Title: '**Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer**'

Author list:
Jinyoung Byun[1,2†],
Younghun Han[1,2†],
Yafang Li[1-3†],
Jun Xia[1,4†],
Erping Long[5†],
Jiyeon Choi[5],
Xiangjun Xiao[1],
Meng Zhu[6],
Wen Zhou[1],
Ryan Sun[7],
Yohan Bossé[8],
Zhuoyi Song[1,4],
Ann Schwartz[9,10],
Christine Lusk[9,10],
Thorunn Rafnar[11],
Kari Stefansson[11],
Tongwu Zhang[5],
Wei Zhao[5],
Rowland W Pettit[1],
Yanhong Liu[2,3],
Xihao Li[12],
Hufeng Zhou[12],
Kyle M. Walsh[13],
Ivan Gorlov[1-3],
Olga Gorlova[1-3],
Dakai Zhu[1,2],
Susan M Rosenberg[3,4],
Susan Pinney[14],
Joan E. Bailey-Wilson[15],
Diptasri Mandal[16],
Mariza de Andrade[17],
Colette Gaba[18],
James C. Willey[18],
Ming You[19],
Marshall Anderson[14],
John K. Wiencke[20],
Demetrius Albanes[5],
Stephan Lam[21],
Adonina Tardon[22],
Chu Chen[23],
Gary Goodman[24],
Stig Bojeson[25,26],
Hermann Brenner[27-29],
Maria Teresa Landi[5],
Stephen J. Chanock[5],

Mattias Johansson[30],
Thomas Muley[31,32],
Angela Risch[31-34],
H.-Erich Wichmann[35],
Heike Bickeböller[36],
David C. Christiani[37],
Gad Rennert[38],
Susanne Arnold[39],
John K. Field[40],
Sanjay Shete[7,41],
Loic Le Marchand[42],
Olle Melander[43],
Hans Brunnstrom[43],
Geoffrey Liu[44],
Angeline S. Andrew[45],
Lambertus A. Kiemeney[46],
Hongbing Shen[47],
Shanbeh Zienolddiny[48],
Kjell Grankvist[49],
Mikael Johansson[50],
Neil Caporaso[5],
Angela Cox[51],
Yun-Chul Hong[52],
Jian-Min Yuan[53],
Philip Lazarus[54],
Matthew B. Schabath[55],
Melinda C. Aldrich[56],
Alpa Patel[57],
Qing Lan[5],
Nathaniel Rothman[5],
Fiona Taylor[51],
Linda Kachuri[58],
John S. Witte[59],
Lori C. Sakoda[60],
Margaret Spitz[2],
Paul Brennan[30],
Xihong Lin[12],
James McKay[30],
Rayjean J. Hung[61,62],
Christopher I. Amos[1-3*]

Affiliations:
1. Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX.
2. Section of Epidemiology and Population Sciences, Department of Medicine, Baylor College of Medicine, Houston, TX.
3. Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX.
4. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.
5. Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD.
6. Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, P.R. China.

7. Department of Biostatistics, University of Texas, M.D. Anderson Cancer Center, Houston, TX.
8. Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval, Department of Molecular Medicine, Laval University, Quebec City, Canada.
9. Department of Oncology, Wayne State University School of Medicine, Detroit, MI.
10. Karmanos Cancer Institute, Detroit, MI.
11. deCODE genetics/Amgen Sturlugata 8, 101, Reykjavik, Iceland.
12. Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA.
13. Duke Cancer Institute, Duke University Medical Center, Durham, NC.
14. University of Cincinnati College of Medicine, Cincinnati, OH.
15. National Human Genome Research Institute, NIH, Baltimore, MD.
16. Louisiana State University Health Sciences Center, New Orleans, LA.
17. Mayo Clinic, College of Medicine, Rochester, MN.
18. The University of Toledo College of Medicine and Life Sciences, University of Toledo, Toledo, OH.
19. Center for Cancer Prevention, Houston Methodist Research Institute, Houston, TX.
20. Department of Neurological Surgery, The University of California, San Francisco, San Francisco, CA.
21. Department of Integrative Oncology, BC Cancer, Vancouver, British Columbia, Canada.
22. Public Health Department, University of Oviedo, ISPA and CIBERESP, Asturias, Spain.
23. Program in Epidemiology, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA.
24. Swedish Cancer Institute, Seattle, WA.
25. Department of Clinical Biochemistry, Herlev Gentofte Hospital, Copenhagen University Hospital, Denmark.
26. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
27. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany.
28. Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany.
29. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany.
30. Section of Genetics, International Agency for Research on Cancer, World Health Organization, Lyon, France.
31. Division of Cancer Epigenomics, DKFZ – German Cancer Research Center, Heidelberg, Germany.
32. Translational Lung Research Center Heidelberg (TLRC-H), German Center for Lung Research (DZL), Heidelberg, Germany.
33. Department of Biosciences and Medical Biology, Allergy-Cancer-BioNano Research Centre, University of Salzburg, Austria.
34. Cancer Cluster Salzburg, Salzburg, Austria.
35. Institute of Epidemiology, Helmholtz Center, München, Germany.
36. Department of Genetic Epidemiology, University Medical Center, Georg-August-University Göttingen, Germany.
37. Department of Epidemiology, Harvard T.H.Chan School of Public Health, Boston, MA.
38. Clalit National Cancer Control Center at Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel.
39. University of Kentucky, Markey Cancer Center, Lexington, Kentucky.
40. Roy Castle Lung Cancer Research Programme, Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool, United Kingdom.
41. Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX.
42. Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI.
43. Faculty of Medicine, Lund University, Lund, Sweden.
44. University Health Network- The Princess Margaret Cancer Centre, Toronto, CA.
45. Departments of Epidemiology and Community and Family Medicine, Dartmouth College, Hanover, NH.
46. Radboud University Medical Center, Nijmegen, The Netherlands.
47. Department of Epidemiology and Biostatistics, Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, School of Public Health, Nanjing Medical University, Nanjing, P.R. China.

48. National Institute of Occupational Health, Oslo, Norway.
49. Department of Medical Biosciences, Umeå University, Umeå, Sweden.
50. Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden.
51. Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK.
52. Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea.
53. UPMC Hillman Cancer Center and Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA.
54. Department of Pharmaceutical Sciences, College of Pharmacy, Washington State University, Spokane, Washington.
55. Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL.
56. Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN.
57. American Cancer Society, Inc., Atlanta, Georgia.
58. Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA.
59. Department of Epidemiology and Population Health, Stanford University, Stanford, CA.
60. Division of Research, Kaiser Permanente Northern California, Oakland, CA.
61. Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada.
62. Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Canada.

[†]These authors have equal contributions.

*Corresponding Author: Christopher I. Amos, Ph.D., Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA, Chris.Amos@bcm.edu

Abstract

To identify new susceptibility loci to lung cancer among diverse populations, we performed cross-ancestry genome-wide association studies in European, East Asian, and African populations and discovered five loci that have not been previously reported. We replicated 26 signals and identified 10 new lead associations from previously reported loci. Rare-variant associations tended to be specific to populations, but even common-variant associations influencing smoking behavior, such as those with *CHRNA5* and *CYP2A6*, showed population specificity. Fine-mapping and eQTL colocalization nominated several candidate variants and susceptibility genes such as *IRF4* and *FUBP1*. DNA damage assays of prioritized genes in lung fibroblasts indicated that a subset of these genes, including the pleiotropic gene *IRF4*, potentially exert effects by promoting endogenous DNA damage.

INTRODUCTION

Lung cancer is a multifactorial disease driven by environmental exposures, especially cigarette smoking, inherited germline genetic variants, and an accumulation of somatic genetic events[1]. Although genome-wide association studies (GWAS) have identified many significant contributing risk loci, the genetic underpinnings of lung cancer according to population variations remain incompletely understood[2-6]. Most GWAS have focused on genetically homogeneous case-control studies from European-ancestry populations[7]. Multi-ancestry studies have been useful in examining the heritability and genetic architecture of complex traits and diseases in diverse populations[7-10]. Multi-population genome-wide meta-analysis (GWMA) has been used to boost statistical power by increasing the total study sample size[8]. In addition, cross-ancestry analysis can improve association signal detection for low-frequency and rare alleles if they are more frequent in one population and help pinpoint functional variants when there is variability in linkage disequilibrium (LD) between functional variants and marker alleles across populations[11]. Consistency in allelic effects across populations can further support causal inference[9,10,12].

In the past two decades, approximately 40 lung cancer susceptibility loci directly influencing lung cancer risk have been identified by GWAS[2,3,13,14]. Array-based and family-based heritability estimates of lung cancer attributable to genetic factors range from 8-21%[1,6,15-17]. Population differences in the incidence of lung cancer suggest underlying heterogeneity in lung cancer etiology among human populations. Building on the recently completed OncoArray lung cancer GWAS[14,18-23] with additional earlier GWAS data sets[24-28], we performed a cross-ancestry discovery GWMA comprising 35,732 cases and 34,424 controls to comprehensively characterize common and rare lung cancer genetic susceptibility loci across multiple ancestral populations (**Table 1, Supplementary table 1**). The significant cross-ancestry single nucleotide polymorphisms (SNPs) identified in discovery analyses were validated by combining the initial cross-ancestry GWMA discovery and independent external validation datasets, adding 25,315 cases and 912,813 controls (**Supplementary table 1**)[16,29-34]. By combining GWMA summary-level data across populations of diverse ancestries, we refined loci that detect associations with lung cancer development[35].

RESULTS

Cross-ancestry GWMA of lung cancer. We included 70,156 individuals from 12 studies of diverse ancestry populations in the discovery study (**Table 1, Supplementary table 1**). Most individuals were inferred as having European ancestry (EUR; 74%), with 18% having East Asian (EAS) ancestry and 8% having African ancestry (AFR)[36]. Prior to association

analysis, all samples from the 12 studies were imputed using 32,470 samples from the Haplotype Reference Consortium (HRC)[37] as a reference panel. Detailed quality control processes are described in **Methods**. We conducted ancestry-stratified analyses in European (EUR), East Asian (EAS), and African (AFR) ancestry populations using Firth's logistic regression method,[36] which reduces bias when dealing with imbalanced data, especially in small sample sizes and with rare variants[38,39]. Firth's logistic regression test may be anticonservative for very rare variants with minor allele frequency (MAF) < 0.001, but the overall performance of Firth's test for GWAS with a combined type I error and accuracy was improved compared with that of conventional logistic regression-based Wald, score, and likelihood ratio tests for unbalanced studies with rare variants[39,40]. We then implemented different multi-ancestry meta-analysis methods[9,11,41-43] described in **Methods** and reported in **Table 2 and Supplementary table 2,** because conventional fixed-effect meta-analysis ignores the potential heterogeneity across different populations. We also performed cross-ancestry GWMA to detect additional loci associated with predominant histological types: lung adenocarcinoma (ADE), lung squamous cell carcinoma (SQC), and small cell lung carcinoma (SCC) (**Table 2, Supplementary table 2-3**). There were no detectable genomic inflations for lung cancer (λ_{Lung} = 1.0044) or any histologic subtypes (λ_{ADE}=1.0054; λ_{SQC}=1.0108; λ_{SCC}=1.0097) after adjustment to reflect a standardized sample size of 1,000 cases and 1,000 controls implying that residual population stratification is unlikely to be influencing association statistics within the ancestry-stratified analyses and combined meta-analyses across these diverse populations (**Fig. 1, Supplementary table 4**).

