outcome was a binary variable indicating diagnosis or death due to lung cancer (N = 2468 in the UK Biobank) or healthy control (N = 384,928 in the UK Biobank) (table 1). In addition to the eight multivariable models fit for each *HLA*-I (*HLA-A*, *HLA-B*, *HLA-C*) and *HLA*-II (*HLA-DRB1*, *HLA-DQB1*, *HLA-DQA1*, *HLA-DPB1*, *HLA-DPA1*) locus, we fit two additional models for maximal heterozygosity at *HLA*-I (6 alleles versus <6) and maximal heterozygosity at *HLA*-II (10 alleles versus <10), which is consistent with the original definition of heterozygote advantage (*38*) (table S4). This analysis revealed that *HLA*-II heterozygosity was significantly enriched in controls relative to lung cancer cases, and thus associated with reduced risk of lung cancer. We observed a protective effect for heterozygosity at each *HLA*-II locus and for maximal heterozygosity across all five *HLA*-II loci, but not for *HLA*-I. The effect of heterozygosity was strongest for *HLA-DRB1* [$P = 5.19 \times 10^{-7}$; logistic regression estimate = −0.3; OR = 0.74; OR 95% confidence interval (CI): 0.65 to 0.83] and *HLA-DQB1* ($P = 2.80 \times 10^{-6}$; logistic regression estimate = −0.27; OR = 0.76; OR 95% CI: 0.68 to 0.85) (Fig. 1H). We then repeated these analyses using the subset of individuals in the UK Biobank for whom whole-exomes were available. We observed that the protective effect of heterozygosity was also seen when we used *HLA* genotypes called from exomes that used both HLA*LA and HLA-HD, despite the much smaller sample size of the exome subset (N = 835 cases, 41,708 and 41,618 controls for HLA*LA and HLA-HD, respectively) (fig. S6), suggesting that the effect of *HLA*-II heterozygosity on reduced risk of lung cancer is independent of the genotyping method used. Although genotypes for *HLA-DPA1* were not available in FinnGen, we observed a similar protective effect of overall *HLA*-II heterozygosity (P = 0.006) and for heterozygosity at *HLA-DRB1* (P = 0.02), *HLA-DQA1* (P = 0.004), and *HLA-DQB1* (P = 0.006) (table S5). *HLA-DPB1* heterozygosity did not associate with lung cancer. However, the point estimate was protective and the P value was close to significance (P = 0.089), suggesting that larger sample sizes may clarify the association (Fig. 1I). These results suggest that heterozygosity at *HLA*-II is associated with reduced risk of lung cancer.

We next asked whether *HLA*-II heterozygosity conferred protection against lung cancer risk over time. We computed follow-up times and censoring for all participants in the UK Biobank as the time from the date of first assessment to the date of diagnosis or death due to lung cancer (Materials and methods). We first assessed the effect of smoking on lung cancer risk in the UK Biobank using a multivariable Cox regression analysis, treating smoking status as a positive control for our follow-up time and censoring calculations. As expected, current smokers had the highest risk of developing lung cancer in both the UK Biobank (Fig. 2A and fig. S9A) and FinnGen (Fig. 2B and fig. S9B), followed by former smokers. To define the role of *HLA*-II heterozygosity in mediating lung cancer risk over time, we asked whether heterozygosity afforded additional protection against lung cancer within current, former, and never-smokers, reasoning that *HLA* heterozygosity may account for some of the variability in lung cancer risk among individuals with the dominant risk factor. Thus, we assessed the effect of maximal *HLA*-II heterozygosity (10 alleles versus <10) in the UK Biobank and FinnGen within each smoking category (current/former/never), adjusting for all covariates within each category (i.e., a separate multivariable Cox regression analysis within each smoking category) (fig. S7).

Notably, we found that among former smokers, maximal *HLA*-II heterozygosity was associated with reduced risk of lung cancer (P = 0.006; HR = 0.82; HR 95% CI: 0.71 to 0.94) (Fig. 2C, fig. S9C, and table S6). This result suggests two critical points: (i) that *HLA*-II heterozygosity is associated with reduced lung cancer risk even when adjusting for known clinical and demographic covariates, and (ii) that *HLA*-II heterozygosity accounts for some of the variability in lung cancer risk among smokers. We repeated these analyses in FinnGen and found that maximal *HLA*-II heterozygosity (8 alleles versus <8, because DPA1 genotypes were not available in FinnGen) was associated with reduced risk of lung cancer among current smokers (Fig. 2D, fig. S9D, and table S7). That the protective effect of *HLA*-II heterozygosity was observed in former smokers in the UK Biobank and in current smokers in FinnGen may reflect differences in smoking habits between the two populations, differences in the proportions of current and former smokers among lung cancer cases in each cohort (41.8% current smokers in the UK Biobank, 78.3% current smokers in FinnGen), or a higher proportion

of former smokers misclassified as current smokers in FinnGen as compared with those in the UK Biobank (table 1). We did not observe a significant difference in cancer risk between heterozygous and homozygous never-smokers, suggesting a possible interaction between *HLA*-II heterozygosity and smoking in driving lung cancer risk.

We next assessed the effects of heterozygosity at each *HLA*-II locus on lung cancer risk over time using multivariable Cox regression analyses in the UK Biobank (table S6). Consistent with our earlier logistic regression analyses (Fig. 1H), the strongest effects of heterozygosity were observed for *HLA-DRB1* (former smokers $P = 1.47 \times 10^{-7}$; HR = 0.64; HR 95% CI: 0.55 to 0.76) (Fig. 3A) and *HLA-DQB1* (former smokers $P = 1.71 \times 10^{-6}$; HR = 0.68; HR 95% CI: 0.58 to 0.79) (Fig. 3B), with a protective effect observed in both current and former smokers. Given the linkage disequilibrium between these two loci, we performed Cox regression analyses testing the effect of *HLA-DRB1* heterozygosity among individuals homozygous

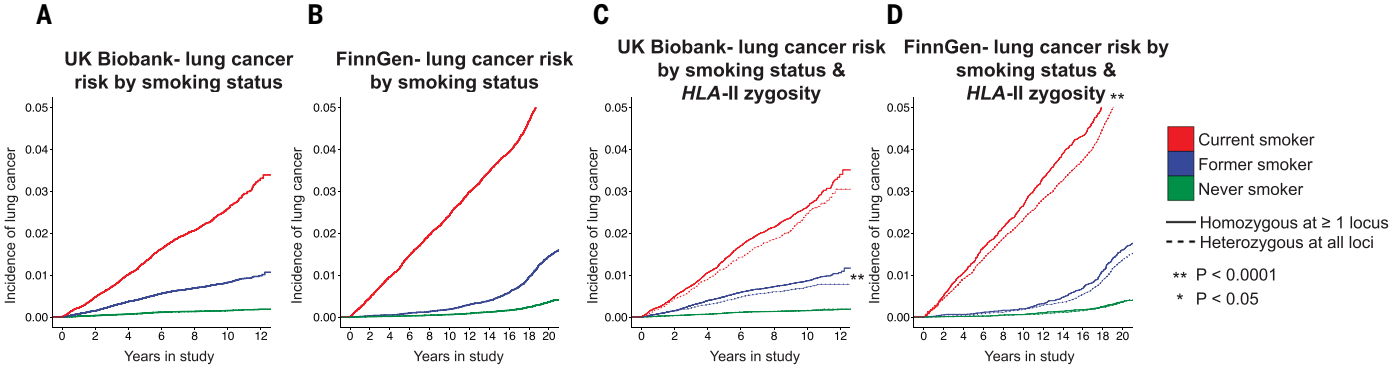


Fig. 2. Maximal *HLA*-II heterozygosity is associated with reduced lung cancer incidence among smokers in UK Biobank and FinnGen. (**A**) Effect of smoking status (current/former/never) on lung cancer incidence in UK Biobank. (**B**) Effect of smoking status (current/former/never) on lung cancer incidence in FinnGen. (**C**) Association of maximal *HLA*-II heterozygosity (10 distinct alleles at *HLA-DRB1*, *-DQB1*, *-DQA1*, *-DPB1*, and *-DPA1*) with reduced lung cancer incidence among former smokers in UK Biobank. Heterozygous individuals are denoted by dotted lines; solid lines denote homozygous individuals. (**D**) Association of maximal *HLA*-II heterozygosity (8 distinct alleles at *HLA-DRB1*, *-DQB1*, *-DQA1*, and *-DPB1* because *DPA1* genotypes were unavailable in FinnGen) with reduced lung cancer incidence among former smokers in FinnGen. Heterozygous individuals are denoted by dotted lines; solid lines denote homozygous individuals. Plots with 95% CI are shown in fig. S9. All *P* values were calculated with multivariable Cox regression.

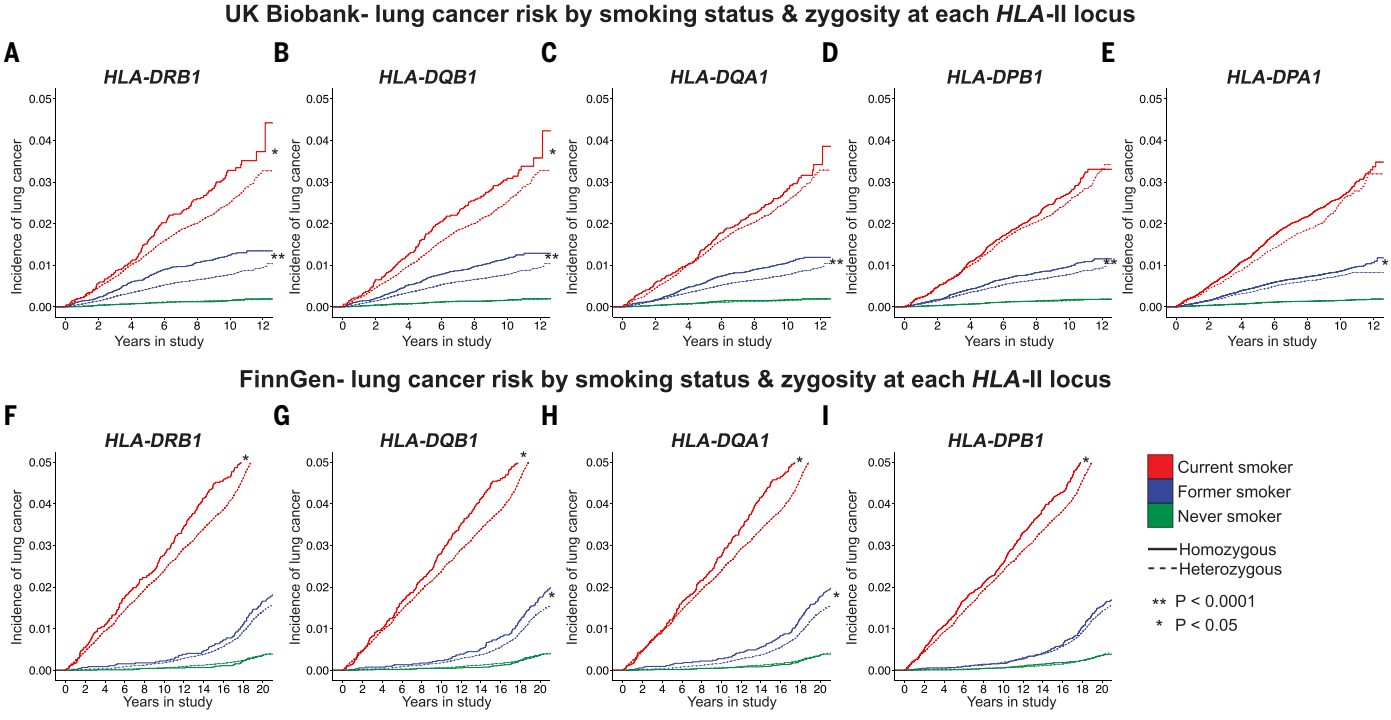


Fig. 3. Heterozygosity at individual *HLA*-II loci is associated with reduced lung cancer incidence among smokers in UK Biobank and FinnGen. (**A** to **E**) Association of heterozygosity at the indicated *HLA*-II locus with reduced lung cancer incidence among current and former smokers in the UK Biobank. Dotted lines denote heterozygous individuals; solid lines represent homozygous individuals. (**F** to **I**) Association of heterozygosity at the indicated *HLA*-II locus with reduced lung cancer incidence among smokers in FinnGen. Dotted lines denote heterozygous individuals; solid lines represent homozygous individuals. Plots with 95% CI are shown in fig. S9. All *P* values were calculated with multivariable Cox regression.

at *HLA-DQB1*. This analysis confirmed that the effect of *HLA-DRB1* heterozygosity on reduced lung cancer risk was stronger than that of *HLA-DQB1* heterozygosity in the UK Biobank (fig. S8, A to C). Genetic variation in *HLA-DRB1* has been strongly linked to changes in the peripheral T cell receptor (TCR) repertoire (*63*) and risk of autoimmune diseases (*64*, *65*). In both cohorts, heterozygosity at both *HLA-DRB1* and *HLA-DQB1* was associated with reduced risk as compared with homozygosity at both (fig. S8, A to F). Whereas the protective effects of *HLA-DRB1* and *HLA-DQB1* heterozygosity were observed in both current and former smokers, the protective effects of heterozygosity at *HLA-DQA1* (Fig. 3C), *HLA-DPB1* (Fig. 3D), and *HLA-DPA1* (Fig. 3E) were observed only in former smokers (fig. S9, E to I). In general, we replicated the protective effects of heterozygosity at each of the *HLA*-II loci among smokers in FinnGen (table S7), with *HLA-DQB1* and *HLA-DQA1* mediating the strongest effects in both current and former smokers (Fig. 3, F to I; fig. S8, D to F; and fig. S9, J to M). Although we did not detect any effect of *HLA*-II heterozygosity in never-smokers, the lack of such an association may be due to power, given the much lower number of never-smokers compared to current and former smokers in both cohorts. To examine whether the protective effect of heterozygosity was driven by the presence or absence of individual *HLA*-I or *HLA*-II alleles, we performed multivariable Cox regression analyses in the UK Biobank and FinnGen, controlling for all alleles associated with lung cancer risk, and found that all heterozygosity signals remained significant even after adjusting for the effects of individual alleles, either when testing all individual alleles (fig. S10, A to C, and tables S8 and S9) or when testing those enriched in individuals fully heterozygous or homozygous at *HLA*-II (fig. S10, D to G, and tables S10 and S11). We also observed similar results using Cox regression models unadjusted for any covariates (tables S12 and S13) and when adding up to 20 genetic ancestry principal components to the multivariable Cox regression model (tables S14 and S15). Collectively, these analyses underscore the robustness of the association between *HLA*-II heterozygosity and reduced risk of lung cancer.

