OXFORD

*Human Molecular Genetics*, 2025, 1–11
https://doi.org/10.1093/hmg/ddaf059
**Original Article**

# Genome-wide association study for lung cancer in 6531 African Americans reveals new susceptibility loci

Jinyoung Byun[1,2,3,†,*], Younghun Han[1,2,3,†], Jiyeon Choi[4], Ryan Sun[5], Vikram R. Shaw[1], Catherine Zhu[1], Xiangjun Xiao[1], Christine Lusk[6,7], Hoda Badr[2], Hyun-Sung Lee[8], Hee-Jin Jang[8], Yafang Li[1,2,3], Hyeyeun Lim[2], Erping Long[9], Yanhong Liu[2], Linda Kachuri[10], Kyle M. Walsh[11], John K. Wiencke[12], Demetrius Albanes[4], Stephen Lam[13], Adonina Tardon[14], Marian L. Neuhouser[15], Matt J. Barnett[15], Chu Chen[15], Stig Bojesen[16,17], Hermann Brenner[18,19,20], Maria Teresa Landi[4], Mattias Johansson[21], Angela Risch[22,23,24], H-Erich Wichmann[25], Heike Bickeböller[26], David C. Christiani[27], Gad Rennert[28], Susanne Arnold[29], John K. Field[30], Sanjay Shete[5,31], Loic Le Marchand[32], Geoffrey Liu[33], Angeline S. Andrew[34], Shanbeh Zienolddiny[35], Kjell Grankvist[36], Mikael Johansson[37], Neil Caporaso[4], Fiona Taylor[38], Philip Lazarus[39], Matthew B. Schabath[40], Melinda C. Aldrich[41], Alpa Patel[42], Xihong Lin[43], Krista A. Zanetti[44], Curtis C. Harris[45], Stephen Chanock[4], James McKay[21], Ann G. Schwartz[6,7], Rayjean J. Hung[46,47], Christopher I. Amos[1,2,3,*], INTEGRAL-ILCCO Consortium

[1]Institute for Clinical and Translational Research, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX, 77030, United States
[2]Section of Epidemiology and Population Sciences, Department of Medicine, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX, 77030, United States
[3]University of New Mexico Comprehensive Cancer Center, 1201 Camino de Salud NE, Albuquerque, NM, 87102, United States
[4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, 9615 Medical Center Drive, Rockville, MD, 20850, United States
[5]Department of Biostatistics, University of Texas, M.D. Anderson Cancer Center, 7007 Bertner Ave, Houston, TX, 77030, United States
[6]Department of Oncology, Wayne State University School of Medicine, 4100 John R, Detroit, MI, 48201, United States
[7]Karmanos Cancer Institute, 4100 John R Street, Detroit, MI, 48201, United States
[8]Systems Onco-Immunology Lab, David Sugarbaker Division of Thoracic Surgery, Michael E. DeBakey Department of Surgery, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX, 77030, United States
[9]State Key Laboratory of Respiratory Health and Multimorbidity, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100005, China
[10]Department of Epidemiology and Population Health, Stanford University, 300 Pasteur Drive, Stanford, CA, 94305, United States
[11]Duke Cancer Institute, Duke University Medical Center, 20 Duke Medicine Cir, Durham, NC, 27701, United States
[12]Department of Neurological Surgery, The University of California, San Francisco, 400 Parnassus Ave, San Francisco, CA, 94143, United States
[13]Department of Integrative Oncology, University of British Columbia, 675 West 10th Ave, Vancouver, BC V5Z 1L3, Canada
[14]Public Health Department, University of Oviedo, and Health Research Institute of Asturias, ISPA, Av. del Hospital Universitario, s/n, 33011 Oviedo, Asturias, Spain
[15]Program in Cancer Prevention, Public Health Sciences Division, Fred Hutchinson Cancer Center, 1100 Fairview Ave N, Seattle, WA, 98109, United States
[16]Department of Clinical Biochemistry, Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen, Denmark
[17]Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen, Denmark
[18]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Deutsches Krebsforschungszentrum (DKFZ), Im Neuenheimer Feld 280, 69120, Heidelberg, Germany
[19]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Im Neuenheimer Feld 581, 69120, Heidelberg, Germany
[20]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120, Heidelberg, Germany
[21]Section of Genetics, International Agency for Research on Cancer, World Health Organization, 25 avenue Tony Garnier, CS 90627, 69366 LYON CEDEX 07, France
[22]Translational Lung Research Center Heidelberg (TLRC-H), German Center for Lung Research (DZL), Im Neuenheimer Feld 156, 69120, Heidelberg, Germany
[23]Division of Cancer Epigenomics, DKFZ-German Cancer Research Center, Im Neuenheimer Feld 280, D-69120, Heidelberg, Germany
[24]Department of Biosciences and Medical Biology, Center for Tumor Biology and Immunology, University of Salzburg and Cancer Cluster Hellbrunner Strasse 34, Salzburg, 5020, Austria
[25]Helmholtz-Munich Institute of Epidemiology, Ingolstädter Landstraße 1, Neuherberg, 85764, Germany
[26]University Medical Center Göttingen, Institute of Genetic Epidemiology, Humboldtallee 32, 37073 Göttingen, Germany
[27]Department of Environmental Health and Epidemiology, Harvard T.H.Chan School of Public Health, 665 Huntington Avenue, Building 1, Boston, MA, 02115, United States
[28]Clalit National Cancer Control Center at Carmel Medical Center and Technion Faculty of Medicine, Mikhal St 7, Haifa, 3436212, Israel
[29]University of Kentucky, Markey Cancer Center, 800 Rose Street, Lexington, KY, 40536, United States
[30]Institute of Translational Medicine, University of Liverpool, the Sherrington Building, Ashton St, Liverpool, L69 3GE, United Kingdom
[31]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX, 77030, United States
[32]Epidemiology Program, University of Hawaii Cancer Center, 701 Ilalo Street, Honolulu, HI, 96813, United States
[33]University Health Network- The Princess Margaret Cancer Centre, 610 University Ave, Toronto, ON M5G 2M9, Canada
[34]Departments of Epidemiology and Community and Family Medicine, Dartmouth College, 1 Rope Ferry Road, Hanover, NH, 03755, United States
[35]National Institute of Occupational Health, Gydas Vei 8, 0363, Oslo, Norway
[36]Department of Medical Biosciences, Umeå University, 901 87 Umeå, Sweden
[37]Department of Radiation Sciences, Oncology, Umeå University, 901 87 Umeå, Sweden
[38]Sheffield Teaching Hospitals Foundation Trust, 8 Beech Hill Road, Sheffield, S10 2SB, United Kingdom
[39]Department of Pharmaceutical Sciences, College of Pharmacy, Washington State University, 412 East Spokane Falls Blvd, PBS 130, Spokane, WA, 99202, United States

**Received:** August 8, 2024. **Revised:** March 31, 2025. **Accepted:** April 9, 2025

[40]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, 12902 USF Magnolia Drive, Tampa, FL, 33612, United States
[41]Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, 1161 21st Ave S, Nashville, TN, 37232, United States
[42]American Cancer Society, Inc., 270 Peachtree Street NW, Atlanta, GA, 30303, United States
[43]Department of Biostatistics, Harvard TH Chan School of Public Health, 655 Huntington Avenue, Boston, MA, 02115, United States
[44]Office of Nutrition Research, Division of Program Coordination, Planning, and Strategic Initiatives, Office of the Director, National Institutes of Health, 6705 Rockledge Drive, Bethesda, MD, 20817, United States
[45]Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, 37 Convent Dr, Bethesda, MD, 20892, United States
[46]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, 600 University Ave, Toronto, ON M5G 1X5, Canada
[47]Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, Ontario, M5T 3M7, Canada
*Corresponding authors. Christopher I. Amos and Jinyoung Byun, University of New Mexico Comprehensive Cancer Center, Albuquerque, NM, 87131, United States. Email: ciamos@salud.unm.edu; jbyun@salud.unm.edu
†Jinyoung Byun and Younghun Han made equal contributions.

## Abstract

Despite lung cancer affecting all races and ethnicities, disparities are observed in incidence and mortality rates among different ethnic groups in the United States. Non-Hispanic African Americans had a high incidence rate of lung cancer at 55.8 per 100 000 people, as well as the highest death rate at 37.2 per 100 000 people from 2016 to 2020. While previous genome-wide association studies (GWAS) have identified over 45 susceptibility risk loci that influence lung cancer development, few GWAS have investigated the etiology of lung cancer in African Americans. To address this gap in knowledge, we conducted GWAS of lung cancer focused on studying African Americans, comprising 2267 lung cancer cases and 4264 controls. We identified three loci associated with lung cancer, one with lung adenocarcinoma, and four with lung squamous cell carcinoma in this population at the genomic-wide significance level. Among them, three novel loci were identified near *VWF* at 12p13.31 for overall lung cancer and *GACAT3* at 2p24.3 and *LMAN1L* at 15q24.1 for lung squamous cell carcinoma. In addition, we confirmed previously reported risk loci with known or new lead variants near *CHRNA5* at 15q25.1 and *CYP2A6* at 19q13.2 associated with lung cancer and *TRIP13* at 5p15.33 and *ERC1* at 12p13.33 associated with lung squamous cell carcinoma. Further multi-step functional analyses shed light on biological mechanisms underlying these associations of lung cancer in this population. Our study highlights the importance of ancestry-specific studies for the potential alleviation of lung cancer burden in African Americans.

