Supplementary Online Content

Campbell JD, Lathan C, Sholl L, et al. The mutational landscape of lung cancers from black and white populations. *JAMA Oncology*. Published online January 19, 2017. doi:10.1001/jamaoncol.2016.6108

eMethods

Additional cohort information

Independent pathological assessment was performed by a board certified Anatomic Pathologist with thoracic pathology expertise in the Department of Pathology, Brigham and Women's Hospital. Never-smokers were defined as adults who had smoked less than 100 cigarettes in their lifetime. A former smoker was defined as an adult who had smoked at least 100 cigarettes in his or her lifetime but who had quit smoking at the time of interview.

Hybrid capture and sequencing of cancer genes

DNA was extracted and sonicated to 250 bp following Covaris FFPE DNA Extraction & Purification protocol and further purified using Agencourt AMPure XP beads. Libraries were made with sample-specific barcodes, quantified using qPCR, and the libraries were pooled in equimolar concentrations to a total of 500 ng for OncoPanel enrichment (Agilent SureSelect) for hybrid capture of 502 cancer-related genes. Libraries were sequenced on an Illumina HiSeq2500 to a mean depth-of-coverage of 208X (range 0.50-682X). Tumors with >30X sequencing depth over >80% of targeted bases were analyzed.

Alignment

Pooled sample reads were de-multiplexed and sorted using Picard (http://broadinstitute.github.io/picard/command-line-overview.html). Reads were aligned to the reference sequence b37 edition from the Human Genome Reference Consortium using BWA using the following parameters "-q 5 -l 32 -k 2 -o 1". Duplicate reads were marked using Picard tools. The Genome Analysis Tool Kit (GATK2) was used for local realignment of reads around indels and base quality score recalibration (BQSR).

Mutation calling

MuTect and SomaticIndelDetector (http://www.broadinstitute.org/cancer/cga/indelocator) were used to identify somatic single nucleotide variants (SNVs) and short insertions or deletions (indels), respectively. Mutation annotation was performed using Oncotator.

Copy number and rearrangements calling

Copy number alterations were called using ReCapSeg (http://gatkforums.broadinstitute.org/categories/cancer-tools) using default parameters. ReCapSeg performs tangent normalization against a "panel-of-normals" to remove systematic noise and technical artifacts. We used a panel of 49 samples from normal tissue and three normal cells lines profiled with the same OncoPanel platform.

Rearrangements were detected using BreaKmer (https://github.com/ccgd-profile/BreaKmer) and confirmed by manual review of sequences.

Inference of mutational signatures

Non-negative matrix factorization (NMF) was used to deconvolute a *M x N* matrix of mutation catalogues into a *M x K* matrix of mutational processes and an *K x N* matrix of mutational exposures (where *N* is the number of tumors, *M* is the number of mutation types, and *K* is the number of estimated mutational processes). Code to run NMF was obtained from http://www.mathworks.com/matlabcentral/fileexchange/38724 and run using the nnmf function from the MATLAB Statistics Toolbox. We used 6 mutation types with 16 different trinucleotide contexts for a total of 96 mutational states. The number of possible signatures was varied from 1 to 10 and signature stability was assessed via sampling as previously described [ref]. The cosine correlation was used to determine similarity between signatures derived in this dataset and signatures from the COSMIC database.

Principle Component Analysis

For the PCA analysis, we used 3,517 SNV sites called by MuTect that were found in more than 1% of African American or European populations in the ExAC database. Variants with an alternate allele fraction less than 0.95 were considered heterozygous while variants with an alternate allele fraction greater than 0.95 were considered homozygous for the alternate allele. Let X_{ij}be the matrix of variants with SNV i and tumor j, where $i = 1 \mathinner{..} M$ and $j = 1 \mathinner{..} N$. Each entry in the matrix was coded as 0, 1, or 2 corresponding to homozygous or for the reference, heterozygous, or homozygous for the alternate allele, respectively. Each SNV was centered and scaled according to equation 1:

(Eq. 1)

$$X'_{ij} = \frac{X_{ij} - \mu_i}{\sqrt{p_i(1 - p_i)}}$$

where $p_i = \mu_i/2$ and $\mu_i = \frac{1}{N}\Sigma_{j=1}^{N} X_{ij}$. PCA was performed on the scaled matrix X'_{ij} using the prcomp function in the R statistical computing language.