We next used the UK Biobank data to estimate the lifetime risk of lung cancer by age 80 (*53*) among individuals heterozygous or homozygous at *HLA*-II by using age as the timescale (fig. S11), analogous to prior studies (*53*, *66*) (table S6). We observed notable differences in lifetime risk between smokers heterozygous and homozygous at *HLA*-II in the UK Biobank; for example, among current smokers, homozygosity at *HLA-DRB1* was associated with a 13.92% lifetime risk of lung cancer as compared with a 10.81% risk among current smokers heterozygous at *HLA-DRB1*, representing an excess risk of 3.11% (fig. S12). We observed similar trends in FinnGen (table S7), with 26.3% lifetime risk attributed to *HLA-DRB1* homozygotes as compared with 22.0% for *HLA-DRB1* heterozygotes (fig. S13).

To evaluate the potential effect of *HLA*-II heterozygosity on reduced lung cancer risk in comparison with genetic predisposition conferred by other loci of the genome, we applied a recently developed polygenic risk score (PRS) for lung cancer (*67*) to both the UK Biobank (fig. S14) and FinnGenn (fig. S15). We evaluated two forms of the PRS, starting in the UK Biobank: one without SNPs in the MHC region ("PRS no MHC"; fig. S14A) and one with SNPs in the MHC region ("PRS w/MHC"; fig. S14B). The HRs for *HLA-DRB1* homozygosity were comparable to the HRs for both versions of the PRS (homozygosity *DRB1* HR = 1.36 compared to PRS no MHC HR = 1.57, compared to PRS w/ MHC HR = 1.42). As expected, these results suggest that *HLA*-II homozygosity is associated with lung cancer risk, but not to the same extent as a genome-wide PRS, which includes many more loci. We next asked whether *HLA*-II heterozygosity remains independently associated with lung cancer risk even after adjusting for the effect of the PRS. This was the case regardless of which version of the PRS was used; none of the *HLA*-I loci showed significant heterozygosity effects when adjusting for continuous PRS (fig. S14, C and D). Even among individuals with high PRS, *HLA*-II homozygosity was able to further stratify lung cancer risk (fig. S14, E to N). In particular, even among individuals with high genome-wide PRS, *HLA*-II homozygosity conferred up to 8.2% additional lifetime risk in current smokers and 2.1% additional lifetime risk in former smokers in the UK Biobank (fig. S14, O and P). We repeated these analyses in FinnGen and observed similar results (fig. S15). In FinnGen, the combination of PRS high and *HLA*-II homozygosity conferred up to 9.0% additional lifetime risk in current smokers and up to 2.89% additional lifetime risk among former smokers (fig. S15, O and P). These analyses show that *HLA*-II heterozygosity is a critical and independent factor associated with reduced risk of lung cancer, even among smokers and individuals with high genome-wide genetic predisposition.

Although we did not observe significance for *HLA*-I heterozygosity in our logistic regression analyses (Fig. 1), we tested the effects of maximal *HLA*-I heterozygosity and heterozygosity at each *HLA*-I locus (fig. S16, A to D, and table S6) on lung cancer risk over time using multivariable Cox regression analyses as performed for *HLA*-II. We observed an unexpected significant effect of maximal *HLA*-I heterozygosity and heterozygosity at *HLA-C* among former smokers in the UK Biobank, but these results did not replicate in FinnGen (fig. S16, E to H, and table S7). These data suggest that further studies, perhaps at larger sample sizes, are required to clarify the effect of *HLA*-I heterozygosity on lung cancer risk.

We further performed subgroup multivariable Cox regression analyses in the UK Biobank to test the effect of *HLA*-II heterozygosity in individual lung cancer subtypes, when available. First, the effect of smoking alone was significant in all three histologies evaluated and was strongest in small cell and squamous carcinoma (fig. S17A), consistent with prior reports (*1*). Furthermore, our analyses revealed that the protective effect of *HLA*-II heterozygosity was observed in small cell carcinoma, squamous carcinoma, and adenocarcinoma, with most effects observed in former smokers, which is consistent with the combined analyses (fig. S17). Moreover, the HRs for *HLA*-II heterozygosity were lower in squamous (HR range for significant loci: 0.49 to 0.66) and small cell (HR range for significant loci: 0.47 to 0.56) than in adenocarcinoma (HR range for significant loci: 0.69 to 0.76) (table S16). Thus, our data may suggest that the protective effect of *HLA*-II heterozygosity could be attenuated by smoking; i.e., the protective effect is strongest in lung cancer subtypes, in which the effect of tobacco is also the strongest. We also observed significant associations between *HLA*-II heterozygosity and reduced risk of squamous carcinoma in FinnGen (fig. S18 and table S17).

We also evaluated the effect of *HLA*-II evolutionary divergence (HED)—a quantitative measure of the antigen-presentation capacity of an individual's *HLA*-II allomorphs that is captured by measuring the molecular distance between the peptide binding grooves of each allele (*41*, *47*)—on lung cancer risk in the UK Biobank. Among former smokers, we observed a protective effect of HED at *HLA-DRB1* and *HLA-DQB1* against lung cancer risk when treating HED as a continuous variable and adjusting for all covariates (fig. S19A). We repeated these analyses in FinnGen and replicated the HED association for *HLA-DRB1* among both current and former smokers (fig S19B). These data suggest that granular differences between amino acids within *HLA*-II peptide binding grooves may be associated with a reduced risk of lung cancer.

Fine-mapping implicates amino acid heterozygosity within the *HLA*-II peptide binding groove in reducing lung cancer risk

To explore the relationship between *HLA*-II heterozygosity, antigen presentation, and lung cancer risk, we sought to perform fine-mapping analyses using the amino acid sequences of the peptide binding groove of *HLA*-II alleles (Materials and methods). Fine-mapping of the peptide binding groove of the MHC has been conducted previously to directly implicate antigen presentation in HIV-1 control (*68*, *69*). To

examine the effect of amino acid polymorphisms in the peptide-binding groove on lung cancer risk, we first collected the amino acid sequences of the peptide binding grooves of all *HLA-DRB1* and *DQB1* alleles in the UK Biobank because heterozygosity at *DRB1* and *DQB1* mediated the strongest protective effects against lung cancer in the UK Biobank. We then defined the polymorphic positions within the peptide-binding groove through sequence entropy analysis in order to narrow the peptide binding groove to a core set of polymorphic positions to test for association with lung cancer risk. Our entropy analysis revealed that roughly 33% of the amino acid positions were polymorphic (fig. S20A). We analyzed the average C-α distances between bound peptides and *HLA-DRB1* and *HLA-DQB1* protein residues using peptide-MHC crystal structure data from the Protein Data Bank (PDB) (Materials and methods and table S18), which showed that polymorphic residues are significantly closer to bound peptides than are monomorphic residues ($P < 0.01$ for *HLA-DQB1*; $P < 0.0001$ for *HLA-DRB1*) (fig. S20, B and C), suggesting their relevance in peptide presentation.

We next tested the set of polymorphic positions defined through sequence entropy analysis for association with lung cancer risk using multivariable logistic regression in the UK Biobank. In standard MHC fine-mapping analysis, amino acid positional diversity within individual *HLA* alleles is tested for disease associations or quantitative traits (e.g., HIV-1 viral load). Because our interest is in heterozygosity, we adapted MHC fine-mapping to test heterozygosity at each position within the peptide-binding groove, defined as two different amino acids at a particular position. For each polymorphic position in the peptide binding groove, we fit a multivariable logistic regression model incorporating heterozygosity at that position along with smoking status. The outcome variable was binary, representing lung cancer case or control as defined in earlier analyses. This analysis revealed that five positions within the *DRB1* peptide-binding groove and seven positions within that of *DQB1* remained significant after multiple testing corrections with the Benjamini-Hochberg method (Fig. 4, A and B, and table S19). Several of the significant positions have been previously implicated in other diseases—e.g., P70 in *DRB1*, previously associated with rheumatoid arthritis (*70*)—in addition to smoking and Parkinson's disease (*71*). We also observed the effect of P57 in *DQB1*, previously associated with type 1 diabetes (*72*). A stepwise regression analysis in the UK Biobank incorporating all covariates yielded significance for P31 and P70 in *HLA-DRB1* and P14 in *HLA-DQB1* (Fig. 4, A and B). The significant positions were found to be a median of 7.04 Å (*HLA-DQB1*) and 9.04 Å (*HLA-DRB1*) (both in the 21st percentile) to bound peptides through quantification and visual inspection of peptide-MHC crystal PDB structures (Fig. 4, C and D). We repeated these analyses in FinnGen and found that we replicated four associations from the UK Biobank, including *HLA-DRB1* P70 (fig S21). Altogether, these data implicate antigen presentation together with heterozygosity at both the population level (through differences in allele identity across individuals) and at the molecular level (through differences in amino acid sequence at particular positions in *HLA* peptide-binding grooves) in reduced lung cancer risk. Although the association of *HLA*-II heterozygosity with reduced lung cancer risk in the longitudinal Cox regression analyses implies that variation within the peptide-binding groove should also be associated with reduced lung cancer risk, the principal contribution of our fine-mapping analyses is that heterozygosity of specific amino acid positions within the peptide-binding groove are themselves associated with reduced lung cancer risk.

Single-cell RNA sequencing of the adjacent normal lung reveals that smoking drives up-regulation of *HLA*-II and proinflammatory pathways in alveolar macrophages

Our data suggest that *HLA*-II heterozygosity and smoking interact through antigen presentation to modulate lung cancer risk. One possible explanation for this phenomenon is increased neoantigen presentation due to an elevated mutation rate induced by smoking. A complementary hypothesis is that smoking alters the lung microenvironment to create an inflammatory milieu that favors antigen presentation by the *HLA*-II allomorphs. To define the molecular effects of smoking on the lung microenvironment, we analyzed scRNA-seq data from three lung cancer studies profiling the adjacent normal lung (*73*–*75*). Despite these prior studies and others assessing the effect of smoking on the lung tumor microenvironment (*76*), the effect of smoking on the normal lung microenvironment is unclear. We hypothesized that smoking might modulate the expression of the *HLA*-II genes in relevant immune-cell subsets; such modulation of *HLA*-II gene expression may promote antigen presentation within an inflammatory milieu created in response to tissue damage by smoking.

We first analyzed scRNA-seq data from the matched adjacent normal lung of 27 individuals (N = 19 smokers, 8 never-smokers) who underwent surgical resection for lung cancer from Leader *et al.* (*73*) (Fig. 5, A to C). We used cell-type annotations as specified in the original study and noted a large compartment of myeloid cells and alveolar macrophages (Fig. 5A). We first asked whether smoking induces changes in the proportion of cell types in the healthy lung. We directly compared cell-type prevalence in smokers versus never-smokers, accounting for the compositional nature of the data using a Dirichlet multinomial regression adjusting for clinical covariates, as used in prior studies (*77*, *78*). This analysis revealed an enrichment of alveolar macrophages (C25) in smokers (Dirichlet multinomial $P = 0.03$) (Fig. 5D and table S20). Differential expression analysis performed among all individuals comparing C25 with all other macrophage clusters showed that the C25 alveolar macrophages cluster markedly up-regulated the *HLA*-II genes (Fig. 5E and table S21), in addition to other inflammatory markers such as *IFI6* and *ISG15*. Although high expression of the *HLA*-II genes was also observed in C55, C25 was the only macrophage cluster (and the only cluster overall) with significantly different prevalence between smokers and never-smokers. To examine granular differences in cell state, we performed differential expression analysis within C25 between smokers and never-smokers (table S22). This analysis revealed that *HLA-DRB1* was up-regulated on smoker C25 cells, in addition to other inflammatory genes related to the innate immune response (*CXCL8*, *ISG15*, *DEFB1*, and *IFITM3*) (Fig. 5F). Moreover, unbiased pathway analysis of the differentially expressed genes confirmed enrichment of proinflammatory pathways in smoker C25 cells compared with never-smoker C25 cells (Fig. 5G and tables S22 and S23). Additionally, we queried an independent dataset of scRNA-seq data from the normal lung from Travaglini *et al.* (*74*), for which cluster annotations and smoking status were available. Although limited in sample size, this analysis demonstrated a twofold enrichment of alveolar macrophages in a smoker compared with two never-smokers (Fig. 5H). Gene set enrichment analysis (GSEA) with the differentially expressed genes from Leader *et al.* C25 and the macrophage cluster from Travaglini *et al.* as input showed that genes defining the Travaglini *et al.* macrophage cluster were enriched in C25 (fig. S22A), suggesting that the clusters were similar across datasets. Alveolar macrophages can act as antigen-presenting cells (*79*); thus, our data indicate that smoking may increase antigen presentation and inflammatory responses by *HLA*-II–high alveolar macrophages.