*Keywords*: Lung cancer; Genome-wide association study; African American; Post-GWAS; Polygenic risk score

## Introduction

Lung cancer has a high impact on African Americans, among other populations, in the United States [1]. Both African American men and women had a high rate of developing lung cancer at 58.2 per 100 000 people and the highest death rate at 33.4 per 100 000 people from 2016 to 2020 [1]. In 2020, non-Hispanic African American men had the highest incidence rate at 60.6 per 100 000 people and the highest death rate at 45.7 per 100 000 people.

While previous genome-wide association studies (GWAS) of lung cancer have discovered over 45 risk loci [2], a large proportion of the heritability of lung cancer remains unexplained. Moreover, most single population-based GWAS have been performed in Europeans. Although a few East-Asian ancestry-specific genetic studies [3–5] and two cross-ancestry studies [6, 7] have been conducted so far, non-European ancestry-specific GWAS of lung cancer has been underrepresented [8]. African Americans have a higher lung cancer incidence and poorer lung cancer survival rate, while they smoke fewer cigarettes per day and have different genetic aspects of nicotine metabolism compared to Europeans, which implies unknown potentially complex risk factors [9–11]. Few comprehensive GWAS of lung cancer focused on studying African Americans and have been conducted in a modest sample size, confirming a few loci previously identified in European and/or Asian populations but have not identified new loci specific to African Americans [2, 8, 10]. More recently, the Million Veteran Program (MVP) performed a GWAS of overall lung cancer in 2438 African American cases and identified a new locus specific to this population due to differences in allele frequency [12]. Given differences in genetic architectures, including allele frequencies and linkage disequilibrium in African Americans compared to Europeans, GWAS focusing on this population can potentially identify new loci as well as new variants from known susceptibility loci in lung cancer.

We leveraged the existing large-scale cross-ancestry GWAS datasets [6], which were not fully explored, for ancestry-specific analyses to investigate genetic susceptibility loci associated with lung cancer and histological subtypes in African-ancestry populations. We conducted genome-wide association analyses and genome-wide meta-analysis (GWMA) on a dataset comprising 2267 lung cancer cases and 4264 controls of African Americans to identify common and low-frequency genetic susceptibility variants associated with lung cancer and its subtypes. We then performed multi-step functional annotation on the newly identified genetic loci to prioritize risk variants specific to African-ancestry and to understand the biological mechanisms affecting lung cancer susceptibility in this population. Specifically, we used Functional Mapping and Annotation of GWAS platform (FUMA GWAS) [13], Functional Annotation of Variants—Online Resource platform (FAVOR) [14], ProteomicsDB [15], STRING [16], and RegulomeDB [17]. We also surveyed the pleiotropic effect of the sentinel SNP on various human traits of each locus identified in lung cancer GWMA in this population.

## Results

### Identification of lung cancer risk loci in African Americans

Leveraging the recent cross-ancestry GWAS in lung cancer [6], we selected 6531 African Americans from three studies, the OncoArray Consortium of Lung Study (ONCO), Lung Cancer and Smoking Phenotypes in African American Cases and Controls (AA-NCI), and INflammation, Health, Ancestry and Lung Epidemiology study (INHALE), to perform GWAS in this population (Table 1). Associations of 15 462 133, 15 424 778, and 14 438 297 genetic variants for overall lung cancer (Lung), lung adenocarcinoma (ADE), and lung squamous cell carcinoma (SQC) were subjected to meta-analysis using METASOFT [18] to achieve

**Table 1.** Sample characteristics of genome-wide association studies in African Americans.

| Strata | STUDY 1 | | | | | | STUDY 2 | | TOTAL | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ONCO | | AA-NCI | | Total | | INHALE | | | |
| Cases | Controls | Cases | Controls | Cases | Controls | Cases | Controls | Cases | Controls | |
| Lung cancer | 269 | 286 | 1704 | 3460 | 1973 | 3746 | 294 | 518 | 2267 | 4264 |
| ADE | 117 | 286 | 723 | 3460 | 840 | 3746 | 170 | 518 | 1010 | 4264 |
| SQC | 55 | 286 | 394 | 3460 | 449 | 3746 | 68 | 518 | 517 | 4264 |

**Table 2.** Top associations in ancestry-specific and cross-ancestry lung cancer and histological subtype analyses.

| Strata | SNP | Cytoband | Position | Nearest Gene | Allele | EAF (Study 1; Study 2; GWMA) | OR (Study 1; Study 2; GWMA) | P value (Study 1; Study2; GWMA) | I2 | P_Q |
|---|---|---|---|---|---|---|---|---|---|---|
| **Lung[1]** | **rs216859** | **12p13.31** | **6 087 531** | ***VWF*** | **C_T** | **0.72; 0.67; 0.71** | **1.30; 1.34; 1.30** | **$4.88 \times 10^{-8}$; $1.69 \times 10^{-2}$; $2.67 \times 10^{-9}$** | **0.00** | **0.81** |
| Lung[2] | rs17486278 | 15q25.1 | 78 867 482 | *CHRNA5* | C_A | 0.30; 0.32; 0.30 | 1.31; 1.55; 1.34 | $9.01 \times 10^{-10}$; $1.18 \times 10^{-4}$; $1.27 \times 10^{-12}$ | 49.20 | 0.16 |
| Lung[2] | rs144437384 | 19q13.2 | 41 351 563 | *CYP2A6* | A_G | 0.05; 0.06; 0.05 | 0.58; 0.44; 0.56 | $4.39 \times 10^{-7}$; $1.98 \times 10^{-3}$; $4.98 \times 10^{-9}$ | 0.00 | 0.35 |
| ADE[2] | rs17486278 | 15q25.1 | 78 867 482 | *CHRNA5* | C_A | 0.29; 0.31; 0.29 | 1.33; 1.55; 1.36 | $8.32 \times 10^{-7}$; $1.33 \times 10^{-3}$; $6.92 \times 10^{-9}$ | 2.81 | 0.31 |
| **SQC[1]** | **rs6431749** | **2p24.3** | **16 166 988** | ***MYCN, GACAT3*** | **T_C** | **0.73; 0.73; 0.73** | **0.66; 0.54; 0.64** | **$4.88 \times 10^{-8}$; $3.13 \times 10^{-3}$; $8.15 \times 10^{-10}$** | **0.00** | **0.38** |
| SQC[2] | rs115287843 | 5p15.33 | 945 063 | *TRIP13, LOC100506688* | A_G | 0.01; 0.02; 0.02 | 3.39; 4.25; 3.53 | $4.66 \times 10^{-7}$; $5.11 \times 10^{-3}$; $8.83 \times 10^{-9}$ | 0.00 | 0.69 |
| SQC[2] | rs113048688 | 12p13.33 | 1 262 177 | *ERC1* | A_G | 0.01; 0.02; 0.01 | 4.14; 5.10; 4.37 | $4.39 \times 10^{-6}$; $1.86 \times 10^{-3}$; $3.09 \times 10^{-8}$ | 0.00 | 0.73 |
| **SQC[1]** | **rs115735578** | **15q24.1** | **75 112 459** | ***LMAN1L*** | **A_T** | **0.01; 0.01, 0.01** | **4.43; 3.19; 4.23** | **$4.36 \times 10^{-8}$; $8.43 \times 10^{-2}$; $1.04 \times 10^{-8}$** | **0.00** | **0.65** |

Nearest gene (reference NCBI build37) is given as locus label and includes all the genes +/− 200 kb of the genomic risk SNP; Allele, effect allele_other allele; EAF, effect allele frequency for AA-NCI and ONCO (Study 1), INHALE (Study 2), Fixed-effect Genome-Wide Meta-Analysis (GWMA); OR, odds ratio effect size for AA-NCI and ONCO (Study 1), INHALE (Study 2), Fixed-effect GWMA (GWMA); I2, I^2 Heterogeneity Index (0–100); P_Q, P-value for Cochrane's Q statistic; 1, newly identified susceptibility loci (also shown in bold); 2, new lead variant from a previously reported loci in other populations.

optimal statistical power, with fixed effects model based on inverse-variance-weighted effect size. We performed a meta-analysis on all the available samples, rather than implementing a two-stage discovery-replication approach, since Skol et al. demonstrated this approach as generally more powerful [19]. The study workflow shown in Fig. 1 summarizes the steps from data preparation to subsequent analyses in the present study. There was no evidence of genomic inflations for Lung ($\lambda_{Lung} = 1.013$) or any histologic subtypes, ADE and SQC ($\lambda_{ADE} = 0.992$; $\lambda_{SQC} = 0.989$), implying that residual population stratification is unlikely to influence association statistics within individual genome-wide association analyses and combined meta-analysis (Fig. 2). GWMA identified three, one, and four susceptibility loci associated with Lung, ADE, and SQC at the genome-wide significance level of $P \leq 5 \times 10^{-8}$, respectively. Top GWMA results in this population are reported in Table 2, and genomic loci associated with lung cancer risk at a suggestive significance level of $P \leq 10^{-5}$ for lung cancer overall and adenocarcinoma and squamous carcinoma specific strata are reported in Supplementary Tables 1–3. Genomic regional association plots for the top genetic variants discovered in GWMA are shown in Supplementary Fig. 1.

