Genetic Ancestry Contributes to Somatic Mutations in Lung Cancers from Admixed Latin American Populations

Jian Carrot-Zhang1,2,3, Giovanny Soca-Chafre4, Nick Patterson2,3, Aaron R. Thorner1, Anwesha Nag1, Jacqueline Watson4,2, Giulio Genovese2,3, July Rodriguez2, Maya K. Gelbard1, Luis Corrales-Rodriguez6,7, Yoichiro Mitsuishi8, Gavin Ha9, Joshua D. Campbell10, Geoffrey R. Oxnard1, Oscar Arrieta4,11, Andres F. Cardona5,12, Alexander Gusev1,2,13, and Matthew Meyerson1,2,3

ABSTRACT

Inherited lung cancer risk, particularly in nonsmokers, is poorly understood. Genomic and ancestry analysis of 1,153 lung cancers from Latin America revealed striking associations between Native American ancestry and their somatic landscape, including tumor mutational burden, and specific driver mutations in EGFR, KRAS, and STK11. A local Native American ancestry risk score was more strongly correlated with EGFR mutation frequency compared with global ancestry correlation, suggesting that germline genetics (rather than environmental exposure) underlie these disparities.

SIGNIFICANCE: The frequency of somatic EGFR and KRAS mutations in lung cancer varies by ethnicity, but we do not understand why. Our study suggests that the variation in EGFR and KRAS mutation frequency is associated with genetic ancestry and suggests further studies to identify germline alleles that underpin this association.

See related commentary by Gomez et al., p. 534.

INTRODUCTION

Lung cancer causes more than 1.7 million deaths per year worldwide (1) and kills more people than any other malignancy in Latin America (2). Lung adenocarcinoma is the most common subtype of lung cancer that is typically driven by genomic alterations of genes in the receptor tyrosine kinase (RTK)/RAS/RAF pathway (3), often allowing effective therapeutic targeting by RTK inhibitors and other pathway inhibitors. It is well known, but mysterious, that the frequency of somatic EGFR mutations is higher in lung adenocarcinomas from patients in East Asia (∼45%) compared with lung adenocarcinomas from patients in Europe or patients of European (EUR) and/or African (AFR) descent in North America (∼10%; refs. 4–8). In Latin American countries, the frequency of somatic EGFR mutations in lung adenocarcinoma ranges from roughly 14% in Argentina, to 25% to 34% in Colombia, Brazil, and Mexico, to 51% in Peru (refs. 9–11; Fig. 1A). Moreover, recent genomic studies from East Asian (EAS) and AFR populations have suggested different distributions of tumor mutation
Results

To explore the landscape of somatic cancer mutation in lung cancers from Latin America and to assess the influence of germline ancestry of genetically admixed patient populations on these somatic alterations, we performed genomic analysis of 601 lung cancer cases from Mexico and 552 from Colombia, including 499 self-reported nonsmokers (Supplementary Table S1). Next-generation sequencing targeting a panel of 547 cancer genes plus intronic regions of 60 cancer genes (16) was used to identify single-nucleotide variants (SNV), insertions/deletions (indels), SCNA, and gene fusions. This gene panel covers all currently known lung cancer drivers (3), which are the focus of this work. Because we do not have matched germline samples, we applied a custom script to identify known, hotspot lung cancer driver mutations for the full 1,153 samples to ensure the sensitivity for low-coverage samples, as well as to avoid potential germline contamination (Methods). We found that 552 (48%) of all samples harbored oncogenic mutations in EGFR, KRAS, BRAF, ERBB2, or MET, or fusions in ALK, ROS1, or RET; 785 of 1,153 samples harbored at least one detectable alteration in a broader set of known lung cancer driver genes also including TP53, STK11, KEAP1, SMARCA4, SETD2, MYC, and MDM2 (Fig. 2; Supplementary Tables S2 and S3). The detected mutation frequencies of EGFR and KRAS were 30% and 10%, respectively, in the tested lung cancer samples from Mexican patients, and 23% and 13%, respectively, in the tested lung cancer samples from Colombian patients. SCNA analysis (see Methods) identified 9% and 2% cases with high-level amplifications in MYC and MDM2, respectively. We did not observe novel amplification or deletion peaks in this Latin American lung cancer cohort as assessed by GISTIC analysis.