The cross-ancestry GWMA across three intercontinental populations identified 40 associations, including 15 associations for overall lung cancer, and 14, 9, and 2 associations for ADE, SQC, and SCC, respectively, in the discovery study at Bonferroni-corrected genome-wide significance level of $P\_BE^1 < 1.25\times10^{-8}$, where P_BE^1 represented a P-value from random binary-effects (BE) model[42] using METASOFT in the discovery study (**Fig. 1** and **Table 2**). These conditions allow for genome-wide discovery across overall lung cancer and the three histological subtypes. The nine new loci passing the nominal genome-wide significance level of $P\_BE^1 < 5.00\times10^{-8}$ are reported in **Table 2**. Top ancestry-specific GWMA results in the discovery study showed only nine cross-ancestry variants (20% of all loci), defined as those variants with M-values[42] greater than 0.9 across all three population models, indicating posterior probability that an effect exists in each population model under the assumption of heterogeneity (**Supplementary table 2**). For thirty-two cross-ancestry variants (71%), the significant effects stemmed from at least two ancestry populations, and thirty-one associations were significant in the EUR ancestry. Genomic cross-ancestry loci associated with lung cancer susceptibility with a $P \leq 10^{-5}$ are

reported in **Supplementary table 5-8**. Ancestry-specific and cross-ancestry genomic regional association plots for the top new cross-ancestry genetic variants discovered in our discovery study are shown in **Supplementary figure 1**.

Validation of cross-ancestry GWMA. To validate our cross-ancestry GWMA findings from the discovery dataset, we combined data from 938,128 individuals from seven studies consisting of a large, population-based cohort, the UK Biobank (UKBB)[29] lung cancer cases and controls and other published summary-level data[16,32-34] (**Table 1, Supplementary table 1**). Validation data include ancestry-specific summary-level data of European ancestry (UKBB, FinnGen, deCODE, SPAIN, INHALE, KPRB/GERA), East Asian ancestry (Nanjing), and African ancestry (INHALE). These data were not included in the discovery study. We used a validation analysis[44] approach that further evaluated findings from the discovery phase. Individual-level data for lung cancer GWAS were integrated into the discovery phase, so validation analysis using summary-level data let us support or reduce evidence of cross-ancestry associations from the discovery study[44]. We conducted cross-ancestry GWMA of validation datasets for the 45 suggestive associations identified from discovery analyses ($P\_BE^1 < 5\times10^{-8}$). We validated 11 associations for overall lung cancer, 11 for ADE, 4 for SQC, and 1 for SCC. In the validation study, to evaluate nine new susceptibility loci, we used a Bonferroni-corrected significance level of $P\_BE^2 < 5.55\times10^{-3}$ (=0.05/9 newly identified SNPs) (**Table 2**). Complementary validation analyses where we apply METASOFT to ancestry-specific summary-level datasets for all 45 SNPs are provided in **Supplementary table 9**.

During validation analysis, we combined three ancestry-specific discovery summary-level datasets (EUR1, EAS1, AFR1) with three ancestry-specific validation datasets (EUR2, EAS2, AFR2) using METASOFT for the 45 association signals identified in the discovery study (**Supplementary table 2, Supplementary table 10**). The combined cross-ancestry GWMA of discovery and validation studies supported 41 associations with lung cancer risk at the Bonferroni-corrected genome-wide significance ($P \leq 1.25\times10^{-8}$, adjusting for 3 histological subtypes and overall lung cancer) after post-imputation quality control (**Table 2, Supplementary table 10**). Eighteen association signals showed effects across at least five studies with an M value ≥ 0.9, indicating the genetic effect is likely to be present in each study tested in the cross-study meta-analysis (**Supplementary figure 2**)[42]. **Table 2** presents all the risk loci based on cross-ancestry GWMA with sentinel variants from the discovery study at the nominal genome-wide significance level of $P\_BE^1 < 5\times10^{-8}$, P-values from a binary random-effect meta-analysis model of the validation study (P_BE^2) and P-values from a combined model of discovery and validation summary-level data (P_BE^C).

New and known associations in previously reported loci. Our cross-ancestry GWMA identified 9 association signals for overall lung cancer, 10 for ADE, 6 for SQC, and 1 for SCC that were identified in previous ancestry-specific studies of EUR[2,14,18], EAS[25] or AFR[27,45] populations and cross-ancestry studies of EUR and EAS populations[34] at a Bonferroni-corrected genome-wide significance level of $P\_BE^C < 1.25\times10^{-8}$ (**Table 2**, **Supplementary table 2-3**). These include rare, larger-effect variants in *BRCA2* (rs11571815) and *CHEK2* (rs17879961) genes initially identified in EUR populations and a variant in the *ATM* (rs56009889) gene in Ashkenazi Jewish populations; these variants are significant when combined with EAS and AFR populations. Further, our study identified ten new associated variants (four for overall lung cancer, five for ADE, and one for SQC) within ±500kb of a previously reported locus with a new lead SNP (an r^2 with the previously reported lead SNP < 0.6 in a 1000 Genomes ALL populations[46] or not reaching a genome-wide significance level of 5×10^{-8} in the previous study[47]) at $P\_BE^C < 1.25\times10^{-8}$. These include new variants from well-established loci at 5p15.33, 6p21.32, and 9p21.3, among others. An intergenic variant rs9374662 between *ROS1* and *DCBLD1* on 6q22.1 showed strong genetic signals in both EUR and EAS populations and was associated with ADE and overall lung cancer risk. A few intronic variants in *DCBLD1* have previously been associated with lung[18] and colorectal cancer[48]. Additionally, well-known variants associated with smoking behaviors, rs55781567 in *CHRNA5* on 15q25.1 and rs56113850 in *CYP2A6* on 19q13.2 were also substantial in overall lung cancer, ADE, and SQC[49,50], but showed variable associations with risk among non-European populations. Another well-established lung cancer risk-associated variant, rs2853677 in *TERT* on 5p15.33 demonstrated allelic homogeneity showing the more consistent effects across three intercontinental populations[51] (**Supplementary figure 2**). These new and known signals across lung cancer histology suggest that our analytic approach can robustly detect previously reported and additional signals within the known GWAS loci.

Identification of new susceptibility loci in cross-ancestry GWMA. In addition to new and known signals in the previously reported loci, cross-ancestry GWMA identified five new susceptibility loci, including two in overall lung cancer, one in ADE, one in SQC, and one in SCC at the Bonferroni-corrected genome-wide significance level of $P\_BE^C < 1.25\times10^{-8}$ (**Table 2**). LocusZoom regional plots for the cross-ancestry genetic variants newly identified in our discovery study are shown in **Extended Data figure 1**. Among them is rs9865715 in *CYP8B1* on 3p22.1, a low-frequency missense variant (allele frequency (AF) = 0.99, 1.00, and 0.95 in EUR, EAS, and AFR, respectively) with a binary effect model P-value, $P\_BE^C=3.53\times10^{-10}$. Newly identified association of an intronic variant, rs12203592, was detected in *IRF4* on 6p25.3. rs12203592 has been previously associated with numerous pigmentation traits[52], multiple blood cell traits[53-55],

squamous cell carcinoma of the skin[56-58], and smoking cessation[59] implying an important pleiotropy with this variant. We identified an intronic variant, rs17534632, in *PPIL6* on 6q21, which was associated with lung cancer risk in EUR and EAS populations at a nominal genome-wide significance level of 5×10^{-8} ($P\_BE^{C}=3.61\times10^{-8}$). Associations in *PPIL6* with blood traits have been reported in blood traits in EUR populations[60] and cross-ancestry and ancestry-specific studies[53].

The histological subtype-stratified analysis identified three new susceptibility loci achieving Bonferroni-corrected genome-wide significance and three additional new variants/loci at a nominal genome-wide significance level of 5×10^{-8} (**Table 2, Supplementary table 2)**. For SCC, we detected a new association signal for a rare missense variant, rs141178913 on *IL17RC* at 3p25.3 (AF = 0.001, 1.2×10^{-4}, and 2.8×10^{-4} in EUR, EAS, and AFR, respectively; $P\_BE^{C}=2.35\times10^{-9}$). Another association at rs191133092 near HLA complex group 15 (*HCG15*) was detected in AFR population ($P\_BE^{C}=1.56\times10^{-8}$; $P_{AFR1}=6.64\times10^{-7}$). For ADE, one additional cross-ancestry locus was identified at rs268864, an intronic variant of *ACTR2* on 2p14 showing the ancestry-specific association signals in both EUR and EAS populations ($P\_BE^{C}=1.60\times10^{-16}$; $P_{EUR1}=1.13\times10^{-7}$; $P_{EAS1}=0.07$; $P_{EUR2}=4.42\times10^{-4}$; $P_{EAS2}=3.56\times10^{-7}$). For SQC, one newly identified intronic variant, rs2041742, in ncRNA *LINC01122* on 2p16.1 was detected in EAS and AFR populations ($P\_BE^{C}=1.48\times10^{-11}$; $P_{EAS1}=1.97\times10^{-12}$; $P_{AFR1}=0.09$). Another intronic variant, rs6757055 in *IKZF2* on 2q34 was identified in EAS only ($P\_BE^{C}=4.54\times10^{-8}$; $P_{EAS1}=1.51\times10^{-10}$).