Among the significant loci, two loci for overall lung cancer (15q25.1 and 19q13.2), one for ADE (15q25.1), and two for SQC (5p15.33 and 12p13.33) were reported in previous GWAS [2, 6, 8, 10] (Table 2), including two loci (5p15.33 and 15q25.1) reported in African Americans [10]. Confirmation of these loci was either with already reported or newly identified lead variants (Table 2, Fig. 2, Supplementary Tables 1–3). Among them were the loci implicated in smoking behaviors (Ever and Never smokers) [10], such as rs17486278 and rs55781567 within *CHRNA5* at 15q25.1 and rs144437384, in *CYP2A6* at 19q13.2. [2]. It has been observed that the intronic variant, rs2853677, in *TERT* at 5p15.33, has consistent effects across diverse populations [6] and was previously reported in African Americans [10]. In our study, this variant displayed an association with overall lung cancer and ADE at a suggestive genome-wide significance level of $P < 10^{-5}$ (Supplementary Tables 1 and 2). Further, we identified two new lead variants associated with SQC at two known loci: rs115287843, between *TRIP13* and *LOC100506688* at 5p15.33 and rs113048688 in *ERC1* at 12p13.33 (Table 1, Supplementary Table 3, Supplementary Fig. 1f–g). The new lead variants were defined to be within ±250 kb of a previously reported variant with an $r^2 < 0.6$ (1000 Genomes AFR population) [20]. These two new lead variants for SQC showed low allele frequencies in our dataset (EAF = 0.02 for rs115287843 and 0.01 for rs113048688) but monoallelic in the 1000 Genomes European populations.

Notably, we identified three novel susceptibility loci, including one in Lung (12p13.31) and two in SQC (2p24.3 and 15q24.1) at the genome-wide significance level (Table 1, Fig. 2, Supplementary Fig. 1). From the locus at 12p13.31, an intronic variant, rs216859,

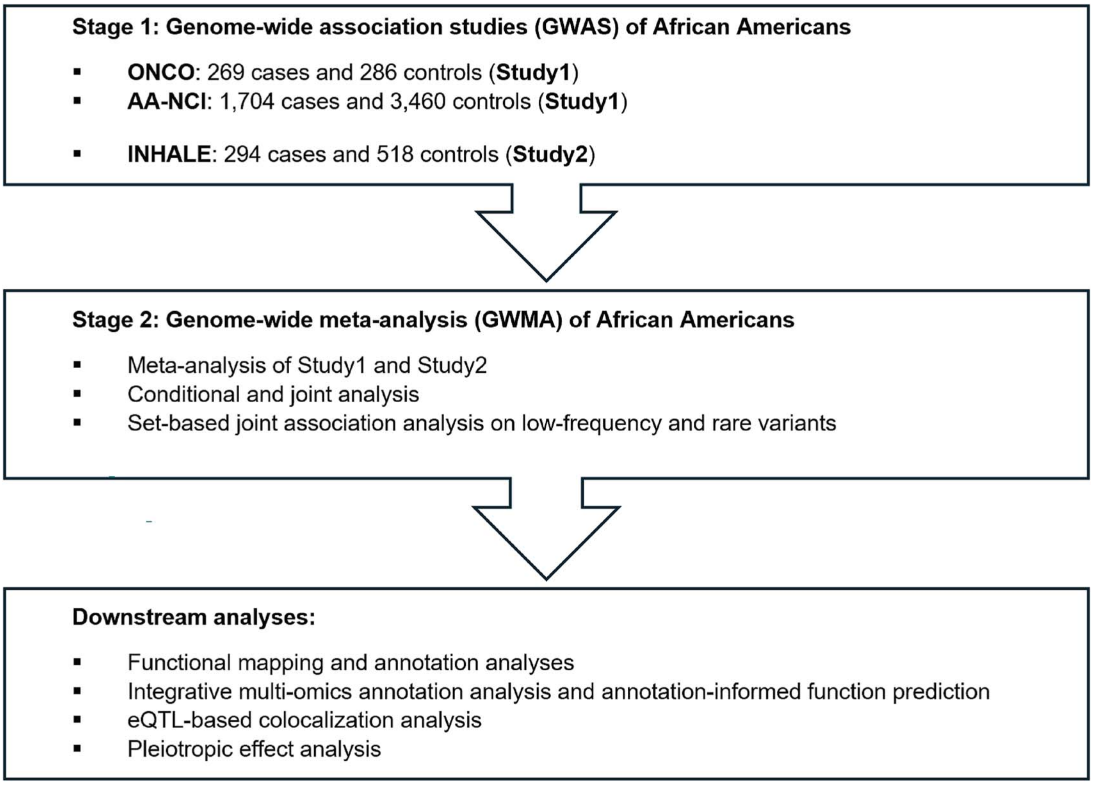
Stage 1: Genome-wide association studies (GWAS) of African Americans
ONCO: 269 cases and 286 controls (Study1)
AA-NCI: 1,704 cases and 3,460 controls (Study1)
INHALE: 294 cases and 518 controls (Study2)
Stage 2: Genome-wide meta-analysis (GWMA) of African Americans
Meta-analysis of Study1 and Study2
Conditional and joint analysis
Set-based joint association analysis on low-frequency and rare variants
Downstream analyses:
Functional mapping and annotation analyses
Integrative multi-omics annotation analysis and annotation-informed function prediction
eQTL-based colocalization analysis
Pleiotropic effect analysis

**Figure 1.** Study design. ONCO, OncoArray consortium of lung study; AA-NCI, lung cancer and smoking phenotypes in African American cases and controls; INHALE, inflammation, health, ancestry and lung epidemiology study.

in *VWF* was associated with overall lung cancer at genome-wide significance (allele frequency (AF) = 0.71, Odds Ratio (OR) = 1.30, and a GWMA fixed-effect model *P*-value, $P = 2.67 \times 10^{-9}$). Two other novel susceptibility loci, 2p24.3 and 15q24.1, were specific to SQC, where we detected an intergenic variant rs6431749 between *MYCN* and *GACAT3* at 2p24.3 (AF = 0.73, OR = 0.64, $P = 8.15 \times 10^{-10}$) and rs115735578 in *LMAN1L* at 15q24.1 (AF = 0.01, OR = 4.23, $P = 1.04 \times 10^{-8}$) (Table 1, Supplementary Fig. 1e–h). Importantly, lead variants from all three novel loci showed either the highest minor allele frequencies in African populations (rs216859 and rs6431749) or were rare in African populations but monoallelic in non-African populations (rs115735578) among the 1000 Genomes populations. We also explored the risk variants identified in our study for European and East Asian populations using two large multi-population GWAS [6, 12]. Lung locus in *CHRNA5* at 15 q25.1 tagged by rs17486278 shows the same direction of the allelic effect in European [6] and MVP African populations [12] (Supplementary Table 4).

In addition to identifying the top associations in African-ancestry GWMA of lung cancer, we searched secondary independent association signals at each locus using conditional joint analysis. While we did not observe any independent variants conditional on the lead SNPs identified at the genome-wide significance level of $P < 5 \times 10^{-8}$, we report candidate suggestive associations at the nominal significance level of $P < 10^{-3}$ (Supplementary Table 5).

