

Genetic Ancestry and Somatic Mutations in Lung Adenocarcinoma: Insights from Real-World Clinico-Genomic Data

Author Information

Brooke Rhead Ph.D.; Yannick Pouliot, Ph.D.; Justin Guinney, PhD.;
and Francisco M. De La Vega, D.Sc.*

Tempus AI, Inc., Chicago IL 60654, USA

*Corresponding author:

Francisco M. De La Vega, D.Sc.

Tempus AI, Inc.

600 West Chicago Avenue

Suite 510

Chicago, IL 60654

E-mail: Francisco.DeLaVega@tempus.com

Abstract

Background: Lung cancer presents a significant global health challenge, with disparities in incidence and outcomes across races and ethnicities. These disparities underscore the need to explore the molecular landscapes of lung cancer in relation to ancestry. Here, we leverage data from a real-world clinico-genomic database to discover associations between molecular profiles and genetic ancestry or race/ethnicity categories.

Methods: We utilized data from a cohort of 13,196 primarily late-stage non-small cell lung adenocarcinoma (LUAD) patients, sequenced with the Tempus xT NGS 648-gene panel, of which normal tissue was also sequenced for 6,520 cases. Genetic ancestry proportions were estimated using ancestry informative markers. Race and ethnicity categories were imputed using an ancestry-backed method, resulting in the assignment of 568 Hispanic/Latino, 892 non-Hispanic (NH) Asian, 1,581 NH Black, and 10,063 NH White individuals. Multiple imputation addressed missing data on smoking status. Logistic regression models assessed associations between ancestry proportions and somatic variants in 23 LUAD-related genes, adjusting for a false discovery rate of 5%. Analyzed mutations included copy number alterations, gene fusions, protein-altering SNVs and indels, and actionable or predicted driver mutations.

Results: Our analysis confirmed previously reported associations, such as a positive correlation between East Asian (EAS) ancestry and *EGFR* (OR per doubling ancestry=1.1) and a negative correlation with *KRAS* driver mutations (OR=0.96), while European ancestry exhibited the opposite relationship (OR=0.93 and 1.08, correspondingly; all $p < 0.0001$). We also verified a positive association with *EGFR* driver mutations (OR=2) and a negative one with *KRAS* (OR=0.46; $p < 0.001$) among Hispanic/Latino patients and American Indigenous (AMR) genetic ancestry (OR=1.03 and 0.97, correspondingly; $p < 0.05$). Novel associations were identified between African (AFR) and South Asian (SAS) ancestries and LUAD genes. Some associations are explained by differences in smoking status (e.g., *ATM* and *ALK* fusions), while others persist even after adjusting for smoking (e.g., *EGFR*, *KRAS*, and *CDKN2A* copy-number alterations). Notably, we identified a positive association between EAS ancestry and the imputed NH Asian category with driver mutations in *CTNNB1* (OR=1.05 and 2.2, respectively; $p < 0.01$), independent of smoking. These mutations are rare in NH White patients (2.4%) but are more prevalent in never-smoker NH Asian patients with predominant EAS ancestry (8.5%).

Conclusion: This study underscores the value of clinico-genomic databases in revealing associations between LUAD mutational profiles and genetic ancestry, shedding light on lung cancer disparities. Identification of a previously unappreciated association between EAS with *CTNNB1*, a potential biomarker for spindle assembly checkpoint kinase (*TTK*) inhibitors effectiveness and prognosis in LUAD, emphasizes the value of studying diverse populations in cancer research, paving the way for more equitable lung cancer treatments.

Introduction

Lung cancer is a multifaceted disease, shaped by a mix of environmental exposures and genetic predispositions.¹ While smoking is the foremost causal factor in lung cancer development, decreasing rates of smoking and associated mortality² have highlighted the rising incidence of lung cancer in never-smokers.³ This shift accentuates the need to explore lung cancer's contributing factors beyond smoking.

Non-small cell lung adenocarcinoma (LUAD), the most common lung cancer subtype,⁴ is distinguished by unique molecular and epidemiological characteristics. LUAD features diverse genetic alterations, such as mutations in *EGFR*, *KRAS*, and *ALK* genes,⁴ which have been reported to differ across race and ethnicity (R/E) groups.^{5–10} The incidence and mortality rates of LUAD also vary markedly among different R/E and regions.^{11–13}

Disparities in lung cancer incidence and outcomes across R/E and gender suggests a complex interaction between genetic vulnerability and environmental factors. For instance, never-smoker women of East Asian descent have a higher prevalence of LUAD with *EGFR* mutations.³ Black men face disproportionately higher lung cancer rates and worse outcomes than their White counterparts, suggesting that socio-economic factors, healthcare access, and potential genetic differences contribute to lung cancer disparities.^{9,12} The distinct molecular profiles of lung cancer in Hispanic/Latino populations, such as the higher prevalence of *KRAS* mutations,⁶ further emphasize the importance of examining lung cancer's molecular landscapes in the context of genetic ancestry and race/ethnicity to discover new insights for targeted treatments and personalized care.

Historically, the exploration of lung cancer's molecular profiles in relation to genetic ancestry and R/E categories have been hampered by limited diversity in research cohorts, small sample sizes with minimal minority representation,¹⁴ and reliance on broad and overlapping US federal R/E categories.¹⁵ These categories, which group genetically and geographically diverse populations together (e.g. conflate East and South Asians in “Asian”), and obscure the potential range of genetic admixtures (e.g. in Blacks and Hispanic or Latinos), limit the understanding of lung cancer's complexities.^{16,17}

Real-world clinical genomics databases that aggregate de-identified data from patients undergoing clinical testing offer a valuable resource for overcoming these limitations.^{18–20} Such databases are growing rapidly, reflecting the increasing adoption of tumor profiling and liquid biopsies in treatment guidelines.¹⁹ Despite healthcare access disparities, the representation of minority groups in these databases has improved,^{21,22} providing a rich source of multimodal molecular data for investigating molecular associations with race, genetic ancestry, and other clinical factors.

However, challenges such as significant missingness in R/E data^{21,23} and gaps in clinical data, like smoking history,²⁴ persist. To bridge these gaps, our study leverages the Tempus clinico-

genomic database to explore the associations between LUAD molecular profiles and genetic ancestry or R/E categories, inferring continental-level genetic ancestry from molecular data.

Our approach has revealed both known and novel associations between somatic mutations and genetic ancestry and R/E groups. Notably, we uncovered population-based associations with biomarkers of drug effectiveness and prognosis in LUAD, demonstrating the importance of studying diverse populations to identify new therapeutic strategies and insights that may help address healthcare disparities.

Methods

Patient cohort

We obtained records for 13,196 cancer patients diagnosed with LUAD from the de-identified Tempus clinico-genomic database, which includes genomic and clinical data from cancer patients that underwent tumor profiling using the Tempus xT assay as part of their healthcare. Briefly, Tempus xT is a targeted, tumor-normal-matched DNA panel that detects single-nucleotide variants, insertions and/or deletions, and copy number variants in 648 genes, as well as chromosomal rearrangements in 22 genes with high sensitivity and specificity.^{25,26} Selection criteria included tumor profiling with the Tempus xT assay (v2-v4) from 2018 to 2022. For patients with multiple independent test results, we selected the results corresponding to the first collection date.

Genetic ancestry estimation

We determined proportions of continental genetic ancestry employing a supervised variant of the ADMIXTURE algorithm for global genetic ancestry inference,²⁷ following methodologies outlined in prior research.^{22,28} We estimated ancestry proportions across five major super-populations—Africa (AFR), American Indigenous (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS)—utilizing a custom set of 654 ancestry informative markers (AIMs) previously identified in the targeted sequencing regions of the Tempus xT assay.²² Reference allele frequencies for these AIMs were derived from the 1,000 Genomes Project,²⁹ the Human Genome Diversity Project,^{30,33} and the Simons Genome Diversity Project³¹ databases. Specifically for the AMR super-population, we omitted the admixed "AMR" group from the 1,000 Genomes Project, opting instead for allele frequencies of Native American individuals from the alternate sources to enhance accuracy in reflecting American Indigenous population similarities.

Imputation of race and ethnicity

The categories of race and ethnicity in real-world data (RWD) follow the guidelines set by the US Office of Management and Budget.¹⁵ Yet, these classifications can complicate analyses due to the separate questions for race and ethnicity, leading to the practical approach of flattening these categories into distinct, non-overlapping categories used in this study.³² Hispanic or Latino, non-Hispanic (NH) Asian, NH Black, and NH White. We imputed these race and ethnicity categories based on genetic ancestry proportions, utilizing a boosted logistic regression machine learning algorithm as outlined in earlier research.³³ Individuals without at least a 50% probability of belonging to one of the race/ethnicity groups were categorized as "No Call"

(0.7%). Prior publications have validated the accuracy of this approach using data from the Tempus clinico-genomic database (correct rate of 96% and weighted error of 0.9%).³³

Statistical analyses

Somatic mutations in genes previously associated with LUAD^{34–3637–39} present on the Tempus xT gene panel with a mutation in at least one percent of patients in the selected cohort were tested for association with genetic ancestry proportions and race and ethnicity categories. Five mutation types were tested separately: protein-altering SNVs and indels, gene fusions, SCNAs, mutations with an OncoKB classification for any cancer type as therapeutic level one or two or resistance level one,³⁷⁴⁰ and LUAD driver mutations as predicted by the boostDM algorithm.³⁵

A directed acyclic graph (DAG) helped us to visualize the confounders and causative factors affecting LUAD incidence, particularly the relationships among R/E or genetic ancestry, smoking, social determinants of health (SDOH) and other environmental exposures, genetic predispositions, and somatic mutations (Supplementary Figure 1). The DAG positions smoking as a key mediator in the pathway from race/ethnicity to somatic mutations, indicating that the relationship between genetic ancestry and mutation rates is modulated, in part, by smoking behaviors. Associations were determined using likelihood ratio tests (LRTs) comparing full and nested logistic regression models. Three analyses were performed for each gene and mutation type: (1) univariable analysis, with the full model consisting of an indicator for the presence of somatic mutations in the gene as the dependent variable and either genetic ancestry proportions or imputed race and ethnicity category as the independent variables, and an intercept-only nested model; (2) complete case analysis adjusted for smoking status, with the same models as in the univariable analysis but with smoking status included as an additional independent variable in both the full and nested models, with only patients with known smoking status included; and (3) multiple imputation analysis adjusted for smoking status, using the same models as complete case analysis but including all patients and utilizing multiple imputed values for smoking status and pooling results (see below). Compositional data analysis methods were used to enable inclusion of all five genetic ancestry proportions in the same models. Specifically, genetic ancestry proportions were transformed into an isometric log ratio representation using the *pivotCoord* function in the *robCompositions* R package.³⁸ Models including imputed race and ethnicity categories used NH White as the reference category. For each mutation type, LRT p-values were corrected for the number of genes tested using the Benjamini-Hochberg method to control the false discovery rate at 5%. For any association where the corrected LRT p-value was <0.05, all logistic regression p-values <0.05 for a specific genetic ancestry proportion or imputed race and ethnicity group from the full model were considered associated.

Multiple imputation

Multivariate imputation by chained equations (MICE) was performed in R using the *mice* package.³⁹ Variables included as predictors in imputation models were smoking status, gender, age quartile at collection of tumor specimen, U.S. census division of patients' home address state derived from 3-digit Zip Code (Pacific, Mountain, West North Central, West South Central, East North Central, East South Central, New England, Middle Atlantic, or South Atlantic),

genetic ancestry proportions transformed to isometric log ratio pivot coordinates, tumor grade, cancer stage, tumor mutational burden (TMB, in mutations/megabase), assay version, and indicators for the presence of mutations in the genes tested: (1) actionable mutations in *ATM*, *BRAF*, *CTNNB1*, *EGFR*, *KRAS*, *NF1*, *PIK3CA*, *RBM10*, *STK11*, and *TP53*; (2) *CDKN2A*, *EGFR*, *ERBB2*, *KRAS*, and *MET* SCNAs; and (3) gene fusions with *ALK* (Supplemental Figure 2). The following methods were used for variables with missingness: smoking status (46% missing), logistic regression; age quartile at collection of tumor specimen (19% missing), multinomial regression; U.S. census division (19% missing), random forests; tumor grade (64% missing), multinomial regression; cancer stage (32% missing), multinomial regression; TMB (0.05% missing), predictive mean matching. Ten datasets were imputed, each with 20 iterations of the MICE algorithm. Plots of the mean and standard deviation of each variable with missingness were examined to assess convergence of the MICE algorithm, and distributions of imputed values were compared to measured values to assess the quality of imputations. Likelihood ratio test statistics from the imputed datasets were combined using the *D3* function in the *mice* package,⁴¹ and logistic regression test statistics were combined using the *pool* function in the *mice* package using the “Rubin 1987” pooling rule.⁴⁰ We observed that MICE iterations converge well in our data (Supplementary Figure 4), that the multiply imputed categorical variables align with observed data (Supplementary Figure 5), and that the multiply imputed smoking status is distributed similarly across R/E categories, albeit with a tendency toward over-imputation of current/former smoking status in Hispanic/Latino and NH Asian (Supplementary Figure 6).

Results

Patient characteristics

Our study examined a cohort of LUAD patients totaling 13,196 individuals. Patients were categorized by imputed R/E into Hispanic or Latino (N=568), NH Asian (N=891), NH Black (N=1,581), and NH White (N=10,063), with 93 patients not included in any of these categories (**Table 1**). While most patients have majority European genetic ancestry (EUR), there is also a substantial number of patients with African (AFR), American Indigenous (AMR) and East Asian (EAS) ancestries, while patients of South Asian ancestry (SAS) are less represented in our cohort (**Figure 1**). We also observed a conflation of EAS and SAS ancestries in the NH Asian category—a well-known challenge. A notable majority of patients (75%) were former or current smokers, especially within the NH Black (84%) and NH White (79%) groups. Ages at specimen collection and diagnosis were, on average, in the mid- to late-sixties across groups, with minor variations across R/E. Gender distribution maintained a balance across groups, albeit with a modest female majority overall in the cohort. In this study, we also analyzed a sub-cohort of patients that had paired tumor and normal (T/N) specimens sequenced (Supplementary Table 1).