Ancestry effects on somatic cancer genomes are under-studied (17, 18), and few genomic data sets have been developed from patients with lung cancer with mixed ancestry. One potential source of samples for analysis of ancestry effects is discarded tissue or nucleic acids, left over after the clinical analysis of cancer samples. Here, we developed an analytic pipeline (https://github.com/carrotzhang/ancestry-from-panel) that offers the advantage of simultaneous measurement of global and local ancestry from sequencing tumor DNA only, without requiring matched germline samples (Supplementary Fig. S1A–S1C). Briefly, we called the geno-type of germline single-nucleotide polymorphisms (SNP) using on-target and off-target reads from the sequencing panel and measured global ancestry based on principal component analysis (PCA; ref. 19) of the germline SNPs, in which principal components (PC) 1, 2, and 3 captured prominently the axis of AFR, EUR, and NAT ancestry, respectively (Supplementary Fig. S2A). We then imputed missing SNPs using an external haplotype reference panel (15) and assigned local ancestry to each genomic region (20), based on the imputed variants. We validated our approach to ancestry analysis by performing whole-genome sequencing on a subset of 44 tumor samples, and SNP genotyping on 12 paired tumor and normal samples (Supplementary Fig. S2B and S2C). We found a high accuracy of our off-target reads approach based on panel sequencing of tumor DNA, by comparing tumor and normal ancestry estimations (Pearson r > 0.99), and by comparing panel sequencing to whole-genome sequencing (Pearson r > 0.96). As panel testing of cancer genes has emerged as a practical diagnostic tool in clinical care, our
Genetic Ancestry Affects Somatic Alterations in Lung Cancers

Figure 2. Somatic genome analysis of lung cancers from Mexico and Colombia. Comutation plot displays alterations in known activators of the RTK/RAS/RAF pathway, tumor suppressor genes, and significantly amplified genes. *, Indicates that oncogenic mutations in EGFR and KRAS as well as truncating mutations in STK11 are associated with NAT ancestry, but other somatic alterations are not. Correlations with mutations are controlled for TMB.

We conclude that NAT ancestry was associated with genomic alterations by reanalyzing large, existing somatic sequencing datasets without matched germline sequencing data.

Having obtained data on both somatic alteration and genetic ancestry, our next step was to assess the correlation of these features, using multivariable regression controlling for self-reported smoking status and country of sample collection (see Methods). As previous work focused on differences between populations, these associations with ancestry within a single admixed population provide more direct evidence of a putative genetic cause. First, we found a significant anti-correlation between TMB and PC3 representing the NAT ancestry ($P \approx 9 \times 10^{-7}$, $\text{coef.} = -0.02$), in line with previous study of lung cancers from EAS patients (12); no correlation was found with the total SCNA burden or with aneuploidy.

Evaluation of ancestry–mutation association, adjusting for sample-specific TMB, in each gene from Fig. 2 showed that NAT ancestry was positively correlated with mutations in EGFR (FDR corrected $P \approx 9 \times 10^{-6}$, $\text{coef.} = 0.005$) and anticorrelated with mutations in KRAS (FDR corrected $P = 9 \times 10^{-3}$, $\text{coef.} = -0.007$), and mutations in STK11 (FDR corrected $P = 7 \times 10^{-4}$, $\text{coef.} = -0.013$), in line with previous studies focusing on EAS patients (12, 21). Each feature (TMB, EGFR, KRAS, or STK11 mutation) was independently associated with NAT ancestry in a joint model (Supplementary Table S4). In never-smoking patients, the TMB and NAT ancestry association was stronger in EGFR-mutant ($P = 0.002$, $\text{coef.} = -0.301$) than in EGFR–wild-type ($P = 0.038$, $\text{coef.} = -0.013$). Moreover, the interaction of EGFR and NAT ancestry was significantly associated with TMB ($P = 0.04$, $\text{coef.} = -0.22$) in a joint model (TMB × NAT ancestry + EGFR + EGFR* NAT ancestry), suggesting that the association of TMB and NAT ancestry is different in EGFR-mutant and EGFR–wild-type samples. Furthermore, we demonstrated that NAT ancestry was predominately associated with oncogenic, driver mutations in EGFR, but not with nononcogenic, passenger mutations (Supplementary Fig. S3). In addition, we did not observe SCNA of any lung cancer driver gene associated with ancestry (Methods). The observed correlations held in separate analyses of the Mexican and Colombian cohorts (Supplementary Fig. S4A and S4B).