## SNP-set analysis based on the top SNPs

In light of the identification of low-frequency variants as a new locus or new lead variants from known loci for SQC in African-ancestry GWMA, we further performed SNP-set test to facilitate the detection of an aggregated effect of low- and high-frequency variants at the locus level. For each of the three loci tagged by a low-frequency variant from the GWMA of SQC, we extracted all SNPs within 1 Mb of rs115287843 (near *TRIP13* at 5p15.33), rs113048688 (in *ERC1* at 12p13.33), and rs115735578 (in *LMAN1L* at 15q24.1) and applied the aggregated Cauchy association test (ACAT) [21] to the summary statistics. We observed significant aggregated signals for all three loci tagged by a low-frequency variant for SQC, rs115287843 at 5p15.33 ($P = 3.40 \times 10^{-9}$), rs113048688 at 12p13.33 ($P = 2.16 \times 10^{-8}$), and rs115735578 at 15q24.1 ($P = 8.10 \times 10^{-9}$). ACAT results at the locus level were concordant with the GWMA results at the variant level with slightly lower p-values but did not show significant improvement (Table 2).

## Pleiotropic effects of risk variants associated with lung cancer in African Americans

We conducted a phenome-wide association study (PheWAS) look-up to explore pleiotropic effects of lung cancer risk-associated variants identified in African-ancestry GWMA. 536 of the 3302 unique traits in the GWASATLAS database [22] were associated with the lead variants from eight loci (rs216859, rs17486278, and rs144437384 for Lung; rs17486278 for ADE; rs6431749, rs115287843, rs113048688, and rs115735578 for SQC) at the nominally significant level of 0.05. A total of 11 traits (type 2 diabetes (T2D), body mass index (BMI), estimated glomerular filtration rate (eGFR), HbA1c, hip circumference, waist circumference, right cerebellum white matter, schizophrenia, FEV1, FVC, PEF) were more likely to show a pleiotropic effect on variants associated with lung cancer development in our study (Supplementary Table 6). A PheWAS of rs17486278 showed a strong association with lung function, including FEV1 and FVC.

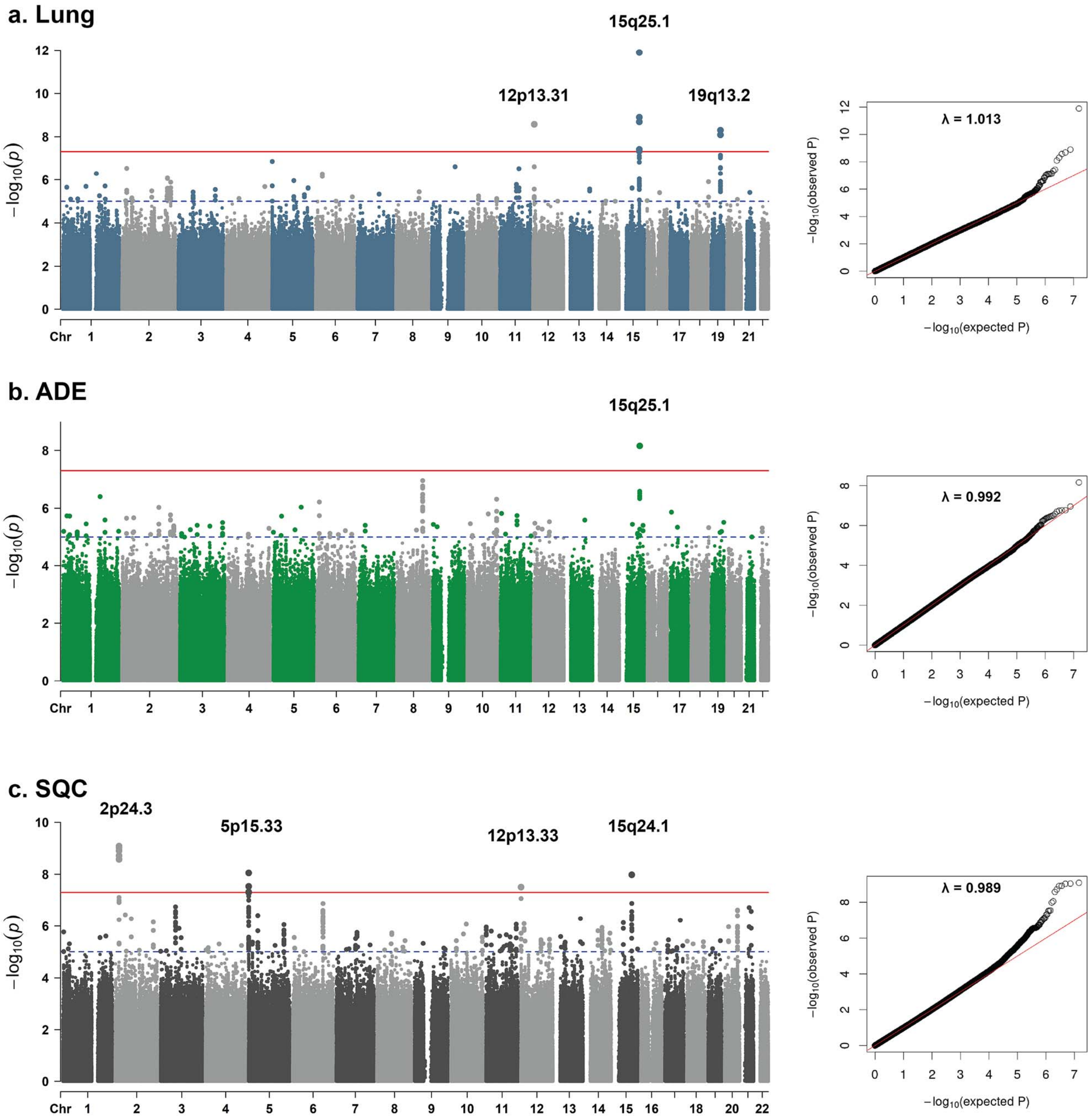

**Figure 2.** Manhattan plots and quantile-quantile plots of the GWMA for lung cancer in the African descent population. (a) lung cancer: 2267 cases and 4264 controls. (b) ADE: 1010 cases and 4264 controls. (c) SQC: 517 cases and 4264 controls. The x-axis represents chromosomal location, and the y-axis represents the $-\log_{10}$($P$ value). The gene annotations for newly identified loci are in black. The top and bottom horizontal lines denote the Bonferroni-corrected genome-wide significant two-sided $P$ value of $P = -\log_{10}(1.25 \times 10^{-8})$ and a suggestive significant $P$ value of $P = -\log_{10}(1.00 \times 10^{-5})$, respectively.

The pleiotropic effects of the rs216859 and rs6431749 revealed nominal association with T2D, rs17486278 and rs6431749 with schizophrenia, and rs17486278, rs113048688, and rs115735578 with BMI. PheWAS findings on SNPs in Table 2 extracted from GWASAATLAS [22] are shown in Supplementary Table 6 and Supplementary Fig. 2.

## Functional annotation of lung cancer-associated variants in African Americans

We evaluated functional relevance of the eight lead variants using the FAVOR platform (v2.0) [14] to identify possible epigenetic and evolutionarily conserved functions. The database contained full information on the lead variants from six loci. A selection of annotation values of these variants is plotted in Fig. 3. Variant epigenomic scoring of representative annotations (e.g. H3K27ac, H3K4me1, and H3K4Me3 chromatin modification peaks), as well as annotation principal component derived from multiple epigenomic features are shown as their relative strength among all the variants in the genome. These scoring highlighted variants with potential transcriptional function, where rs17486278 (*CHRNA5*) at 15q25.1 and rs115735578 (*LMAN1L*) at 15q24.1 showed the highest levels of epigenetic functionality among six variants. Similarly, variant conservation scoring integrating multiple relevant annotations indicated that rs144437384 (*CYP2A6*) at 19q13.2 showed the highest conservation annotation values.

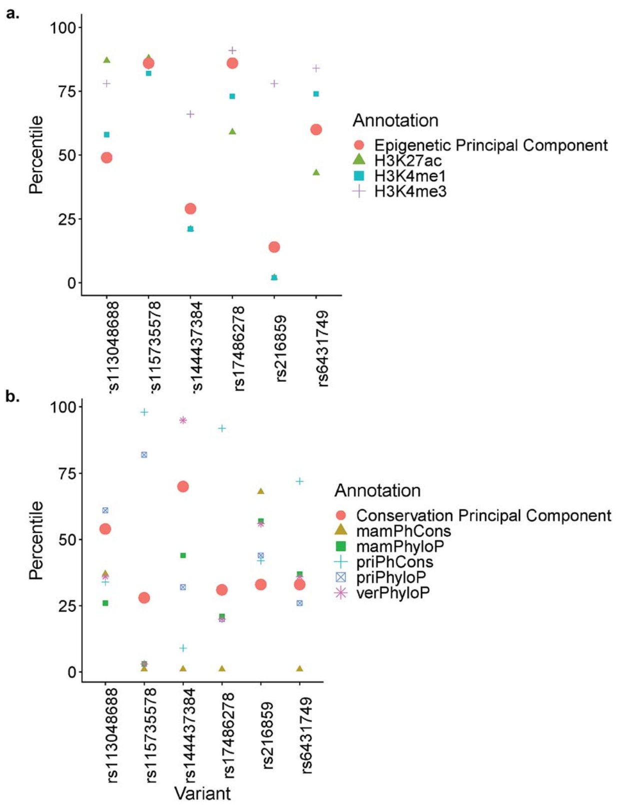

**Figure 3.** Integrative multi-omics functional annotation analyses using the FAVOR platform. (a) plot for epigenetic annotation values. (b) plot for evolutionary conservation annotations. The percentiles indicate the proportion of variants with lower values across the genome; higher values indicate greater functionality. The large filled circle represents each principal component that integrates all epigenetic or conservation annotation values, including some that are not shown.