Smoking status, age at specimen collection, and age at diagnosis all varied significantly across R/E categories ($p < 0.001$), suggesting potential differences in disease exposures and disease course prior to sequencing (**Table 1, Supplementary Figure 1**). Of patients with known cancer stage, 63% had stage 4 cancer, consistent with expected utilization of tumor profiling in cancer

care. This percentage was higher in the Hispanic or Latino (72%) and NH Asian (69%) categories ($p < 0.001$), pointing to a higher prevalence of advanced disease at sequencing. TMB count also varied by R/E ($p < 0.005$), with NH Black patients displaying a higher TMB count in the complete cohort (encompassing tumor only, TO, plus tumor-normal matched, T/N, sequencing modalities). However, TMB count in T/N cases, where misclassifications of germline variants as somatic is avoided, shows Hispanic/Latino and NH Asian exhibiting a lower burden ($p < 0.001$; Supplementary Table 1).

Compared to never smokers, former or current smokers varied considerably by R/E, stated race, and stated ethnicity ($p < 0.001$), more male (48% vs. 32%), and had an earlier cancer stage, higher tumor grade, and higher TMB (average 5.0 vs. 2.3) (Supplementary Tables 2 and 3). There were no R/E differences in age at collection or diagnosis, assay version, microsatellite instability (MSI) status, or availability of normal tissue for sequencing. Compared to patients without available smoking status, patients whose smoking status was known varied by R/E ($p = 0.019$) and stated race ($p = 0.01$), were younger at time of specimen collection and at diagnosis (mean 67 vs. 71 in both cases; $p < 0.001$), female (56% vs. 53% $p < 0.001$), more likely to be sequenced on an earlier xT assay version (82% xT.v4 vs. 90%; $p < 0.001$), had a slightly lower TMB count (4.2 on average vs. 4.6; $p < 0.001$), and were more likely to have normal tissue available for sequencing (54% vs. 44%). There was no difference in cancer stage, tumor grade, MSI status, or stated ethnicity.

We observed that smoking correlates with TMB, gender, and tumor grade (Supplementary Figure 2A) and mutations in genes such as *KRAS*, *EGFR*, and *ALK* fusions (Supplementary Figure 2B; Supplementary Figure 3A), as previously reported in the literature.⁴ Moreover, smoking status was associated with R/E ($p < 0.001$; see **Table 1**) and affects frequency of somatic mutations in genes (Supplementary Figure 3B).

Associations between genetic ancestry and somatic mutations

We tested for associations between continental genetic ancestry proportions and somatic alterations in genes with known oncogenic properties in LUAD (Supplementary Table 4, see Methods for selection criteria).

A univariable analysis revealed associations between genetic ancestry and somatic mutation patterns in several LUAD genes (**Table 2**). We confirmed a previously reported association between EAS with an increased frequency of driver mutations in *EGFR* (OR per doubling ancestry=1.1, $p < 0.0001$), and a decreased frequency in *STK11* (OR=0.95, $p < 0.01$) and *KRAS* (OR=0.95; $p < 0.0001$), which remain associated after adjusting for smoking status (**Figure 2**). These associations were also observed when restricting to OncoKB actionable variants (Level 1, 2 and R1; Supplementary Figure 8). Additionally, we observed a positive association between *CDKN2A* SCNAs (OR=1.03; $p = 0.002$), and a decrease in driver mutations in *BRAF* (OR=0.95; $p = 0.004$) as EAS ancestry increases that persisted after adjustments for smoking. An association between EAS and increased driver mutations in *CTNNB1* was also observed (OR=1.04; $p = 0.009$), which is attenuated by adjusting for multiple imputed smoking status

(**Figure 2**), but not when adjusting for smoking in case-complete analysis (Supplemental Table 5, Supplemental Figure 7).

Associations were also identified between American Indigenous (AMR) ancestry, with a decreased frequency of driver mutations in *TP53* (OR=0.98; p=0.006) and *KRAS* (OR=0.97; p=0.002), although both become non-significant when controlling for smoking status (**Figure 2**). Previous studies have also shown an association between Native American ancestry and increased actionable variants in *EGFR*.⁶ Although we did not observe this association using boostDM predicted driver variants, we were able to replicate using OncoKB actionable variants (OR=1.04; p=0.0030), as well as with protein-altering variants in general (restricted to T/N cases; OR=0.97; p=0.001; Supplementary Figure 7). While the association of AMR with actionable variants in *EGFR* appears to be modulated by smoking, the association with protein altering variants remains significant after adjusting for smoking, either in the case-complete or multiple imputation analyses. Further, we find a significant association between AMR ancestry and *ALK* gene fusions (OR=1.06; p=0.003), which disappeared following adjustment for smoking (**Figure 2**).

For AFR ancestry, we observed decreased driver mutations in *RBM10* (OR=0.96), *EGFR* (OR=0.94) and fewer *ALK* fusions (OR=0.96) and *CDKN2A* SCNAs (OR=0.98), and increased driver mutations in *BRAF* (OR=1.04), *NF1* (OR=1.04), *STK11* (OR=1.03) and *TP53* (OR=1.02, all p<0.05; **Figure 2**). All but *BRAF* and *RBM10* appear to be influenced by smoking (**Figure 2**). There were also associations with SAS ancestry showing decreased SCNAs in *CDKN2A* (OR=0.98) and driver mutations in *NF1* (OR=0.93) and *TP53* (OR=0.98, all p<0.05), with the latter being attenuated by smoking adjustment (**Figure 2**). Further, we observed associations between EUR ancestry and increased mutations in *KRAS* (1.08) and *ATM* (OR=1.07) that were not influenced by smoking, as well as increased driver mutations in *NF1* (OR=1.02), *RBM10* (OR= 1.05), and *STK11* (OR=1.05), and decreased mutations in *CTNNB1* (OR=0.95), and *ALK* fusions (OR=0.9; all p<0.05), which lost significance after adjusting for smoking status (**Figure 2**).

Finally, we also identified significant gene associations when considering any protein-altering variants regardless of their driver/actionable status and restricted to T/N cases only. These include associations between EAS and decreased number of variants in *KEAP1* (OR=0.94) and *SMARCA4* (OR=0.95), which appear to be independent of smoking status, while EUR showed the opposite effect (OR=1.05 and OR=1.05, correspondingly; all p<0.05), although attenuated by smoking (Supplementary Figure 8).

Associations between imputed race and ethnicity categories and somatic mutations

In addition to the associations with genetic ancestry proportions, we conducted analyses with imputed R/E categories, using NH White as the reference group. Consequently, these analyses do not reveal associations specific to the NH White group. Results of all R/E LRT and association tests are given in Supplementary Table 6 and Files 4-6. Generally, we observed a strong concordance between AFR ancestry and NH Black, EAS and NH Asian, and AMR and Hispanic/Latino categories for associations with driver mutations in *BRAF*, *CTNNB1*, *EGFR*,

KRAS, *STK11*, and *TP53*, as well as SCNAs in *CDKN2A* and *ALK* gene fusions (**Figure 3**). However, some discrepancies were noted; for instance, the associations between Hispanic/Latino and an increased or decreased number of driver mutations in *EGFR* and *KRAS*, respectively, were robust against MICE smoking adjustments. Notably absent were associations between Hispanic/Latino (seen with AMR) or NH Asian (seen with SAS) with mutations in *TP53*, and the association between SAS and decreased mutations in *NF1* did not emerge here (potentially because NH Asian conflates East and South Asians, with the latter being underrepresented in our cohort). Additionally, we found associations between Hispanic/Latino and an increased occurrence of *ALK* gene fusions (OR=2.9), and decreased driver mutations in *RBM10* (OR=0.43; both $p < 0.01$). The analysis of OncoKB actionable and short protein-altering mutations essentially mirrors the results with driver mutations and genetic ancestry, with a few additions: associations between NH Black and increased protein altering mutations in *ALK* (OR=1.51) and *KEAP1* (OR=1.24); between Hispanic/Latinos and reduced mutations in *KEAP1* (OR=0.29) and *STK11* (OR=0.52); and between NH Asian and reduced mutations in *KEAP1* (OR=0.27) and *STK11* (OR=0.2; all p -value <0.3 Supplementary Figure 9).

Distribution of mutations in CTNNB1 across imputed R/E categories

We aimed to further characterize the associations found with driver mutations in the β -Catenin gene, *CTNNB1*, and EAS and NH Asians. Initially, we observed that this association was identified using boostDM predicted drivers and total protein-altering mutations, and with both genetic ancestry and the imputed NH Asian category. Although the association is attenuated with smoking in multiple imputations for predicted driver mutations (see Table 2), the association with genetic ancestry remains unaffected after adjustment for smoking in protein-altering variants (Supplemental Figure 8), as well as for associations found with the imputed NH Asian category with predicted drivers (Figure 3) or protein-altering mutations (Supplemental Figures 8 and 9). Thus, our results suggest that smoking does not totally explain the increased burden of mutations in *CTNNB1* in EAS ancestry or NH Asian patients.

We examined the proportion of patients harboring a predicted driver mutation in *CTNNB1* across imputed R/E categories, distinguishing individuals with East Asian (EAS) and South Asian (SAS) ancestries within the NH Asian group. Figure 4 demonstrates that while the overall fraction of NH White patients in the cohort with such mutations is low (2.4%), albeit slightly higher in never-smoker NH White (3.6%). In contrast, this fraction is higher overall in NH Asian-EAS patients (5%; Fisher exact p -value <0.001) and even higher in never-smokers NH Asian-EAS patients (8.6%; $p=0.001$). It also appears in a considerable number of never-smoker Hispanic/Latino (3.8%) and NH Black (5.6%) patients (Supplementary Table 7).

Discussion

In recent years, advances in genomic technologies have enabled a deeper understanding of the molecular underpinnings of lung cancer. However, the relationship between the molecular profiles of lung cancer and genetic ancestry or R/E categories remains relatively underexplored due to limited diversity in research studies. Previous studies in this area have been limited by a)

a lack of diversity in research cohorts (e.g. TCGA⁴¹); b) smaller sample sizes; and c) reliance on US government-mandated R/E categories,¹⁵ which can be problematic. Considering these limitations, real-world clinical genomics databases represent a valuable resource for advancing our understanding of disparities and their molecular correlates in cancer research. Furthermore, the volume of data stored by such databases is expanding rapidly, ensuring that despite existing healthcare access disparities, minority groups are represented in substantially larger numbers compared to traditional research cohorts. Nevertheless, real-world data can have several shortcomings, notably substantial missingness in R/E data. The extent of this missingness is complex and varies across sources, ranging from 30-70%,²³ but can largely be attributed to issues in data transmission and collection rather than simply patient abstention.^{42,43}

The present study aimed to analyze data from a real-world clinico-genomic database to provide insights into the associations between lung cancer molecular profiles and genetic ancestry or R/E categories. To address the challenges typically associated with RWD, we implemented several strategies. Firstly, we inferred continental-level genetic ancestry from molecular data obtained during testing,²² allowing us to eschew categorical analysis in favor of logistic regression methods.³⁸ We directly assessed the association between genetic ancestry proportions and somatic mutations of a specific type in LUAD genes via likelihood ratio tests (LRTs). P-values were examined if a positive association was identified. Compared to a strategy that forgoes LRTs and examines all logistic regression p-values directly, our approach reduces the multiple testing burden, ensuring high statistical power while minimizing the risk of type I errors. Additionally, when analyzing genetic ancestry proportions, it is essential to address the fact that they sum to one, resulting in collinearity (e.g., in admixed NH Black patients, it is common for AFR ancestry to increase as EUR ancestry decreases). We thus employed compositional logistic regression methods, applying an isometric log-ratio transform to include all proportions in the same model.³⁸ This enabled us to identify associations with increased EUR ancestry, whereas in R/E analysis (see below), NH White is consistently used as the control, thereby not permitting the elucidation of these effects. To address the issue of missing clinical data, we utilized a multiple imputation strategy (MICE³⁹) and performed an extensive assessment of the results to understand the accuracy of imputed values. This approach enabled us to increase power and to explore smoking's role in our findings. For completeness, and because multiple imputation and complete case analysis are subject to different types of bias, we provided complete case analysis results from patients with known smoking status only (cf. Supplementary Materials). Finally, for a categorical analysis involving R/E, we employed a previously developed ancestry-backed R/E imputation method demonstrated to be highly accurate within the Tempus data set.³³

We identified associations between genetic ancestry or imputed R/E categories with different types of variants in LUAD-related genes. Our analysis included protein-altering short variants, SCNAs, rearrangements associated with gene fusions, actionable variants listed in OncoKB³⁷ (Levels 1 & 2 and L1), and somatic driver variants for LUAD predicted by the boostDM algorithm.³⁵ The boostDM algorithm evaluates and categorizes all possible single base substitutions in cancer genes according to their tumorigenic potential, drawing on a comprehensive analysis of mutations observed in a vast collection of sequenced tumors and

annotating each site with relevant mutational features. This approach allowed us to compile a more comprehensive and potentially less biased list of driver mutations for genes for which boostDM models are available. Some of the associations we found with EAS, such as those with *CDKN2A* and *BRAF* persist after adjustment for smoking status. These associations are particularly noteworthy as the global decline in smoking has led to a higher prevalence of non-smoker LUAD, which is disproportionately represented in patients of East Asian ancestry.

One notable finding is the association between driver mutations in *CTNNB1* (the β -Catenin gene) and EAS ancestry. β -Catenin acts as a crucial co-activator in the oncogenic Wnt signaling pathway, where aberrations often lead to oncogenesis.⁴⁴ Somatic mutations in *CTNNB1*, especially in exon 3, are implicated in this process, causing stabilization and accumulation of β -Catenin in cells and activating the Wnt/ β -Catenin signaling pathway to increase cell proliferation.⁴⁵ It has been reported that cancer cell lines with activating mutations in the *CTNNB1* gene are five times more sensitive to inhibitors of the spindle assembly checkpoint kinase (*TTK*), which are emerging as promising antineoplastic agents.⁴⁶ Thus, *CTNNB1* mutations have been proposed as prognostic drug response biomarkers, potentially enabling the selection of patients most likely to benefit from *TTK* inhibitor therapy in proof-of-concept clinical trials.^{44,47} While *CTNNB1* mutations are prevalent in endometrial cancer and hepatocellular carcinoma,⁴⁶ they have also been suggested as biomarkers for post-operative recurrence-free survival in *EGFR*-mutant LUAD.⁴⁸ However, previous studies have indicated that *CTNNB1* mutations are rare in LUAD, with two large series reporting a frequency of approximately 2%—studies predominantly involving patients of European descent.^{49,50} In contrast, smaller case studies focusing on East Asian patients have indicated a higher frequency of *CTNNB1* mutations in LUAD.^{51,52} Our RWD study, which has a much larger sample size, now confirms the presence of *CTNNB1* driver mutations at appreciable frequencies in East Asian and Hispanic/Latino patients, as well as a statistically significant association with EAS ancestry. These results open the possibility of using *CTNNB1* mutations as biomarkers for the effectiveness of *TKK* inhibitors and prognosis in LUAD.