Previous work in mice has suggested that MHC-II can be expressed on tumor cells and that such expression may be correlated with improved clinical outcomes (*80*). To explore this hypothesis in humans, we obtained a third scRNA-seq dataset (*75*) of 44 patients for whom both immune and epithelial cells were profiled from both tumor and normal lung. Using this dataset, we found that the *HLA*-II genes were expressed most highly on myeloid cells and B cells, which is consistent with their role as antigen-presenting cells. However, we also detected expression of the *HLA*-II genes in small amounts on epithelial cells (Fig. 5I), which is consistent with prior reports (*81*). The *HLA*-II genes were expressed on epithelial cells from

both normal (fig. S22B) and tumor lung (fig. S22C). Differential expression analyses comparing smokers with never-smokers confirmed that, as observed in myeloid cells, *HLA-DRB1* was up-regulated in smokers' normal epithelial cells [both alveolar type 1 (AT1) and AT2] (Fig. 5J and table S24). Our findings are supportive of previous studies demonstrating that lung epithelial cells can present antigen to $CD4^+$ T cells by means of *HLA*-II (*82*). Our results validate earlier observations of MHC-II expression in tumors (*83*) and suggest that alveolar macrophages and epithelial cells may cooperatively respond to tobacco smoking through up-regulation of the *HLA*-II genes and proinflammatory pathways in normal tissues.

To investigate the effects of *HLA*-II heterozygosity on cellular phenotypes in non–small cell lung cancer (NSCLC), we used CIBERSORTx (*84*) to deconvolve cell type–specific expression from bulk RNA-seq data in the TCGA lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) cohorts. We performed exploratory analyses assessing the effect of *HLA*-II heterozygosity *HLA*-II expression in specific cell types. This analysis revealed that *HLA*-II heterozygosity drove higher expression of the *HLA*-II genes in intratumoral dendritic cells in both LUAD and LUSC (fig. S23) and increased TCR clonality (fig. S24, A and B); in LUSC, we also observed a trend toward higher $CD4^+$ T cell infiltration in *HLA*-II heterozygous individuals (fig. S24, C and D). Altogether, these data suggest that in lung tumors, *HLA*-II heterozygosity is associated with increased expression of *HLA*-II primarily in dendritic cells. Although dendritic cells and other professional antigen-presenting cells are the dominant expressers of *HLA*-II, our data from normal tissues indicate that epithelial cells and alveolar macrophages may also contribute to risk of lung cancer through expression of *HLA*-II.

Our observations of both *HLA*-II expression on epithelial cells from tumor scRNA-seq samples and the effect of *HLA*-II heterozygosity on reduced lung cancer risk prompted us to ask whether lung tumors evade the immune

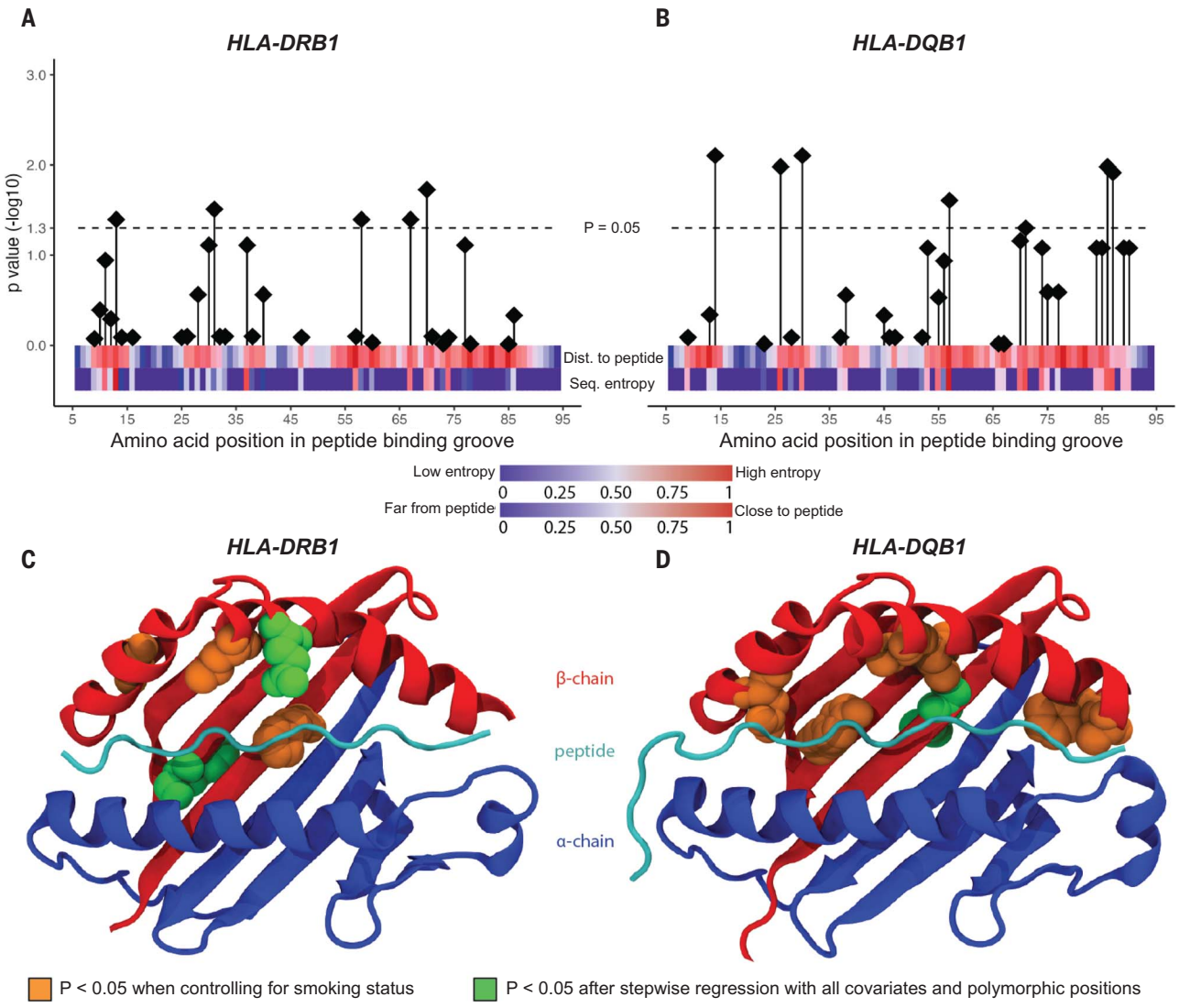


Fig. 4. Heterozygosity fine-mapping and structural analyses of *HLA*-II peptide binding groove amino acid sequences. (**A** and **B**) Associations between heterozygosity at the indicated position of the peptide binding groove of *HLA-DRB1* (A) and *HLA-DQB1* (B), respectively, and lung cancer risk, analyzed using a multivariable logistic regression in UK Biobank and adjusting for smoking status. The dotted line indicates false discovery rate (FDR) $P = 0.05$. Annotation bars indicate polymorphism at the indicated position defined by sequence entropy and distance from peptide, based on analysis of representative peptide-MHC crystal structures. (**C**) Structural visualization of significant amino acid positions from (A) and positions significant after stepwise regression on a representative *HLA-DRB1* crystal structure in complex with bound peptide. (**D**) Structural visualization of significant amino acid positions from (B) and positions significant after stepwise regression on a representative *HLA-DQB1* crystal structure in complex with bound peptide.

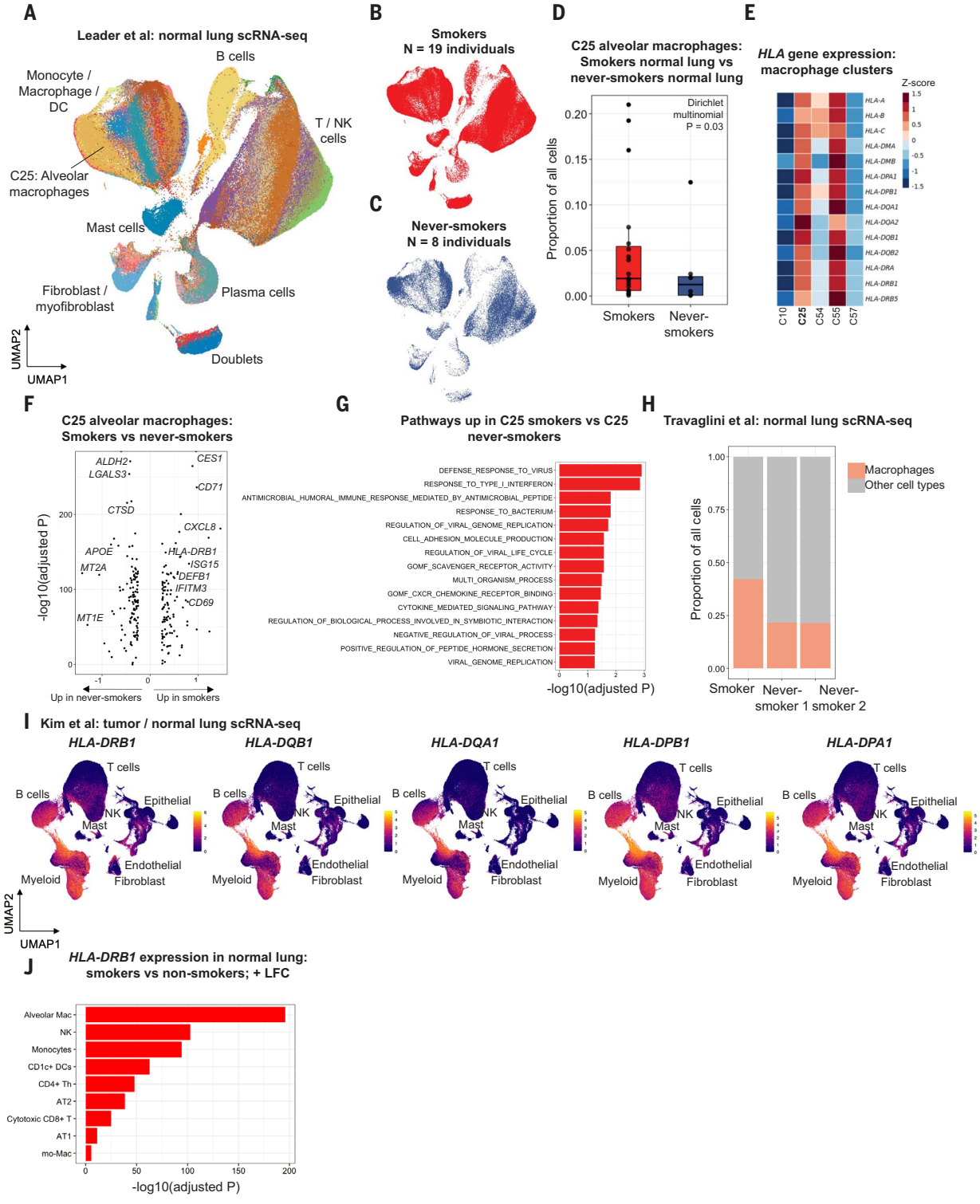


Fig. 5. Tobacco smoking-induced inflammatory programs identified by sc RNA-seq analysis of the normal lung from three independent cohorts. (**A**) Uniform manifold approximation and projection (UMAP) of normal lung scRNA-seq data from Leader *et al*. Broad compartments containing multiple clusters are labeled. (**B**) UMAP of cells from smokers only from Leader *et al*. (**C**) UMAP of cells from never-smokers only from Leader *et al*. (**D**). Increased prevalence of the C25 alveolar macrophage cluster in smokers as compared with never-smokers. Boxplots depict minimum, first quartile, median, third quartile, maximum, and outliers. (**E**) Up-regulation of *HLA*-II genes in C25 as compared with other macrophage clusters from Leader *et al*. (**F**) Differential expression analysis comparing smoker C25 cells with never-smoker C25 cells. (**G**) Pathway analysis performed using differentially expressed genes from (F) as input. (**H**) Enrichment of macrophages in a smoker as compared with two never-smokers in an independent scRNA-seq dataset from Travaglini *et al*. (**I**) Expression of *HLA*-II cells in antigen-presenting cells (B cells and macrophages) and epithelial cells from an independent scRNA-seq dataset from Kim *et al*. containing both tumor and normal lung data. (**J**) Up-regulation of *HLA-DRB1* expression across immune and epithelial cells in smokers as compared with never-smokers from Kim *et al*.

system through LOH of the *HLA*-II genes. Such LOH events would dampen the tumor's ability to present *HLA*-II–restricted neoantigens, thus evading recognition by $CD4^+$ T cells, which would provide further evidence for the importance of *HLA*-II expression on epithelial cells for tumor immune surveillance. Although prior work has estimated that roughly 40% of lung cancers exhibit allele-specific LOH at the *HLA*-I genes (*32*), the presence and extent of *HLA*-II LOH in cancer remain unknown. To investigate whether *HLA*-II LOH occurs in lung cancer, we adapted LOHHLA (*32*), originally developed to compute allele-specific *HLA*-I loss in cancer, to evaluate allele-specific loss of *HLA*-II genes by using exome sequencing data from LUAD (N = 486) and LUSC (N = 450) from TCGA (Materials and methods). We observed that *HLA*-II LOH was just as prevalent as *HLA*-I LOH in NSCLC; in particular, we observed rates of 24 and 38% *HLA*-II LOH in LUAD and LUSC, respectively. For *HLA*-I LOH, we observed rates of 24 and 37% in LUAD and LUSC (Fig. 6A). Unequivocally, these data suggest that *HLA*-II LOH, which was previously uncharacterized, is widespread in NSCLC. To validate the observed rates of *HLA*-II LOH discovered in TCGA, we obtained whole-genome sequencing from two independent cohorts of patients with NSCLC: the PCAWG cohort (*85*) (N = 83) and the Hartwig Medical Foundation cohort (*86*) (N = 657). We adapted the methodology from Martínez-Jiménez *et al*. that was used to call *HLA*-I LOH (*33*) to call *HLA*-II LOH in these two additional cohorts. In PCAWG, 19 to 34% of patients had *HLA*-II LOH; comparably, 26 to 28% of patients in the Hartwig cohort had *HLA*-II LOH (Fig. 6A). Although the rates of *HLA*-II LOH are comparable to those of *HLA*-I LOH, our analysis demonstrates across three independent cohorts and two independent algorithms that *HLA*-II LOH is as prevalent in NSCLC as *HLA*-I LOH is. Our data show that *HLA*-I LOH is often accompanied by *HLA*-II LOH, suggesting that loss of both loci may be important for tumor evolution.