Conditional analyses on the top cross-ancestry associations. Along with identifying the top association signals in multi-ancestry case-control meta-analyses of lung cancer, we investigated secondary association signals at each locus having multiple associated SNPs based on cross-ancestry GWMA findings with a stringent imputation quality score ≥ 0.8. Stepwise iterative conditional analysis was performed by conditioning on the primary associated SNP at each locus to test if any other SNPs are significantly associated until no SNP associated with lung cancer or any subtype remained[61]. We first implemented conditional analysis on the identified cross-ancestry lead SNP per population using GCTA v1.93[61] (--cojo-cond) and then meta-analyzed them across populations. Additional associations identified from conditional analyses are reported in **Supplementary table 11**. Conditioning on intronic variant rs2853677 of *TERT*[62] revealed one new independent variant in *CLPTM1L*[18,63-65] (rs31487, $P=3.91\times10^{-32}$), and two in *TERT*[18,34,66-68] (rs7705526, $P=4.21\times10^{-11}$; rs72709458, $P=1.70\times10^{-10}$). By conditioning on rs55781567 in *CHRNA5* of 15q25.1, we discovered two independent variants, rs576982 in *CHRNA5* ($P=4.71\times10^{-14}$) and rs28654165 near *IREB2* (P= 0.005), for lung cancer susceptibility and another two independent variants, rs113352275 in *PSMA4* ($P=1.09\times10^{-9}$) and rs6495350 in *MORF4L1* (P=0.002), for SQC.

Fine-mapping and functional annotation of candidate variants. To nominate candidate variants from each locus for further follow-up, we performed fine-mapping of cross-ancestry GWAS loci. We first performed cross-ancestry analysis using MANTRA[11] to obtain Bayes factors for the variants passing our criteria ($P < 10^{-4}$ in initial logistic regression, ±250kb of lead variant) while accounting for heterogeneity between different ancestries. Based on the cumulative Bayes factors within each locus[69,70], we identified 715 variants falling into the 99% credible set across 45 GWAS loci (median 10 per locus, ranging 1-178 per locus; **Supplementary table 12**, **Methods**).

To functionally characterize the prioritized variants, we performed an integrated variant functional annotation approach[71] using the Functional Annotation of Variants-Online Resource (FAVOR) platform (http://favor.genohub.org/), by incorporating the Multi-dimensional Annotation Class Integrative Estimator (MACIE)[72,73] (**Supplementary table 13**). MACIE is a generalized linear mixed model designed to predict regulatory and evolutionarily conserved SNPs using 36 genome-wide annotations. Out of 715 variants within the 99% credible set, 105 unique variants across 27 GWAS loci (median 4 per locus, ranging 1–22 per locus) displayed a marginal probability > 0.9 for either "regulatory" or "conserved" functional features. For example, from the ADE locus in *ACTR2* at 2p14 tagged by rs268864, 3 of 23 variants found in the 99% credible set (rs10116, rs268882, and rs72822431) were predicted to have regulatory potential (marginal probability > 0.98). From another ADE locus in *IRF4*, at 6p25.3 and tagged by rs2316515, 4 of 10 variants within the 99% credible set (rs1050979, rs2316515, rs7768807, and rs872071) displayed regulatory potential (marginal probability > 0.99). Using a cross-ancestry fine-mapping followed by MACIE analysis we provide a prioritization of lung cancer GWAS loci and candidate variants for follow up with functional genomics experiments.

Prioritization of candidate genes from cross-ancestry GWMA. To map susceptibility genes underlying the lung cancer GWAS associations, we performed expression quantitative trait locus (eQTL)-based analyses to identify allelic-specific effects on gene expression. Using the Functional Mapping and Annotation (FUMA) platform[74], we surveyed an overlap between GWAS variants (variants that are linked with 38 lead SNPs from the multi-ancestry meta-analysis, $r^2 > 0.6$ in 1000G ALL populations, phase 3) and significant eQTL variants in lung cancer-relevant tissue types. Based on the predictions that risk variants may exert their effects via circulation, lung, immune system, and brain regions (potentially underlying smoking behavior), we defined eight lung cancer-relevant tissues as whole blood (n=670), lung (n=515), EBV-transformed lymphocytes (n=147), cortex (n=206), frontal cortex (n=175), hypothalamus (n=170), cerebellum (n=209), and nucleus accumbens basal ganglia (n=202). A total of 285 unique eQTL genes showed an overlap with GWAS variants (**Supplementary table 14**). To prioritize candidate susceptibility genes, we performed colocalization analyses using

eQTL summary statistics of eight lung cancer-relevant tissues from the Genotype-Tissue Expression (GTEx) v8. Accounting for the heterogenous LD in our multi-ancestry GWAS population and eQTL population in GTEx v8, we applied an LD-dependent (eCAVIAR) approach using a European LD matrix (accounting for 74% of GWAS and 85% of eQTL populations) as well as an LD-independent approach (coloc) to avoid spurious colocalizations (**Methods**). Based on the concordance between eCAVIAR and coloc (posterior probability of > 0.01 and 0.8, respectively), a substantial proportion of lung cancer risk-associated variants (20 of 38 variants, 52.6%) colocalized with eQTL of at least one gene from a tissue (**Supplementary table 15-17**). A total of 48 candidate genes, including three from the new cross-ancestry GWAS loci (*IRF4, MICAL1,* and *AK9*), were identified as potential susceptibility genes contributing to lung cancer risk (**Supplementary table 17**). Based on histological subtypes, colocalization identified 23, 23, 17, and 2 candidate genes for overall lung cancer, ADE, SQC, and SCC, respectively. For 11 of 38 risk variants, colocalizations were detected with multiple genes including 9 HLA genes for 3 MHC loci, highlighting the roles of cellular immune response. When excluding genes from MHC loci, these colocalized genes and the genes from FUMA results, including non-GTEx datasets, were significantly enriched in pathway affecting AMPK and calcium signaling, cell stress, and injury (**Supplementary tables 18-19**). Notably, the known cross-ancestry locus *FUBP1* and a new cross-ancestry locus *IRF4* displayed the highest probability scores from both eCAVIAR and coloc in more than one tissue types (lung, cerebellum, or whole blood) (**Supplementary table 17; Fig. 2a–d**). Concurrent alterations of *FUBP1* and *PTEN* have been shown to promote breast cancer through a global effect on alternative splicing[75]. Elevated expression of *FUBP1* was reported in multiple cancer types including non-small cell lung cancer[76]. The lung cancer risk-associated A allele of the candidate colocalizing SNP, rs34517439 is correlated with lower *FUBP1* expression in normal lung and brain tissues (GTEx, **Supplementary figure 3**). *IRF4* has pleiotropic roles in immune cell differentiation[77] and pigmentation phenotypes[78] and is a master regulator of aberrant gene expression networks in multiple myeloma[79]. The candidate colocalizing SNP, rs12203592, is the lead SNP in both GWAS of overall lung cancer and eQTL of lung and whole blood. It has been shown to be a functional SNP, displaying an allelic binding to TFAP2A driving an allelic enhancer function in both primary melanocytes and blood cells[78,80]. Importantly, the direction of the allelic effect in skin tissue is opposite of that in lung and whole blood tissues. The lung cancer risk-associated T allele is correlated with higher *IRF4* expression in lung (**Supplementary figure 3**) and whole blood tissues but lower expression in skin tissues (GTEx v8) and primary melanocytes[78,81,82].

Dysregulation of lung cancer risk genes promotes DNA damage. To further characterize top candidate susceptibility genes, we performed cell-based DNA damage assays on a subset of candidate genes prioritized from the eQTL colocalization analysis (**Fig. 2a–d, Supplementary table 20**). DNA damage from cellular intrinsic processes promotes mutations and cancer[83]. An increasing number of genes are shown to promote DNA damage via various direct and indirect mechanisms[14,84,85]. We hypothesized that a fraction of the GWAS-nominated risk-associated genes could promote lung cancer by increasing endogenous DNA damage and genome instability[14]. Of the 48 colocalized genes, we prioritized 17 genes that were amenable to test in our system (**Supplementary table 20**). We also included two candidate genes from the GWAS loci based on the eQTL findings from non-GTEx datasets identified through FUMA analysis (**Supplementary table 20**). We performed knockdown and/or overexpression of candidate genes, mimicking allelic expression changes associated with lung cancer (**Supplementary table 20**) followed by assessment of DNA damage as evidenced by γH2AX levels in immortalized, untransformed human lung fibroblasts. Of 19 genes, 7 over-expressing (*IRF4, AK9, CYP21A2, DCBLD1, SECISBP2L, CCDC97,* and *FUBP1*) (**Fig. 3a, Supplementary table 20**) and 2 knocked down (*PPIL6* and *ACTR2*) genes displayed significantly increased DNA damage (**Figure 2F, Supplementary tables 20-23**). The validation rates for candidates with increased DNA damage is higher than a set of over-expressed human genes[84] (P = 0.0197, Fisher exact test) or knockdown of randomly selected genes (P=0.0286, Fisher exact test, **Supplementary table 24**). The 12 genes showing either increased or decreased DNA damage phenotype when over-expressed or knockdown included the top 5 colocalized genes with the highest probability scores (*FUBP1, IRF4, SECISBP2L, CCHCR1,* and *CYP2A6*) (**Supplementary table 20**) and 4 genes from new multi-ancestry meta-analysis loci (*IRF4*, *ACTR2, PPIL6,* and *AK9*) (**Fig. 3b, Supplementary table 20**). *ACTR2* increased DNA damage when knocked down ($P = 9.89\times10^{-3}$, effect = 0.966) and reduced when overproduced ($P = 6.46\times10^{-5}$, effect = -1.826) (**Fig. 3a** and **Supplementary table 20**), suggesting a potential protective role. Conversely, *AK9* significantly reduced median DNA damage when knocked down ($P = 2.16\times10^{-4}$, effect = -1.816; **Supplementary table 20**) and increased when overproduced ($P = 4.84\times10^{-5}$, effect = 1.865) indicating a DNA damage-promoting role (**Fig. 3a and Supplementary table 20**). Notably, IRF4 ($P = 2.32\times10^{-3}$, effect = 1.320) increased DNA damage when overproduced. Given that the lung cancer risk allele is correlated with higher *IRF4* levels in lung tissue (**Supplementary figure 3a**), an endogenous DNA damage-promoting role of IRF4 in lung cells could support the evidence of a potential mechanism contributing to lung cancer risk (**Fig. 3a**). Altogether, we expanded the list of genes associated with DNA damage in lung cancer and assigned a known cancer-promoting phenotype (DNA damage) to many lung cancer risk genes.

DISCUSSION

We conducted cross-ancestry GWMA of lung cancer involving 51,961 individuals of EUR ancestry, 12,434 of EAS ancestry, and 5,766 of AFR ancestry and validated the findings with 910,609 individuals of EUR ancestry, 26,640 of EAS ancestry, and 879 of AFR ancestry. GWAS have identified approximately 40 loci[2] associated with lung cancer by analyzing populations of relatively homogeneous ancestry background. However, most findings are biased toward EUR-ancestry studies because multi-ancestry GWMA has not been feasible due to limited genotyping data in other populations. In this study, we identified and validated five new cross-ancestry SNPs associated with overall lung cancer, ADE, SQC, and SCC at the Bonferroni-corrected genome-wide significance level of 1.25×10^{-8}.