Next, we estimated the regulatory potential of top hits identified in African-ancestry GWMA using RegulomeDB [17]. RegulomeDB provides a probability ranging from 0 to 1, with 1 being most likely a regulatory variant (Supplementary Table 7). We observed that four SNPs, rs17486278 (*CHRNA5*), rs6431749 (*MYCN*), rs115287843 (*TRIP13*), and rs115735578 (*LMAN1L*) were predicted to be regulatory variants with probability $>0.6$. Among them, rs17486278 in *CHRNA5* at 15q25.1 displayed the highest probability of 0.87. (Supplementary Table 7). The findings of rs17486278 in *CHRNA5* and rs115735578 in *LMAN1L* from FAVOR and RegulomeDB suggested that these two variants are more likely involved in the epigenetic regulation of transcription.

## Candidate gene prioritization from African-ancestry GWMA

To identify susceptibility genes from the eight risk loci (12p13.31, 15q25.1, 19q13.2 for Lung; 15q25.1 for ADE; 2p24.3, 5p15.33, 12p13.33, 15q24.1 for SQC), we used FUMA GWAS functional annotation tool [13]. First, we defined independent significant variants from each of these loci as reported in Supplementary Table 8 utilizing FUMA GWAS [13]. With this mapping, FUMA GWAS prioritized 146 unique genes based on position, variant functional score, eQTL, and chromatin interactions (Supplementary Table 9). Based on the position mapping of deleterious coding variants, two genes (*AC010145.4* on chromosome 2 and *TRIP13* on chromosome 5 for SQC) displayed the highest combined annotation-dependent depletion (CADD) score (posMapMaxCADD = 14.59 and 16.45, respectively; a CADD score $\geq 12.37$ is considered deleterious) [23]. Based on the chromatin interaction mapping, 124 genes were nominated of which one gene (*FBXL14* on chromosome 12 in SQC) was based on the chromatin interaction in the lung tissue (Supplementary Table 9). Multiple connections for potential target genes based on chromatin interactions or eQTLs within lung cancer risk loci were shown for chromosomes 2, 5, 12, 15, and 19, respectively (Supplementary Fig. 3). Among them, eight genes were mapped by both eQTLs and chromatin interactions including *RP11-650 L12.2*, *CHRNA3*, and *CHRNB4* on 15q25.1 for ADE (Supplementary Fig. 3d) and *ZDHHC11*, *BRD9*, *TRIP13*, *RP11-661C8.2*, and *RP11-43F13.3* (Supplementary Fig. 3f, Supplementary Table 9) on the 5p15.33 for SQC.

## Colocalization of GWMA and eQTL signals at lung cancer risk loci in African Americans

To map susceptibility genes from the lung cancer risk loci in African Americans, we further performed eQTL colocalization analysis. We surveyed an overlap between GWMA variants within ±100 kb windows of each lead variant (Table 2) and significant eQTL variants in multiple tissue types from the GTEx v8. We also included 26 unique eQTL variants that overlapped with African-ancestry GWMA variants using FUMA GWAS (Supplementary Table 10). To prioritize candidate susceptibility genes, we performed colocalization analyses using eQTL summary statistics of multiple tissues from the GTEx v8. A total of six unique candidate genes were identified as potential susceptibility genes (Supplementary Table 11). Based on histological subtypes, eQTL-based colocalization has identified three and one candidate genes for ADE and SQC, respectively. Notably, previously implicated candidate susceptibility genes, *PSMA4* and *CHRNA3*, on chromosome 15 displayed the higher probability scores ($PPH4 > 0.95$) from coloc analysis in brain tissues (hypothalamus, cortex, and cerebellum) that are potentially associated with smoking behavior as well as esophageal muscular tissue. We did not observe any candidate SNPs colocalizing in lung tissue with $PPH4 > 0.65$.

Finally, we explored the protein–protein interaction (PPI) network associated with each nearest gene identified from eight African-ancestry GWMA loci. We observed that four proteins (VWF and CYP2A6 for Lung and TRIP13 and ERC1 for SQC) were highly enriched for PPI at $FDR < 0.05$. Each prioritized gene of *VWF*, *CYP2A6*, *MYCN*, *TRIP13*, *ERC1* reported a PPI P-value of $2.03 \times 10^{-12}$, $1.00 \times 10^{-16}$, 0.0498, $2.18 \times 10^{-5}$, and 0.0166, respectively (Supplementary Table 12). Visualization of the PPI network for each gene using interaction data based on ProteomicsDB highlighted pathways such as PI3K-Akt signaling pathway, platelet activation, focal adhesion kinase, neuroactive ligand-receptor interaction, various metabolism-associated pathways, NF-kappa B signaling pathway, and transcriptional misregulation in cancer (Supplementary Fig. 4). Subsequently, we scrutinized the PPI network associated with 146 unique genes prioritized from FUMA GWAS. Among them, 65 proteins remained for PPI network analysis, with PPI P-value of $1.00 \times 10^{-16}$. The PPI network for 32 proteins after hiding disconnected nodes in the network is displayed in Supplementary Fig. 5, which highlights various metabolic processes, response to nicotine, chemical carcinogenesis, pathway in cancer, and epoxygenase P450 pathway (Supplementary Table 13).

## Evaluation of multi-ancestry polygenic risk variants of lung cancer in AA

We evaluate the robustness of the multi-ancestry polygenic risk scores (PRS) in AA by utilizing two recent multi-ancestry GWAS summary statistics for lung cancer [6, 12]. The summary for PRS

**Table 3.** Odds ratios (ORs) and 95% confidence intervals (CIs) for the association between the polygenic risk scores and risks of lung cancer in AA.

| | Adjusted for age and sex | | Adjusted for age, sex, and smoking status | |
|---|---|---|---|---|
| | OR(CI) | P value | OR(CI) | P value |
| Continuous | 1.24 (1.17–1.31) | 5.24E-15 | 1.23 (1.16–1.30) | 5.81E-13 |
| Decile1 | 1.00 (reference) | | 1.00 (reference) | |
| Decile2 | 1.16 (0.90–1.50) | 2.52E-01 | 1.22 (0.93–1.60) | 1.44E-01 |
| Decile3 | 1.07 (0.82–1.38) | 6.20E-01 | 1.11 (0.85–1.45) | 4.48E-01 |
| Decile4 | 1.14 (0.88–1.47) | 3.29E-01 | 1.17 (0.90–1.53) | 2.45E-01 |
| Decile5 | 1.32 (1.02–1.70) | 3.37E-02 | 1.34 (1.03–1.74) | 3.14E-02 |
| Decile6 | 1.45 (1.13–1.87) | 3.67E-03 | 1.51 (1.16–1.97) | 2.33E-03 |
| Decile7 | 1.42 (1.10–1.83) | 6.86E-03 | 1.40 (1.08–1.83) | 1.21E-02 |
| Decile8 | 1.42 (1.10–1.83) | 6.89E-03 | 1.40 (1.08–1.83) | 1.21E-02 |
| Decile9 | 1.82 (1.42–2.33) | 2.84E-06 | 1.80 (1.39–2.34) | 9.94E-06 |
| Decile10 | 2.21 (1.73–2.84) | 3.92E-10 | 2.27 (1.75–2.96) | 7.63E-10 |

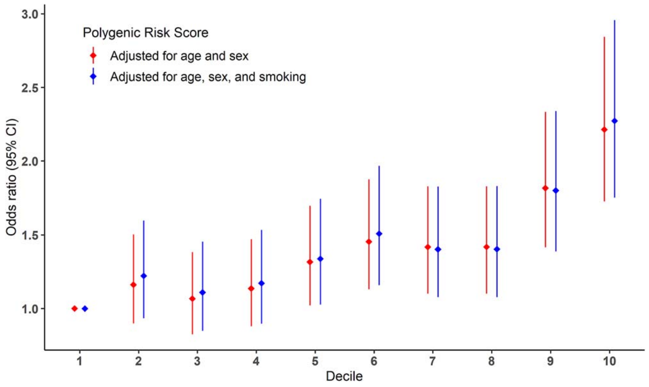

**Figure 4.** Analysis of odd ratios (ORs) for AA-specific lung cancer by polygenic risk score (PRS) decile. ORs are shown for two models: one adjusted for age and sex, and a second adjusted for age, sex, and smoking. Within each decile group, the ORs for the age- and sex-adjusted model are shown on the left, and those for the model additionally adjusted for smoking are shown on the right. Data are presented as ORs with 95% confidence intervals (CIs).