We next sought to investigate the effects of germline *HLA*-II heterozygosity and *HLA*-II LOH on the tumor mutational landscape and immunopeptidome. In TCGA LUAD, *HLA*-II heterozygosity had no effect on tumor mutational burden (TMB) (fig. S25A) but was associated with a larger predicted neopeptide repertoire (fig. S25B), suggesting that *HLA*-II heterozygosity specifically affects MHC-bound mutations. Next, we asked whether *HLA*-II LOH affected both TMB and the neopeptide repertoire. Notably, tumors with *HLA*-II LOH had a higher TMB than did individuals without LOH (fig. S25C) and a larger neopeptide repertoire at baseline (LOH pre-loss compared with tumors with no LOH) (Fig. 6B). This result suggests that *HLA*-II LOH is selected for in lung cancer through preferential loss of *HLA*-II alleles with larger neopeptide repertoires. Moreover, we found that LOH of *HLA-DRB1* was associated with lower expression of *HLA*-II in NSCLC epithelial cells (fig. S24E), suggesting that *HLA*-II LOH may affect both the tumor immunopeptidome and microenvironment. We next repeated these analyses in the TCGA LUSC cohort; as in LUAD, germline *HLA*-II heterozygosity was not associated with TMB (fig. S26A) but was associated with a larger neopeptide repertoire (fig. S26B). Whereas in LUSC, *HLA*-II LOH was not associated with TMB (fig. S26C), we observed that tumors with LOH at *HLA-DPB1* and *HLA-DPA1* had higher neopeptide repertoires at baseline (pre-LOH) compared with those with no *HLA*-II LOH, again suggesting selection for *HLA*-II LOH in lung cancer (Fig. 6C). To investigate the properties of peptides lost through *HLA*-II LOH, we calculated peptide hydrophobicity, previously shown to be a critical determinant of neoantigen immunogenicity (*87*–*90*). In both LUAD and LUSC, peptides lost through *HLA-DRB1* LOH tended to be more hydrophobic than those that were not lost (fig. S25D and fig. S26D). Collectively, these analyses demonstrate that *HLA*-II LOH is as prevalent as *HLA*-I LOH in lung cancer and affects the dynamics of the tumor immunopeptidome.

Discussion

Here we show that *HLA*-II heterozygosity is associated with reduced risk of lung cancer, which may account for the variability in lung cancer risk among current and former smokers. Through analysis of genetic epidemiological data from two large-scale population cohorts and of multimodal genomic data, our study suggests an immunogenetic basis for lung cancer risk. Our data underscore the role of immunosurveillance in protecting against lung cancer. We propose that the immune system, consisting of immunogenetic and cellular diversity, comprises the foundation of tumor rejection and initiation (*12*–*16*), together with replicative and hereditary defects and environmental exposures as proposed by Tomasetti and Vogelstein (*91*, *92*).

Our study represents a multimodal interrogation of the influence of *HLA* heterozygosity on lung cancer risk. The combination of orthogonal approaches, including epidemiological, genetic, and transcriptomic analyses, suggests several complementary mechanisms that may explain the association of *HLA*-II heterozygosity with reduced risk of lung cancer. Heterozygosity at *HLA*-II may lead to increased diversity of smoking-related antigens in developing tumors, which could be presented by alveolar macrophages—which express inflammatory markers in response to tissue damage by smoking—or by dendritic cells, for recognition by $CD4^+$ T cells. It is also possible that antigens could be presented to T cells by pre-cancerous epithelial cells, such as AT1 or AT2 cells. The importance of *HLA*-II expression on epithelial and tumor cells is underscored by our finding of widespread *HLA*-II LOH in lung cancer. We show that *HLA*-II LOH favors the loss of alleles with larger neopeptide repertoires, underscoring the importance of the *HLA*-II loci in lung cancer. Further investigation is required to clarify the exact mechanisms by which *HLA*-II heterozygosity reduces lung cancer risk, including clarification of whether $CD4^+$ T cells themselves clear early neoplastic cells or facilitate $CD8^+$ T cell–mediated clearance. Altogether, our data are in agreement with an increasing body of evidence suggesting that $CD4^+$ T cells and MHC-II are critical in the immune response to cancer (*83*, *93*–*95*).

Although our study revealed an *HLA*-II heterozygote advantage in reducing lung cancer risk, examples of *HLA*-II heterozygote advantage have been shown previously for other diseases, including ulcerative colitis (*96*) and hepatitis B infection (*39*). However, given the many prior examples of *HLA*-I heterozygote advantage, e.g., in individuals with HIV (*38*) and metastatic cancer (*40*–*46*), it is notable that we did not observe a robust association between *HLA*-I heterozygosity and lung cancer risk in our study. Statistical power may influence the observed associations; we conducted a power analysis down-sampling the number of lung cancer cases in UK Biobank and found that the number of cases required to observe significance for heterozygosity varied even across the individual *HLA*-II loci (fig. S27); accordingly, perhaps larger cohorts are needed to observe a signal at *HLA*-I. In addition, the effects of *HLA*-II versus *HLA*-I heterozygosity on cancer risk may depend on the cancer type. To explore this question, we investigated the effects of *HLA* heterozygosity on risk of 16 other cancer types in the UK Biobank and FinnGen (fig. S28). This analysis revealed that *HLA*-II heterozygosity was associated with reduced risk of multiple additional solid tumor types in either the UK Biobank or FinnGen. The strongest effects of both *HLA*-I and *HLA*-II heterozygosity were observed in lymphoma in both cohorts, motivating further investigation of immunogenetic mechanisms of blood cancer risk (*97*, *98*). However, further work is needed to clarify the differences between immunosurveillance mediated by *HLA*-I and *HLA*-II in early tumor development, including the development of refined models in other cancer types incorporating disease-specific covariates.

GWAS have strongly implicated *HLA*-II alleles in risk of autoimmune diseases (*99*–*101*); our fine-mapping analyses identified positions within the peptide-binding groove of *HLA*-II alleles that were previously identified by fine-mapping of the MHC in autoimmune disease. The varying roles of *HLA*-II heterozygosity in

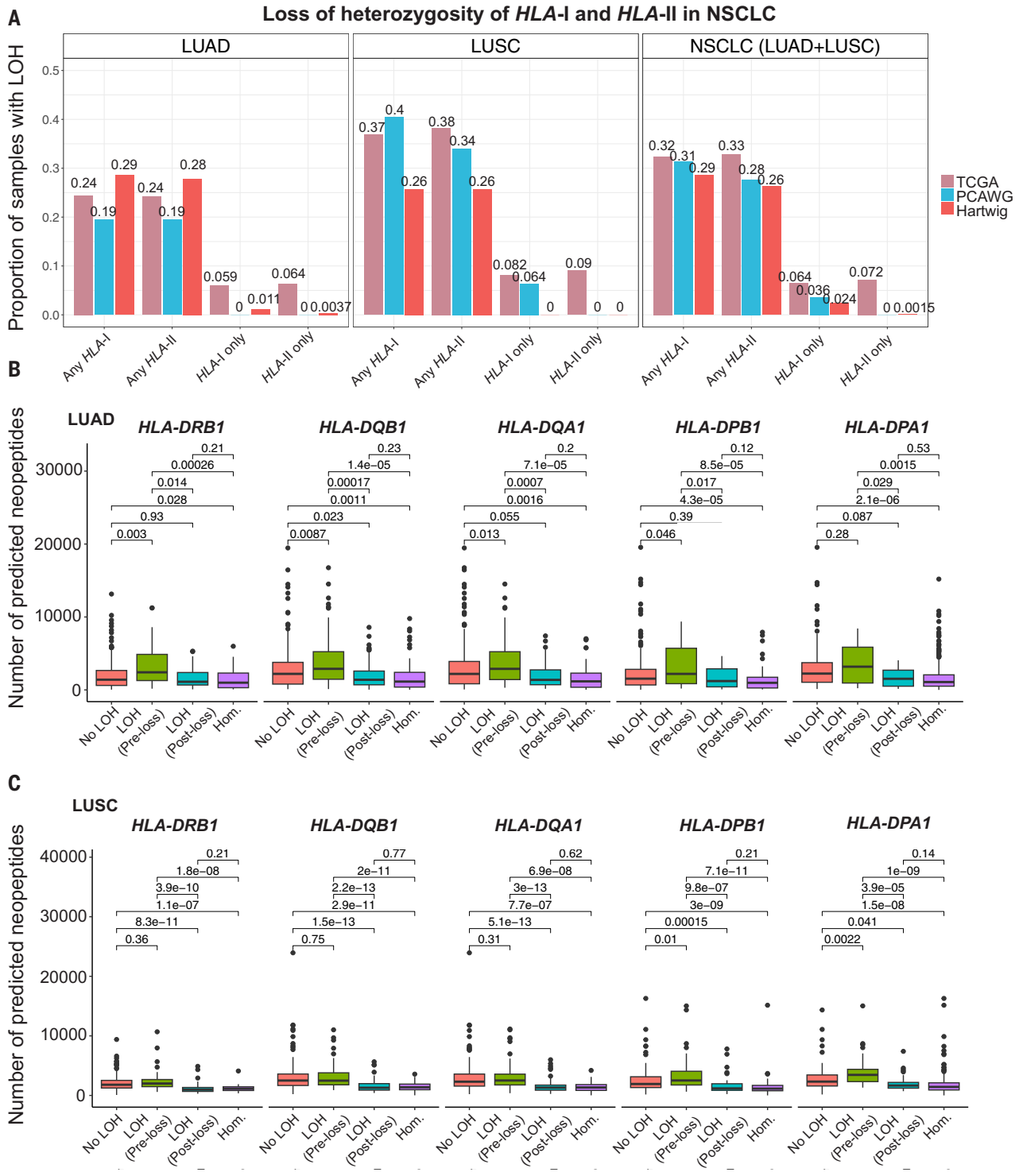


Fig. 6. *HLA*-I and *HLA*-II loss of heterozygosity and immunopeptidome dynamics in lung cancer. (**A**) Rates of loss of heterozygosity (LOH) at *HLA*-I and *HLA*-II across multiple independent large lung cancer cohorts. *HLA* LOH at all eight *HLA* loci in TCGA was calculated with LOHHLA. The proportion of individuals with loss at any class *HLA*-I (any one or more of *HLA-A/B/C*) or any class *HLA*-II locus (any one or more of *HLA-DRB1/DQB1/DQA1/DPB1/DPA1*) was determined for LUAD (*HLA*-I, *N* = 458; *HLA*-II, *N* = 465), LUSC (*HLA*-I, *N* = 416; *HLA*-II, *N* = 381), and for the full NSCLC cohort (*HLA*-I, *N* = 874; *HLA*-II, *N* = 846), and displayed as the mean across six LOHHLA coverage filters (5 to 30 in increments of 5). For individuals evaluated at ≥1 *HLA*-I locus and ≥1 *HLA*-II locus, LOH at only *HLA*-I was defined as LOH at one or more *HLA*-I loci but no *HLA*-II loci, and vice versa for *HLA*-II only LOH (LUAD *N* = 437, LUSC *N* = 347). For PCAWG and Hartwig, *HLA*-I and *HLA*-II LOH were determined by using the Hartwig Medical Foundation analytical pipeline (*33*). Loss at any *HLA*-I locus and any *HLA*-II locus was calculated similarly for the full NSCLC cohort (TCGA= 784; Hartwig, *N* = 657; PCAWG, *N* = 83). A subset of samples in Hartwig and PCAWG were specifically annotated by histology (LUAD or LUSC); for these samples, rates within each histology were also calculated (Hartwig LUAD, *N* = 273; Hartwig LUSC, *N* = 35; PCAWG LUAD, *N* = 36; PCAWG LUSC, *N* = 47). All other samples in Hartwig and PCAWG are labeled in the original

metadata as NSCLC and are presented in the rightmost panel NSCLC (LUAD+LUSC), which includes samples with and without histology annotation. (**B**) Dynamics of the predicted neopeptide repertoire in TCGA LUAD and (**C**) in TCGA LUSC, in tumors with and without *HLA*-II LOH. The neopeptide repertories of heterozygous patients unaffected by LOH are indicated by the red boxes. The neopeptide repertoires of patients with LOH at the specified locus before accounting for peptide loss and after accounting peptide loss due to the LOH event are indicated by the green and blue boxes, respectively. Homozygous patients without LOH are indicated by the purple boxes. Boxplots in (B) and (C) depict minimum, first quartile, median, third quartile, maximum, and outliers. Numbers above boxplots in (B) and (C) indicate *P* values computed with two-sided Wilcoxon test.

cancer, infectious disease, and autoimmunity should certainly be investigated further. *HLA*-II heterozygosity may also be associated with risk of viral or bacterial infections; these risks, in turn, could be exacerbated by one's smoking status. We recommend that such studies should combine longitudinal and lifetime risk data available in large biobanks with mechanistic analyses investigating the various genomic effects of *HLA* heterozygosity, which together represent key advances of our study in contrast to traditional GWAS or other studies investigating the effect of *HLA* diversity on cancer risk (*98*).