Lung carcinogenesis is a complex process that involves the acquisition of genetic mutations and epigenetic changes that alter cellular processes, such as proliferation and differentiation[86]. Lung cancer development also seems to have distinct population and geographical differences. Many studies have uncovered aspects of lung cancer pathogenesis[87], but identifying new genetic variants associated with lung cancer remains challenging due to small effect sizes and the contribution of cigarette smoking. To date, a limited number of lung cancer-specific genes have been detected. Quantifying the genomic architecture of lung cancer risk is important to better understanding its pathogenesis. Therefore, improved elucidation of genetics in lung carcinogenesis is critical. For instance, better understanding the genomic diversity of oncogenes, tumor suppressor genes, or specific alterations across diverse ancestry populations can help in designing population-specific targeted therapies. Additionally, deciphering the shared genetic variants underlying lung cancer predisposition in populations of diverse ancestry can help refine risk prediction models for individuals at high-risk across ancestral populations and potentially identify variants associated with lung cancer susceptibility in admixed populations and across ancestral groups. Our investigation of ancestry-specific and cross-ancestry associations with lung cancer and specific histological subtypes resulted in several key findings.

First, we confirmed nine cross-ancestry genomic risk associations for overall lung cancer, ten for ADE, six for SQC, and one for SCC, while concurrently identifying an additional four new genome-wide significant risk variants in previously reported loci for overall lung cancer, five for ADE, and one for SQC. These results highlight the utility of multi-population meta-analysis in identifying and fine-mapping new signals.

Second, we discovered ancestry-specific effects of common and rare variants on lung cancer risk among EUR, EAS, and AFR populations. A common variant, rs9374662 located between *ROS1* and *DCBLD1*, displayed a strong association

with lung cancer and ADE in both EUR and EAS populations at the Bonferroni-corrected genome-wide significance level and a suggestive association in AFR population although sample size of AFR individuals was smaller than those of other populations. While the sample size of AFR ancestry in this study is limited, we were able to identify a few new association signals in AFR population. As presented in **Supplementary figure 1**, a rare variant rs141178913 in *IL17RC* displayed a strong association with SCC risk among EUR and AFR populations ($P\_BE^1_{EUR1}=5.81\times10^{-9}$; $P\_BE^1_{AFR1}=0.01$) with both effect allele frequencies < 0.0001. Another population-specific association with SCC was observed at low-frequency variation rs191133092 near *HCG15* in AFR population ($P\_BE^1_{AFR1}=6.64\times10^{-7}$) with effect allele frequency of 0.01. Our cross-ancestry fine-mapping considering heterogeneity between different populations nominated a small number of variants at each locus that were further annotated using functional databases. Future studies using statistical approaches incorporating the complexity of cross-population LD structure followed by experimental validation of candidate variants is warranted.

Third, we identified and prioritized candidate lung cancer susceptibility genes, inferred biological pathways identified through eQTL colocalization analyses, and performed cell-based DNA damage assays. Colocalization analyses showed candidate genes were frequently part of biological pathways involving immune response and cellular stress response. 63% (12/19) of candidate genes prioritized based on eQTL colocalization displayed a significantly altered level of endogenous DNA damage in lung fibroblasts. This high positivity rate highlights the efficiency of our prioritization scheme of using eQTL colocalization in relevant tissue types. The results also indicated that a high proportion of candidate lung cancer susceptibility genes, including those from new cross-ancestry loci (*IRF4, ACTR2, PPIL6*, and *AK9*), could have roles in promoting or preventing endogenous DNA damage.

One limitation of this study was that some candidate associations evaluated in the validation study failed to reach the Bonferroni-corrected significance of $P\_BE^2<5.55\times10^{-3}$. This could be due to a lack of power because replication datasets with similar study and analytical design features, such as imputation panel, availability of low-frequency and rare variants, and data for SCC were lacking, or some of these findings could be false positive associations. Future studies will help to resolve any possible false positive results, despite stringent quality control steps we have implemented. Thus, there is a need to include more individuals of underrepresented populations, so we can further characterize the genetic contribution to lung cancer development and provide better insight into genetic architecture of ancestry-specific and cross-ancestry lung cancer etiology.

Overall, our cross-ancestry meta-analysis of population-specific GWAS across multi-ancestry populations has helped elucidate the etiology and mechanisms of lung cancer. Understanding the genetic architecture of lung cancer predisposition will help reveal how lung cancer develops and could assist in identifying new susceptibility biomarkers for better risk evaluation directed at early detection and diagnosis, targeted therapy, and improved preventive measures.

Acknowledgements

Our study was supported by the National Institutes of Health (NIH) for Integrative Analysis of Lung Cancer Etiology and Risk (U19CA203654) and Sequencing Familial Lung Cancer (R01CA243483). C.I.A. is a Research Scholar of the Cancer Prevention Research Interest of Texas (CPRIT) award (RR170048). Functional studies were partially supported by NIH grants (R01CA250905 (S.M.R), CPRIT RR170048 (C.I.A) and DP1-AG072751 (S.M.R)). This project was supported by the Cytometry and Cell Sorting Core at Baylor College of Medicine with funding from the CPRIT Core Facility Support Award (CPRIT RP180672) and the NIH (CA125123 and RR024574) as well as the assistance of Joel M. Sederstrom. The Resource for the Study of Lung Cancer Epidemiology in North Trent (ReSoLuCENT) study was funded by the Sheffield Hospitals Charity, Sheffield Experimental Cancer Medicine Centre, and Weston Park Hospital Cancer Charity. F.T. was supported by a clinical PhD fellowship funded by the Yorkshire Cancer Research/Cancer Research UK Sheffield Cancer Centre. D.M. was supported by Department of Health and Human Services contracts HHSN26820100007C, HHSN268201700012C, and 75N92020C00001. J.E.B was supported by the Intramural Research Program of the National Human Genome Research Institute, NIH. R.W.P. was supported by NIH T32ES027801. J.X. was supported by the National Institute of Environmental Health Sciences of the NIH under Award Number K99ES033259. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH. We want to acknowledge the participants and investigators of INTEGRAL-ILCCO Consortium, Genetic Epidemiology of Lung Cancer Consortium (GELCC), FinnGen study and Kaiser Permanente Research Bank (KPRB) Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort study.

Author Contributions: J.B., Y.H., Y.L., and C.I.A. conceived and designed the study. Y.H. and X.X. acquired the data. Y.H., E.L., J.C., and X.X. performed the analysis. J.X. performed experimental validation. M.Z., W.Z., R.S., A.S., C.L., T.R., L.K., L.S. provided substantial support on validation study. J.B., Y.H., Y.L., J.X., E.L., J.C., and C.I.A. interpreted the results. J.B., J.X., E.L., and J.C. wrote the first draft of the manuscript. J.B., Y.H., J.X., E.L., J.C., and X.X. provided

supplementary materials. J.B. and C.I.A. provided supervision and contributed to analyses. All authors reviewed and commented on the manuscript and approved the final version of the manuscript.

Competing Interests Statement: All authors declare no competing interests.

Table 1. Sample characteristics of the study populations by ancestry.

	Discovery GWAS sample population								Validation GWAS sample population							
	EUR		EAS		AFR		Total	Total	EUR		EAS		AFR		Total	Total
Strata	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
Lung	26,683	25,278	7,062	5,372	1,987	3,774	35,732	34,424	11,680	898,929	13,316	13,324	319	560	25,315	912,813
ADE	9,791	23,173[a]	4,630	5,372	844	3,774	15,265	32,319	3,095	436,443	8,755	13,324	186	560	12,036	450,327
SQC	6,107	23,173[a]	1,292	5,372	451	3,774	7,850	32,319	1,607	365,037	3,857	13,324	75	560	5,539	378,921
SCC	2,267	23,173[a]	99	5,372	116	3,774	2,482	32,319	1,268	365,282	-	-	29	560	1,297	365,842

[a]Number of individuals included in the corresponding histology-specific analysis with histological information. Lung, Overall lung cancer; ADE, Lung adenocarcinoma; SQC, Lung squamous cell carcinoma; SCC, Small cell lung carcinoma; EUR, European; EAS, East Asian; AFR, African.