construction is shown in Supplementary Fig. 6. An increase in the PRS decile is associated with statistically significantly higher risks of lung cancer in AA, adjusted for demographic (age and sex) and lifestyle (smoking status) factors (Table 3, Fig. 4). A risk gradient is also observed across the decile of the PRS, such that individuals in the highest decile of the PRS show an over 2-fold higher risk of lung cancer in AA (OR = 2.21 adjusted for age and sex; 2.27 adjusted for age, sex, and smoking status) compared to those in the lowest decile of the PRS (Table 3). The model that only considers age and sex has a low area under the curve (AUC) of 0.581, with a 95% confidence interval (CI) of (0.566–0.600). In contrast, the combined model, which includes age, sex, smoking status, and PRS, shows moderate predictive ability with an AUC of 0.691 (95% CI: 0.678–0.705). This combined model is more effective in distinguishing between lung cancer cases and controls in African Americans (AA), as presented in Table 4 and Fig. 5.

## Discussion

We conducted an African-ancestry GWMA of lung cancer involving 6531 individuals of African ancestry. While most previously published findings mainly focused on European-ancestry studies [6], our GWMA of African-ancestry has identified eight loci associated with lung cancer at the genome-wide association level of $P < 5 \times 10^{-8}$.

**Table 4.** Performance of models for distinguishing lung cancer cases from controls.

| Model | AUC (95% CI) |
|---|---|
| Age + Sex | 0.581 (0.566–0.600) |
| Age + Sex + PRS | 0.605 (0.590–0.620) |
| Age + Sex + Smoking | 0.671 (0.657–0.685) |
| Age + Sex + Smoking + PRS | 0.691 (0.678–0.705) |

PRS, polygenic risk scores; AUC, the area under the receiver operating characteristic curve; CI, confidence interval.

Lung tumorigenesis is a complex process, including acquiring genetic mutations and epigenetic alterations in cellular processes. Lung cancer GWAS have shown shared and distinct genetic architectures across different populations [6, 24]. Since most GWAS signals have shown small effect sizes and can be affected by confounding, identifying new genetic association signals remains challenging. To date, a limited number of risk variants have been identified in African-ancestry populations. Elucidating the genomic architecture of lung cancer risk in African Americans is critical to better understanding lung cancer development in this population. In addition, the shared genetic variants underlying lung cancer predisposition in African Americans can help refine risk prediction profiling for individuals at high risk in other African-ancestry admixed populations [6, 25]. Our investigation of lung cancer and specific histological subtypes in African Americans provided several key findings.

We have confirmed two African-ancestry susceptibility loci for Lung, one for ADE, and two for SQC. These associations are new independent genome-wide significant risk variants in previously reported loci. An intronic variant associated with smoking behaviors, rs144437384, in *CYP2A6* at 19q13.2 is more frequent in African Americans (GNOMAD AF = 0.05) compared to other populations (GNOMAD AF = 0 for both Europeans and East Asians). These findings highlight the utility of African-ancestry GWMA in conducting further downstream analyses to examine the biological and functional mechanisms underlying ancestry-specific lung cancer.

In this study, we have surveyed the genetic effects of common and low-frequency (rare) variants on lung cancer development in African descent populations. A common intronic variant, rs17486278, located in *CHRNA5*, showed a strong association with

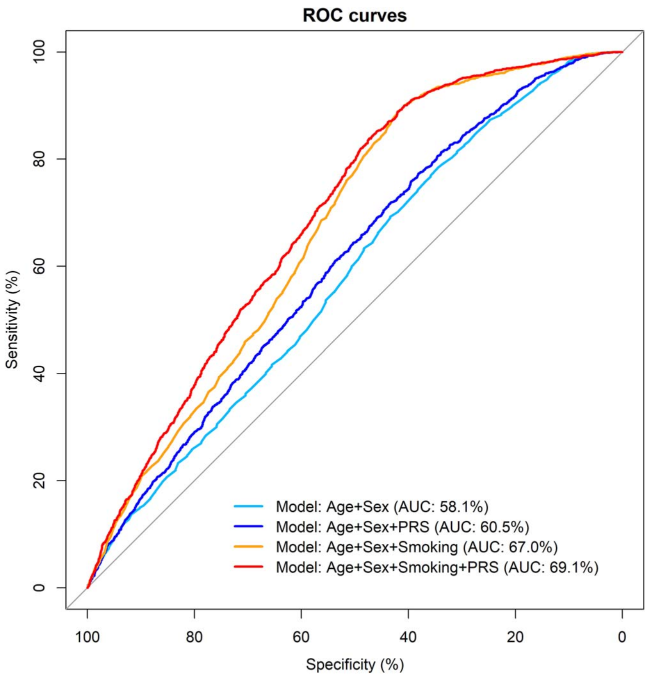

**Figure 5.** Prediction accuracy of lung cancer models in African American individuals, shown using ROC curves. The models, ordered from highest to lowest performance (top to bottom curves), are: (1) PRS combined with demographic and lifestyle covariates (age, sex, smoking, and PRS), (2) demographic and lifestyle covariates only (age, sex, and smoking), (3) PRS combined with demographic covariates (age, sex, and PRS), and (4) demographic covariates only (age and sex). A diagonal line representing 50% prediction accuracy (random chance) is included for reference. Each model is represented by a distinct color, as indicated in the legend. The curves illustrate the comparative predictive performance of the models.

lung cancer and ADE in a single population of African descent and has been reported in multi-population studies of Europeans and East Asians for lung cancer [26] and Europeans and African Americans for lung function-related traits [27]. Although the sample size in SQC GWAS was small, we identified four new association signals in African Americans (two from novel and two from previously reported loci). Since these low-frequency variants are more common in African Americans, we were able to identify them in our study populations despite smaller sample sizes compared to other populations. Table 2 and Fig. 2c show that we observed one common and three low-frequency variants. A common intergenic variant, rs6431749, near *MYCN* and *GACAT3*, displayed strong evidence of pleiotropic effects on various traits relevant to cardiovascular, endocrine, immunological, metabolic, neoplastic, psychiatric, and neurological domains. A low-frequency intergenic variant, rs115287843 near *TRIP13* and *LOC100506688*, was reported to affect the metabolically-related traits, eGFR and elevated fasting glucose level. Two low-frequency intronic variants, rs113048688 in *ERC1* and rs115735578 in *LMAN1L*, were involved in the genetic contribution to the immunological domain traits, platelets and HbA1c, and BMI.

We have identified and prioritized lung cancer risk genes in African Americans. Our findings also predicted biological pathways that are associated with these genes, and we performed integrative trans-omics functional annotation analyses. Through PPI network analyses, we found candidate genes that were frequently altered in various biological pathways, including acetylcholine receptor activity, immune response, cellular response to acetylcholine, regulation of cellular response to stress, regulation of DNA repair, and metabolic process. Furthermore, integrative multi-omics functional analyses using FAVOR and RegulomeDB suggested that two African-ancestry risk-associated variants, rs17486278 (ADE) and rs115735578 (SQC) at 15q25.1 and 15q24.1, are more likely involved in regulatory mechanisms.

Finally, we assessed the performance of PRS based on the multi-ancestry genome-wide significant SNPs for individuals in our study population. We observed that individuals at the top decile of the PRS are at a statistically significantly higher risk of lung cancer than those in the lowest decile. The PRS results demonstrate that the risk prediction model for lung cancer combining non-genetic and genetic data may have limited clinical utility for discriminating lung cancer cases from controls in AA using SNPs derived from multi-population GWAS. The risk assessment model, including age, sex, smoking status, and PRS, shows better discriminatory ability and provides an incremental improvement in the AUC. This implies that genetic predictors in combination with information on demographic and lifestyle factors, may improve risk stratification efforts for lung cancer.

A limitation of this study is that we utilized PheWAS data obtained from multiple populations, not only the African-ancestry population. However, we believe the PheWAS survey on lung cancer can provide valuable insights into the pleiotropic effects on the disease risk. The interpretation of our eQTL colocalization analysis is also limited as the GTEx eQTL dataset is predominantly of European populations. Given the low allele frequencies of some of the loci specific to this population, a larger eQTL dataset using lung tissues from African ancestry populations might be needed.