To better understand the relationship of ancestry and exposure-induced mutagenesis in the risk of developing lung adenocarcinomas through the activation of the RTK/RAS/RAF pathway, we tested the ancestry associations with RTK/RAS/RAF pathway oncogene alterations adjusting for mutational signatures (Supplementary Table S5). The positive correlation of NAT ancestry with $EGFR$ mutation ($OR = 1.23$ in every $10\%$ increase of NAT ancestry; $95\% CI, 1.12–1.35$), and negative correlation of NAT ancestry with $KRAS$ mutation ($OR = 0.85$ in every $10\%$ increase of NAT ancestry; $95\% CI, 0.77–0.95$) remained significant (Fig. 2A and B). The association of NAT ancestry with $EGFR$ mutation was also observed in an analysis restricted to patients who reported themselves as never-smokers ($OR = 1.46$ in every $10\%$ increase of NAT ancestry; $95\% CI, 1.25–1.70$; Supplementary Fig. S5). This association was also observed in an analysis restricted to patients who reported themselves as smokers ($OR = 1.45$ in every $10\%$ increase of NAT ancestry; $95\% CI, 1.25–1.70$; Supplementary Fig. S5). The association of NAT ancestry with $KRAS$ mutation rate increased with the proportion of smoking signature ($OR = 1.27$ in every $10\%$ increase of smoking signature; $95\% CI, 1.04–1.56$; Fig. 2B). When including smokers, $KRAS$ mutation rate increased with the proportion of smoking signature ($OR = 1.27$ in every $10\%$ increase of smoking signature; $95\% CI, 1.04–1.56$; Fig. 3B). The ancestry effect on $KRAS$ mutations in reported never-smokers tended toward significance but was not significant ($P = 0.08$) in this study, perhaps due to sample size ($n = 387$). Moreover, the interaction of smoking signature and NAT ancestry did not modify the effect size of ancestry on $KRAS$ ($P = 0.34$; Methods). Gender and the APOBEC signature were not associated with mutations of any lung cancer oncogenes. Age of diagnosis was negatively associated with the risk of $ALK$-translocated cases ($P = 3 \times 10^{-5}$, $OR = 0.97$ in every $10\%$ increase of age; $95\% CI, 0.96–0.99$). Together, we conclude that NAT ancestry was associated with genomic...
differences in Latin American patients with lung adenocarcinoma that are independent of smoking activity.

To assess whether the observed association with EGFR and KRAS mutations is due to NAT ancestry itself or to an environmental exposure/socioeconomic status related to NAT ancestry, we next investigated the influence of local ancestry. Previous work has shown that associations between local ancestry and phenotype (while accounting for global ancestry) provide evidence of a genetically driven phenotype, as local ancestry is not expected to be causally associated with environmental exposure or socioeconomic status (22, 23). We used RFMix (20) to map local ancestry, producing 5,425 genomic regions with an assignment of AFR, EUR, or NAT ancestral population for each parental chromosome (Supplementary Table S6–S7). We performed a multivariable logistic regression of NAT ancestry for each genomic region correlating with the EGFR-mutant or KRAS-mutant samples, controlling for the global ancestry (Methods). We did not identify any region where correlation reached genome-wide significance of $P < 5 \times 10^{-5}$ (Fig. 4A; Supplementary Fig. S6).