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

Strata	SNP	Cytoband	Position	Nearest Gene	Allele	EAF (EUR1; EAS1; AFR1)	OR (EUR1; EAS1; AFR1)	P_BE[1]	P_BE[2]	P_BE[C]
Lung Cancer										
New#	**rs9865715***	**3p22.1**	**42917047**	**CYP8B1**	**G_A**	**0.993; 0.999; 0.951**	**1.64; 1.24; 0.87**	**1.64×10^{-10}**	**0.904**	**3.53×10^{-10}**
New#	**rs12203592***	**6p25.3**	**396321**	**IRF4**	**T_C**	**0.143; 0.010; 0.042**	**1.12; 0.88; 0.96**	**2.85×10^{-8}**	**1.97×10^{-5}**	**3.33×10^{-12}**
New#	rs17534632	6q21	109740101	PPIL6	T_C	0.204; 0.020; 0.041	1.10; 1.09; 0.99	8.78×10^{-9}	0.166	3.61×10^{-8}
New	rs34102154*	6p21.32	32572106	HLA-DRB1, HLA-DQA1	A_G	0.157;0.231;0.213	0.90; 0.94; 0.93	7.51×10^{-10}	0.048	1.88×10^{-9}
New	rs1885281*	10q25.2	114492898	VTI1A	A_G	0.059; 0.290; 0.191	1.01; 1.24; 1.13	9.21×10^{-13}	4.02×10^{-10}	1.23×10^{-20}
New	rs11607355*	11q23.3	118093547	JAML, AMICA1	T_C	0.502; 0.551; 0.253	0.94; 0.90; 0.87	2.79×10^{-10}	8.38×10^{-12}	2.58×10^{-20}
New	rs2413932*	15q21.1	49383481	SECISBP2L, COPS2	T_C	0.734; 0.752; 0.619	1.09; 1.10; 1.06	1.53×10^{-12}	3.49×10^{-9}	8.31×10^{-20}
Lung adenocarcinoma										
New#	**rs268864***	**2p14**	**65489742**	**ACTR2**	**A_G**	**0.843; 0.805; 0.899**	**0.88; 0.94; 0.87**	**3.41×10^{-8}**	**5.04×10^{-10}**	**1.60×10^{-16}**
New#	rs2316515	6p25.3	410848	IRF4	G_A	0.582; 0.549; 0.388	1.11; 0.98; 1.09	3.21×10^{-9}	0.563	1.82×10^{-8}
New	rs3129860*	6p21.32	32401079	LOC101929163, HLA-DRA	G_A	0.868; 0.875; 0.938	0.89; 0.79; 1.03	2.09×10^{-11}	0.609	5.56×10^{-11}
New	rs12348845*	9p21.3	21775492	MIR31HG, MTAP	A_G	0.098; 0.256; 0.522	1.21; 1.14; 1.11	1.45×10^{-14}	2.73×10^{-6}	9.01×10^{-19}
New	rs7902587*	10q24.33	105694301	STN1,SLK	T_C	0.105; 0.005; 0.179	1.19; 0.51; 1.19	2.67×10^{-10}	0.124	1.83×10^{-10}
New	rs12265047*	10q25.2	114487925	VTI1A	A_G	0.968; 0.727; 0.701	0.90; 0.80; 0.85	1.61×10^{-12}	4.40×10^{-13}	1.06×10^{-23}
New	rs75031349*	20q13.33	62314054	RTEL1-TNFRSF6B	G_A	0.084; 0.028; 0.168	0.84; 0.79; 0.90	1.19×10^{-8}	0.190	1.04×10^{-8}
Lung squamous cell carcinoma										
New#	**rs2041742***	**2p16.1**	**59086026**	**LINC01122**	**G_A**	**0.947; 0.981; 0.987**	**1.03; 0.54; 0.60**	**2.03×10^{-12}**	**0.846**	**1.48×10^{-11}**
New#	rs6757055	2q34	213999410	IKZF2	A_C	0.950; 0.969; 0.938	1.04; 0.59; 1.00	3.01×10^{-9}	0.941	4.54×10^{-8}
New	rs9267123*	6p21.33	31427395	LINC01149, HCP5	C_G	0.116; 0.021; 0.038	1.42; 1.14; 1.42	2.93×10^{-18}	0.071	2.50×10^{-18}
Small cell lung cancer										
New#	**rs141178913***	**3p25.3**	**9970073**	**IL17RC**	**G_C**	**$0.001; 1.2\times10^{-4}; 2.8\times10^{-4}$**	**5.36; NA; 76.69**	**2.37×10^{-9}**	**NA**	**2.37×10^{-9}**
New#	rs191133092	6p22.1	28932985	LINC01556, HCG15	T_A	1.4×10^{-4}; 2.8×10^{-5}; 0.014	12.56; NA; 5.30	1.52×10^{-8}	NA	1.52×10^{-8}

Nearest gene (reference NCBI build37) is given as locus label and includes all the genes +/- 200kb of the genomic risk SNP; Asterisks * indicate the SNPs at the Bonferroni-corrected significance level of $P\_BE^C \leq 1.25E10\text{-}8$; Allele, effect allele_other allele; EAF, effect allele frequency for European (EUR1), East Asian (EAS1), African (AFR1) population in discovery study, respectively; OR, odds ratio effect size for EUR1, EAS1, and AFR1 population, respectively; P_BE^1, P_BE^2, and P_BE^C, P-value from binary random-effect meta-analysis model in cross-ancestry model of discovery study, validation study, and cross-ancestry combined model of discovery and validation studies using METASOFT, respectively; New#, new susceptibility loci discovered in this study; New, new lead variant from a previously reported loci with $r^2 < 0.6$.

Figure Legends

Figure 1 Manhattan plots and quantile-quantile plots of the GWAS meta-analysis for lung cancer in the cross-ancestry analyses. (a) Lung carcinoma: 35,732 cases and 34,424 controls. (b) Lung adenocarcinoma: 15,359 cases and 32,558 controls. (c) Lung squamous cell carcinoma: 7,896 cases and 32,558 controls. (d) Small cell lung carcinoma: 2,499 cases and 32,558 controls. The x-axis represents chromosomal location, and the y-axis -$\log_{10}$(P-value). The gene annotation for newly identified loci are in blue. The red horizontal line denotes the Bonferroni-corrected genome-wide significant two-sided P-value of $P = -log_{10}(1.25 \times 10^{-8})$. P-values are based on random binary-effects meta-analysis of three ancestry-specific summary statistics adjusted for principal components and study sites using Firth test.

Figure 2. Functional validation of the prioritized genes from cross-ancestry lung cancer GWAS. (a, c) eQTL signals in GTEx v8 lung tissues (n = 515) for *IRF4* (a) and *FUBP1* (c) colocalize with those of overall lung cancer GWAS by eCAVIAR (CLPP = 0.976 for rs12203592 and 1.000 for rs34517439) and coloc (PPH_4 = 0.979 for rs12203592 and 0.996 for rs34517439). Pearson correlation is shown between log-transformed *P* values of eQTL (y-axis) and GWAS (x-axis). Variants are color-coded based on the LD R^2 (1000 Genomes, EUR, phase 3) with the candidate variants (red dots). Variants with imputation quality > 0.6 were plotted in this region. (b, d) Regional association plots of eQTL (blue shadow) and GWAS (green shadow) within +/− 100kb of rs12203592 (b) and rs34517439 (d) are presented. The horizontal line indicates Bonferroni-corrected genome-wide significant *P*-value for GWAS (1.25×10^{-8}) and genome-wide empirical *P*-value threshold for eQTL of *IRF4* (1×10^{-4}) or *FUBP1* (1.8×10^{-4}). UCSC genes tracks are displayed as the full mode in this region.

Figure 3. Dysregulation of cross-ancestry lung cancer GWAS-nominated risk genes promotes DNA damage. (a, b) A flow-cytometric screen for lung cancer DNA damageome genes and proteins. (a) Overproduction screen. Upper: assay scheme, N-terminal EmGFP fusions of lung cancer risk genes were transiently overproduced for 72hours, followed by DNA damage detection using flow cytometry. Lower: normalized γH2AX level of each of the overproduction candidate (GFP positive cells). *FUBP1* (Representative histogram shown in the upper right corner), *CCDC97, IRF4, DCBLD1, SECISBP2L, CCDC97, CYP21A2*, and *AK9* promote DNA damage when overexpressed. Gating strategy is shown in Extended Data figure 2 (a-d). All candidates are normalized to the median γH2AX intensity of GFP^+ Tubulin (Tub) overproducing cells. mean ± SEM, n>=6. Two sample two-sided t-test assuming equal variances, * P < 0.00263 after Bonferroni correction, exact P-values in Supplementary table 20. (b) siRNA knockdown screen identifies *PPIL6* as loss-of-function DNA damageome gene. Upper: assay scheme, siRNAs targeting several lung cancer risk genes were transfected for 72 hours to achieve knockdown, followed by DNA damage measurements by flow cytometry. Lower left: normalized DNA damage for each siRNA knockdown. γH2AX-high cells are quantified using a threshold described in the methods, and gating strategy is shown in Extended Data figure 2 (e-g). All candidates are normalized to non-targeting (NT) pooled siRNAs. mean ± SEM, n>=6. Two sample t-test assuming unequal variances, * P < 0.0125 after Bonferroni correction, exact P-values in Supplementary table 20.

References

1. Sampson, J.N. *et al.* Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. *J Natl Cancer Inst* **107**, djv279 (2015).
2. Bosse, Y. & Amos, C.I. A Decade of GWAS Results in Lung Cancer. *Cancer Epidemiol Biomarkers Prev* **27**, 363-379 (2018).
3. Park, S.L., Cheng, I. & Haiman, C.A. Genome-Wide Association Studies of Cancer in Diverse Populations. *Cancer Epidemiol Biomarkers Prev* **27**, 405-417 (2018).
4. Popejoy, A.B. & Fullerton, S.M. Genomics is failing on diversity. *Nature* **538**, 161-164 (2016).
5. Rosenberg, N.A. *et al.* Genome-wide association studies in diverse populations. *Nat Rev Genet* **11**, 356-66 (2010).
6. Schabath, M.B., Cress, D. & Munoz-Antonia, T. Racial and Ethnic Differences in the Epidemiology and Genomics of Lung Cancer. *Cancer Control* **23**, 338-346 (2016).
7. Asimit, J.L., Hatzikotoulas, K., McCarthy, M., Morris, A.P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur J Hum Genet* **24**, 1330-6 (2016).
8. Conti, D.V. *et al.* Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nat Genet* **53**, 65-75 (2021).

9. Magi, R. *et al.* Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum Mol Genet* **26**, 3639-3650 (2017).
10. Li, Y.R. & Keating, B.J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med* **6**, 91 (2014).
11. Morris, A.P. Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* **35**, 809-22 (2011).
12. Marigorta, U.M. & Navarro, A. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet* **9**, e1003566 (2013).
13. Wang, J. *et al.* Genetic predisposition to lung cancer: comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies. *Sci Rep* **7**, 8371 (2017).
14. Bossé, Y. *et al.* Transcriptome-wide association study reveals candidate causal genes for lung cancer. *Int J Cancer* **146**, 1862-1878 (2020).
15. Kanwal, M., Ding, X.J. & Cao, Y. Familial risk for lung cancer. *Oncol Lett* **13**, 535-542 (2017).
16. Rashkin, S.R. *et al.* Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat Commun* **11**, 4423 (2020).
17. Jiang, X. *et al.* Shared heritability and functional enrichment across six solid cancers. *Nat Commun* **10**, 431 (2019).
18. McKay, J.D. *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).
19. Amos, C.I. *et al.* The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev* **26**, 126-135 (2017).
20. Li, Y. *et al.* Genome-wide interaction study of smoking behavior and non-small cell lung cancer risk in Caucasian population. *Carcinogenesis* **39**, 336-346 (2018).
21. Li, Y. *et al.* Genetic interaction analysis among oncogenesis-related genes revealed novel genes and networks in lung cancer development. *Oncotarget* **10**, 1760-1774 (2019).
22. Ji, X. *et al.* Identification of susceptibility pathways for the role of chromosome 15q25.1 in modifying lung cancer risk. *Nat Commun* **9**, 3221 (2018).
23. Ji, X. *et al.* Protein-altering germline mutations implicate novel genes related to lung cancer development. *Nat Commun* **11**, 2220 (2020).
24. Byun, J. *et al.* Genome-wide association study of familial lung cancer. *Carcinogenesis* **39**, 1135-1140 (2018).
25. Lan, Q. *et al.* Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat Genet* **44**, 1330-5 (2012).
26. Kachuri, L. *et al.* Fine mapping of chromosome 5p15.33 based on a targeted deep sequencing and high density genotyping identifies novel lung cancer susceptibility loci. *Carcinogenesis* **37**, 96-105 (2016).
27. Zanetti, K.A. *et al.* Genome-wide association study confirms lung cancer susceptibility loci on chromosomes 5p15 and 15q25 in an African-American population. *Lung Cancer* **98**, 33-42 (2016).
28. Wang, Y. *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nature genetics* **46**, 736-741 (2014).
29. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
30. Wang, Y. *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet* **46**, 736-41 (2014).
31. Truong, T. *et al.* Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. *J Natl Cancer Inst* **102**, 959-71 (2010).
32. Zuber, V. *et al.* Pleiotropic Analysis of Lung Cancer and Blood Triglycerides. *J Natl Cancer Inst* **108**(2016).
33. Watza, D. *et al.* COPD-dependent effects of genetic variation in key inflammation pathway genes on lung cancer risk. *Int J Cancer* **147**, 747-756 (2020).
34. Dai, J. *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* **7**, 881-891 (2019).