In conclusion, we have performed African-ancestry genome-wide association analyses and various multi-omics functional annotation analyses for lung cancer and specific histological subtypes. We have identified eight genome-wide significant variants and prioritized potential causal variants that may influence lung cancer development in African Americans. In addition, we have highlighted several *in silico* functional approaches to map and prioritize the variants identified. We also utilized integrative functional annotation platforms to functionally characterize the prioritized genes, including coding and non-coding genes, which provide numerous functional information on variants. Our GWMA of African Americans has helped elucidate the etiology and biological mechanisms of lung cancer susceptibility. Understanding the African Americans-specific genetic architecture of lung cancer predisposition will help reveal how lung cancer develops in African descent populations and could assist in identifying new susceptibility biomarkers for better risk evaluation directed at early detection and diagnosis, targeted therapy, and improved preventive measures for African Americans and other admixed populations.

## Materials and methods

### Genome-wide association studies of African ancestry populations

We included three African-ancestry GWAS [6] of lung cancer, including (i) the OncoArray Consortium of Lung Study (ONCO) with 269 cases and 286 controls (mean age at diagnosis for cases = 65.5, mean age at enrollment among controls = 67.1) [28], (ii) Lung Cancer and Smoking Phenotypes in African American Cases and Controls (AA-NCI) with 1704 cases and 3460 controls (mean age at diagnosis for cases = 63.6, mean age at enrollment among controls = 58.8) [10, 29], and (iii) INflammation, Health, Ancestry and Lung Epidemiology study (INHALE) with 294 cases and 518 controls (mean age at diagnosis for cases = 62.9, mean age at enrollment among controls = 59.6) [30, 31] (Table 1). In brief, genotyping for each study was performed using the Infinium OncoArray-500 K for the ONCO GWAS [28], the Illumina HumanHap 1 M-Duo chip for the NCI-AA GWAS [10], and the Illumina Multi-Ethnic GWAS/Exome Array for INHALE GWAS [30], respectively. Full details for these GWAS have been reported elsewhere [6, 10, 28, 30]. A total of 1010 cases of lung adenocarcinoma (ADE) and 517 cases of lung squamous cell carcinoma (SQC) were defined based on available histological information [6].

## Quality control and imputation

We implemented FastPop [32] to infer genetic similarity of 6531 individuals with 505 samples comprising known ancestries from CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), CHB (Han Chinese in Beijing, China), and YRI (Yoruba in Ibadan, Nigeria) from HapMap 3 using a total of 2042 ancestry informative markers. All samples were imputed using 32 470 samples from the Haplotype Reference Consortium (HRC r1.1) as a reference panel through the Sanger imputation service. Further details regarding the imputation methods and quality control steps can be found in our previous publication [6].

## Statistical analysis for African-ancestry GWAS and GWMA

For this study, we followed the same design as we previously described [6]. Genome-wide analyses were performed for the merged ONCO and AA-NCI data in Study 1. We subsequently received genome-wide association data from the INHALE study and analyzed it as Study 2. We performed an association analysis on the combined data of the ONCO and AA-NCI samples (Study1), and on INHALE data (Study2), adjusting for age, sex, study, and the first four principal components (PCs) using SNPTEST with option EM [33]. Principal component analysis using PLINK (v1.9 option –pca) was run to calculate the principal components as covariates in the additive genetic model. Instead of a two-stage discovery-replication approach, we opted for a meta-analysis on the available samples, which has been shown to be generally more powerful [19, 34]. To ensure optimal statistical power, we used fixed-effects GWMA with inverse-variance weighting using METASOFT [35] and further filtered with P-value for Cochrane's Q statistics ≥ 0.01.

## Conditional and joint analysis on lead SNPs identified in GWMA

Conditional and joint analysis has been used to search secondary genome-wide significant association signals at a particular locus, involving association analysis conditioning on the sentinel GWMA SNP for each locus within a particular genomic region, followed by a stepwise procedure of selecting additional SNPs, one by one, according to their conditional P values (P). Such a strategy can enable us to discover more than two independently associated SNPs at a locus. We adopted a genome-wide stepwise selection procedure to detect SNPs based on conditional P values using GCTA v1.94 (−cojo-cond) [36]. Conditional analysis of each associated locus was performed within a standard region of a 1 Mb window centered on the lead SNP, the most strongly associated SNP in lung cancer. LD patterns were estimated using best-guess genotype data in 5164 African participants from AA-NCI data as reference (6). Conditional association analysis was performed, including the lead SNP as a covariate. Any SNP showing a conditional association with $P < 5 \times 10^{-8}$ was considered an independent signal and was repeated until no SNP with $P < 10^{-3}$ remained in any of the genomic regions explored.

## Characterization of genomic susceptibility loci prioritized by functional mapping and annotation

We utilized the Functional Mapping and Annotation of GWAS platform (FUMA GWAS) [13] (v1.4.1; https://fuma.ctglab.nl/) to prioritize risk variants associated with lung cancer in African-ancestry populations. To identify genomic risk loci, we used the SNP2GENE function with default thresholds. In SNP2GENE, independent significant SNPs were defined as those with a $P \leq 5 \times 10^{-8}$ and with independence at r2 < 0.6 in the GWMA results. Lead SNPs were determined based on a pairwise r2 < 0.1 among these significantly independent SNPs. Genomic risk loci were identified by merging the LD blocks (r2 > 0.6) of independent significant SNPs that are close to each other within 250 kb. These lead SNPs are crucial in capturing the genetic association signal specific to each locus. The genetic data of the African population in 1000G phased 3 was used as a reference to estimate LD. Finally, susceptibility variants from GWMA were prioritized and mapped by functional annotation such as positional, expression quantitative trait loci (eQTL), and chromatin interaction mappings using the SNP2GENE function with default settings.

## Colocalization between GWMA and GTEx eQTL association signals

The Genotype-Tissue Expression (GTEx v8) database includes data from 49 normal tissues from 838 donors. GTEx eQTL association data for variants within ±100 kb windows of the lead variants presented in the African-ancestry GWMA were extracted. Colocalization between the five African-ancestry GWMA associations within the newly identified loci and eQTL signals were calculated using the coloc package (v5.1.0; https://cran.rproject.org/web/packages/coloc/) [37]. To account for the heterogeneous LD in our African-ancestry GWMA and eQTL population in the GTEx v8 and avoid potentially spurious colocalizations due to the violation of common LD assumption in genome-wide associations, eQTL signals, and ancestry-specific LD matrix, we applied the LD-independent approach using coloc package [37].

## Survey of prioritized variants in phenome-wide association database

We examined the pleiotropic effects of SNPs identified in lung cancer GWMA in African Americans on various complex human traits. We surveyed a phenome-wide association study (PheWAS) database of GWASATLAS (https://atlas.ctglab.nl/) consisting of 4756 GWAS summary statistics from 473 unique studies across 3302 unique traits and 28 domains [22]. We extracted the pleiotropic traits with the minimum univariate P-value of 0.05 across all analyzed traits available in GWASATLAS.

## Integrative multi-omics annotation analysis and annotation-informed function prediction

To functionally characterize the top lead variants associated with lung cancer susceptibility in African-ancestry populations, we utilized the Functional Annotation of Variants—Online Resource (FAVOR v2.0) platform [14] (https://favor.genohub.org/), which includes integrative multi-omics functional annotation information for all possible 8 812 917 339 single nucleotide variants (SNVs) across the human genome and 79 997 898 observed insertions and deletions (INDELs) from the Trans-Omics for Precision Medicine (TOPMed) BRAVO variant set (Genome Reference Consortium

Human Build 38). We examined plausible functional roles for top SNPs by integrating genome-wide annotation values for each SNP, with the goal of identifying variants that possess epigenetic function or evolutionary conserved function. Example of annotations include maximum H3K27Ac values across multiple tissues as well as the phastCons conservation score. Percentiles for each attribute are obtained through comparison against all variants in the genome [38].

We further utilized RegulomeDB, which provides functional context to genetic variants, prioritizes functionally important variants within the non-coding regions of the human genome, and provides a prediction score that is interpreted as the probability of the variant of interest being of real functional significance [17, 24].

We then exploited web-based functional annotation databases to provide insights into the biological and molecular mechanisms underlying lung cancer in African descent populations. We utilized ProteomicsDB (https://www.proteomicsdb.org/protein), a multi-omics resource that includes proteomics, transcriptomics, and phenomics data for multi-organisms and facilitates protein–protein interaction (PPI) and protein-drug interaction analyses [15]. We also analyzed gene-based functional enrichment using the STRING database (STRING v12.0; https://string-db.org/) [16]. We selected the setting options with the maximum number of interactions as 20 and the highest confidence score of 0.9 to examine the functional enrichment of numerous pathways by the gene.