An important consideration with respect to replication of our heterozygosity results is the major compositional differences between the UK Biobank and FinnGen. Although both datasets represent population-scale cohorts with longitudinal follow-up data for cancer risk analyses, an essential difference between the two cohorts is the previously described healthy volunteer bias in the UK Biobank (*102*). This healthy volunteer bias results in lower rates of cancer incidence in the UK Biobank as compared with that found in the general population, and lower rates of smoking as well, which may explain in part the relatively low number of small cell lung cancer cases included in the cohort. For these reasons, we sought to validate our results in FinnGen, which recruited healthy volunteers in addition to individuals diagnosed with particular diseases (*51*, *52*), and is thus more representative of the general population in terms of disease incidence. The preponderance of individuals with disease in FinnGen may in part explain the higher number of lung cancer cases in FinnGen as compared with that found in the UK Biobank despite the smaller total sample size of FinnGen (183,163 individuals compared with 391,182 individuals). Overall, we believe that the broad replication of the protective effect of *HLA* heterozygosity discovered in the UK Biobank represents a clinically relevant effect, given how different the two cohorts are with respect to demographics and healthy volunteer bias.

Our study nominates population-level immunogenetic variation as a factor underlying the risk of lung cancer. A greater understanding of immunogenetic determinants of cancer risk, including genetic variation in *HLA* and other immune genes and pathways commonly associated with autoimmune and infectious diseases, may foster the development of improved strategies for cancer prevention. Our study suggests that current or former smokers homozygous at *HLA*-II could be considered at an earlier age for low-dose computed tomographic (LDCT) screening, which may reduce lung cancer mortality (*103*). Whether the combination of genotype-driven risk assessment and LDCT reduces lung cancer mortality as compared with either method alone should be comprehensively investigated in a future prospective clinical trial.

Materials and methods

Assembly of lung cancer cases and controls from the UK Biobank

We used data from the UK Biobank, last accessed in May 2022. Full details on participant recruitment and demographic characteristics are included in Sudlow *et al.* (*104*) and Bycroft *et al.* (*48*). We defined lung cancer cases by the ICD-10 codes C34.0, C34.1, C34.2, C34.3, C34.8, and C34.9. The histological subtypes of lung cancer analyzed were small cell carcinoma, squamous cell carcinoma, and adenocarcinoma. Any individuals with cancer diagnosed prior to the start of the UK Biobank recruitment period were excluded from analysis; to mitigate the possibility of including such individuals even after this filter, we started the follow-up time 1 year after the date of assessment, which did not affect the association results.

We defined an event as diagnosis or death due to lung cancer. For individuals with an event, their follow-up time was computed as the date of assessment to the earliest of either the date of lung cancer diagnosis or date of death due to lung cancer. If participants died due to lung cancer and were not diagnosed previously, their follow-up time was computed as the days between the assessment date and the death due to lung cancer. All other individuals were designated as event 0 (healthy controls in this study). For individuals with event 0, follow-up time was calculated as the date of assessment to the earliest date of death, date lost to follow-up, or date of the end of cancer follow-up as reported by the UK Biobank for assessment centers in England, Wales, and Scotland. For lung cancer subtype-specific analyses, we assigned event 1 if diagnosed or died due to the subtype of interest. We assigned event 0 for all other individuals, and the follow-up time was calculated as days from the assessment date to the earliest diagnosis or death due to lung cancer if not diagnosed previously. We further collected age at first assessment, smoking status (current/former/never), BMI at first assessment, Townsend deprivation index, assessment center (England/Scotland/Wales), sex, and five principal genetic ancestry components for multivariable Cox regression analyses. We used age at death or end of follow-up for multivariable analyses using age as the time scale.

Assembly of FinnGen lung cancer cases and controls

The FinnGen study (https://www.finngen.fi/en) (*49*, *51*–*53*) is a public-private partnership including Finnish universities, biobanks, and hospital districts together with several pharmaceutical companies founded in the year 2017. The aim is to collect national Health Records (EHRs) and genetic data from 500,000 Finns by the end of 2023. The study participants include patients with acute and chronic diseases, healthy volunteers, and population collections. Version R8 consists of ~340,000 individuals (~190,000 females and 150,000 males). For our analysis, we used the genetic data and national EHR data from FinnGen participants. Specifically, hospital data from inpatient, outpatient, cancer, and cause of death registries were used. Overlapping information from different registries for individuals was removed prior to analyses.

We defined an individual as a case for lung cancer (event 1) if that person was diagnosed with lung cancer or the primary cause of death was lung cancer. We defined lung cancer using the ICD version 9 and 10 from the hospital data. ICD-9 code 162 and ICD-10 codes C34.0, C34.1, C34.2, C34.3, C34.8, and C34.9 were used in the definition of lung cancer. For lung cancer subtype-specific analyses, individuals were defined as a case if they were diagnosed with histology of interest or died due to histology of interest. Morphology codes used for defining specific subtypes from the cancer registry: non–small cell lung cancer, adenocarcinoma code 8140; non–small cell lung cancer, squamous 8070; and small cell lung cancer code 8041. All other individuals in the FinnGen cohort were defined as controls (event 0). Some of the registry data in FinnGen dates back to the 1950s, and different registries included in FinnGen have different start dates. We started the follow-up time calculations in FinnGen from 1 January 2000 onward for all individuals included in the analyses since this is the earliest date for which complete cancer follow-up information was available for FinnGen participants. If an individual was diagnosed with lung cancer, follow-up time ended at the earliest date of diagnosis or death due to lung

cancer or lung cancer subtype in subtype-specific analyses. For controls, the follow-up time ended at earliest of death, loss to follow-up, or the end of the follow-up period (03/24/2021). Individuals without BMI or smoking status information were removed from the analyses to obtain complete data for all individuals included in the analyses. In addition, individuals who had any cancer prior to 01/01/2000 or who were born after 01/01/2000 were removed from the analyses.

HLA imputation and genotyping from whole-exome data

Imputed *HLA* genotypes (data field 22182) were downloaded from the UK Biobank data showcase. These four-digit *HLA* alleles were estimated using HLA*IMP:02 from associated SNPs and is described in more detail in the flagship UK Biobank publication (*48*). The dataset is composed of rows of absolute posterior probability values (PPV) for every possible *HLA*-I allele (i.e. A0101, A0201) and *HLA*-II allele (i.e., DRB10301, DRB10501) per each individual, coded as a value between 0 and 2. We identified the *HLA* haplotypes for each individual from this table and filtered the dataset for individuals with high-quality *HLA* imputation at the eight classical *HLA*-I and *HLA*-II loci (*HLA-A*, *B*, *C*, *DRB1*, *DQB1*, *DQA1*, *DPB1*, *DPA1*) using the following criteria: for individuals with two distinct alleles at a given locus (i.e. two alleles for that locus had a nonzero value and all other alleles were 0), we applied a threshold of 0.7 as recommended by the UK Biobank, where individuals were considered positive for a specific allele if the PPV was ≥0.7 and ≤1. Individuals with a PPV between 1.5 and 2 for a specific allele at a given locus were marked as having two copies of the allele. Individuals with two nonzero alleles where one or both had a PPV <0.7, or one nonzero allele between 1 and 1.5, at any of the eight *HLA* loci, were excluded. Next, we calculated heterozygosity per *HLA* locus. An individual was considered heterozygous at a specific locus if they had two distinct alleles at the four-digit resolution and homozygous if they had two copies of an allele at that locus. Individuals were considered fully heterozygous at *HLA*-I if they had six distinct alleles (two for each of *HLA-A*, -B, and -C) and homozygous if they had less than six distinct alleles (homozygous at one, two, or three of the *HLA*-I loci). Individuals were considered fully heterozygous at *HLA*-II in the UK Biobank if they had ten distinct alleles (two for each of *HLA-DRB1*, *-DQB1*, *-DQA1*, *-DPB1*, and *-DPA1*) and homozygous if they had less than 10 distinct alleles (homozygous at one or more of the *HLA*-II loci). We note that genotypes for *HLA-DPA1* were not available in FinnGen and fully heterozygous at *HLA*-II was defined as eight distinct alleles.

To support the fidelity of the genotypes used in our study, we further downloaded whole-exome sequencing data from UK Biobank for a subset of 835 lung cancer cases and 42,107 controls. We typed the classical *HLA*-I and *HLA*-II genes using two independent methods, HLA*LA (*59*), and HLA-HD (*60*). We used these genotypes from exome sequencing data for comparisons against the imputed *HLA* genotypes provided by the UK Biobank. We also repeated our multivariable logistic regression analysis assessing the effect of heterozygosity at the indicated locus together with all covariates in the subset of UK Biobank individuals with whole-exome sequencing data. We report the logistic regression estimates and corresponding *P* values for the logistic regression performed with HLA*LA genotypes and with HLA-HD genotypes on the lung cancer cases and controls described above.

Population HLA allele frequency data

We obtained *HLA* allele frequency data for the UK and Finland from the AFND (http://www.allelefrequencies.net/). For the comparison of allele frequencies between the UK Biobank and the UK, we took the mean allele frequency across England, Scotland, and Wales. For comparisons of allele frequencies between the UK Biobank and FinnGen, we excluded *HLA-DPA1*, as genotypes for *HLA-DPA1* were not available in FinnGen. For comparisons of either UK Biobank or FinnGen to their respective population allele frequencies, we limited analyses to alleles present in the AFND, under the assumption that some alleles may be missing from AFND (i.e., as opposed to being at 0% allele frequency in the population).

Statistical analyses involving HLA heterozygosity, divergence, individual alleles, and clinical data

To assess the effect of *HLA* heterozygosity on lung cancer risk, we performed multivariable logistic regression analyses separately for each locus incorporating age at assessment, sex, smoking status (current/former/never), body mass index (BMI), Townsend deprivation index [a measure of socioeconomic deprivation previously associated with outcomes in the UK Biobank (*62*)], assessment center (England/Scotland/Wales), and five genetic ancestry principal components to account for population structure. For analyses that involved follow-up time (using days in the study or age at death or end of follow-up as the timescale), we used a multivariable Cox regression analysis within each smoking group (current/former/never), incorporating heterozygosity at the indicated locus (or maximal heterozygosity at *HLA*-I or *HLA*-II) together with all covariates as described above. HED was calculated according to Chowell *et al.* and Pierini *et al.* (*41*, *47*) and tested as a continuous variable with all covariates using multivariable Cox regression. To investigate the association of individual *HLA* alleles with lung cancer risk, we first assembled all *HLA*-I and *HLA*-II alleles present in the UK Biobank at MAF 1% or higher. We then tested each allele independently, coding it as 0/1/2, in a multivariable Cox regression controlling for all covariates. We then selected all alleles with nominal $P < 0.05$ and included them together in a multivariable Cox regression together with heterozygosity (e.g., at *HLA-DRB1*) to assess whether the effect of heterozygosity is independent of the effects of individual *HLA*-I and *HLA*-II alleles. Significance was denoted as multivariable (logistic or Cox) regression $P < 0.05$. We estimated lifetime risk by age 80 in the UK Biobank and age 90 in FinnGen, given the higher age of lung cancer cases in FinnGen compared to the UK Biobank (Fig. 1, F and G). Lifetime risk was estimated by as described in Mars *et al.* and Palomaki *et al.* (*53*, *66*). We estimated lifetime risk up to age 80 in the UK Biobank and 90 in FinnGen, as lung cancer cases were significantly older in FinnGen compared to the UK Biobank.

Polygenic risk score analyses

We applied a weighted lung cancer PRS from Hung *et al.* (*67*), developed in Europeans, to the UK Biobank using PRSice-2 (*105*), and FinnGen using the CS-PRS pipeline (https://github.com/FINNGEN/CS-PRS-pipeline). Importantly, this PRS was not developed in either the UK Biobank or FinnGen, rendering it appropriate for application to these two cohorts. We constructed two versions of the PRS—one without SNPs in the MHC region, and one with SNPs in the MHC region. For Cox regression models including continuous PRS, we scaled the PRS to mean 0 and standard deviation 1 prior to regression modeling. High PRS was defined as greater than the top quartile. We also ran Cox regression analyses combining PRS with *HLA* homozygosity by first splitting individuals into high (greater than the top quartile) and low (all others) PRS, and then splitting the high PRS group into heterozygous and homozygous *HLA* groups.

Power analysis

We conducted a power analysis in the UK Biobank to estimate the number of cases required to observe significance of *HLA*-II heterozygosity while adjusting for all covariates. For each locus, we down-sampled the number of cases while holding the number of controls constant, sweeping the sample size from 100 to 2500 in increments of 100. We randomly selected subsets of the indicated sample size 10 times and fit a multivariable Cox regression with each subset, testing heterozygosity at the indicated locus together with all covariates as described above.

Heterozygosity and structural fine-mapping

HLA-II structures for 96 *HLA-DRB* and 35 *HLA-DQB* alleles were identified using the Immune

Epitope Database (*106*) and extracted from RCSB Protein Data Bank (*107*). Structures were then inspected using Visual Molecular Dynamics (VMD) (*108*) and processed to contain only a single biological assembly. All pairwise Euclidean distances between the C-α atoms of beta chain residues and the C-α atoms of the bound peptide were calculated, and the minimum measured distance was selected for each residue-peptide pair. This process was repeated for each of the 131 structures, and the average distance with respect to each residue and *HLA* gene was determined. Amino acid sequences corresponding to the peptide binding groove for 74 *DRB1* and *DQB1* alleles appearing in the UK biobank dataset were extracted from the IGMT database (*109*). Positional sequence entropy was calculated by grouping the sequences by *HLA* gene (i.e. *HLA-DRB1* or *HLA-DQB1*), and Shannon entropy was calculated with respect to position (*110*).