Our study's methodology, robust in uncovering associations between genetic ancestry and mutational profiles in LUAD, faces several limitations worth noting. First, RWD may display ascertainment bias, as patients undergoing tumor profiling testing are likely those with later-stage cancer. This bias is exacerbated by disparities in insurance coverage and healthcare access, which may be confounded with R/E. Second, our approach to imputing R/E is limited to mutually exclusive categories, currently excluding Native American/Alaskan Natives—who are likely misclassified as "Hispanic/Latino"—and Hawaiian/Pacific Islanders, often misclassified as "NH Asian." As the database expands to include more such patients, we plan to refine our methods to also impute these categories. Additionally, although multiple imputation is a vital tool for addressing missing data, it has its imperfections and potential biases. The directed acyclic graph in Supplementary Figure 1 highlights the roles of social determinants of health (SDOH), environmental, and genetic exposures in LUAD development, and considers how access to healthcare and screening affects cancer diagnosis, thus underscoring the complexity of these interacting factors. A significant limitation of our analysis is the lack of data on SDOH, which undeniably affects both the incidence and outcomes of lung cancer. This absence hinders our

full understanding and integration of these factors into our analysis. Most importantly, it is essential to interpret our findings within the context that our methods establish associations rather than causality. This distinction necessitates cautious interpretation of our results, and underscores the importance of further research to clarify the intricate connections between genetic ancestry, mutational profiles, and lung cancer.

Conclusion

Our methodology allowed us to identify both known and novel associations between somatic mutations and genetic ancestry or R/E groups. Our findings suggest that driver mutations in *CTNNB1* characterize a subgroup of mainly never-smoker LUAD patients more prevalent in East Asian populations, a potential biomarker of drug effectiveness. This result demonstrates how studies within diverse populations can aid the identification of new therapeutic approaches and provide insights that may ultimately help explain and address healthcare disparities.

Acknowledgements

We would like to thank Yan Liu (Tempus), Carlos D. Bustamante and Alex Ioannidis (Stanford University) for statistical and methodology discussions and advice. We also acknowledge Rafael Esteller, Nick Rigan, and Arvind Prasad from the Tempus Lens team for their superb assistance in procuring de-identified data and correcting data problems needed for this work. We thank Frank Nothhaft (formerly Tempus) for support with data access and Joel Dudley (formerly Tempus) for encouraging us to pursue this research. We thank Vanessa Nepomuceno from the Tempus Publications team for copy editing the manuscript.

Funding

This study was funded by Tempus AI, Inc.

Author Information

Authors and Affiliations

Tempus AI, Inc, Chicago, IL, USA

Brooke Rhead, Yannick Pouliot, Justin Guinney, and Francisco M. De La Vega

Corresponding author

Francisco M. De La Vega.

Contributions

Brooke Rhead: Methodology, data analysis, visualization, writing - review, editing. Yannick Pouliot: Data procurement, curation, analysis, methodology, writing - review, editing. Justin Guinney: Methodology, review, editing. Francisco De La Vega: Conceptualization, resources, supervision, writing - review, editing. All authors reviewed and suggested edits for the final version of the manuscript. The authors read and approved the final manuscript.

Ethics declarations

Ethics approval and consent to participate.

All analyses were performed using de-identified data; The need for Institutional Review Board Approval was exempted by the IRB of Advarra, Inc., protocol no: Pro00042950, on April 15, 2020.

Consent for publication

Not applicable.

Competing Interests

B.R., Y.P., J. G., and F.M.D.L. are employees and have received stock options from Tempus AI, Inc.

References

1. Bade BC, Cruz CSD. Lung Cancer 2020 Epidemiology, Etiology, and Prevention. *Clin Chest Med*. 2020;41(1):1-24. doi:10.1016/j.ccm.2019.10.001
2. Cornelius ME, Loretan CG, Jamal A, et al. Tobacco Product Use Among Adults – United States, 2021. *Morb Mortal Wkly Rep*. 2023;72(18):475-483. doi:10.15585/mmwr.mm7218a1
3. LoPiccolo J, Gusev A, Christiani DC, Jänne PA. Lung cancer in patients who have never smoked — an emerging disease. *Nat Rev Clin Oncol*. 2024;21(2):121-146. doi:10.1038/s41571-023-00844-0
4. Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. *Nature*. 2018;553(7689):446-454. doi:10.1038/nature25183
5. Shi H, Seegobin K, Heng F, et al. Genomic landscape of lung adenocarcinomas in different races. *Front Oncol*. 2022;12:946625. doi:10.3389/fonc.2022.946625
6. Carrot-Zhang J, Soca-Chafre G, Patterson N, et al. Genetic Ancestry Contributes to Somatic Mutations in Lung Cancers from Admixed Latin American Populations. *Cancer Discov*. 2021;11(3):591-598. doi:10.1158/2159-8290.cd-20-1165
7. Gimbrone NT, Sarcar B, Gordian ER, et al. Somatic Mutations and Ancestry Markers in Hispanic Lung Cancer Patients. *J Thorac Oncol*. 2017;12(12):1851-1856. doi:10.1016/j.jtho.2017.08.019
8. Adib E, Nassar AH, Alaiwi SA, et al. Variation in targetable genomic alterations in non-small cell lung cancer by genetic ancestry, sex, smoking history, and histology. *Genome Med*. 2022;14(1):39. doi:10.1186/s13073-022-01041-x
9. Campbell JD, Lathan C, Sholl L, et al. Comparison of Prevalence and Types of Mutations in Lung Cancers Among Black and White Populations. *JAMA Oncol*. 2017;3(6):801. doi:10.1001/jamaoncol.2016.6108
10. Graham RP, Treece AL, Lindeman NI, et al. Worldwide Frequency of Commonly Detected EGFR Mutations. *Arch Pathol Lab Med*. 2017;142(2):163-167. doi:10.5858/arpa.2016-0579-cp
11. Kumar R, Castellero F, Bhandari S, Malapati S, Kloecker G. The Hispanic Paradox in Non-Small Cell Lung cancer. *Hematology Oncol Stem Cell Ther*. Published online 2021. doi:10.1016/j.hemonc.2021.02.004
12. Theik NWY, Uribe CC, Alvarez A, Muminovic M, Ruez LE. Diversity and Disparities in Lung Cancer Outcomes Among Minorities. *Cancer J*. 2023;29(6):323-327. doi:10.1097/ppo.0000000000000689
13. Bonner SN, Curley R, Love K, Akande T, Akhtar A, Erhunmwunse L. Structural Racism and Lung Cancer Risk. *JAMA Oncol*. 2024;10(1):122-128. doi:10.1001/jamaoncol.2023.4897
14. Spratt DE, Chan T, Waldron L, et al. Racial/Ethnic Disparities in Genomic Sequencing. *Jama Oncol*. 2016;2(8):1070. doi:10.1001/jamaoncol.2016.1854
15. Budget O of M and. Standards for the classification of federal data on race and ethnicity. *Fed Reg*. 1997;62(210):58782. <https://www.federalregister.gov/documents/2016/09/30/2016-23672/standards-for-maintaining-collecting-and-presenting-federal-data-on-race-and-ethnicity>
16. Gelfand MJ, Denison EE. Moving beyond the West vs. the rest: Understanding variation within Asian groups and its societal consequences. *Proc Natl Acad Sci*. 2020;117(10):5100-5102. doi:10.1073/pnas.2000930117
17. Bryc K, Velez C, Karafet T, et al. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc National Acad Sci*. 2010;107(Supplement 2):8954-8961. doi:10.1073/pnas.0914618107
18. Royce TJ, Zhao Y, Ryals CA. Improving Diversity in Clinical Trials by Using Real-world Data to Define Eligibility Criteria. *JAMA Oncol*. 2023;9(4):455-456. doi:10.1001/jamaoncol.2022.7170
19. Chakravarty D, Johnson A, Sklar J, et al. Somatic Genomic Testing in Patients With Metastatic or Advanced Cancer: ASCO Provisional Clinical Opinion. *J Clin Oncol*. 2022;40(11):1231-1258. doi:10.1200/jco.21.02767
20. Booth CM, Karim S, Mackillop WJ. Real-world data: towards achieving the achievable in cancer care. *Nat Rev Clin Oncol*. 2019;16(5):1-14. doi:10.1038/s41571-019-0167-7
21. Snow T, Snider J, Comment L, et al. Comparison of Population Characteristics in Real-World Clinical Oncology Databases in the US: Flatiron Health-Foundation Medicine Clinico-Genomic Databases, Flatiron Health Research Databases, and the National Cancer Institute SEER Population-Based Cancer Registry. *medRxiv*. Published online 2023:2023.01.03.22283682. doi:10.1101/2023.01.03.22283682
22. Miyashita M, Bell JSK, Wenric S, et al. Molecular profiling of a real-world breast cancer cohort with genetically inferred ancestries reveals actionable tumor biology differences between European ancestry and African ancestry patient populations. *Breast Cancer Res*. 2023;25(1):58. doi:10.1186/s13058-023-01627-2
23. Studna A. Executive Roundtable: The Rise of RWD in Clinical Research. *Applied Clinical Trials*. Published May 17, 2023. Accessed July 16, 2023. <https://www.appliedclinicaltrials.com/view/executive-roundtable-the-rise-of-rwd-in-clinical-research>
24. Patel N, Miller DP, Snavelly AC, et al. A Comparison of Smoking History in the Electronic Health Record With Self-Report. *Am J Prev Med*. 2020;58(4):591-595. doi:10.1016/j.amepre.2019.10.020
25. Beaubier N, Bontrager M, Huether R, et al. Integrated genomic profiling expands clinical options for patients with cancer. *Nat Biotechnol*. 2019;37(11):1351-1360. doi:10.1038/s41587-019-0259-z

26. Beaubier N, Tell R, Huether R, et al. Clinical validation of the Tempus xO assay. *Oncotarget*. 2018;9(40):25826-25832. doi:10.18632/oncotarget.25381
27. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19(9):1655-1664. doi:10.1101/gr.094052.109
28. Rhead B, Hein DM, Pouliot Y, Guinney J, Vega FMDL, Sanford NN. Association of Genetic Ancestry with Molecular Tumor Profiles in Colorectal Cancer. *medRxiv*. Published online 2023:2023.07.12.23292571. doi:10.1101/2023.07.12.23292571
29. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
30. Bergström A, McCarthy SA, Hui R, et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*. 2020;367(6484):eaay5012. doi:10.1126/science.aay5012
31. Mallick S, Li H, Lipson M, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 2016;538(7624):201-206. doi:10.1038/nature18964
32. Flanagan A, Frey T, Christiansen SL, Committee AM of S. Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals. *JAMA*. 2021;326(7):621-627. doi:10.1001/jama.2021.13304
33. Rhead B, Haffener PE, Pouliot Y, Vega FMDL. Imputation of race and ethnicity categories using genetic ancestry from real-world genomic testing data. *Pac Symp Biocomput Pac Symp Biocomput*. 2023;29:433-445. doi:10.1142/9789811286421_0033
34. Collisson EA, Campbell JD, Brooks AN, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511(7511):543-550. doi:10.1038/nature13385
35. Muiños F, Martínez-Jiménez F, Pich O, Gonzalez-Perez A, Lopez-Bigas N. In silico saturation mutagenesis of cancer genes. *Nature*. 2021;596(7872):428-432. doi:10.1038/s41586-021-03771-1
36. Chalela R, Curull V, Enríquez C, Pijuan L, Bellosillo B, Gea J. Lung adenocarcinoma: from molecular basis to genome-guided therapy and immunotherapy. *J Thorac Dis*. 2017;9(7):2142-2158. doi:10.21037/jtd.2017.06.20
37. Chakravarty D, Gao J, Phillips S, et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol*. 2017;2017(1):1-16. doi:10.1200/po.17.00011
38. Templ M, Hron K, Filzmoser P. Compositional Data Analysis. Published online 2011:341-355. doi:10.1002/9781119976462.ch25
39. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*. 2011;45(3). doi:10.18637/jss.v045.i03
40. Rubin DB. Multiple Imputation for Nonresponse in Surveys. *Wiley Ser Probab Stat*. Published online 2023. doi:10.1002/9780470316696
41. Spratt DE, Chan T, Waldron L, et al. Racial/Ethnic Disparities in Genomic Sequencing. *Jama Oncol*. 2016;2(8):1070. doi:10.1001/jamaoncol.2016.1854
42. Srivastav A, Robinson-Ector K, Kipp C, Strompolis M, White K. Who declines to respond to the reactions to race module?: findings from the South Carolina Behavioral Risk Factor Surveillance System, 2016–2017. *BMC Public Heal*. 2021;21(1):1703. doi:10.1186/s12889-021-11748-y
43. Cabrerós I, Agniel D, Martino SC, Damberg CL, Elliott MN. Predicting Race And Ethnicity To Ensure Equitable Algorithms For Health Care Decision Making. *Heal Aff*. 2022;41(8):1153-1159. doi:10.1377/hlthaff.2022.00095
44. Schie EH van, Amerongen R van. Aberrant WNT/CTNNB1 Signaling as a Therapeutic Target in Human Breast Cancer: Weighing the Evidence. *Frontiers Cell Dev Biology*. 2020;8:25. doi:10.3389/fcell.2020.00025
45. Peifer M, Polakis P. Wnt Signaling in Oncogenesis and Embryogenesis—a Look Outside the Nucleus. *Science*. 2000;287(5458):1606-1609. doi:10.1126/science.287.5458.1606
46. Kim S, Jeong S. Mutation Hotspots in the β -Catenin Gene: Lessons from the Human Cancer Genome Databases. *Mol Cells*. 2019;42(1):8-16. doi:10.14348/molcells.2018.0436
47. Zaman GJR, Roos JADM de, Libouban MAA, et al. TTK Inhibitors as a Targeted Therapy for CTNNB1 (β -catenin) Mutant Cancers. *Mol Cancer Ther*. 2017;16(11):2609-2617. doi:10.1158/1535-7163.mct-17-0342
48. Kim Y, Ahn B, Yoon S, et al. An oncogenic CTNNB1 mutation is predictive of post-operative recurrence-free survival in an EGFR-mutant lung adenocarcinoma. *PLOS ONE*. 2023;18(6):e0287256. doi:10.1371/journal.pone.0287256
49. Sequist LV, Heist RS, Shaw AT, et al. Implementing multiplexed genotyping of non-small-cell lung cancers into routine clinical practice. *Ann Oncol*. 2011;22(12):2616-2624. doi:10.1093/annonc/mdr489
50. Mäki-Nevala S, Sarhadi VK, Rönty M, et al. Hot spot mutations in Finnish non-small cell lung cancers. *Lung Cancer*. 2016;99:102-110. doi:10.1016/j.lungcan.2016.06.024
51. Montpréville VT de, Lacroix L, Rouleau E, et al. Non-small cell lung carcinomas with CTNNB1 (beta-catenin) mutations: A clinicopathological study of 26 cases. *Ann Diagn Pathol*. 2020;46:151522. doi:10.1016/j.anndiagpath.2020.151522
52. Zhou C, Li W, Shao J, Zhao J, Chen C. Analysis of the Clinicopathologic Characteristics of Lung Adenocarcinoma With CTNNB1 Mutation. *Frontiers Genetics*. 2020;10:1367. doi:10.3389/fgene.2019.01367

Tables

Table 1. Cohort characteristics by imputed race and ethnicity category. Last column shows the results of statistical tests (as defined in footnote) for differences of the characteristics by imputed race and ethnicity.