## Set-based joint association analysis on low-frequency and rare variants

With improved imputation resources, existing GWAS enable the discovery of low-frequency and rare genetic variations, making substantial contributions to the study of missing heritability and to new genetic discoveries for complex human traits and diseases [21]. Specially, it has been important to apply methods that can increase statistical power for detecting associations between low-frequency or rare variants and complex traits [21, 39]. We implemented the aggregated Cauchy association test (ACAT) [21] to perform a powerful set-based association test that complements single variant analysis in GWAS. The ACAT utilizes all summary statistics of variants in a region to test whether the entire region is associated with disease. The ACAT approach is useful for aggregating multiple low-frequency or rare variants that may not reach statistical significance by themselves.

## Polygenic risk score

To evaluate the robustness of multi-ancestry PRS associated with lung cancer susceptibility, we retrieved 45 and 86 SNPs at $P \leq 5 \times 10^{-8}$ identified by Byun et al. and Gorman et al., who recently published the two large multi-ancestry lung cancer GWAS [6, 12]. We constructed the PRS for cases and controls by summing the risk allele counts as additive genotype components (0,1,2) for 55 variants (Supplementary Table 14) after properly harmonizing two studies weighted by their effect sizes (ORs) extracted from the corresponding GWAS (Supplementary Fig. 6). For each individual in AA, we summed the weighted risk allele counts (PLINK option –score). We then standardized PRS by mean and standard deviation of control samples.

# Acknowledgements

We want to acknowledge the participants and investigators of the INTEGRAL-ILCCO Consortium.

# Supplementary data

Supplementary data is available at *HMG Journal* online.

*Conflict of interest statement:* The authors declare no competing interest.

# Funding

Our study was supported by the National Institutes of Health (NIH) for U19CA203654 and R01CA243483 (CIA); R03CA277197 (JB); R01CA141769 (AGS); R03CA282953 (YLiu); U01CA253560 (MCA); U01CA063673, UM1CA167462, U01CA167462, and N01PC35142 (MN). CIA is a Research Scholar of the Cancer Prevention Research Interest of Texas (CPRIT) award (RR170048).

# Data availability

NCI study of African Americans and imputed OncoArray Consortium of Lung Study using HRC reference panel in this study are publicly available in dbGaP at phs001210.v1.p1 and phs001273.v4.p2, respectively. INflammation, Health, Ancestry, and Lung Epidemiology study are available upon request from Ann G. Schwartz.

# Ethics approval and consent to participate

All participants provided informed consents according to protocols that were evaluated by the local Ethics Committee/Institutional Review Boards of the contributing study centers. All methods were performed in accordance with the ethical guidelines of the 1975 Declaration of Helsinki. All contents in the present study were approved by Baylor College of Medicine internal review boards.

# References

1. U.S. Cancer Statistics Working Group, U.S. Cancer Statistics Data Visualizations Tool, based on 2021 submission data. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute. In: 1999-2019, www.cdc.gov/cancer/dataviz, released in June 2022., in press.
2. Long E, Patel H, Byun J. *et al.* Functional studies of lung cancer GWAS beyond association. *Hum Mol Genet*in press., ddac140 2022;**31**:R22–R36.
3. Dai J, Huang M, Amos CI. *et al.* Genome-wide association study of INDELs identified four novel susceptibility loci associated with lung cancer risk. *Int J Cancer* 2020;**146**:2855–2864.
4. Cheng Y, Jiang T, Zhu M. *et al.* Risk assessment models for genetic risk predictors of lung cancer using two-stage replication for Asian and European populations. *Oncotarget* 2017;**8**: 53959–53967.
5. Blechter B, Wong JYY, Agnes Hsiung C. *et al.* Sub-multiplicative interaction between polygenic risk score and household coal use in relation to lung adenocarcinoma among never-smoking women in Asia. *Environ Int* 2021;**147**:105975.
6. Byun J, Han Y, Li Y. *et al.* Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. *Nat Genet* 2022;**54**: 1167–1177.
7. Dai J, Lv J, Zhu M. *et al.* Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. *Lancet Respir Med* 2019;**7**:881–891.

8. Bosse Y, Amos CI. A decade of GWAS results in lung cancer. *Cancer Epidemiol Biomarkers Prev* 2018;**27**:363–379.
9. Haiman CA, Stram DO, Wilkens LR. *et al.* Ethnic and racial differences in the smoking-related risk of lung cancer. *N Engl J Med* 2006;**354**:333–342.
10. Zanetti KA, Wang Z, Aldrich M. *et al.* Genome-wide association study confirms lung cancer susceptibility loci on chromosomes 5p15 and 15q25 in an African-American population. *Lung Cancer* 2016;**98**:33–42.
11. Chen LS, Saccone NL, Culverhouse RC. *et al.* Smoking and genetic risk variation across populations of European, Asian, and African American ancestry–a meta-analysis of chromosome 15q25. *Genet Epidemiol* 2012;**36**:340–351.
12. Gorman BR, Ji SG, Francis M. *et al.* Multi-ancestry GWAS meta-analyses of lung cancer reveal susceptibility loci and elucidate smoking-independent genetic risk. *Nat Commun* 2024;**15**:8629.
13. Watanabe K, Taskesen E, van Bochoven A. *et al.* Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017;**8**:1826.
14. Zhou H, Arapoglou T, Li X. *et al.* FAVOR: functional annotation of variants online resource and annotator for variation across the human genomebioRxiv. 2022; in press., 2022.2008.2028.505582.
15. Lautenbacher L, Samaras P, Muller J. *et al.* ProteomicsDB: toward a FAIR open-source resource for life-science research. *Nucleic Acids Res* 2022;**50**:D1541–D1552.
16. Szklarczyk D, Kirsch R, Koutrouli M. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 2023;**51**:D638–D646.
17. Boyle AP, Hong EL, Hariharan M. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012;**22**:1790–1797.
18. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* 2011;**88**:586–598.
19. Skol AD, Scott LJ, Abecasis GR. *et al.* Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006;**38**:209–213.
20. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 2015;**31**:3555–3557.
21. Liu Y, Chen S, Li Z. *et al.* ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am J Hum Genet* 2019;**104**:410–421.
22. Watanabe K, Stringer S, Frei O. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* 2019;**51**:1339–1348.
23. Kircher M, Witten DM, Jain P. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–315.
24. Li Y, Xiao X, Li J. *et al.* Lung cancer in ever- and never-smokers: findings from multi-population GWAS studies. *Cancer Epidemiol Biomarkers Prev* 2024;**33**:389–399.
25. Chen T, Pham G, Fox L. *et al.* Genomic insights for personalised care in lung cancer and smoking cessation: motivating at-risk individuals toward evidence-based health practices. *EBioMedicine* 2024;**110**:105441.
26. Sakaue S, Kanai M, Tanigawa Y. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* 2021;**53**:1415–1424.
27. Lutz SM, Cho MH, Young K. *et al.* A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet* 2015;**16**:138.
28. Amos CI, Dennis J, Wang Z. *et al.* The OncoArray consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiol Biomarkers Prev* 2017;**26**:126–135.
29. Mitchell KA, Shah E, Bowman ED. *et al.* Relationship between west African ancestry with lung cancer risk and survival in African Americans. *Cancer Causes Control* 2019;**30**:1259–1268.
30. Watza D, Lusk CM, Dyson G. *et al.* COPD-dependent effects of genetic variation in key inflammation pathway genes on lung cancer risk. *Int J Cancer* 2020;**147**:747–756.
31. Schwartz AG, Lusk CM, Wenzlaff AS. *et al.* Risk of lung cancer associated with COPD phenotype based on quantitative image analysis. *Cancer Epidemiol Biomarkers Prev* 2016;**25**:1341–1347.
32. Li Y, Byun J, Cai G. *et al.* FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. *BMC Bioinf* 2016;**17**:122.
33. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;**11**:499–511.
34. Nagel M, Jansen PR, Stringer S. *et al.* Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat Genet* 2018;**50**:920–927.
35. Han B, Eskin E. Interpreting meta-analyses of genome-wide association studies. *PLoS Genet* 2012;**8**:e1002555.
36. Yang J, Lee SH, Goddard ME. *et al.* GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**:76–82.
37. Wallace C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet* 2020;**16**:e1008720.
38. Hassan MM, Li D, Han Y. *et al.* Genome-wide association study identifies high-impact susceptibility loci for hepatocellular carcinoma in North America. *Hepatologyin press* 2024;**80**:87–101.
39. Butler-Laporte G, Povysil G, Kosmicki JA. *et al.* Exome-wide association study to identify rare variants influencing COVID-19 outcomes: results from the host genetics initiative. *PLoS Genet* 2022;**18**:e1010367.

*Human Molecular Genetics*, 2025, 1–11
https://doi.org/10.1093/hmg/ddaf059
Original Article