Positional heterozygosity was first described at the patient level. For each patient, individual amino acid positions of the peptide binding groove were queried for heterozygosity across both haplotype alleles with respect to gene (*HLA-DRB1* or *HLA-DQB1*). For example, if residue 5 of *HLA-DRB1* binding pocket was being considered for a given patient, then the amino acid residues corresponding to residue 5 would be extracted from the *HLA-DRB1* allele sequences defined by the patient genotype and compared. In cases where the compared positions were the same amino acid across both alleles, that position would be assigned a 0, while positions showing a mismatch in amino acid residues would be assigned a 1. The process was repeated for all patients and with respect to binding pocket residues which showed a nonzero entropy across the entire allele set.

Sixty individual logistic regression models specific to each polymorphic binding pocket position (30 for *HLA-DRB1* and 30 for *HLA-DQB1*) were fit where the selected residue and smoking status were included as independent variables and detected lung cancer events within the time frame of the study as the dependent variable. The process was repeated with respect to each polymorphic position and locus. Positions with implications in lung cancer risk were identified as those that produced logistic regression models with significant *P* values for parameter estimates. Prior to selection, parameter *P* values were corrected for multiple testing using the Benjamini-Hochberg (FDR) method with respect to the number of polymorphic positions within each locus.

A stepwise logistic regression investigated the concerted impact of positions with significant association with lung cancer risk within each locus. The stepwise analysis was performed for each *HLA*-II allele where all positions that showed individual association with lung cancer within that locus were combined with all additional covariates described in the gene-level heterozygous logistic regression analysis. Stepwise regressions were performed using the stepAIC function from the R "MASS" package allowing for both forward and backward searches. Polymorphic positions in the final model following stepwise regression with a significant parameter estimate were chosen. Select positions were visualized using VMD.

Single-cell RNA-sequencing analysis

We collected three publicly available scRNA-seq datasets (*73*–*75*). The datasets were selected based on the availability of normal lung data, smoking status, and profiling of either immune cells, epithelial cells, or both. We refer to the lungs profiled here as "normal" rather than "healthy" because they were collected in studies investigating lung cancer or lung pathologies (Travaglini *et al.*). We used cluster definitions exactly as specified in the original studies. To assess changes in cell type prevalence between smokers and never-smokers, we used the Dirichlet multinomial regression, which accounts for the compositional nature of the data. We controlled for age, sex, stage, and diagnosis. Significance was assessed at $P < 0.05$. We used MAST (*111*) to perform differential expression analyses comparing C25 to all other macrophage clusters or between smokers and never-smokers within C15, adjusting for the same covariates used in the Dirichlet multinomial regression. We used log-normalized counts for heatmaps and marker plots depicting *HLA*-II gene expression. All data were analyzed in Seurat (*112*) in R using clinical and cell-type data exactly as specified in the original studies.

Effect of HLA-II heterozygosity on the tumor-immune microenvironment

To perform exploratory analyses assessing the effect of *HLA*-II heterozygosity on the tumor-immune microenvironment, we first ran CIBERSORTx (*84*) on TCGA LUAD and LUSC cases to infer cell type–specific expression of the *HLA*-II genes in intratumoral epithelial and dendritic cells. To assess the effect of *HLA*-II heterozygosity or *HLA*-II LOH (coded as a binary variable- 1 or 0) on *HLA*-II expression in these cell types, we used a linear model with covariates analogous to those used to map expression quantitative trait loci. Specifically, we used age, sex, 20 genetic ancestry principal components, and 20 gene expression principal components as covariates, and tested these against the inverse rank-normalized CIBERSORTx values. The purpose of including genetic ancestry PCs is to control for population structure; the purpose of using gene expression PCs is to control for technical and latent sources of variation in the RNA-seq data, as performed in previous QTL studies. Genetics ancestry PCs were obtained from Carrot-Zhang *et al.* (*113*). The number of patients with all covariates available was 424 in LUAD and 438 in LUSC. We adopted a Bonferroni significance threshold of $P < 0.1$ for significance given the exploratory nature of the analysis. Given the limited number of tests performed (5 *HLA*-II heterozygosity variables; 3 to 5 *HLA*-II genes inferred in each cell type), we also examined nominally significant associations at $P < 0.05$. We considered genes such as *DQB2* given the previously described difficulty in assigning reads to the *HLA* genes in RNA-seq data. We obtained TCR clonality and $CD4^+$ T cell infiltration estimates from Thorsson *et al.* (*114*). As these estimates are also obtained from RNA-seq, we used the same model described above to assess their dependence on *HLA*-II heterozygosity. For associations with TCR clonality and $CD4^+$ T cell levels, we considered $P < 0.05$ as significant, given the limited number of tests performed (i.e., only a single variable, TCR clonality or $CD4^+$ T cell levels, were tested).

HLA-II loss of heterozygosity for TCGA

We adapted the LOHHLA algorithm (*32*) to determine LOH at all *HLA* class I and class II genes in TCGA) (https://www.cancer.gov/tcga). Whole-exome sequencing BAM files, aligned to hg38 and from tumor- and blood-derived normal samples, were downloaded from the GDC portal (*115*) for all samples with LUAD or LUSC. Tumor ploidy and purity values determined using ABSOLUTE in Hoadley *et al.* (*116*) were also downloaded from GDC. We used *HLA*-I haplotypes called by OptiType (*117*) and *HLA*-II haplotypes called using HLA-HD from Marty-Pyke *et al.* (*35*). Reference *HLA* fasta files were downloaded from IMGT version 3.50, separately for each of the eight *HLA* loci (--HLAfastaLoc). hg38 coordinates used are as follows: *HLA-A*: chr6:29941260-29945884, *HLA-B*: chr6:31353872-31357187, *HLA-C*: chr6:31268749-31272092, *HLA-DRB1*: chr6:32578775-32589848, *HLA-DQB1*: chr6:32659467-32666657, *HLA-DQA1*: chr6:32637406-32655272, *HLA-DPB1*: chr6:33075990-33089696, and *HLA-DPA1*:chr6:33064569-33080748; and the script was run independently for each of the eight *HLA* loci, extracting the relevant *HLA* locus.

Before running LOHHLA, all 963 samples with complete genomic information available (495 LUAD and 468 LUSC)—including both tumor and normal exome sequencing, full *HLA* haplotype identified, and tumor ploidy information—were evaluated for coverage at each of the eight *HLA* loci and samples with normal or tumor coverage below the bottom fifth percentile of the coverage distribution for a given locus were excluded. A locus was then considered "available" for LOHHLA if it was heterozygous at that locus and it passed our coverage filter. This resulted in 936 distinct samples (486 LUAD, 450 LUSC) heterozygous

and with sufficient coverage at ≥1 *HLA* locus. For a given locus of each sample, we ran LOHHLA at six independent coverage filters (--minCoverageFilter 5 through 30 inclusive, at intervals of 5). We used bedtools version 2.29, samtools version 1.9, picard version 2.2.4, jellyfish version 2.3.0, GATK version 4.3.0.0, and NovoAlign version 3.09.02. A sample was annotated as having LOH at a given *HLA* locus if minor allele copy number was < 0.5, and allelic imbalance *P* value < 0.01 (*32*). We excluded samples for which LOH was unable to be determined at any of the sample's available *HLA* loci, or for which there were less than five mismatch sites with sufficient coverage, consistent with the warning message output by LOHHLA for alleles with only five supportive mismatch sites. After filtering and running LOHHLA, the number of individuals with LOH calls at each individual locus was as follows: *HLA-A*: 383 LUAD, 359 LUSC; *HLA-B*: 395 LUAD, 371 LUSC; *HLA-C*: 365 LUAD, 351 LUSC; *HLA-DRB1*: 285 LUAD, 226 LUSC; *HLA-DQB1*: 369 LUAD, 253 LUSC; *HLA-DQA1*: 387 LUAD, 320 LUSC; *HLA-DPB1*: 289 LUAD, 243 LUSC; *HLA-DPA1*: 128 LUAD, 136 LUSC. Our analysis cohort consisted of all individuals who had at least one locus with an LOH call. A sample was considered having LOH at "any class I" locus if *HLA* LOH was detected at any one or more of the three *HLA*-I genes, and at "any class II" locus if *HLA* LOH was detected at any one or more of the five *HLA*-II genes. In total, we were able to determine *HLA*-I LOH for 874 samples (458 LUAD and 416 LUSC) and *HLA*-II for 846 samples (465 LUAD and 381 LUSC). Frequency of *HLA*-I LOH and *HLA*-II LOH among the sample set was displayed as the mean of *HLA*-I LOH or *HLA*-II LOH frequencies across the six cutpoints. Within samples that were evaluated for LOH at ≥1 *HLA*-I loci and ≥1 *HLA*-II loci (N = 784 for full cohort, LUAD N = 437, LUSC N = 347), we determined the frequency of individuals with LOH at only *HLA*-I, defined as LOH at one or more *HLA*-I loci but none of the available *HLA*-II loci and only *HLA*-II, defined as LOH at one or more *HLA*-II loci but none of the available *HLA*-I loci.

HLA-II *loss of heterozygosity validation in the PCAWG and Hartwig Medical Foundation cohorts*

To validate our *HLA* LOH analyses performed in the TCGA, we curated two additional independent cohorts and evaluated *HLA*-II LOH using an adapted version of the methodology used for *HLA*-I LOH in Martínez-Jiménez *et al.* (*33*). The first dataset used for validation was the PCAWG cohort (*85*), consisting of 83 NSCLC samples (36 LUAD and 47 LUSC). The second dataset used for validation was the Hartwig Medical Foundation cohort (*86*), consisting of 657 samples (273 LUAD, 47 LUSC, remaining not annotated by histology). For each sample, LOH at each of the three *HLA*-I loci and each of the five *HLA*-II loci were determined. LOH of the *HLA*-I and *HLA*-II loci were defined as those events with a minor allele copy number lower than 0.5 and a major allele copy number greater than 0.5 as provided by PURPLE (https://github.com/hartwigmedical/hmftools/blob/master/purple/README.md). As in the TCGA analysis, samples with LOH at any one or more of the three *HLA*-I loci were annotating as having loss at "any *HLA*-I", and samples with LOH at any one or more of the five *HLA*-II loci were annotating as having loss at "any *HLA*-II." The frequency of *HLA*-I and *HLA*-II loss among the full cohort was calculated, as well as the frequencies within each histology for samples for which histology annotation was available. Mutually exclusive LOH classification for each sample was also identified as above, where samples with loss at "*HLA*-I only" had LOH at ≥1 *HLA*-I but none of the *HLA*-II loci, and vice versa for "*HLA*-II only." Rates for "*HLA*-I only" and "*HLA*-II only" were similarly determined within the full cohort as the proportion of samples with only *HLA*-I or only *HLA*-II loss out of the full cohort, and similarly within each histology.

HLA-II *neopeptide prediction*

Mutation annotation format (MAF) files for patients in the LUAD and LUSC TCGA cohorts were downloaded from the GDC portal (*115*). Patient MAFs were then filtered for missense mutations that were assigned to proteins characterized in the canonical UniProt human reference proteome (*118*). Next, a sliding window, ranging from 13 to 21 amino acids, was used to extract all possible *HLA*-II neopeptides bearing a given mutation in a patient-specific manner. Binding-affinity rank scores for each potential neopeptide were then assessed for patient *HLA-DP*, *HLA-DQ*, and *HLA-DR* alleles (*119*). Following prediction, potential neopeptides were assigned to the allele with the highest predicted affinity and filtered to only included peptides with a predicted affinity rank score of 10% or better (*120*).

HLA-II *neopeptide repertoire analysis*

Neopeptide repertoire size was defined as the number of distinct potential neopeptides assigned to a given locus. Prior to repertoire size analysis, several locus specific filtering steps were performed. First, patients without MAF files or with uncertain LOH status (i.e., failures in LOHHLA analysis for heterozygous patients) were filtered out. Next, in cases where MHC-II binding predictions failed for a particular allele, that patient would be removed from comparisons for that locus only. Lastly, patients without TMB information— obtained from Niknafs *et al.* (*121*)—were also removed. This resulted in the following number of patients being considered for each locus in the LUSC cohort: *DRB1*: 258, *DQA1*: 339, *DQB1*: 346, DPA1: 385, *DPB1*: 292. Similarly, the following number of patients were considered at each locus in the LUAD cohort: *DRB1*: 323, *DQA1*: 420, *DQB1*: 423, *DPA1*: 449, *DPB1*: 347. To assess the impact of *HLA* LOH on repertoire size, peptides assigned to alleles determined to be lost in a patient with LOH were labeled as such to assist in downstream sorting. Peptide hydrophobicity was calculated using the "Peptides" R package for heterozygous patients with and without LOH. Following peptide hydrophobicity calculation, peptides were grouped by patient and loss status (Loss or No Loss) and averaged.

Statistical analysis

We used a Spearman correlation for analyses comparing allele frequencies across cohort and populations. *P* values in logistic regression analyses were derived using the Wald statistics. We used the log-rank test to estimate statistical significance for multivariable Cox regression analyses. For analyses comparing distances between peptide and amino acid positions using crystal structure data for monomorphic and polymorphic positions, *P* values were computed using a two-sided Wilcoxon test. Unless otherwise noted, significance was denoted at $P < 0.05$. Comparisons of cluster prevalence between smokers and nonsmokers using scRNA-seq data were conducted using a Dirichlet multinomial regression. All analyses were conducted in RStudio with R version 3.6.1.