We next evaluated whether local ancestries across multiple sub-genome-wide significance threshold regions ($5 \times 10^{-5} < P < 0.05$) were associated with the somatic mutation phenotype by constructing a polygenic ancestry score. This approach is

---

**Figure 3.** Targetable LUAD driver genes associated with genetic ancestry. A, The percentage of NAT germline ancestry is positively correlated with the percentage of somatic EGFR mutations, and negatively correlated with the percentage of somatic KRAS mutations. Color bar represents the number of samples in the NAT ancestry percentage range. B, Association of targetable LUAD driver genes with NAT ancestry, mutational signature, and gender (n = 705; left), and the association in never-smokers only (n = 387; right). Multivariable logistic regression P-values are shown, with NAT ancestry percentage, gender, smoking, and APOBEC signature as covariates. Red dots represent $P < 0.05$. Lines represent 95% confidence intervals.

**Figure 4.** Germline local ancestry in association with somatic EGFR and KRAS mutations. A, Genome-wide association of local NAT ancestry with EGFR (left) and KRAS (right). “NAT ancestry high” indicates positive association, whereas “NAT ancestry low” indicates negative association. Red line indicates $P = 0.05$. Orange line indicates genome-wide significance threshold ($P < 5 \times 10^{-5}$). B, Association of local ancestry risk score with somatic EGFR or KRAS mutations, controlling for global ancestry (proportion of overall NAT ancestry).
Genetic Ancestry Affects Somatic Alterations in Lung Cancers

Genetic Ancestry Affects Somatic Alterations in Lung Cancers populations. mutations in the Latin American to harbor KRAS ancestry may modulate the evolution of lung cancer tumors finding suggests that one or more genetic loci specific to NAT Latin American patients. comprehensively characterize lung cancer genomes from lung cancer driver genes. Future studies will be needed to hotspot mutations and protein-truncating mutations in and the use of panel sequencing data, we tested only known, try may thus be even more significant than we have observed. The ancestry and TMB association was independent of smoking-related mutational processes, and there- mutations in a joint model with global ancestry and somatic mutations in EGFR locus and somatic mutations in EGFR (P = 0.8). Moreover, when including the local ancestry of the EGFR locus as a covariate, the association of EGFR mutations and the local ancestry risk score remained significant (P = 4 × 10⁻⁶). Our finding suggests that one or more genetic loci specific to NAT ancestry may modulate the evolution of lung cancer tumors to harbor EGFR or KRAS mutations in the Latin American populations.

DISCUSSION

In summary, the genomic landscape of lung adenocarcinomas is strikingly varied in Latin American patients with mixed ancestries. In our study of 1,153 lung cancers, we demonstrated that NAT ancestry was correlated with somatic driver alterations, including EGFR and KRAS mutations that can be effectively targeted by small-molecule inhibitors to prolong survival (4, 25), and TMB and STK11 that are potential prognostic biomarkers in patients with lung cancer (26, 27). The ancestry and TMB association was independent of smoking-related mutational processes, and therefore further investigation on the impact of ancestry-related TMB differences on the response to checkpoint inhibitors is needed (28). Of note, our TMB estimates may be susceptible to germline contaminations due to the lack of matched normal samples. If germline variants specific to the Mexican or Colombian population could not be sufficiently filtered (due to smaller germline reference panels), individuals with higher NAT ancestry would have more germline contamination, and the anticorrelation between TMB and NAT ancestry may thus be even more significant than we have observed. Furthermore, due to the lack of matched germline samples and the use of panel sequencing data, we tested only known, hotspot mutations and protein-truncating mutations in lung cancer driver genes. Future studies will be needed to comprehensively characterize lung cancer genomes from Latin American patients.

Disparities in access to genetic testing have been observed among Hispanic patients with lung cancer in the United States (29). Our study shows that while controlling for global ancestry, local ancestry is associated with mutations in EGFR and KRAS, providing the first example, to our knowledge, of a germline influence, which may or may not act together with ancestry-specific environmental exposure, on targetable somatic events in lung cancer. These findings highlight the importance of providing somatic genetic testing for Latin American patients with lung cancer with admixed ancestries. Given the limited sample size, we could not determine the precise risk loci for EGFR and KRAS mutation by local ancestry mapping. As low-dose CT scans have enabled lung cancer screening that can significantly reduce lung cancer mortality (30, 31), we believe that the identification of germline allele(s) predisposing to the development of EGFR-mutant or KRAS-mutant lung adenocarcinoma from large, Latin American lung cancer sample sets may improve our understanding of the biological causes of EGFR and KRAS mutations and the evolutionary processes in lung cancers. Such findings could therefore shed light on prevention and early-detection strategies for lung cancer in Latin America and beyond, particularly for nonsmokers.