Characteristic	Hispanic or Latino, N = 568 ¹	NH Asian, N = 891 ¹	NH Black, N = 1,581 ¹	NH White, N = 10,063 ¹	No Call, N = 93 ¹	p-value ²
Smoking status						<0.001
Former or current smoker	157 (46%)	162 (34%)	731 (84%)	4,256 (79%)	39 (70%)	
Never smoked tobacco	184 (54%)	317 (66%)	142 (16%)	1,128 (21%)	17 (30%)	
Unknown	227	412	708	4,679	37	
Age at specimen collection						<0.001
Unknown	66 (57, 74)	68 (60, 75)	66 (60, 72)	69 (62, 76)	66 (55, 72)	
Unknown	96	197	323	1,931	22	
Age at diagnosis						<0.001
Unknown	66 (57, 74)	67 (59, 75)	65 (60, 72)	69 (62, 76)	66 (55, 72)	
Unknown	123	215	401	2,412	29	
Gender						0.085
Female	324 (57%)	480 (54%)	850 (54%)	5,514 (55%)	39 (42%)	
Male	244 (43%)	411 (46%)	731 (46%)	4,549 (45%)	54 (58%)	
Assay version						>0.9
xT.v4	488 (86%)	758 (85%)	1,372 (87%)	8,589 (85%)	80 (86%)	
xT.v2	36 (6.3%)	59 (6.6%)	96 (6.1%)	680 (6.8%)	7 (7.5%)	
xT.v3	44 (7.7%)	74 (8.3%)	113 (7.1%)	794 (7.9%)	6 (6.5%)	
Cancer stage						<0.001
Stage 1	27 (6.7%)	59 (10%)	131 (12%)	895 (13%)	6 (8.7%)	
Stage 2	29 (7.2%)	36 (6.2%)	64 (6.0%)	553 (8.0%)	4 (5.8%)	
Stage 3	58 (14%)	85 (15%)	174 (16%)	1,180 (17%)	10 (14%)	
Stage 4	290 (72%)	403 (69%)	695 (65%)	4,281 (62%)	49 (71%)	
Unknown	164	308	517	3,154	24	
Tumor grade						0.8
Low	24 (12%)	32 (12%)	54 (10%)	424 (12%)	3 (7.7%)	
Medium	77 (39%)	98 (36%)	214 (41%)	1,397 (38%)	12 (31%)	
High	98 (49%)	146 (53%)	256 (49%)	1,858 (51%)	24 (62%)	
Unknown	369	615	1,057	6,384	54	

MSI status						0.4
Low/Stable	567 (100%)	888 (100%)	1,573 (100%)	10,032 (100%)	92 (99%)	
High	1 (0.2%)	3 (0.3%)	7 (0.4%)	29 (0.3%)	1 (1.1%)	
Unknown	0	0	1	2	0	
TMB count (mutations/mb)						
	3.1 (1.7, 5.4)	3.4 (1.9, 5.4)	5.8 (3.1, 9.2)	4.2 (2.5, 6.9)	3.9 (2.7, 6.9)	<0.001
Unknown	0	0	1	5	0	
Tumor/normal tissue status						0.045
Tumor and normal	291 (51%)	433 (49%)	727 (46%)	5,020 (50%)	49 (53%)	
Tumor only	277 (49%)	458 (51%)	854 (54%)	5,043 (50%)	44 (47%)	
Stated race						<0.001
American Indian or Alaska Native	8 (3.6%)	0 (0%)	0 (0%)	10 (0.2%)	0 (0%)	
Asian	2 (0.9%)	364 (87%)	1 (0.1%)	6 (0.1%)	2 (5.3%)	
Black or African American	3 (1.4%)	2 (0.5%)	834 (95%)	4 (<0.1%)	8 (21%)	
Native Hawaiian or Other Pacific Islander	0 (0%)	10 (2.4%)	0 (0%)	2 (<0.1%)	0 (0%)	
Other Race	88 (40%)	30 (7.2%)	24 (2.7%)	181 (3.2%)	6 (16%)	
Race not stated	3 (1.4%)	0 (0%)	2 (0.2%)	15 (0.3%)	1 (2.6%)	
White	118 (53%)	11 (2.6%)	13 (1.5%)	5,439 (96%)	21 (55%)	
Unknown	346	474	707	4,406	55	
Stated ethnicity						<0.001
Hispanic or Latino	233 (89%)	7 (2.8%)	10 (2.3%)	63 (1.8%)	9 (35%)	
Not Hispanic or Latino	28 (11%)	246 (97%)	423 (98%)	3,403 (98%)	17 (65%)	
Unknown	307	638	1,148	6,597	67	

¹ n (%); Median (IQR)

² Pearson's Chi-squared test; Wilcoxon rank sum test; Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

Figures

Figure 1. Ancestry proportions for patients in cohort by imputed race and ethnicity. Vertical bars represent the global continental ancestry admixture proportions for each patient, color-coded according to the scale provided in the legend. Patients are categorized by imputed race and ethnicity on the x-axis and are further sorted within each group by increasing values of their predominant continental ancestry.

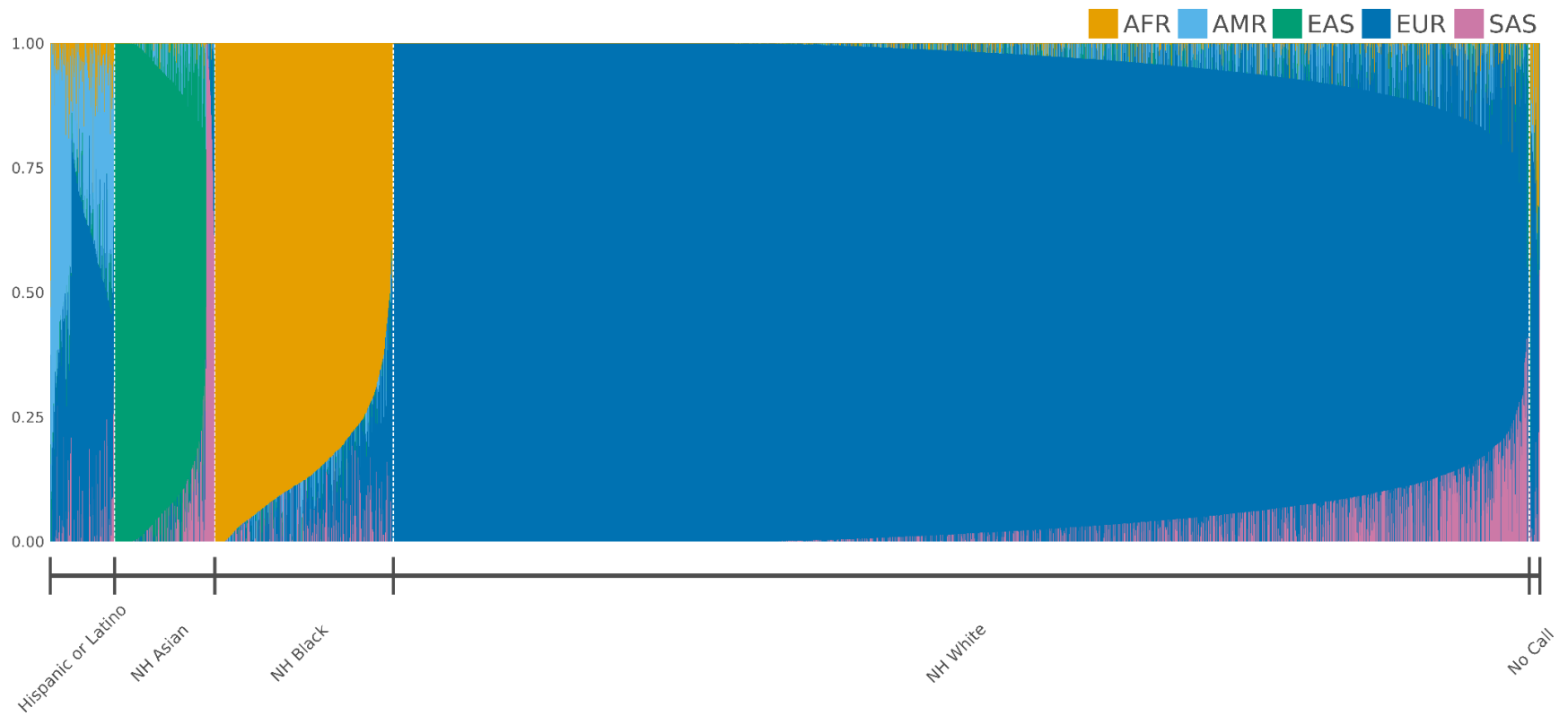


Figure 2. Forest plots for associations between genetic ancestry and SCNAs, gene fusions, and driver somatic mutations. The odds ratios (depicted as circles) and 95% confidence intervals (represented by horizontal lines) are shown for LUAD genes that met our criteria and displayed significant likelihood ratio tests (LRT) after adjusting for multiple testing at least in the univariable analysis. Colors represent ancestries as per color legend, full circles indicate significant logistic regression results, while empty circles denote cases where the odds ratio did not reach significance. Panel A, significant findings from univariable analyses without adjusting for smoking (n=13,196). Panel B, results after adjusting for smoking status derived from multiple imputation (n=13,196).

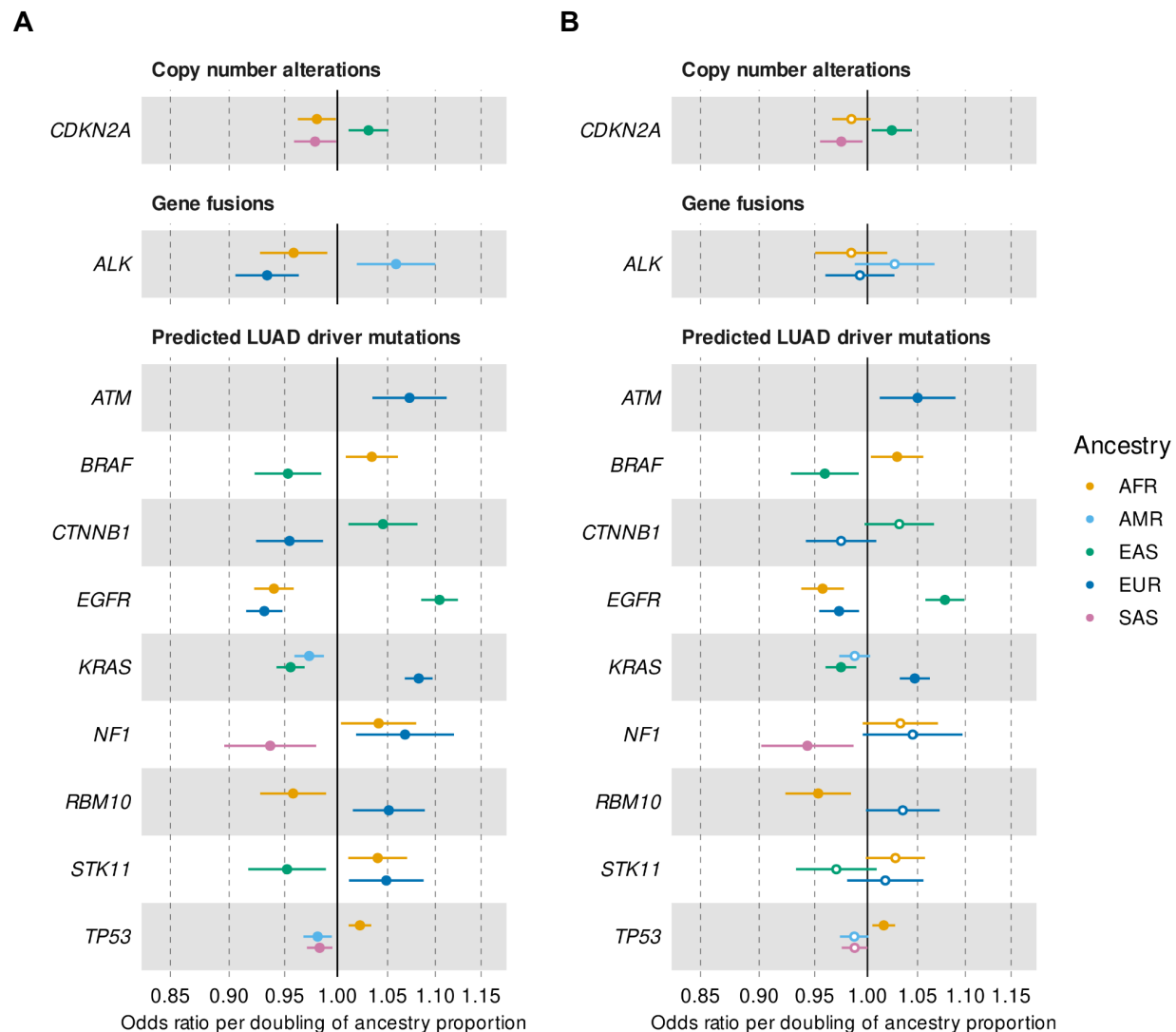


Figure 3. Forest plots for associations between imputed R/E categories and somatic mutations. Odds ratios (circles) and 95% confidence intervals (horizontal lines) for imputed race and ethnicity associations with somatic mutations using NH White as the reference group. LUAD genes that met our criteria and displayed significant likelihood ratio tests (LRT) after adjusting for multiple testing in at least in the univariable analysis are shown. Full circles indicate significant logistic regression results, while empty circles denote cases where the odds ratio did not reach significance. Panel A, all significant results from univariable analyses without adjustment for smoking (N=13,103). Panel B, results in the multiple imputation analyses adjusted for smoking status (N=13,103).