REFERENCES AND NOTES

1. R. S. Herbst, J. V. Heymach, S. M. Lippman, Lung cancer. *N. Engl. J. Med.* **359**, 1367–1380 (2008). doi: 10.1056/NEJMra0802714; pmid: 18815398
2. R. L. Siegel, K. D. Miller, H. E. Fuchs, A. Jemal, Cancer statistics, 2022. *CA Cancer J. Clin.* **72**, 7–33 (2022). doi: 10.3322/caac.21708; pmid: 35020204
3. H. Sung *et al.*, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **71**, 209–249 (2021). doi: 10.3322/caac.21660; pmid: 33538338
4. K. Yoshida *et al.*, Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020). doi: 10.1038/s41586-020-1961-1; pmid: 31996850
5. R. Doll, A. B. Hill, Smoking and carcinoma of the lung; preliminary report. *BMJ* **2**, 739–748 (1950). doi: 10.1136/bmj.2.4682.739; pmid: 14772469
6. X. Dai *et al.*, Health effects associated with smoking: A Burden of Proof study. *Nat. Med.* **28**, 2045–2055 (2022). doi: 10.1038/s41591-022-01978-x; pmid: 36216941
7. E. Long, H. Patel, J. Byun, C. I. Amos, J. Choi, Functional studies of lung cancer GWAS beyond association. *Hum. Mol. Genet.* **31** (R1), R22–R36 (2022). doi: 10.1093/hmg/ddac140; pmid: 35776125
8. W. Hill *et al.*, Lung adenocarcinoma promotion by air pollutants. *Nature* **616**, 159–167 (2023). doi: 10.1038/s41586-023-05874-3; pmid: 37020004
9. J. Malhotra, M. Malvezzi, E. Negri, C. La Vecchia, P. Boffetta, Risk factors for lung cancer worldwide. *Eur. Respir. J.* **48**, 889–902 (2016). doi: 10.1183/13993003.00359-2016; pmid: 27174888
10. P. B. Bach *et al.*, Variations in lung cancer risk among smokers. *J. Natl. Cancer Inst.* **95**, 470–478 (2003). doi: 10.1093/jnci/95.6.470; pmid: 12644540
11. A. Corthay, Does the immune system naturally protect against cancer? *Front. Immunol.* **5**, 197 (2014). doi: 10.3389/fimmu.2014.00197; pmid: 24860567

12. P. Ehrlich, Über den jetzigen Stand der Chemotherapie. *Ber. Dtsch. Chem. Ges.* **42**, 17–47 (1909). doi: 10.1002/cber.19090420105
13. M. Burnet, Cancer: A biological approach. III. Viruses associated with neoplastic conditions. IV. Practical applications. *BMJ* **1**, 841–847 (1957). doi: 10.1136/bmj.1.5023.841; pmid: 13413231
14. M. Burnet, Immunological factors in the process of carcinogenesis. *Br. Med. Bull.* **20**, 154–158 (1964). doi: 10.1093/oxfordjournals.bmb.a070310; pmid: 14168097
15. L. Thomas, On immunosurveillance in human cancer. *Yale J. Biol. Med.* **55**, 329–333 (1982). pmid: 6758376
16. F. M. Burnet, The concept of immunological surveillance. *Prog. Exp. Tumor Res.* **13**, 1–27 (1970). pmid: 4921480
17. R. D. Schreiber, L. J. Old, M. J. Smyth, Cancer immunoediting: Integrating immunity's roles in cancer suppression and promotion. *Science* **331**, 1565–1570 (2011). doi: 10.1126/science.1203486; pmid: 21436444
18. E. A. Engels *et al*., Spectrum of cancer risk among US solid organ transplant recipients. *JAMA* **306**, 1891–1901 (2011). doi: 10.1001/jama.2011.1592; pmid: 22045767
19. P. R. Burch, Leucocyte phenotypes in Hodgkin's disease. *Lancet* **296**, 771–772 (1970). doi: 10.1016/S0140-6736(70)90244-8; pmid: 4195991
20. K. E. Schratz *et al*., T cell immune deficiency rather than chromosome instability predisposes patients with short telomere syndromes to squamous cancers. *Cancer Cell* **41**, 807–817.e6 (2023). doi: 10.1016/j.ccell.2023.03.005; pmid: 37037617
21. E. Billerbeck *et al*., Mouse models of acute and chronic hepacivirus infection. *Science* **357**, 204–208 (2017). doi: 10.1126/science.aal1962; pmid: 28706073
22. A. R. Marderstein *et al*., Demographic and genetic factors influence the abundance of infiltrating immune cells in human tissues. *Nat. Commun.* **11**, 2213 (2020). doi: 10.1038/s41467-020-16097-9; pmid: 32371927
23. N. McGranahan *et al*., Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016). doi: 10.1126/science.aaf1490; pmid: 26940869
24. M. Yarchoan, A. Hopkins, E. M. Jaffee, Tumor Mutational Burden and Response Rate to PD-1 Inhibition. *N. Engl. J. Med.* **377**, 2500–2501 (2017). doi: 10.1056/NEJMc1713444; pmid: 29262275
25. D. Chowell *et al*., Improved prediction of immune checkpoint blockade efficacy across multiple cancer types. *Nat. Biotechnol.* **40**, 499–506 (2022). doi: 10.1038/s41587-021-01070-8; pmid: 34725502
26. R. M. Samstein *et al*., Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206 (2019). doi: 10.1038/s41588-018-0312-8; pmid: 30643254
27. N. A. Rizvi *et al*., Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015). doi: 10.1126/science.aaa1348; pmid: 25765070
28. J. D. McKay *et al*., Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* **49**, 1126–1132 (2017). doi: 10.1038/ng.3892; pmid: 28604730
29. A. Ferreiro-Iglesias *et al*., Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. *Nat. Commun.* **9**, 3927 (2018). doi: 10.1038/s41467-018-05890-2; pmid: 30254314
30. Y. Bossé, C. I. Amos, A Decade of GWAS Results in Lung Cancer. *Cancer Epidemiol. Biomarkers Prev.* **27**, 363–379 (2018).
31. P. Parham, T. Ohta, Population biology of antigen presentation by MHC class I molecules. *Science* **272**, 67–74 (1996). doi: 10.1126/science.272.5258.67; pmid: 8600539
32. N. McGranahan *et al*., Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **171**, 1259–1271.e11 (2017). doi: 10.1016/j.cell.2017.10.001; pmid: 29107330
33. F. Martínez-Jiménez *et al*., Genetic immune escape landscape in primary and metastatic cancer. *Nat. Genet.* **55**, 820–831 (2023). doi: 10.1038/s41588-023-01367-1; pmid: 37165135
34. R. Marty *et al*., MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell* **171**, 1272–1283.e15 (2017). doi: 10.1016/j.cell.2017.09.050; pmid: 29107334
35. R. Marty Pyke *et al*., Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell* **175**, 416–428.e13 (2018). doi: 10.1016/j.cell.2018.08.048; pmid: 30245014
36. D. J. Penn, K. Damjanovich, W. K. Potts, MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11260–11264 (2002). doi: 10.1073/pnas.162006499; pmid: 12177415
37. J. Arora *et al*., HLA Heterozygote Advantage against HIV-1 Is Driven by Quantitative and Qualitative Differences in HLA Allele-Specific Peptide Presentation. *Mol. Biol. Evol.* **37**, 639–650 (2020). doi: 10.1093/molbev/msz249; pmid: 31651980
38. M. Carrington *et al*., HLA and HIV-1: Heterozygote advantage and B*35-Cw*04 disadvantage. *Science* **283**, 1748–1752 (1999). doi: 10.1126/science.283.5408.1748; pmid: 10073943
39. M. R. Thursz, H. C. Thomas, B. M. Greenwood, A. V. Hill, Heterozygote advantage for HLA class-II type in hepatitis B virus infection. *Nat. Genet.* **17**, 11–12 (1997). doi: 10.1038/ng0997-11; pmid: 9288086
40. D. Chowell *et al*., Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science* **359**, 582–587 (2018). doi: 10.1126/science.aao4572; pmid: 29217585
41. D. Chowell *et al*., Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. *Nat. Med.* **25**, 1715–1720 (2019). doi: 10.1038/s41591-019-0639-4; pmid: 31700181
42. A. M. Goodman *et al*., MHC-I genotype and tumor mutational burden predict response to immunotherapy. *Genome Med.* **12**, 45 (2020). doi: 10.1186/s13073-020-00743-4; pmid: 32430031
43. K. Cuppens *et al*., HLA-I diversity and tumor mutational burden by comprehensive next-generation sequencing as predictive biomarkers for the treatment of non-small cell lung cancer with PD-(L)1 inhibitors. *Lung Cancer* **170**, 1–10 (2022). doi: 10.1016/j.lungcan.2022.05.019; pmid: 35689896
44. S. Takahashi *et al*., Impact of germline HLA genotypes on clinical outcomes in patients with urothelial cancer treated with pembrolizumab. *Cancer Sci.* **113**, 4059–4069 (2022). doi: 10.1111/cas.15488; pmid: 35848083
45. J. H. Shim *et al*., HLA-corrected tumor mutation burden and homologous recombination deficiency for the prediction of response to PD-(L)1 blockade in advanced non-small-cell lung cancer patients. *Ann. Oncol.* **31**, 902–911 (2020). doi: 10.1016/j.annonc.2020.04.004; pmid: 32320754
46. M. Montesion *et al*., 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 (2021). doi: 10.1158/2159-8290.CD-20-0672; pmid: 33127846
47. F. Pierini, T. L. Lenz, Divergent Allele Advantage at Human MHC Genes: Signatures of Past and Ongoing Selection. *Mol. Biol. Evol.* **35**, 2145–2158 (2018). doi: 10.1093/molbev/msy116; pmid: 29893875
48. C. Bycroft *et al*., The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018). doi: 10.1038/s41586-018-0579-z; pmid: 30305743
49. M. I. Kurki *et al*., FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023). doi: 10.1038/s41586-022-05473-8; pmid: 36653562
50. J. Ritari, K. Hyvärinen, J. Clancy, J. Partanen, S. Koskela; FinnGen, Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. *NAR Genom. Bioinform.* **2**, lqaa030 (2020). doi: 10.1093/nargab/lqaa030; pmid: 33575586
51. M. I. Kurki *et al*., FinnGen: Unique genetic insights from combining isolated population and national health register data. medRxiv 2022.03.03.22271360 [Preprint] (2022). https://doi:.doi: 10.1101/2022.03.03.22271360
52. K. Borodulin *et al*., Cohort Profile: The National FINRISK Study. *Int. J. Epidemiol.* **47**, 696–696i (2018). doi: 10.1093/ije/dyx239; pmid: 29165699
53. N. Mars *et al*., Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat. Med.* **26**, 549–557 (2020). doi: 10.1038/s41591-020-0800-0; pmid: 32273609
54. S. Jukarainen *et al*., Genetic risk factors have a substantial impact on healthy life years. *Nat. Med.* **28**, 1893–1901 (2022). doi: 10.1038/s41591-022-01957-2; pmid: 36097220
55. F. F. Gonzalez-Galarza *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** (D1), D783–D788 (2020). pmid: 31722398
56. O. D. Solberg *et al*., Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. *Hum. Immunol.* **69**, 443–464 (2008). doi: 10.1016/j.humimm.2008.05.001; pmid: 18638659
57. K. J. Karczewski *et al*., The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). doi: 10.1038/s41586-020-2308-7; pmid: 32461654
58. J. Robinson *et al*., Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLOS Genet.* **13**, e1006862 (2017). doi: 10.1371/journal.pgen.1006862; pmid: 28650991
59. A. T. Dilthey *et al*., HLA*LA-HLA typing from linearly projected graph alignments. *Bioinformatics* **35**, 4394–4396 (2019). doi: 10.1093/bioinformatics/btz235; pmid: 30942877
60. S. Kawaguchi, K. Higasa, M. Shimizu, R. Yamada, F. Matsuda, HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mutat.* **38**, 788–797 (2017). doi: 10.1002/humu.23230; pmid: 28419628
61. N. H. Thuesen, M. S. Klausen, S. Gopalakrishnan, T. Trolle, G. Renaud, Benchmarking freely available HLA typing algorithms across varying genes, coverages and typing resolutions. *Front. Immunol.* **13**, 987655 (2022). doi: 10.3389/fimmu.2022.987655; pmid: 36426357
62. H. M. E. Foster *et al*., The effect of socioeconomic deprivation on the association between an extended measurement of unhealthy lifestyle factors and health outcomes: A prospective analysis of the UK Biobank cohort. *Lancet Public Health* **3**, e576–e585 (2018). doi: 10.1016/S2468-2667(18)30200-7; pmid: 30467019
63. K. Ishigaki *et al*., HLA autoimmune risk alleles restrict the hypervariable region of T cell receptors. *Nat. Genet.* **54**, 393–402 (2022). doi: 10.1038/s41588-022-01032-z; pmid: 35332318
64. S. Raychaudhuri *et al*., Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296 (2012). doi: 10.1038/ng.1076; pmid: 22286218
65. X. Hu *et al*., Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat. Genet.* **47**, 898–905 (2015). doi: 10.1038/ng.3353; pmid: 26168013
66. A. Palomäki *et al*., Lifetime risk of rheumatoid arthritis-associated interstitial lung disease in *MUC5B* mutation carriers. *Ann. Rheum. Dis.* **80**, 1530–1536 (2021). doi: 10.1136/annrheumdis-2021-220698; pmid: 34344703
67. R. J. Hung *et al*., Assessing Lung Cancer Absolute Risk Trajectory Based on a Polygenic Risk Model. *Cancer Res.* **81**, 1607–1615 (2021). doi: 10.1158/0008-5472.CAN-20-1237; pmid: 33472890
68. International HIV Controllers Study, The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551–1557 (2010). 10.1126/science.1195271pmid: 21051598
69. Y. Luo *et al*., A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat. Genet.* **53**, 1504–1516 (2021). doi: 10.1038/s41588-021-00935-7; pmid: 34611364
70. S. W. Scally *et al*., A molecular basis for the association of the *HLA-DRB1* locus, citrullination, and rheumatoid arthritis. *J. Exp. Med.* **210**, 2569–2582 (2013). doi: 10.1084/jem.20131241; pmid: 24190431
71. J. A. Hollenbach *et al*., A specific amino acid motif of *HLA-DRB1* mediates risk and interacts with smoking history in Parkinson's disease. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 7419–7424 (2019). doi: 10.1073/pnas.1821778116; pmid: 30910980
72. K. H. Lee, K. W. Wucherpfennig, D. C. Wiley, Structure of a human insulin peptide-HLA-DQ8 complex and susceptibility to type 1 diabetes. *Nat. Immunol.* **2**, 501–507 (2001). doi: 10.1038/88694; pmid: 11376336
73. A. M. Leader *et al*., Single-cell analysis of human non-small cell lung cancer lesions refines tumor classification and patient stratification. *Cancer Cell* **39**, 1594–1609.e12 (2021). doi: 10.1016/j.ccell.2021.10.009; pmid: 34767762
74. K. J. Travaglini *et al*., A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020). doi: 10.1038/s41586-020-2922-4; pmid: 33208946
75. N. Kim *et al*., Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* **11**, 2285 (2020). doi: 10.1038/s41467-020-16164-1; pmid: 32385277
76. A. Desrichard *et al*., Tobacco Smoking-Associated Alterations in the Immune Microenvironment of Squamous Cell Carcinomas. *J. Natl. Cancer Inst.* **110**, 1386–1392 (2018). doi: 10.1093/jnci/djy060; pmid: 29659925