METHODS

Sample Collection

The protocol of this work was approved by the ethical and scientific committees of the Instituto Nacional de Cancerologia in Mexico City, Mexico, the Foundation for Clinical and Applied Cancer Research in Bogotá, Colombia, and the Dana-Farber Cancer Institute in Boston, Massachusetts, for detecting EGFR mutations and further genomic analysis. Biopsies were collected for histologic diagnosis by the pathology departments of Instituto Nacional de Cancerologia and Foundation for Clinical and Applied Cancer Research.

Library Preparation and Sequencing

Genomic DNA was extracted from fresh-frozen, blood and paraffin-embedded samples by a standard procedure using the Wizard Genomics DNA Kit (Promega) according to the manufacturer’s instructions. DNA was fragmented to 250 bp, and size-selected DNA was ligated to sample-specific barcodes. A custom targeted hybrid capture sequencing platform (OncoPanel) was used to assay genomic alterations in tumor DNA (16). Each library was quantified by sequencing on an Illumina MiSeq nano flow cell (Illumina). Libraries were pooled in equal mass to a total of 500 ng for enrichment using the Agilent SureSelect hybrid capture kit (Agilent Technologies; cat. no. G9611A). Libraries were sequenced on an Illumina HiSeq2500 or HiSeq3000. Pooled samples were demultiplexed using the Picard tools (https://broadinstitute.github.io/picard/). Paired reads were aligned to the hg19 reference genome using BWA (32) with the following parameters “-q 5 -l 32 -k 2 -o 1.” Aligned reads were sorted and duplicate-marked using Picard. In each batch, we sequenced a control DNA sample as a “plate normal.” For a subset of 44 cases, the same libraries were sequenced on Illumina NovaSeq for low-coverage whole-genome sequencing at 1× coverage.

Mutation Analysis

Mutation detection for SNVs was performed using MuTect v1.1.4 (33) in paired mode by pairing each sample to a control DNA sample profiled with the same OncoPanel. SomaticIndelDetector (34) was used for indel calling. Mutations were annotated by variant effect predictor (VEP; ref. 35) and Oncotator (36). Called variants with a frequency greater than or equal to 0.01% that were found in the
Genome Aggregation Database (gnomAD; ref. 37), containing 25,748 exomes, were excluded. TMB was calculated by dividing the total number of coding, nonsilent mutations in an individual by the target size (3 MB). Mutational signatures were called using SignatureAnalyzer (38) with SNVs classified by 96 trinucleotide contexts. Prior studies have shown that mutational signature analysis can be inferred based on on-target readouts from clinical panel sequencing (8, 39). Smoking (COSMIC signature 4) and APOBEC signature (COSMIC signature 2) activities were inferred by the estimated number of mutations in a trinucleotide context associated with each signature.

A custom script (https://github.com/carrotzhang/ancestry-from-panel) was applied to inspect the sites of hotspot driver mutations in EGFR, KRAS, BRAF, ERBB2, MET, and TP53. For each mutation, we counted reads supporting the reference base and the altered base, after filtering out reads with bases mapping quality less than 20 (40). A mutation was called if the total read count was greater than 5, the altered read count was greater than 2, and the mutant allele frequency was greater than 5%. Identified mutations with total coverage lower than 30x were manually inspected using Integrative Genomics Viewer (41).

**Copy Number and Rearrangements**

Read coverage was calculated at 1 MB bins across the genome and was corrected for guanine-cytosine (GC) content and mappability biases using ichorCNA version 0.1.0 (42) using the plate normal as the matched control for each sample. CISTIC version 2.0.22 (43) was applied to identify focal and arm-level SCNAS on ichorCNA-generated copy-number segments, with the high-level amplification defined as log-transformed copy-number ratio greater than 0.7. Rearrangement events were called by Breakmer (44) and filtered on discordant read counts and split read copy-number ratio greater than 0.3. Rearrangement events were repeated 10 times until a local ancestry risk score was generated for each sample:

$$\text{local ancestry risk score} = \sum_{i=1}^{n} A|Z_i|,$$

where $n$ is the number of regions associated with the somatic feature ($P < 0.05$), $A$ is the ancestry of each associated region (NAT ancestry was coded as 1, and EUR or AFR ancestry was coded as 0), and $Z$ is the Z-score of that associated region.