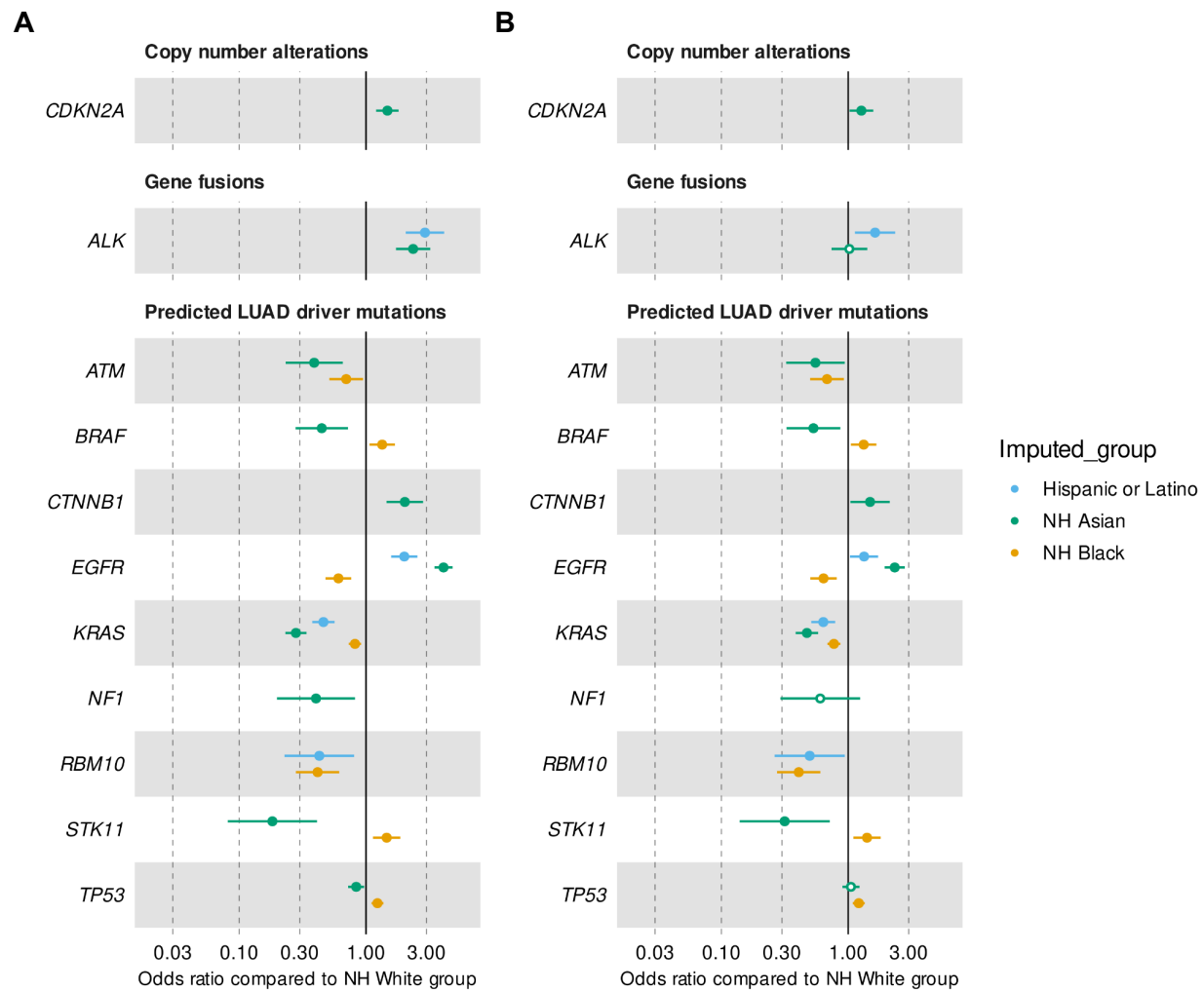
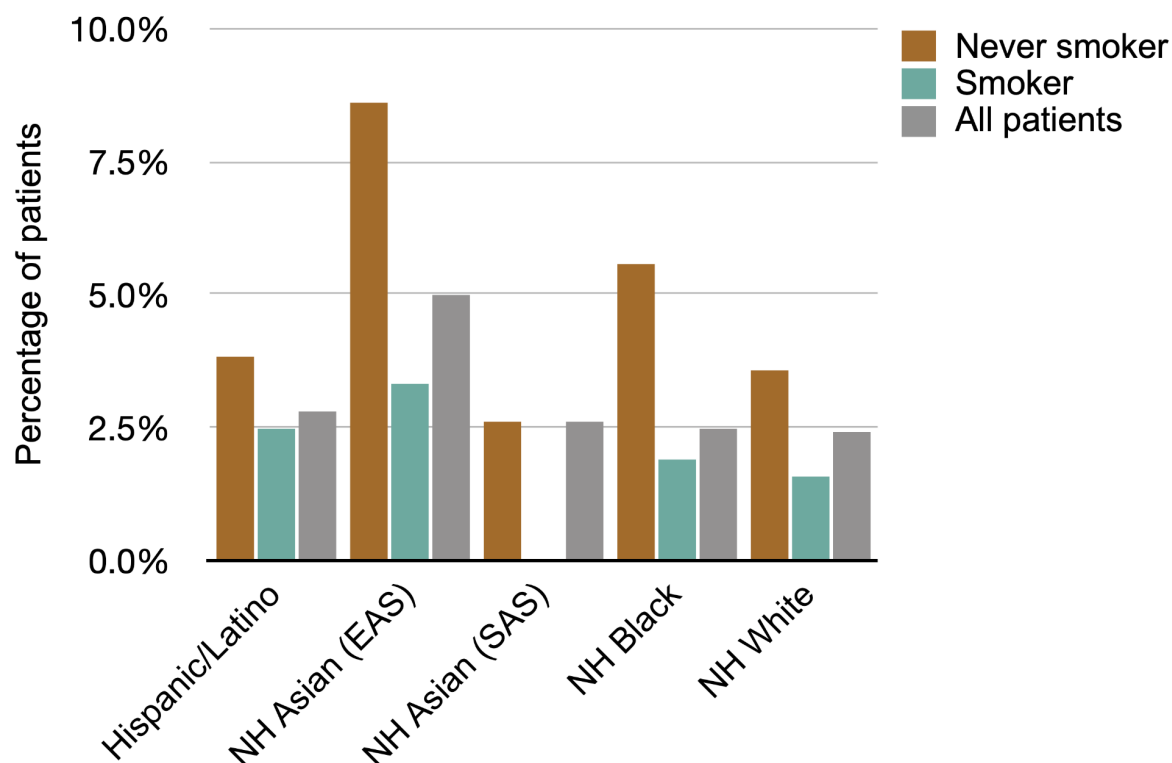


Figure 4. Distribution of driver mutations in *CTNNB1* by R/E and smoking status.

Percentage of patients harboring a driver mutation in the *CTNNB1* gene predicted by boostDM, stratified by smoking status and R/E. Smoking status is color coded as per the legend. “Smoker” refers to “current or former smoker”. Only patients with measured smoking status were included in calculations. Imputed R/E categories are shown in the x-axis. In the case of the NH Asian category, we split patients by their proportions of EAS and SAS: EAS > SAS is indicated as NH Asian (EAS), whereas patients with SAS > EAS are designated as SAS. R/E is statistically significantly associated with the presence of mutations for all patients and never smoker cohorts (Fisher exact test, $p=0.0018$, and $p=0.015$, correspondingly). See Supplementary Table 7 for pair-wise statistical assessment of the differences.



Supplementary Materials

Supplementary Table 1. Characteristics of patients with tumor and normal samples by imputed race and ethnicity category. Last column shows the results of statistical tests (as defined in footnote) for differences of the characteristics by imputed race and ethnicity.

Characteristic	Hispanic or Latino, N = 291 ¹	NH Asian, N = 433 ¹	NH Black, N = 727 ¹	NH White, N = 5,020 ¹	No Call, N = 49 ¹	p-value ²
Smoking status						<0.001
Former or current smoker	100 (52%)	163 (65%)	74 (16%)	610 (21%)	10 (29%)	
Never smoked tobacco	93 (48%)	86 (35%)	376 (84%)	2,309 (79%)	25 (71%)	
Unknown	98	184	277	2,101	14	
Age at specimen collection						<0.001
Unknown	32	68	87	680	7	
Age at diagnosis						<0.001
Unknown	47	79	133	942	10	
Gender						0.8
Female	167 (57%)	238 (55%)	399 (55%)	2,728 (54%)	24 (49%)	
Male	124 (43%)	195 (45%)	328 (45%)	2,292 (46%)	25 (51%)	
Assay version						0.6
xT.v4	238 (82%)	336 (78%)	597 (82%)	3,965 (79%)	38 (78%)	
xT.v2	23 (7.9%)	47 (11%)	60 (8.3%)	504 (10%)	6 (12%)	
xT.v3	30 (10%)	50 (12%)	70 (9.6%)	551 (11%)	5 (10%)	
Cancer stage						0.076
Stage 1	14 (6.2%)	33 (10%)	66 (12%)	466 (12%)	5 (12%)	
Stage 2	13 (5.8%)	20 (6.2%)	25 (4.6%)	261 (7.0%)	3 (7.3%)	
Stage 3	33 (15%)	49 (15%)	88 (16%)	634 (17%)	6 (15%)	
Stage 4	166 (73%)	220 (68%)	362 (67%)	2,369 (64%)	27 (66%)	
Unknown	65	111	186	1,290	8	
Tumor grade						0.5
Low	11 (10%)	18 (11%)	29 (11%)	225 (11%)	0 (0%)	

Medium	43 (39%)	51 (32%)	114 (42%)	734 (37%)	9 (39%)	
High	56 (51%)	88 (56%)	130 (48%)	1,012 (51%)	14 (61%)	
Unknown	181	276	454	3,049	26	
MSI status						0.15
Low/Stable	290 (100%)	432 (100%)	722 (99%)	5,007 (100%)	48 (98%)	
High	1 (0.3%)	1 (0.2%)	4 (0.6%)	13 (0.3%)	1 (2.0%)	
Unknown	0	0	1	0	0	
TMB count (mutations/mb)	1.9 (1.2, 3.9)	1.9 (1.2, 3.4)	3.9 (1.9, 7.7)	3.4 (1.9, 6.1)	3.3 (1.5, 6.9)	<0.001
Unknown	0	0	1	5	0	
Stated race						<0.001
American Indian or Alaska Native	3 (2.4%)	0 (0%)	0 (0%)	4 (0.1%)	0 (0%)	
Asian	0 (0%)	201 (87%)	1 (0.2%)	1 (<0.1%)	1 (4.0%)	
Black or African American	1 (0.8%)	0 (0%)	432 (95%)	2 (<0.1%)	7 (28%)	
Native Hawaiian or Other Pacific Islander	0 (0%)	3 (1.3%)	0 (0%)	1 (<0.1%)	0 (0%)	
Other Race	52 (42%)	17 (7.4%)	11 (2.4%)	73 (2.4%)	4 (16%)	
Race not stated	2 (1.6%)	0 (0%)	1 (0.2%)	7 (0.2%)	0 (0%)	
White	65 (53%)	10 (4.3%)	8 (1.8%)	2,985 (97%)	13 (52%)	
Unknown	168	202	274	1,947	24	
Stated ethnicity						<0.001
Hispanic or Latino	142 (93%)	5 (4.2%)	4 (2.0%)	34 (1.9%)	5 (38%)	
Not Hispanic or Latino	11 (7.2%)	115 (96%)	199 (98%)	1,747 (98%)	8 (62%)	
Unknown	138	313	524	3,239	36	

¹n (%); Median (IQR)

²Pearson's Chi-squared test; Wilcoxon rank sum test; Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates).

Supplementary Table 2. Cohort characteristics by smoking status.

Characteristic	Never smoked tobacco, N = 1,788¹	Former or current smoker, N = 5,345¹	p-value²
Imputed race and ethnicity category			<0.001
Hispanic or Latino	184 (10%)	157 (2.9%)	
NH Asian	317 (18%)	162 (3.0%)	
NH Black	142 (7.9%)	731 (14%)	
NH White	1,128 (63%)	4,256 (80%)	
No Call	17 (1.0%)	39 (0.7%)	
Age at specimen collection			0.3
Unknown	6	12	
Age at diagnosis			>0.9
Unknown	61	165	
Gender			<0.001
Female	1,222 (68%)	2,766 (52%)	
Male	566 (32%)	2,579 (48%)	
Assay version			0.3
xT.v4	1,440 (81%)	4,392 (82%)	
xT.v2	160 (8.9%)	452 (8.5%)	
xT.v3	188 (11%)	501 (9.4%)	
Cancer stage			<0.001
Stage 1	147 (9.9%)	586 (13%)	
Stage 2	109 (7.3%)	333 (7.4%)	
Stage 3	207 (14%)	808 (18%)	
Stage 4	1,027 (69%)	2,794 (62%)	
Unknown	298	824	
Tumor grade			<0.001

Low	119 (16%)	231 (9.4%)	
Medium	328 (45%)	881 (36%)	
High	279 (38%)	1,343 (55%)	
Unknown	1,062	2,890	
MSI status			0.8
Low/Stable	1,780 (100%)	5,325 (100%)	
High	7 (0.4%)	19 (0.4%)	
Unknown	1	1	
TMB count			<0.001
Unknown	1	3	
Tumor/normal tissue status			
Tumor and normal	957 (54%)	2,889 (54%)	0.7
Tumor only	831 (46%)	2,456 (46%)	
Stated race			<0.001
American Indian or Alaska Native	0 (0%)	10 (0.3%)	
Asian	185 (16%)	85 (2.3%)	
Black or African American	94 (8.0%)	521 (14%)	
Native Hawaiian or Other Pacific Islander	5 (0.4%)	2 (<0.1%)	
Other Race	79 (6.7%)	144 (3.9%)	
Race not stated	4 (0.3%)	10 (0.3%)	
White	812 (69%)	2,909 (79%)	
Unknown	609	1,664	
Stated ethnicity			<0.001
Hispanic or Latino	113 (15%)	123 (5.3%)	
Not Hispanic or Latino	665 (85%)	2,195 (95%)	
Unknown	1,010	3,027	

¹ n (%); Median (IQR)

² Pearson's Chi-squared test; Wilcoxon rank sum test; Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

Supplementary Table 3. Cohort characteristics by availability of smoking status.