35. van Rooij, F.J.A. *et al.* Genome-wide Trans-ethnic Meta-analysis Identifies Seven Genetic Loci Influencing Erythrocyte Traits and a Role for RBPMS in Erythropoiesis. *Am J Hum Genet* **100**, 51-63 (2017).
36. Li, Y. *et al.* FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. *BMC Bioinformatics* **17**, 122 (2016).
37. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).
38. Wang, X. Firth logistic regression for rare variant association tests. *Front Genet* **5**, 187 (2014).
39. Ma, C., Blackwell, T., Boehnke, M., Scott, L.J. & Go, T.D.i. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet Epidemiol* **37**, 539-50 (2013).
40. Dey, R. *et al.* Robust meta-analysis of biobank-based genome-wide association studies with unbalanced binary phenotypes. *Genet Epidemiol* **43**, 462-476 (2019).
41. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* **88**, 586-98 (2011).
42. Han, B. & Eskin, E. Interpreting meta-analyses of genome-wide association studies. *PLoS Genet* **8**, e1002555 (2012).
43. Bhattacharjee, S. *et al.* A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am J Hum Genet* **90**, 821-35 (2012).
44. Igl, B.W., Konig, I.R. & Ziegler, A. What do we mean by 'replication' and 'validation' in genome-wide association studies? *Hum Hered* **67**, 66-8 (2009).
45. Spitz, M.R. *et al.* Role of selected genetic variants in lung cancer risk in African Americans. *J Thorac Oncol* **8**, 391-7 (2013).
46. Machiela, M.J. & Chanock, S.J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555-7 (2015).
47. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-d1012 (2019).
48. Schumacher, F.R. *et al.* Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat Commun* **6**, 7138 (2015).
49. Doyle, G.A. *et al.* In vitro and ex vivo analysis of CHRNA3 and CHRNA5 haplotype expression. *PLoS One* **6**, e23373 (2011).
50. Tanner, J.A. *et al.* Novel CYP2A6 diplotypes identified through next-generation sequencing are associated with in-vitro and in-vivo nicotine metabolism. *Pharmacogenet Genomics* **28**, 7-16 (2018).
51. Kang, E.Y. *et al.* Meta-analysis identifies gene-by-environment interactions as demonstrated in a study of 4,965 mice. *PLoS Genet* **10**, e1004022 (2014).
52. Pena-Chilet, M. *et al.* Genetic variants in PARP1 (rs3219090) and IRF4 (rs12203592) genes associated with melanoma susceptibility in a Spanish population. *BMC Cancer* **13**, 160 (2013).
53. Chen, M.H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213 e14 (2020).
54. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214-1231 e11 (2020).
55. Astle, W.J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429 e19 (2016).
56. Liyanage, U.E. *et al.* Combined analysis of keratinocyte cancers identifies novel genome-wide loci. *Hum Mol Genet* **28**, 3148-3160 (2019).
57. Asgari, M.M. *et al.* Identification of Susceptibility Loci for Cutaneous Squamous Cell Carcinoma. *J Invest Dermatol* **136**, 930-937 (2016).
58. Chahal, H.S. *et al.* Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma. *Nat Commun* **7**, 12048 (2016).
59. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet* **51**, 237-244 (2019).

60. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet* **104**, 65-75 (2019).
61. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
62. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3 (2012).
63. Landi, M.T. *et al.* A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet* **85**, 679-91 (2009).
64. McKay, J.D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat Genet* **40**, 1404-6 (2008).
65. Hung, R.J. *et al.* Lung Cancer Risk in Never-Smokers of European Descent is Associated With Genetic Variation in the 5p15.33 TERT-CLPTM1Ll Region. *J Thorac Oncol* **14**, 1360-1369 (2019).
66. Shiraishi, K. *et al.* A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population. *Nat Genet* **44**, 900-3 (2012).
67. Hu, Z. *et al.* A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat Genet* **43**, 792-6 (2011).
68. Hsiung, C.A. *et al.* The 5p15.33 locus is associated with risk of lung adenocarcinoma in never-smoking females in Asia. *PLoS Genet* **6**(2010).
69. Schaid, D.J., Chen, W. & Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491-504 (2018).
70. Cannon, M.E. *et al.* Trans-ancestry Fine Mapping and Molecular Assays Identify Regulatory Variants at the ANGPTL8 HDL-C GWAS Locus. *G3 (Bethesda)* **7**, 3217-3227 (2017).
71. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet* **52**, 969-983 (2020).
72. Sun, R. *et al.* Integration of multiomic annotation data to prioritize and characterize inflammation and immune-related risk variants in squamous cell lung cancer. *Genet Epidemiol* **45**, 99-114 (2021).
73. Li, X. *et al.* A multi-dimensional integrative scoring framework for predicting functional variants in the human genome. *Am J Hum Genet* (2022).
74. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
75. Elman, J.S. *et al.* Identification of FUBP1 as a Long Tail Cancer Driver and Widespread Regulator of Tumor Suppressor and Oncogene Alternative Splicing. *Cell Rep* **28**, 3435-3449 e5 (2019).
76. Singer, S. *et al.* Coordinated expression of stathmin family members by far upstream sequence element-binding protein-1 increases motility in non-small cell lung cancer. *Cancer Res* **69**, 2234-43 (2009).
77. Man, K. *et al.* The transcription factor IRF4 is essential for TCR affinity-mediated metabolic programming and clonal expansion of T cells. *Nat Immunol* **14**, 1155-65 (2013).
78. Praetorius, C. *et al.* A polymorphism in IRF4 affects human pigmentation through a tyrosinase-dependent MITF/TFAP2A pathway. *Cell* **155**, 1022-33 (2013).
79. Shaffer, A.L. *et al.* IRF4 addiction in multiple myeloma. *Nature* **454**, 226-31 (2008).
80. Do, T.N., Ucisik-Akkaya, E., Davis, C.F., Morrison, B.A. & Dorak, M.T. An intronic polymorphism of IRF4 gene influences gene transcription in vitro and shows a risk association with childhood acute lymphoblastic leukemia in males. *Biochim Biophys Acta* **1802**, 292-300 (2010).
81. Zhang, T. *et al.* Cell-type-specific eQTL of primary melanocytes facilitates identification of melanoma susceptibility genes. *Genome Res* **28**, 1621-1635 (2018).
82. Visser, M., Palstra, R.J. & Kayser, M. Allele-specific transcriptional regulation of IRF4 in melanocytes is mediated by chromatin looping of the intronic rs12203592 enhancer to the IRF4 promoter. *Hum Mol Genet* **24**, 2649-61 (2015).
83. Tubbs, A. & Nussenzweig, A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **168**, 644-656 (2017).
84. Xia, J. *et al.* Bacteria-to-Human Protein Networks Reveal Origins of Endogenous DNA Damage. *Cell* **176**, 127-143 e24 (2019).
85. Liu, Y. *et al.* Rare deleterious germline variants and risk of lung cancer. *NPJ Precis Oncol* **5**, 12 (2021).

86. Gomperts, B.N. *et al.* Evolving concepts in lung carcinogenesis. *Semin Respir Crit Care Med* **32**, 32-43 (2011).
87. Miller, Y.E. Pathogenesis of lung cancer: 100 year report. *Am J Respir Cell Mol Biol* **33**, 216-23 (2005).

Methods

Ancestry-specific and cross-ancestry GWAS in lung cancer. There are 101,821 samples from 12 studies: Affymetrix Axiom Array Study (AFFY)[1], the Female Lung Cancer Consortium in Asia (FLCCA)[2], the Genetic Epidemiology of Lung Cancer Consortium (GELCC)[3], the Environment and Genetics in Lung cancer Etiology study (EAGLE)[4,5], Helmholtz-Gemeinschaft Deutscher Forschungszentren Lung Cancer GWAS (GERMAN)[5,6], the International Agency for Research on Cancer (IARC)[5], the Institute of Cancer Research (ICR)[5], MD Anderson Cancer Center Study (MDACC)[5,7], NCI Lung Cancer and Smoking Phenotypes in African-American Cases and Controls (NCI)[8], OncoArray Consortium Lung Study (OncoArray)[5,9], the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO)[5], and Samuel Lunenfeld Research Institute Study (SLRI)[5] (**Supplementary note, Supplementary table 25**). Markers from various genotyping platforms were filtered based on the following criterion: only biallelic marker, call rate ≥ 0.95, and MAF > 0 in each study. Markers were further checked using McCarthy Haplotype Reference Consortium (HRC) imputation preparation and checking tool (v4.2.11, https://www.well.ox.ac.uk/~wrayner/tools/) to make strand, position, ref/alt assignment consistent with HRC reference panel[10]. We conducted imputation of the phased data through Sanger imputation service in a two-stage strategy of pre-phasing and imputation using SHAPEIT2 (v2.r790) and PBWT (2014). The reference panel was HRC (r1.1), which contains 32,470 samples of predominantly European ancestry and about 40 million markers.

There were 2,854,462 common markers with information score of greater than or equal to 0.6 among 12 studies and were further thinned to 193,050 markers based on r-square value of less than or equal to 0.5. The new set of 193,050 markers was used to calculate principal components and pair-wise identity by descent (IBD) values among 101,821 samples in PLINK. An empirical value of IBD of 0.15 was used as a cutoff to define samples' related status, and all related samples were categorized into 15,884 clusters. While priority of sample was quantified by scoring properties such as disease status and study specific measurement such as average imputation information score in each cluster and samples with missing disease status were assigned the lowest priority. Lists of independent or less-independent samples were generated and sorted by the total priority score. 70,639 samples with the highest scores in each cluster were finally generated for analysis through clustering and sampling process (**Supplementary figure 4, Supplementary table 26**).