**Data and Code Availability**

Raw sequencing data from cancer gene panel sequencing are deposited at European Genome-phenome Archive (EGA) under the accession code EGAS000001004752. Ancestry identification method from panel sequencing and custom code used in the analyses are available at https://github.com/jcarrotzhang/ancestry-from-panel. This code and data are also available at https://codeocean.com/capsule/849209/. All other data supporting the findings of this study are available upon reasonable request from the corresponding authors.

**Authors’ Disclosures**

G. Ha reports a patent for methods for genome characterization pending. G.R. Oxnard reports other from Foundation Medicine and Roche outside the submitted work. O. Arrieta reports personal fees from Pfizer, Lilly, Merck, and Bristol Myers Squibb and grants and personal fees from AstraZeneca, Boehringer Ingelheim, and Roche outside the submitted work. A.F. Cardona reports grants, personal fees, and non-financial support from Merck Sharp & Dohme, Boehringer Ingelheim, Roche, Bristol-Myers Squibb, Foundation Medicine, and Astra Zeneca; non-financial support from BioNtech, Roche Biocare, and INQBox; grants and personal fees from Amgen and Bayer; personal fees and non-financial support from Foundation for Clinical and Applied Cancer Research; personal fees from Eisai, Merck Serono, Janssen Pharmaceuticals, Pfizer, Celldex, and Eli Lilly, and other from Illumina, Thermo Fisher, and Roche Diagnostics outside the submitted work. M. Meyerson reports personal fees from OrigMed and grants from Bayer, Janssen, Novo, and grants Onc outside the submitted work; in addition, M. Meyerson has a patent
Genetic Ancestry Affects Somatic Alterations in Lung Cancers

for EGFR mutations for lung cancer diagnosis issued, licensed, and with royalties paid from LabCorp and a patent for EGFR inhibitors pending to Bayer; and was a founding advisor of, consultative to, and equity holder in Foundation Medicine, shares of which were sold to Roche. No disclosures were reported by the other authors.

Authors’ Contributions

J. Carrot-Zhang: Formal analysis, methodology, writing—original draft, writing—review and editing. N. Patterson: Methodology. A.R. Thorner: Resources, writing—review and editing. A. Nag: Resources, writing—review and editing. J. Watson: Resources. G. Genovese: Methodology. J. Rodriguez: Resources. M.K. Gelbard: Resources. L. Corrales-Rodríguez: Resources. Y. Mitsuishi: Resources, writing—review and editing. G. Ha: Methodology, writing—review and editing. J.D. Campbell: Methodology, writing—review and editing. G.R. Oxnard: Resources, writing—review and editing. O. Arrieta: Conceptualization, resources, writing—review and editing. A.F. Cardona: Conceptualization, resources, writing—review and editing. A. Gusev: Supervision, writing—review and editing. M. Meyerson: Conceptualization, supervision, writing—review and editing.

Acknowledgments

We would like to acknowledge the patients from Mexico and Colombia for their participation in this study. This study is supported by the V Foundation Translational Research Award (the Stuart Scott Memorial Fund) and by NCIs grants R35 CA197568, R01 CA116020, and R01 CA227237, as well as an anonymous gift to the Dana-Farber Cancer Institute. M. Meyerson is an American Cancer Society Research Professor. J. Carrot-Zhang holds the CIHR Banting fellowship. J.D. Campbell is funded by the LUNGevity Career Development award. We thank the Center for Cancer Genomics at Dana-Farber Cancer Institute and the Genomic Platform at Broad Institute for their sequencing and genotyping efforts, and we thank Aruna Ramachandran and Kar-Tong Tan for helpful discussions.

Received August 11, 2020; revised October 26, 2020; accepted November 19, 2020; published first December 1, 2020.

REFERENCES


45. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.