Characteristic	Smoking status available, N = 7,133¹	Smoking status missing, N = 6,063¹	p-value²
Imputed race and ethnicity category			0.019
Hispanic or Latino	341 (4.8%)	227 (3.7%)	
NH Asian	479 (6.7%)	412 (6.8%)	
NH Black	873 (12%)	708 (12%)	
NH White	5,384 (75%)	4,679 (77%)	
No Call	56 (0.8%)	37 (0.6%)	
Age at specimen collection			<0.001
Unknown	18	2,551	
Age at diagnosis			<0.001
Unknown	226	2,954	
Gender			0.001
Female	3,988 (56%)	3,219 (53%)	
Male	3,145 (44%)	2,844 (47%)	
Assay version			<0.001
xT.v4	5,832 (82%)	5,455 (90%)	
xT.v2	612 (8.6%)	266 (4.4%)	
xT.v3	689 (9.7%)	342 (5.6%)	
Cancer stage			0.5
Stage 1	733 (12%)	385 (13%)	
Stage 2	442 (7.4%)	244 (8.1%)	
Stage 3	1,015 (17%)	492 (16%)	
Stage 4	3,821 (64%)	1,897 (63%)	
Unknown	1,122	3,045	
Tumor grade			0.4
Low	350 (11%)	187 (12%)	

Medium	1,209 (38%)	589 (38%)	
High	1,622 (51%)	760 (49%)	
Unknown	3,952	4,527	
MSI status			0.2
Low/Stable	7,105 (100%)	6,047 (100%)	
High	26 (0.4%)	15 (0.2%)	
Unknown	2	1	
TMB count	4.2 (2.3, 7.1)	4.6 (2.7, 7.3)	<0.001
Unknown	4	2	
Tumor/normal tissue status			<0.001
Tumor and normal	3,846 (54%)	2,674 (44%)	
Tumor only	3,287 (46%)	3,389 (56%)	
Stated race			0.01
American Indian or Alaska Native	10 (0.2%)	8 (0.3%)	
Asian	270 (5.6%)	105 (4.5%)	
Black or African American	615 (13%)	236 (10%)	
Native Hawaiian or Other Pacific Islander	7 (0.1%)	5 (0.2%)	
Other Race	223 (4.6%)	106 (4.5%)	
Race not stated	14 (0.3%)	7 (0.3%)	
White	3,721 (77%)	1,881 (80%)	
Unknown	2,273	3,715	
Stated ethnicity			0.2
Hispanic or Latino	236 (7.6%)	86 (6.4%)	
Not Hispanic or Latino	2,860 (92%)	1,257 (94%)	
Unknown	4,037	4,720	

¹n (%); Median (IQR)

²Pearson's Chi-squared test; Wilcoxon rank sum test; Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates).

Supplementary Table 4. List of LUAD genes tested for association by mutation type. Genes tested varied depending on availability of data and type detected by assay (e.g. gene fusions and somatic copy number alterations, SCNAs) or available annotations (e.g. LUAD genes included in OncoKB as L1, L2 or R1, and available LUAD boostDM gene models).

Mutation type	Number of genes	Genes
Protein-altering SNV/indels	22	<i>ALK, ARID1A, ATM, BRAF, CDKN2A, CTNNB1, EGFR, ERBB2, KEAP1, KRAS, MET, NF1, PIK3CA, RB1, RBM10, RET, ROS1, SETD2, SMARCA4, STK11, TP53, U2AF1</i>
SCNAs	5	<i>CDKN2A, EGFR, ERBB2, KRAS, MET</i>
Gene fusions	1	<i>ALK</i>
OncoKB actionable	4	<i>BRAF, EGFR, KRAS, PIK3CA</i>
boostDM predicted drivers	10	<i>ATM, BRAF, CTNNB1, EGFR, KRAS, NF1, PIK3CA, RBM10, STK11, TP53</i>

Supplementary Table 5. Summary of likelihood ratio test results for genetic ancestry and SCNA, gene fusion, and predicted driver mutations. Univariable and multiple imputation analyses were performed on N=13,196 patient samples, and complete case analyses were performed on N=7,133 patient samples. Bold indicates statistically significant corrected p-value.

Mutation type	Gene	N tests	N (%) with mutation in all patients	N (%) with mutation in patients with known smoking status	Univariable analysis corrected p-value	Complete Case analysis corrected p-value	Multiple imputation analysis corrected p-value
SCNA	<i>CDKN2A</i>	5	1,275 (10%)	681 (10%)	0.0017	0.2036	0.0319
Fusion	<i>ALK</i>	1	372 (3%)	228 (3%)	8.8e-06	0.7900	0.6876
Driver	<i>ATM</i>	10	503 (4%)	258 (4%)	0.0024	0.1862	0.0845
Driver	<i>BRAF</i>	10	579 (4%)	305 (4%)	0.0089	0.1469	0.0845
Driver	<i>CTNNB1</i>	10	347 (3%)	175 (2%)	0.0040	0.0074	0.2123
Driver	<i>EGFR</i>	10	1,248 (9%)	661 (9%)	5.47E-48	1.01E-07	1.08E-17
Driver	<i>KRAS</i>	10	4,569 (35%)	2,411 (34%)	1.57E-36	5.21E-05	1.57E-10
Driver	<i>NF1</i>	10	284 (2%)	136 (2%)	0.0024	0.1469	0.0686
Driver	<i>RBM10</i>	10	476 (4%)	228 (3%)	0.0062	0.1862	0.0296
Driver	<i>STK11</i>	10	461 (3%)	252 (4%)	0.0021	0.6165	0.2014
Driver	<i>TP53</i>	10	4,980 (38%)	2,739 (38%)	0.0006	0.0142	0.0029

Supplementary Table 6. Summary of likelihood ratio test results for imputed race and ethnicity and SCNA, gene fusion, and predicted driver mutations. Univariable and multiple imputation analyses were performed on N=13,103 patient samples, and complete case analyses were performed on N=7,077 patient samples. Bold indicates statistically significant corrected p-value.

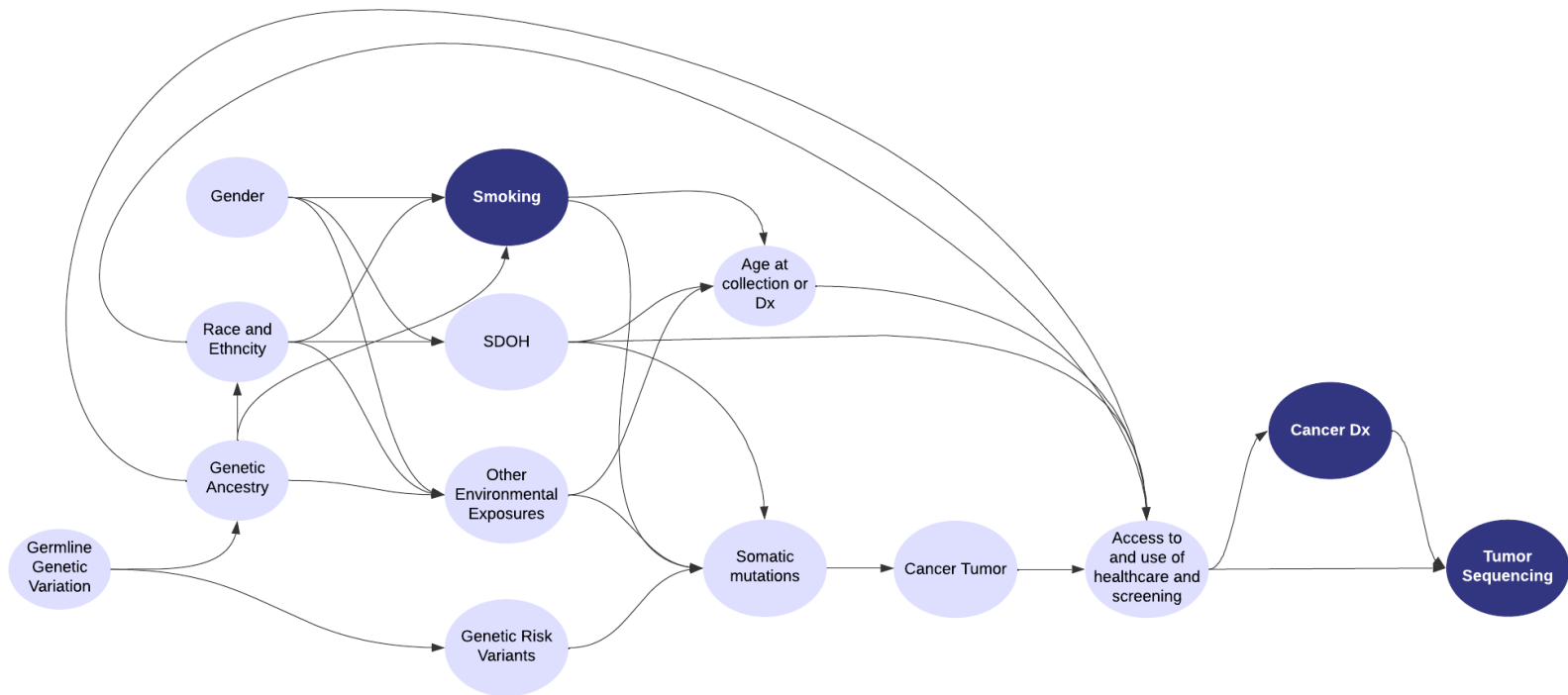
Mutation type	Gene	N tests	N (%) with mutation in all patients	N (%) with mutation in patients with known smoking status	Univariable analysis corrected p-value	Complete Case analysis corrected p-value	Multiple imputation analysis corrected p-value
SCNA	<i>CDKN2A</i>	5	1264 (10%)	675 (10%)	0.0026	0.5944	0.5266
Fusion	<i>ALK</i>	1	368 (3%)	225 (3%)	1.24E-11	0.4187	0.1117
Driver	<i>ATM</i>	10	502 (4%)	257 (4%)	8.50E-05	0.0300	0.0118
Driver	<i>BRAF</i>	10	575 (4%)	302 (4%)	8.50E-05	0.0026	0.0118
Driver	<i>CTNNB1</i>	10	345 (3%)	173 (2%)	0.0022	0.0197	0.3224
Driver	<i>EGFR</i>	10	1242 (9%)	657 (9%)	1.39E-63	8.43E-09	1.13E-21
Driver	<i>KRAS</i>	10	4540 (35%)	2394 (34%)	1.87E-57	5.61E-06	1.33E-16
Driver	<i>NF1</i>	10	278 (2%)	132 (2%)	0.0238	0.0978	0.0999
Driver	<i>RBM10</i>	10	474 (4%)	227 (3%)	1.47E-06	0.0106	5.65E-06
Driver	<i>STK11</i>	10	461 (4%)	252 (4%)	4.29E-09	0.0197	9.27E-05
Driver	<i>TP53</i>	10	4943 (38%)	2714 (38%)	4.11E-05	0.0891	0.0236

Supplementary Table 7. Fisher exact test for differences in the presence of *CTNNB1* driver variants across R/E categories for all patients and stratified by current/former smokers or never smokers. 2-sided group test p-value under strata label; pairwise tests in cells (significant cells highlighted in orange). NH Asian patients were stratified by those with EAS or SAS ancestries.