Inference of ancestry memberships. 2,042 ancestry informative markers shared by 70,639 samples and 505 HapMap2 samples of CEU (EUR), CHB (EAS) and YRI (AFR) ancestry were used to infer ancestry origins using FastPop[11], and then 51,961 samples of EUR origin, 12,434 samples of EAS origin and 5,761 samples of AFR origin were inferred (**Extended Data figure 3**). 15,265 cases of ADE, 7,850 SQC, and 2,482 SCC were defined based on available histological information (**Supplementary figure 4**).

Statistical analysis for ancestry- and cross-ancestry GWAS. About 6 million markers having information score ≥ 0.4 were analyzed using logistic regression method using R glm function (R3.6.0) and markers were filtered for each stratum by population and histological subtypes at $P < 1.0 \times 10^{-4}$. The total number of unique 49,576 markers were analyzed using Firth's logistic regression procedure with the option of "firth" and "maxinter" using SAS (version 9.4). The first 20 principal components and 12 study sites as categorical variables were included in the model. Histological subtype-specific analyses were performed in each racial population including European-, East Asian-, and African-ancestry, respectively. Trans-ancestry genome-wide meta-analyses were further performed using METASOFT with binary random-effects meta-analysis

model[12]. All statistical tests conducted in the multi-ancestry GWMA were two-sided. The METASOFT software provides four different meta-analysis methods; fixed-effects model (FE), conventional random-effects model (RE), new random-effects model optimized to detect associations under heterogeneity (RE2)[13], and a new random binary-effects model optimized to detect associations when some studies have an effect and some studies do not (BE)[12] along with M-values, i.e., the posterior probabilities that the effects exist in the populations being studied. M-values using cross-population information can be simply interpreted. An ancestry-specific study is predicted to have an effect, if M-value ≥ 0.9 and no effect, if M-value ≤ 0.1; all other values have ambiguous predictive power. In addition, we implemented MANTRA with Bayes Factor[14], across diverse populations.

Since the genomic inflation factor (lambda, λ) increases with sample size, we rescaled the observed lambda value (λ_{obs}) to the adjusted one (λ_{adj}) reflecting a standardized sample size of 1,000 cases and 1,000 controls based on the following formula,

$$\lambda_{adj} = 1 + (\lambda_{obs} - 1) \times \frac{\frac{1}{N.cases} + \frac{1}{N.controls}}{\frac{1}{1000} + \frac{1}{1000}}.$$

In addition, we conducted post-imputation quality control using a two-proportion Z-test ("prop.test" in R v3.6.2) for missingness rate between cases and controls on the genotyped samples at the threshold of $|Z| > 9.336$ to minimize false-positive findings from ancestry-specific GWAS. For the East Asian samples, this Z-value corresponded to less than 0.1% of tests and corresponds to a nominal P-value less than 1×10^{-20}, which is conservative but allows for the large number of tests we conducted. We applied the exact test of Hardy-Weinberg equilibrium[15,16] ("HardyWeinberg" package in R v3.6.2) in controls, stratified by ancestry, to reduce the false-positive trans-ancestry association signals, and variants with a mid-P adjustment threshold < $1x10^{-8}$ in controls were excluded.

Validation of ancestry-specific and cross-ancestry GWMA. Validation analyses for thirty-nine novel variants and sixteen new and nineteen known variants in the previously reported loci identified in trans-ancestry GWMA were conducted in seven independent genome-wide association studies comprising 25,315 cases and 912,813 controls (**Table 1**, **Supplementary table 1**). This work includes UK Biobank[17], FinnGen, deCODE study[18,19], SPAIN study[20,21], INHALE study[22], China Nanjing Medical University Lung study[23], and KPRB/GERA study[24].

Ethics statement. All participants provided informed consents according to protocols that were evaluated by the Internal Review Boards (IRB) of the contributing centers. All contents in the present study were approved by Baylor College of Medicine IRB.

Conditional analysis on the cross-ancestry lead SNPs. Conditional analysis has been used as a tool to identify secondary association signals at a locus, involving association analysis conditioning on the primarily associated SNP at the locus to test whether there are any other SNPs significantly associated. A comprehensive strategy is to conduct a conditional analysis, starting with the sentinel trans-ancestry GWMA SNP for each locus, across the whole genome followed by a stepwise procedure of selecting additional SNPs, one by one, according to their conditional P-values. Such a strategy would enable us to detect more than two independently associated SNPs at a locus. We adopted a genome-wide stepwise selection procedure to select SNPs on the basis of conditional P-values in each population using GCTA v1.93 (--cojo-cond)[25] and then meta-analyzed across all three populations. Conditional analysis of each associated locus was performed within a

standard region of 1 Mb-window centered on the lead SNP, which is the most associated SNP in lung cancer. LD patterns were estimated using best guess genotype data in each population consisting of 31,016 Europeans, 3850 East Asians, and 557 Africans, from Oncoarray data as reference. To extract best guess genotype data (FORMAT ID=GT) per population from the variant call format file, we implemented PLINK 1.9 using --vcf onco.vcf \ --keep sample list from each population (for example, CEU.list for European ancestry) \ --make-bed \ --out output file (for example, OncoCEU). Conditional association analysis per population was performed including the lead SNP as covariate. Any SNP showing a conditional association $P < 5 \times 10^{-8}$ was considered as independent signal and was further included in a new round of conditional analysis. This process was repeated until no SNP with $P < 5 \times 10^{-8}$ remained in any of the genomic regions explored.

Fine-mapping of cross-ancestry GWAS loci. As described above, trans-ancestry meta-analysis using MANTRA was performed in each histological subtype following the ancestry-specific analysis using Firth's logistic regression for 49,576 variants. MANTRA facilitates cross-ancestry analysis by assuming that allelic effects of a variant are the same within the cluster of individuals sharing similar ancestry but different between individuals in separate ancestry clusters[14]. To define credible causal variants for each locus, we ranked the variants within +/- 250kb of the lead variant based on their Bayes factors generated from MANTRA analysis. Credible set value for each variant was then calculated by dividing cumulative Bayes factor of ranked variants by the total cumulative Bayes factor in each locus. Variants within the credible set value of 0.99 (including the first one that goes above the cutoff) were defined as 99% credible set[26,27].

Integrative multi-omic annotation analysis. We annotated the fine-mapped variants (within 99% credible set) using Multi-dimensional Annotation Class Integrative Estimator (MACIE)[28] and the Functional Annotation of Variants – Online Resource (FAVOR) platform[29] (http://favor.genohub.org/) for further prioritization. We integrated a variety of variant functional annotations in a generalized linear mixed model (GLMM) approach[28,29]. The Multi-dimensional Annotation Class Integrative Estimator (MACIE) models the regulatory and evolutionary conserved functionality of individual variants using two latent binary classes. Random effects are used to account for correlations among 8 annotations that are modeled as a function of the conserved class as well as 28 annotations that are modeled as a function of the regulatory class. Estimation occurs using an EM algorithm. The fitted model parameters are first found using a training dataset, and then one additional iteration of the EM algorithm is performed using these fitted parameters on the new SNPs of interest identified in this work. The MACIE output is a vector of 2*2 probabilities corresponding to the probabilities of belonging to both functional classes, either one of the classes alone, or neither class. The probabilities necessarily sum to 1. Marginal probabilities of regulatory function or evolutionarily conserved function can be found by simply adding two of the four probabilities. Formulating functionality as a set of multiple characteristics offers a more versatile and more detailed prediction than other integrative methods that produce a one-dimensional score that can be difficult to interpret. Specially, the model treats functionality as an unobserved latent class and predicts (1) the probability of regulatory class only (MACIE01), (2) the probability of evolutionarily conserved class only (MACIE10), (3) the probability of neither class (MACIE00), or (4) the probability of both functional classes (MACIE11). Variants displaying a marginal probability score higher than 0.9 for regulatory function or evolutionary conservation from MACIE analysis were reported together with the detailed annotation obtained from FAVOR analysis.

Combined eQTL-based analysis. We searched significant eQTL genes for high-LD variants from each GWAS locus using functional mapping and annotation (FUMA) platform[30]. We set the LD cutoff at $r^2 > 0.6$ with the lead SNP (1000 Genomes, all populations) and eQTL cut off at FDR < 0.05. Based on the prior knowledge, eQTL datasets for three lung-cancer-relevant tissue types (lung, blood – based on the contribution of inflammation and immune cells, and brain – based on the contribution of smoking behaviors to the etiology) were selected from 4 different studies (eQTLGen, BIOSQTL, BRAINEAC, and GTEx v8) (details in **Supplementary table 14**).

Colocalization between GWAS and GTEx eQTL signals. The GTEx v8 includes data from normal tissues from 838 donors. GTEx eQTL association data for variants within +/− 100kb windows of the lead variants presented in the GWAS were extracted. For those loci overlapping MHC regions with an extended LD and high density of variants, we narrowed down to +/− 10kb windows of the lead variants. Colocalization of the GWAS and eQTL signals were calculated using the LD-dependent (eCAIVAR) and LD-independent approach (coloc). Given that both the study population in GWAS (74%) and eQTL (85%) is mainly European, European LD matrix from 1000 Genomes (phase 3)[31] was incorporated into the LD-dependent (eCAIVAR) approach[32]. We allowed the maximum number of 2 candidate 'causal' SNPs in eCAVIAR. To avoid potentially spurious colocalizations due to the violation of common LD assumption in all three datasets[32] (GWAS, eQTL, LD matrix), we applied the LD-independent approach (coloc)[33] to find concordance with the results from eCAVIAR. In coloc, we used the nominated *P*-value and MAF of complete GWAS population as the inputs. We only considered the colocalizations when both eCAVIAR and coloc suggested plausible posterior probability (CLPP>0.01 and PPH_4>0.8).

Human cell lines, plasmids, and other reagents. MRC5-SV40, an SV40-immortalized human lung fibroblast cell line was maintained in standard Dulbecco's modified Eagle's medium (#11965118, Gibco) with 10% fetal bovine serum (#10438034, Gibco), 2 mM L-glutamine, 100µg/mL penicillin, and 100 µg/mL streptomycin (#10378016, Gibco) as described in [34]. The MRC5-SV40 cell line was authenticated by American Type Culture Collection (ATCC) Short Tandem Repeat (STR) analysis and routinely check to be mycoplasma-free. Gateway entry clones for the following genes: *ACTR2*, *PPIL6*, *AK9*, *CYP21A2*, *VARS2*, *CCHCR1*, *SECISBP2L*, *MPZL3*, and *DCBLD1* were synthesized, sequence-verified, and cloned into pDONR223 by GenScipt. IRF4 (ccsbBroadEn_06459), FUBP1 (IOH14097), CYP2A6 (IOH63274), FLOT1 (IOH4826), SFTA2 (ccsbBroadEn_13655), CCDC97 (ccsbBroadEn_04511), C6orf48 (IOH13777), PSMA4 (ccsbBroadEn_06796), LY6G5B (IOH59693), STN1 (ccsbBroadEn_08997) entry clones were acquired from the Kenneth Scott cDNA ORF library (Baylor College of Medicine). All of the above clones were further subcloned into an N-terminal GFP tagged vector (pcDNA6.2/N-EmGFP-DEST, Invitrogen), using Gateway LR Clonase II Enzyme Mix (# 11791020, Invitrogen). Plasmid transfections were performed using GenJet In Vitro DNA Transfection Reagent Ver. II (# SL100489, SignaGen).