77. C. S. Smillie *et al*., Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714–730.e22 (2019). doi: 10.1016/j.cell.2019.06.029; pmid: 31348891
78. B. Liu *et al*., Temporal single-cell tracing reveals clonal revival and expansion of precursor exhausted T cells during anti-PD-1 therapy in lung cancer. *Nat. Cancer* **3**, 108–121 (2022). doi: 10.1038/s43018-021-00292-8; pmid: 35121991
79. M. Casanova-Acebes *et al*., Tissue-resident macrophages provide a pro-tumorigenic niche to early NSCLC cells. *Nature* **595**, 578–584 (2021). doi: 10.1038/s41586-021-03651-8; pmid: 34135508
80. M. L. Axelrod, R. S. Cook, D. B. Johnson, J. M. Balko, Biological Consequences of MHC-II Expression by Tumor Cells in Cancer. *Clin. Cancer Res.* **25**, 2392–2402 (2019). doi: 10.1158/1078-0432.CCR-18-3200; pmid: 30463850
81. J. E. Wosen, D. Mukhopadhyay, C. Macaubas, E. D. Mellins, Epithelial MHC Class II Expression and Its Role in Antigen Presentation in the Gastrointestinal and Respiratory Tracts. *Front. Immunol.* **9**, 2144 (2018). doi: 10.3389/fimmu.2018.02144; pmid: 30319613
82. A. T. Shenoy *et al*., Antigen presentation by lung epithelial cells directs $CD4^+$ T_{RM} cell function and regulates barrier immunity. *Nat. Commun.* **12**, 5834 (2021). doi: 10.1038/s41467-021-26045-w; pmid: 34611166
83. E. Alspach *et al*., MHC-II neoantigens shape tumour immunity and response to immunotherapy. *Nature* **574**, 696–701 (2019). doi: 10.1038/s41586-019-1671-8; pmid: 31645760
84. A. M. Newman *et al*., Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019). doi: 10.1038/s41587-019-0114-2; pmid: 31061481
85. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, , Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020). doi: 10.1038/s41586-020-1969-6
86. P. Priestley *et al*., Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019). doi: 10.1038/s41586-019-1689-y; pmid: 31645765
87. D. Chowell *et al*., TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1754–E1762 (2015). doi: 10.1073/pnas.1500973112; pmid: 25831525
88. X. Ma *et al*., Functional landscapes of POLE and POLD1 mutations in checkpoint blockade-dependent antitumor immunity. *Nat. Genet.* **54**, 996–1012 (2022). doi: 10.1038/s41588-022-01108-w; pmid: 35817971
89. J. J. A. Calis *et al*., Properties of MHC class I presented peptides that enhance immunogenicity. *PLOS Comput. Biol.* **9**, e1003266 (2013). doi: 10.1371/journal.pcbi.1003266; pmid: 24204222
90. K. M. Wright *et al*., Hydrophobic interactions dominate the recognition of a KRAS G12V neoantigen. *Nat. Commun.* **14**, 5063 (2023). doi: 10.1038/s41467-023-40821-w; pmid: 37604828
91. C. Tomasetti, B. Vogelstein, Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015). doi: 10.1126/science.1260825; pmid: 25554788
92. C. Tomasetti, L. Li, B. Vogelstein, Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017). doi: 10.1126/science.aaf9011; pmid: 28336671
93. R. E. Tay, E. K. Richardson, H. C. Toh, Revisiting the role of $CD4^+$ T cells in cancer immunotherapy-new insights into old paradigms. *Cancer Gene Ther.* **28**, 5–17 (2021). doi: 10.1038/s41417-020-0183-x; pmid: 32457487
94. J. Borst, T. Ahrends, N. Bąbała, C. J. M. Melief, W. Kastenmüller, $CD4^+$ T cell help in cancer immunology and immunotherapy. *Nat. Rev. Immunol.* **18**, 635–647 (2018). doi: 10.1038/s41577-018-0044-0; pmid: 30057419
95. D. Y. Oh *et al*., Intratumoral $CD4^+$ T Cells Mediate Anti-tumor Cytotoxicity in Human Bladder Cancer. *Cell* **181**, 1612–1625.e13 (2020). doi: 10.1016/j.cell.2020.05.017; pmid: 32497499
96. P. Goyette *et al*., High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat. Genet.* **47**, 172–179 (2015). doi: 10.1038/ng.3176; pmid: 25559196
97. M. Greaves, A causal mechanism for childhood acute lymphoblastic leukaemia. *Nat. Rev. Cancer* **18**, 471–484 (2018). doi: 10.1038/s41568-018-0015-6; pmid: 29784935
98. Q.-L. Wang *et al*., Association of HLA diversity with the risk of 25 cancers in the UK Biobank. *EBioMedicine* **92**, 104588 (2023). doi: 10.1016/j.ebiom.2023.104588; pmid: 37148584
99. A. E. Kennedy, U. Ozbek, M. T. Dorak, What has GWAS done for HLA and disease associations? *Int. J. Immunogenet.* **44**, 195–211 (2017). doi: 10.1111/iji.12332; pmid: 28877428
100. V. Matzaraki, V. Kumar, C. Wijmenga, A. Zhernakova, The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**, 76 (2017). doi: 10.1186/s13059-017-1207-1; pmid: 28449694
101. M. M. A. Fernando *et al*., Defining the role of the MHC in autoimmunity: A review and pooled analysis. *PLOS Genet.* **4**, e1000024 (2008). doi: 10.1371/journal.pgen.1000024; pmid: 18437207
102. A. Fry *et al*., Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017). doi: 10.1093/aje/kwx246; pmid: 28641372
103. D. R. Aberle *et al*., Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**, 395–409 (2011). doi: 10.1056/NEJMoa1102873; pmid: 21714641

ACKNOWLEDGMENTS

This work was carried out under UK Biobank application 61123. We acknowledge the participants and investigators of FinnGen study. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie, AstraZeneca UK, Biogen MA, Bristol Myers Squibb (and Celgene Corporation and Celgene International II Sàrl), Genentech, Merck Sharp & Dohme Pfizer, GlaxoSmithKline Intellectual Property Development, Sanofi US Services, Maze Therapeutics, Janssen Biotech, Novartis AG, and Boehringer Ingelheim International GmbH. The following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (https://www.auria.fi/biopankki), THL Biobank (https://www.thl.fi/biobank), Helsinki Biobank (https://www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (https://www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (https://www.ita-suomenbiopankki.fi/en), Central Finland Biobank (https://www.sairaalanova.fi/), Finnish Red Cross Blood Service Biobank (https://www.veripalvelu.fi/verenluovutus/biopankkitoiminta), Terveystalo Biobank (https://www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/) and Arctic Biobank (https://www.oulu.fi/en/university/faculties-and-units/faculty-medicine/northern-finland-birth-cohorts-and-arctic-biobank). All Finnish Biobanks are members of BBMRI.fi infrastructure (https://finbb.fi/en/what-is-finbb). The Finnish Biobank Cooperative—FINBB (https://finbb.fi/en/)—is the coordinator of BBMRI-ERIC operations in Finland. The Finnish biobank data can be accessed through the Fingenious services (https://site.fingenious.fi/en/), managed by FINBB. This work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant ULTR004419 from the National Center for Advancing Translational Sciences. This publication and the underlying study have been made possible partly on the basis of the data that the Hartwig Medical Foundation and the Center for Personalized Cancer Treatment have made available to the study. F.M.J. would like to acknowledge the Cellex Foundation for providing research facilities and equipment. We are grateful to T. A. Chan for his mentorship and helpful discussions. **Funding:** This work was supported by the Alexander and Alexandrine Sinsheimer Foundation (D.C.), Department of Defense (ME220130) (D.C.), the Melanoma Research Alliance (DC), US NIH grant DP5 OD028171 (R.M.S.), the Burroughs Wellcome Fund Career Award for Medical Scientists (R.M.S.), the American Lung Association Lung Cancer Discovery Award (R.M.S.), US NIH grant R01 CA283469 (R.M.S. and D.C.), the Icahn School of Medicine at Mount Sinai Translational Immunology Training Program (T32 AI078892) (M.S.), the NIH Cancer Target Discovery and Development (CTD^2) Network (U01CA282114) (R.M.S., M.M., and D.C.), the Doris Duke Foundation (R.M.S. and M.S.), Department of Defense (CA220766) (R.M.S. and D.C.), the Instrumentarium Science Foundation (H.O. and A.T.), Academy of Finland grant 331671 (N.M.), University of Helsinki HiLIFE Fellows grant 2023-2025 (N.M.), Finska Lakaresallskapet (N.M.), the LUNGevity Foundation (M.E.S. and Z.H.G.), Cancer Moonshot NCI R33 award CA263705-01 (M.E.S. and Z.H.G.), Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) grant 437857095 (T.L.L.), State Agency for Research (Agencia Estatal de Investigación), Severo Ochoa Center of Excellence grant CEX2020-001024-S/AEI/10.13039/501100011033 (F.M.J.), and the CaixaResearch Advanced Oncology Research Programme supported by the "La Caixa" Foundation (F.M.J.).
Author contributions: Conceptualization: D.C., C.K., R.M.S., and H.M.O. Data curation: C.K., D.C., A.T., H.M.O., M.S., R.M.S., and E.W. Investigation: C.K., D.C., A.T., H.M.O., M.S., R.M.S., E.W, S.Y., N.M., V.R., A.C.B., S.E.J., N.V., M.E., Z.H.G., T.L.L., M.M., P.B., and F.M.J. Writing: C.K. and D.C. Statistical analysis: C.K., A.T., M.S., D.C., R.M.S., H.M.O., P.B., and N.M. Supervision: D.C., R.M.S., and H.M.O. **Competing interests:** D.C. and R.M.S. have filed a patent application related to tumor mutational load (17536715). D.C., C.K., and T.L have filed a patent application related to *HLA* class I sequence divergence and cancer therapy (17770259). M.M. serves on the scientific advisory board and holds stock from Compugen, Myeloid Therapeutics, Morphic Therapeutics, Asher Bio, Dren Bio, Nirogy, Oncoresponse, Owkin, Pionyr, OSE and Larkspur. M.M. serves on the scientific advisory board of Innate Pharma, DBV, and Genenta. All other authors declare that they have no competing interests. **Data and materials availability:** All source data for epidemiological analyses can be accessed through applications to the UK Biobank (https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access) and FinnGen R8 (https://www.finngen.fi/en). *HLA* allele amino acid sequences for the fine-mapping analyses can be accessed through IMGT (https://www.ebi.ac.uk/ipd/imgt/hla/alleles/). Datasets used for single-cell RNA-sequencing analyses can be accessed from the corresponding studies listed in the methods section—links to the public repositories containing these data are as follows: Leader *et al*. (https://www.github.com/effiken/Leader_et_al), Travaglini *et al*. (https://www.synapse.org/#!Synapse:syn21041850/wiki/600865), and Kim *et al*. (deposited in GEO with accession ID GSE131907). Whole-exome sequencing and RNA-seq data from the TCGA can be accessed at https://portal.gdc.cancer.gov/. Reanalyzed whole-genome sequencing data from the PCAWG and Hartwig Medical Foundation samples can be accessed from the original study listed in the methods section (Martínez-Jiménez *et al*.) at the following repositories: https://icgc.bionimbus.org/files/5310a3ac-0344-458a-88ce-d55445540120, https://dcc.icgc.org/releases/PCAWG/Hartwig, and https://www.hartwigmedicalfoundation.nl/en/data/data-access-request/.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adi3808
Materials and Methods
Figs. S1 to S28
Tables S1 to S25
References (*104*–*121*)
MDAR Reproducibility Checklist

Submitted 21 April 2023; accepted 16 January 2024
10.1126/science.adi3808