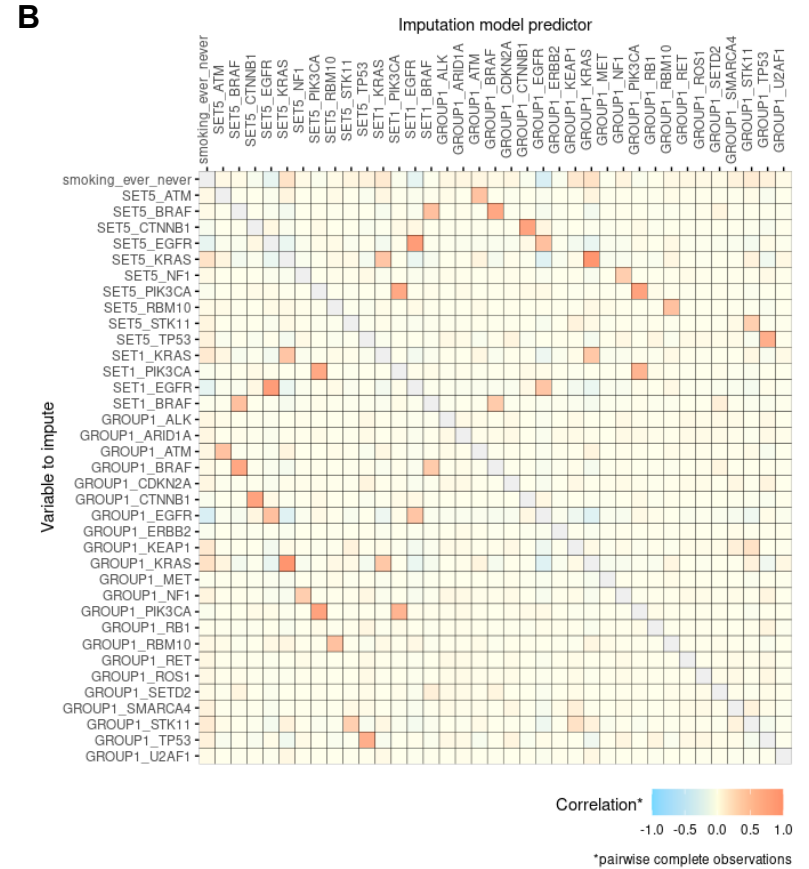
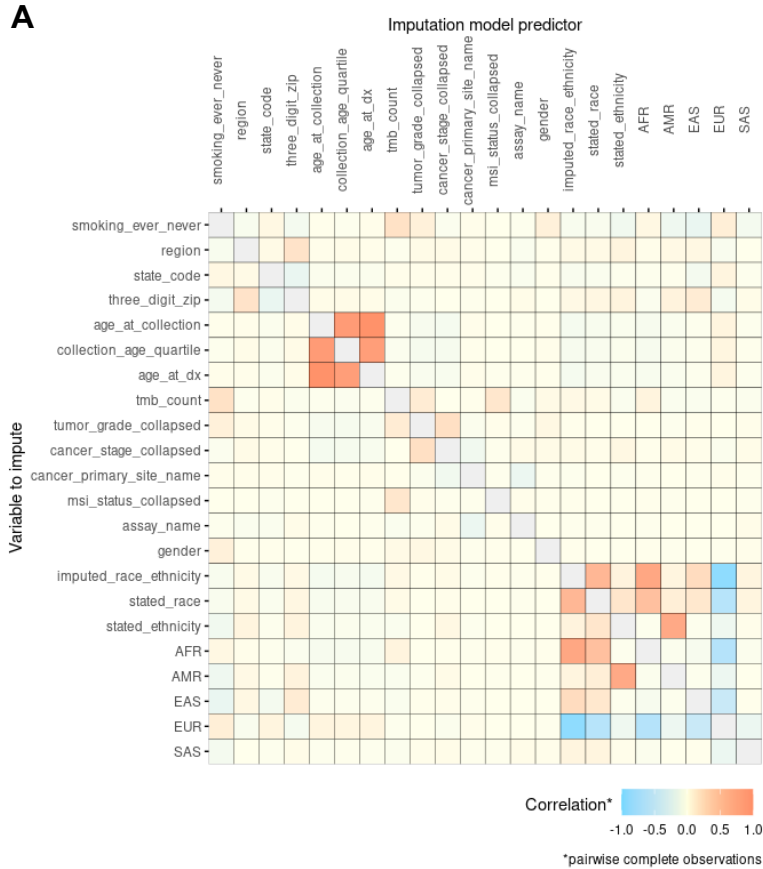
	Hispanic or Latino	NH Asian (EAS)	NH Asian (SAS)	NH Black
All patients (p=0.001836)				
NH Asian (EAS)	0.0534	-	-	-
NH Asian (SAS)	1	0.5727	-	-
NH Black	0.7588	0.0018	0.7187	-
NH White	0.5764	5.51 x 10 ⁻⁰⁵	0.7097	0.7938
Current or former smoker (p=0.3619)				
NH Asian (EAS)	0.7453	-	-	-
NH Asian (SAS)	1	1	-	-
NH Black	0.5410	0.3481	1	-
NH White	0.3304	0.1062	1	0.5327
Never smoker (p=0.01521)				
NH Asian (EAS)	0.0562	-	-	-
NH Asian (SAS)	1	0.3342	-	-
NH Black	0.4398	0.3338	0.6868	-
NH White	0.8338	0.0011	1	0.2458

Supplementary Figures

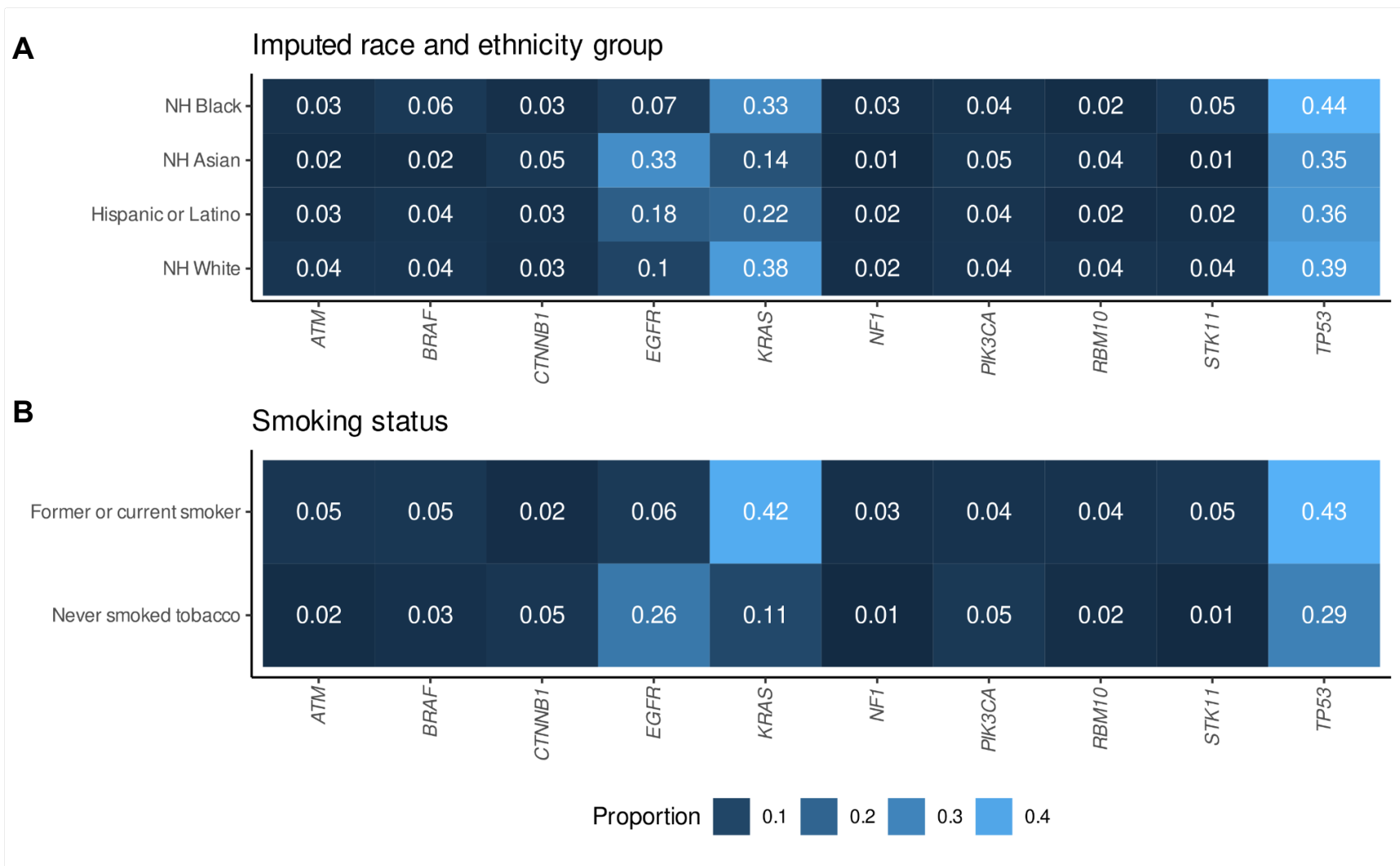
Supplementary Figure 1. Directed Acyclic Graph (DAG) illustrating the causal pathways in lung adenocarcinoma progression. The graph delineates the relationships between race/ethnicity, genetic ancestry, gender, and key variables such as smoking, social determinants of health (SDOH), environmental and genetic exposures, and their collective impact on somatic mutations in genes. The downstream effects on cancer diagnosis, access to healthcare, and tumor sequencing are also represented. Dark blue shapes indicate variables that are controlled for in regression analysis or by selection into the study.



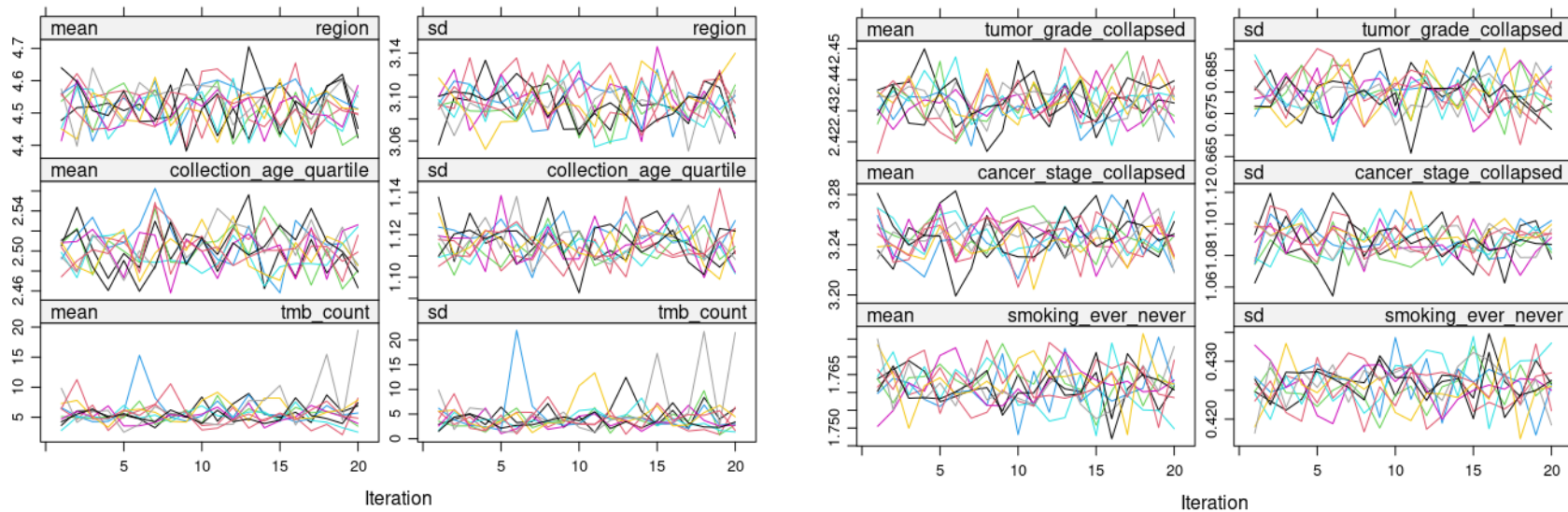
Supplementary Figure 2. Correlation of clinical and somatic mutation data with smoking status. Panel A spearman correlation between clinical variables. Panel B spearman correlation between somatic mutation counts among genes and smoking status. Group 1, short protein altering variants; Group 2, SCNAs; Group 3, gene fusions, Set 1, OncoKb actionable variants; Set 5 BoostDM predicted drivers. Only cases with complete smoking status are included.



Supplementary Figure 3. Proportion of patients with protein-altering variants in LUAD genes with respect to R/E and smoking status. Panel A: Fraction of patients carrying somatic protein-altering variants in selected LUAD genes (cf. Methods) with respect to imputed R/E categories. Panel B: fraction of patients carrying somatic protein-altering variants in selected LUAD genes (cf. Methods) with respect to smoking status (only patients with available smoking status included).

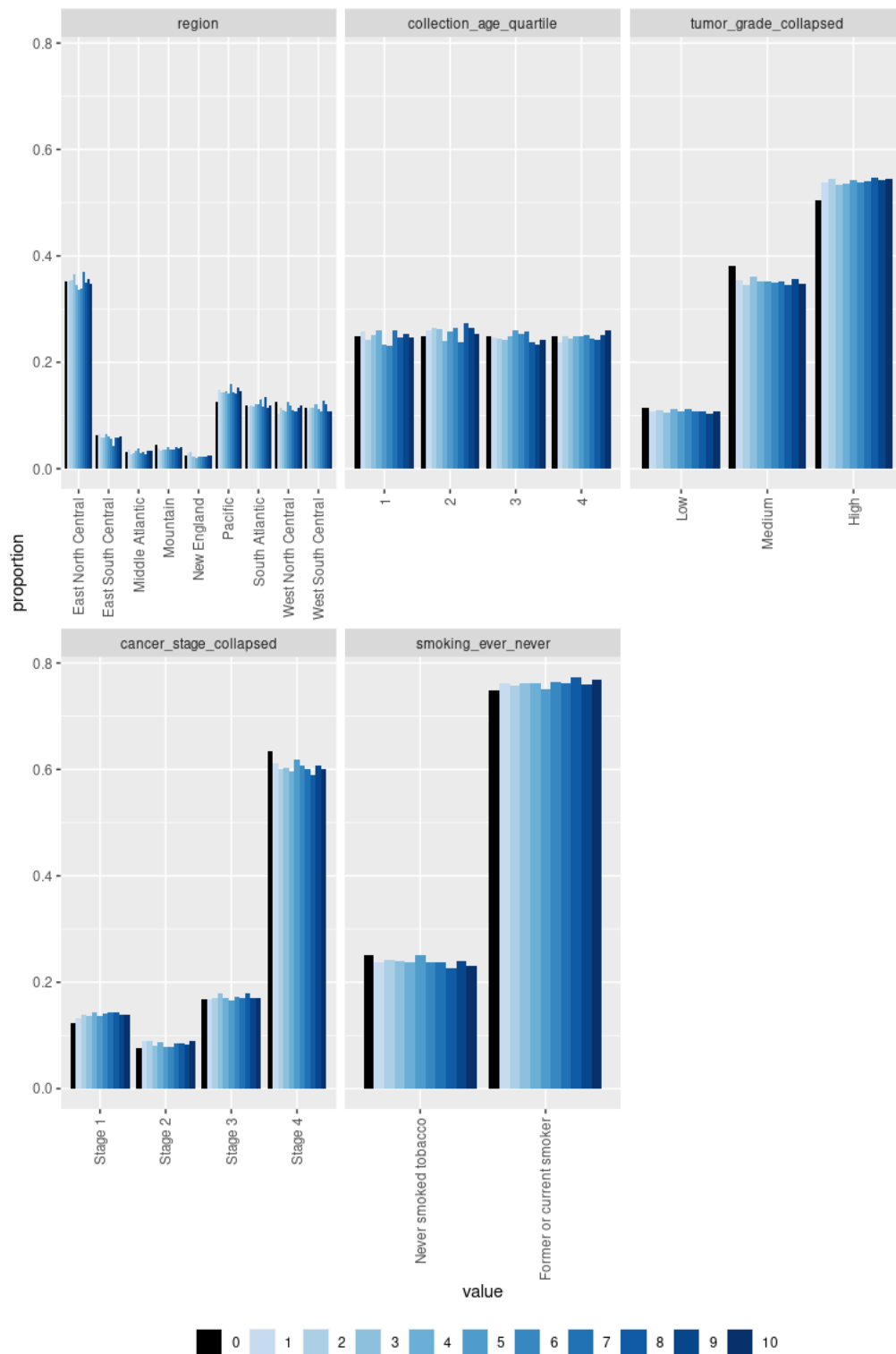


Supplementary Figure 4. Convergence plots for multiple imputation. The mean and standard deviation (SD) for each of the imputed variables at each iteration of multiple imputation using MICE is shown. Variables included: Region, US geographical region of patient home address; Age at collection of tissue; TMB count (mutations/Mb); Tumor grade; Cancer stage; Smoking status, current (smoking), former smoker (ever), or never a smoker (never).