Non-targeting pool siRNA (D-001810-10, Dharmacon) and SMARTpool siRNAs each containing four targeting sequences were purchased from Dharmacon (**Supplementary table 21**). siRNA transfections were carried out with lipofectamine RNAiMax Transfection Reagent (#13778075, Invitrogen), following the manufacturer's recommendations. SMARTpool ON-TARGET*plus* siRNAs were designed and modified for greater specificity and reduce off-targets up to 90%, although

further validation of the phenotype using individual siRNAs or CRISPR/Cas9 editing is required for more in-depth functional studies.

RT-qPCR. Total RNA was extracted by RNeasy mini kit (#74106, Qiagen) from cells 72 hours post siRNA transfection. 300 ng of total RNA from each sample was used to synthesize cDNA by the Superscript III first strand synthesis system (#18080051, Invitrogen). qPCR reactions were performed using iTaq Universal SYBR Green Supermix (#172-5121, Bio-Rad Laboratories) on a QuantStudio 3 Real-Time PCR System (Applied Biosystems). For each gene, three replicates were analyzed and the average threshold cycle (Ct) was calculated. The relative expression levels were calculated with the $2^{-\Delta\Delta Ct}$ method [35]. Primers used in this study are listed in **Supplementary table 22**. siRNA knockdown efficiencies were summarized in **Supplementary table 23**.

Flow-cytometric DNA damage assays. DNA damage assays by flow cytometry[34] were performed as follows: approximately 1 million MRC5-SV40 cells were harvested and prepared for staining 72hours post-transfection with siRNAs or GFP-fusion plasmids. Cells were fixed with 2% formaldehyde for 15 minutes on ice, washed twice in cold-PBS and permeabilized with 0.05% Triton-X for 15 minutes on ice followed by two washes with PBS. The fixed cells were then blocked with 5% BSA-PBS for 30 minutes and stained with γH2AX primary antibody (#05-636, Sigma, 1:750) for 1 hour at room temperature. Cells were further washed three times in 1% BSA-PBS followed by an incubation of Alexa Fluor 647 goat anti-mouse IgG secondary antibody (#A21236, Thermo Fisher, 1:1000) in 5% BSA-PBS for an additional hour at room temperature in the dark, before the flow cytometry acquisition and analyses by a BD LSRFortessa flow cytometer. FCS files were analyzed by FlowJo 10.6 software. For siRNA experiments, cells were collected 72 hours post transfection and mock cells with top 0.5% γH2AX signal were gated as the γH2AX-high cells as previously described [34]. The percentage of γH2AX positive cells in each sample was calculated and normalized to its corresponding non-targeting siRNA control. In addition, median γH2AX intensity for each siRNA knockdown was calculated and normalized to the non-targeting siRNA control.

For overproduction experiments, mock-transfected cells were used to set the gates to determine the GFP and γH2AX positive cells. Mock cells with top 0.5% γH2AX signal were gated as the γH2AX-high cells similar to the siRNA experiments. GFP positive cells were gated and the median γH2AX intensity of each overproducing candidate was calculated and normalized to GFP-Tubulin. In addition, the fraction of the γH2AX positive cells in the GFP positive population for each genotype was calculated and normalized to the fraction of γH2AX positives in the GFP-Tubulin overproducing cells.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data Availability

The following publicly available datasets were used in this work: dbGaP datasets (PLCO study, phs000093.v2.p2; FLCCA study, phs000716.v1.p1; EAGLE study , phs000336.v1.p1; NCI study of African-Americans, phs001210.v1.p1; German, SLRI, IARC and MDACC studies, phs000876.v2.p1; Oncoarray study, phs001273.v3.p2; imputed Oncoarray study using HRC reference panel, phs001273.v4.p2; Affymetrix study, phs001681.v1.p1). The ICR study from the 1958 Birth Cohort from the UK does not allow the general upload of findings. Therefore, this data set is available after request

from Richard Houlston (Richard.Houlston@icr.ac.uk). The individual-level genotype and phenotype data are available through formal application to the UK Biobank (https://www.ukbiobank.ac.uk/). The GWAS summary statistics used in validation study were downloaded from FinnGen (https://finngen.gitbook.io/documentation/v/r5/data-download) and the pan-cancer pleiotropy study (https://github.com/Wittelab/pancancer_pleiotropy). The GWAS summary statistics of the candidate 45 variants identified from the discovery phase were obtained following our request from M.Z. and H.S. (China NJMU lung study), T. R. (deCODE and SPAIN lung study), and A. S. and C. L. (INHALE study) and are available in the supplementary table. The eQTL data from GTEx v8 was obtained from https://www.gtexportal.org/home/datasets. The Icelandic population WGS genetic but not phenotypic data have been deposited at the European Variant Archive under accession code PRJEB15197. (https://www.ebi.ac.uk/ena/browser/view/PRJEB15197?show=analyses). Results from GWMA at $P \leq 10^{-5}$ are available in the supplementary tables. All sequencing reads were mapped to the GRCh37/hg19 human reference genome. More details of data source used in this work are provided in the paper and supplementary tables.

Code Availability

We performed our analyses using the following publicly available software/packages: SHAPE-IT2 (v2.r790; https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html), McCarthy Group Tools (v4.2.11; https://www.well.ox.ac.uk/~wrayner/tools/), PBWT (https://github.com/richarddurbin/pbwt), and Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html#!) were used for imputation and phasing, FastPop (https://github.com/biomedicaldatascience/FastPop4) and KING (v2.0, http://people.virginia.edu/~wc9c/KING/) for population stratification and relatedness analyses, SAS (v9.4, https://www.sas.com/en_us/home.html), R (v3.6.2, https://cran.r-project.org), PLINK (v1.9 and 2.0, https://www.cog-genomics.org/plink/1.9/ and https://www.cog-genomics.org/plink/2.0/), METASOFT and ForestPMPlot (v2.0.1 and v1.0.3, http://genetics.cs.ucla.edu/meta/), and GCTA (v1.93, https://cnsgenomics.com/software/gcta/) were used for data and statistical analyses, FUMA (v1.3.6, https://fuma.ctglab.nl/), FAVOR (https://favor.genohub.org/), GTEx (v8, https://www.gtexportal.org/home/), coloc (v3.2-1, https://cran.r-project.org/web/packages/coloc/), eCAVIAR (v2, http://zarlab.cs.ucla.edu/tag/caviar/), IPA (https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis), ezQTL (v1.0, https://analysistools.cancer.gov/ezqtl/#/home) were used for post-GWAS analyses, and flowjo (v10.6, https://www.flowjo.com) was used for single-cell flow cytometry analysis. MANTRA (version 1) is available as a suite of executables on request from the author (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3460225/pdf/gepi0035-0809.pdf).

Methods-only references

1. Ji, X. *et al.* Protein-altering germline mutations implicate novel genes related to lung cancer development. *Nat Commun* **11**, 2220 (2020).
2. Lan, Q. *et al.* Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat Genet* **44**, 1330-5 (2012).
3. Byun, J. *et al.* Genome-wide association study of familial lung cancer. *Carcinogenesis* **39**, 1135-1140 (2018).
4. Landi, M.T. *et al.* Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. *BMC Public Health* **8**, 203 (2008).

5. McKay, J.D. *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).
6. Landi, M.T. *et al.* A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet* **85**, 679-91 (2009).
7. Ji, X. *et al.* Identification of susceptibility pathways for the role of chromosome 15q25.1 in modifying lung cancer risk. *Nat Commun* **9**, 3221 (2018).
8. Mitchell, K.A. *et al.* Relationship between West African ancestry with lung cancer risk and survival in African Americans. *Cancer Causes Control* **30**, 1259-1268 (2019).
9. Amos, C.I. *et al.* The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev* **26**, 126-135 (2017).
10. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).
11. Li, Y. *et al.* FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. *BMC Bioinformatics* **17**, 122 (2016).
12. Han, B. & Eskin, E. Interpreting meta-analyses of genome-wide association studies. *PLoS Genet* **8**, e1002555 (2012).
13. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* **88**, 586-98 (2011).
14. Morris, A.P. Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* **35**, 809-22 (2011).
15. Wigginton, J.E., Cutler, D.J. & Abecasis, G.R. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**, 887-93 (2005).
16. Graffelman, J. & Moreno, V. The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Stat Appl Genet Mol Biol* **12**, 433-48 (2013).
17. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
18. Wang, Y. *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nature genetics* **46**, 736-741 (2014).
19. Rafnar, T. *et al.* Variants associating with uterine leiomyoma highlight genetic background shared by various cancers and hormone-related traits. *Nat Commun* **9**, 3636 (2018).
20. Truong, T. *et al.* Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. *J Natl Cancer Inst* **102**, 959-71 (2010).
21. Zuber, V. *et al.* Pleiotropic Analysis of Lung Cancer and Blood Triglycerides. *J Natl Cancer Inst* **108**(2016).
22. Watza, D. *et al.* COPD-dependent effects of genetic variation in key inflammation pathway genes on lung cancer risk. *Int J Cancer* **147**, 747-756 (2020).
23. Dai, J. *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* **7**, 881-891 (2019).
24. Rashkin, S.R. *et al.* Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat Commun* **11**, 4423 (2020).
25. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
26. Schaid, D.J., Chen, W. & Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491-504 (2018).
27. Cannon, M.E. *et al.* Trans-ancestry Fine Mapping and Molecular Assays Identify Regulatory Variants at the ANGPTL8 HDL-C GWAS Locus. *G3 (Bethesda)* **7**, 3217-3227 (2017).
28. Sun, R. *et al.* Integration of multiomic annotation data to prioritize and characterize inflammation and immune-related risk variants in squamous cell lung cancer. *Genet Epidemiol* **45**, 99-114 (2021).
29. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet* **52**, 969-983 (2020).
30. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).

31. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
32. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet* **99**, 1245-1260 (2016).
33. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet* **16**, e1008720 (2020).
34. Xia, J. *et al.* Bacteria-to-Human Protein Networks Reveal Origins of Endogenous DNA Damage. *Cell* **176**, 127-143 e24 (2019).
35. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-8 (2001).