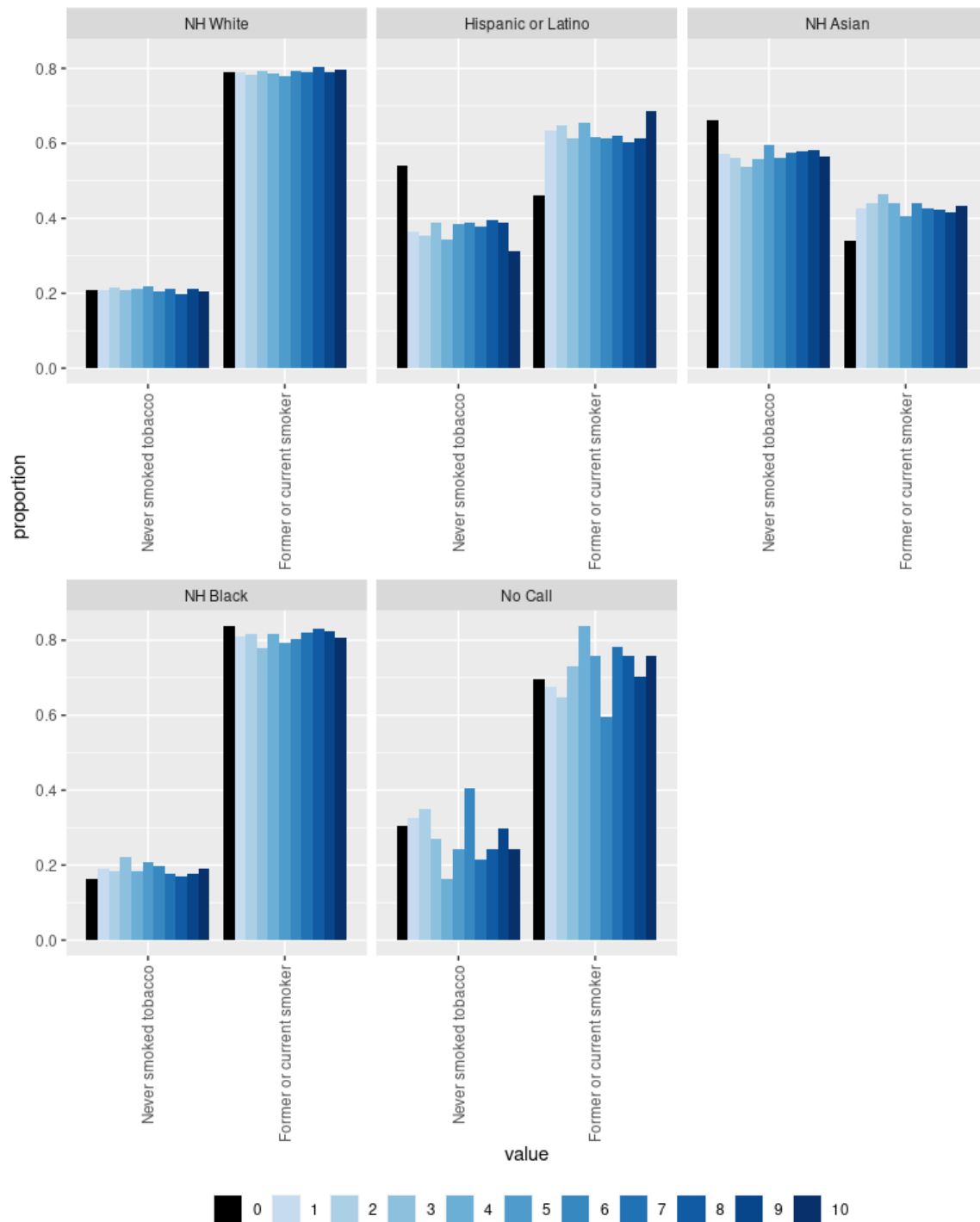


Supplementary Figure 5. Distribution of values of multiply imputed categorical variables.

The first bar (0) shows the distribution of measured categorical variables, remaining bars are data for each of the 10 imputed datasets (1-10). Regions: US geographical regions of patient home address.

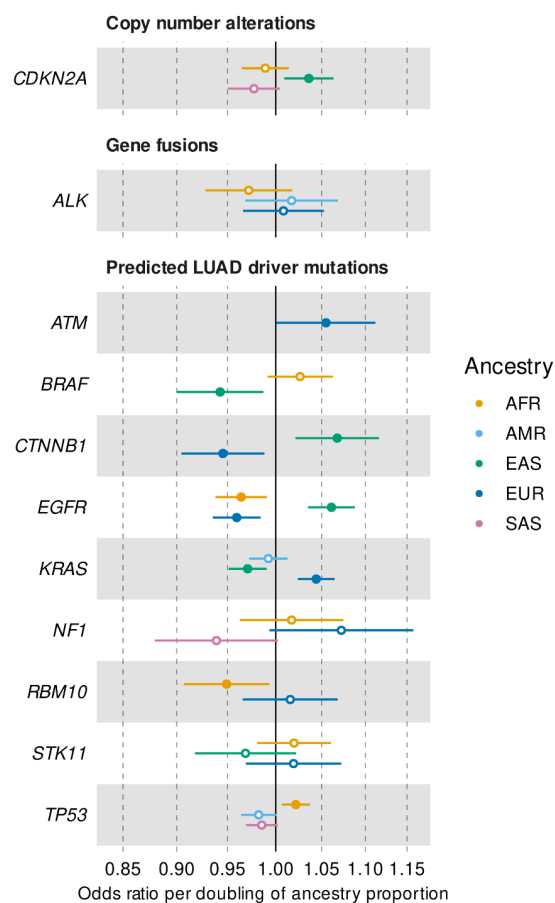


Supplementary Figure 6. Distribution of multiply imputed smoking status by imputed race and ethnicity category. The first bar (0) shows the distribution of measured smoking status, subsequent bars show values for each of the 10 imputed datasets (1-10).

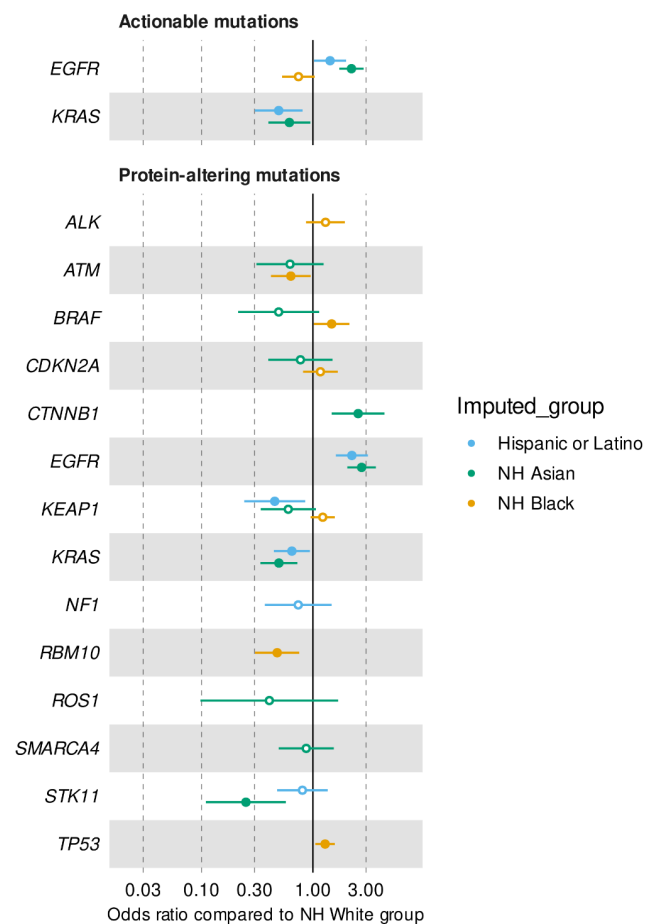


Supplementary Figure 7. Forest plots for associations between genetic ancestry and imputed R/E with SCNAs, gene fusions, and driver somatic mutations adjusted for smoking for complete case analysis (n=7,133). The odds ratios (depicted as circles) and 95% confidence intervals (represented by horizontal lines) are shown for LUAD genes that met our criteria and displayed significant likelihood ratio tests (LRT) after adjusting for multiple testing at least in the univariable analysis. Colors representing ancestries as per color legend, full circles indicate significant logistic regression results, while empty circles denote cases where the odds ratio did not reach significance. Panel A, results for associations with genetic ancestry. Panel B, results for associations with imputed R/E categories using NH White as the reference group.

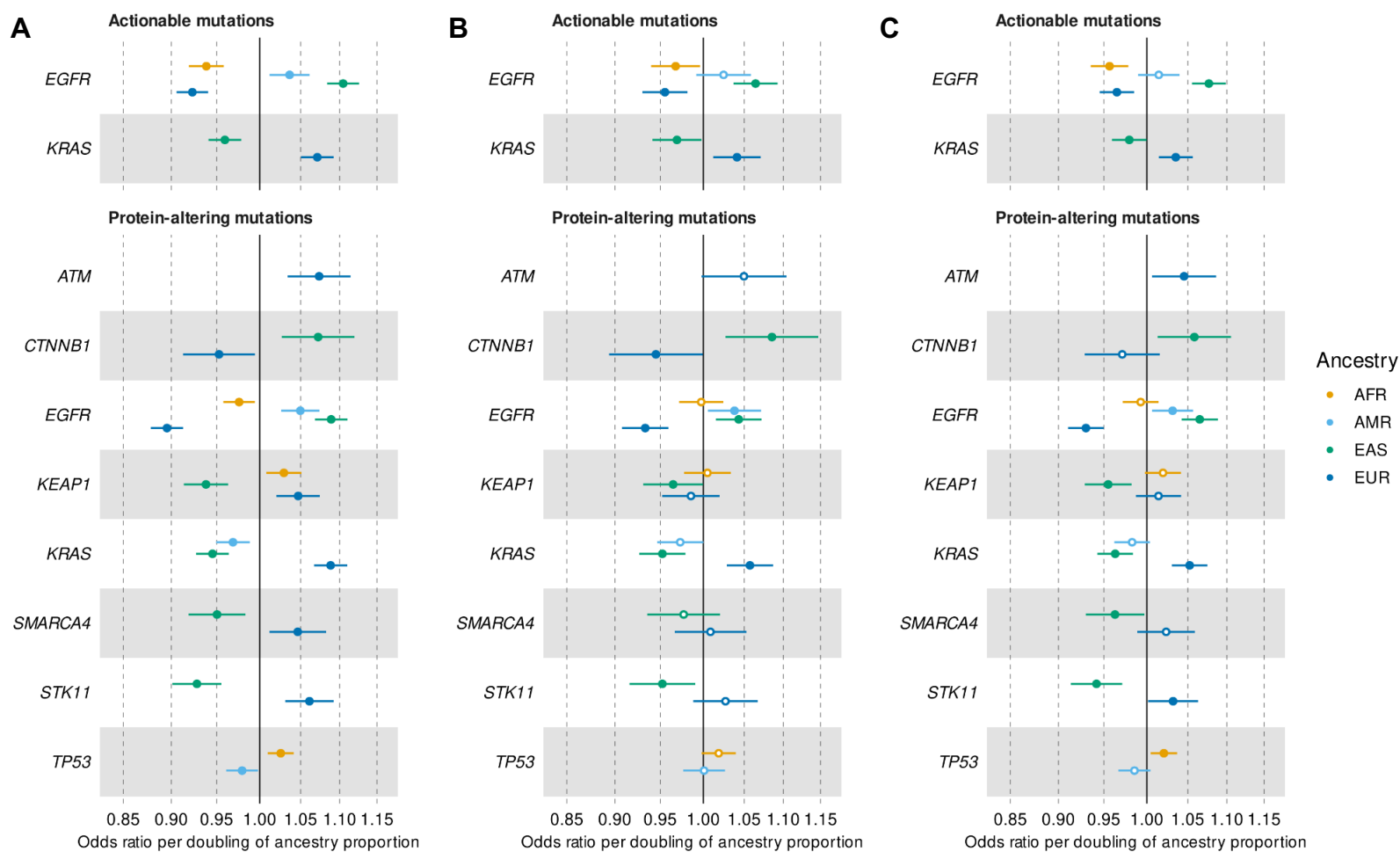
A



B



Supplementary Figure 8. Forest plots for associations between genetic ancestry and actionable and protein-altering somatic mutations. The odds ratios (depicted as circles) and 95% confidence intervals (represented by horizontal lines) are shown for LUAD genes that met our criteria and displayed significant likelihood ratio tests (LRT) at least in the univariable analysis. Colors represent ancestries as per color legend, full circles indicate significant LRT results, while empty circles denote cases where the odds ratio did not reach significance after adjusting for multiple testing. Actionable mutations, OncoKB Level 1, 2, and R1. Panel A, significant findings from univariable analyses without adjusting for smoking (n=13,196). Panel B, results from complete case analyses adjusted for smoking status (n=7,133). Panel C, results after adjusted for smoking status derived from multiple imputation (n=13,196).



Supplementary Figure 9. Forest plots for associations between imputed R/E categories and actionable and protein-altering somatic mutations. Odds ratios (circles) and 95% confidence intervals (horizontal lines) for imputed race and ethnicity associations with somatic mutations using NH White as the reference group. Full circles indicate significant LRT results, while empty circles denote cases where the odds ratio did not reach significance after adjusting for multiple testing. Actionable mutations, OncoKB Level 1, 2, and R1. Panel A, all significant results from univariable analyses without adjustment for smoking. Panel B, results in the complete case analyses adjusted for smoking status. Panel C, results in the multiple imputation analyses adjusted for smoking status.