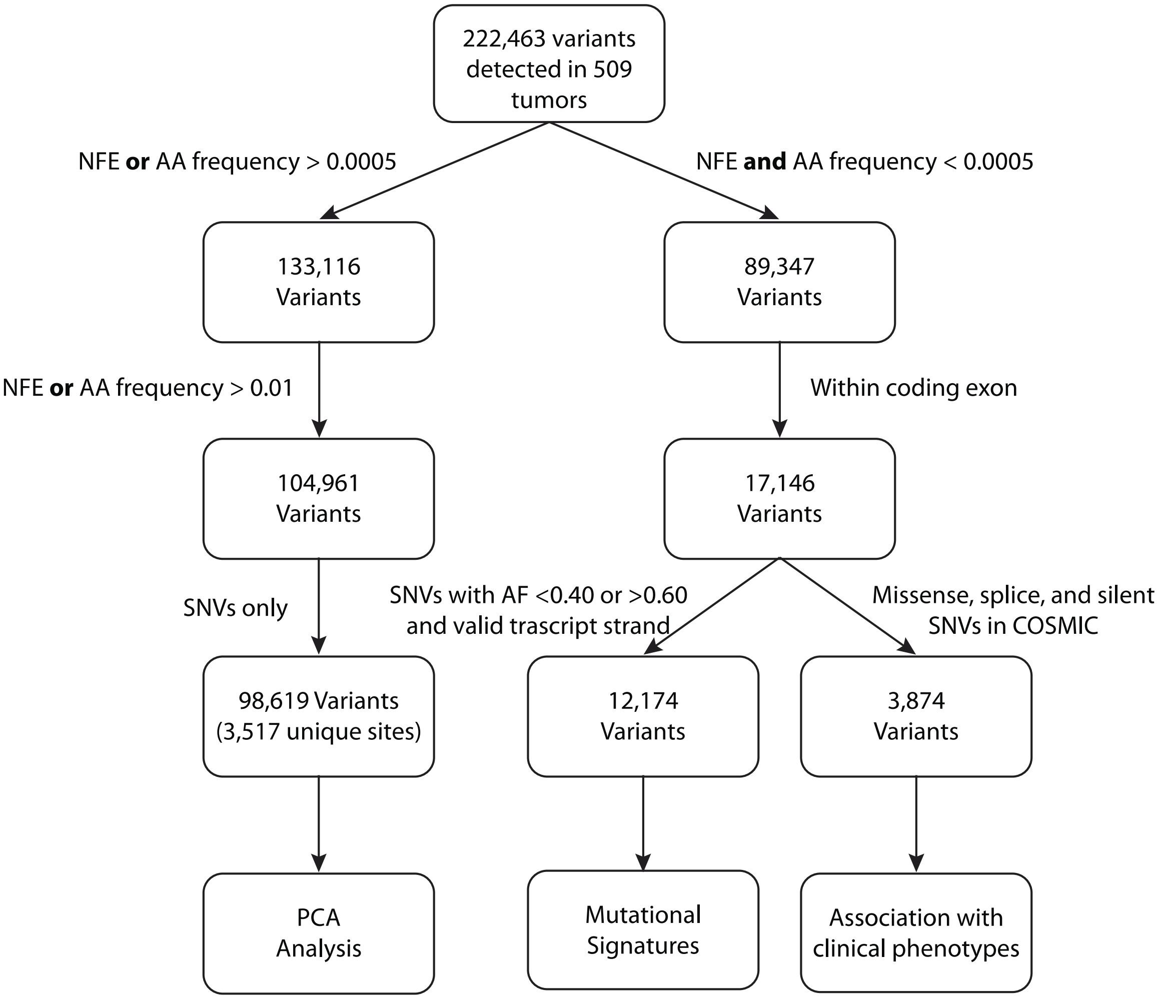
Statistical analysis

For each gene, the Fisher's exact test was used to compare the frequency of mutated tumors between populations. Genes mutated in five or fewer tumors were not considered. Nonparametric tests such as the Wilcoxon rank-sum test (comparison between 2 groups) or the Kruskal-Wallis test (comparison between more than two groups) were used for continuous variables. Correction for multiple hypothesis testing was performed with the Benjamini-Hochberg procedure.

Study power

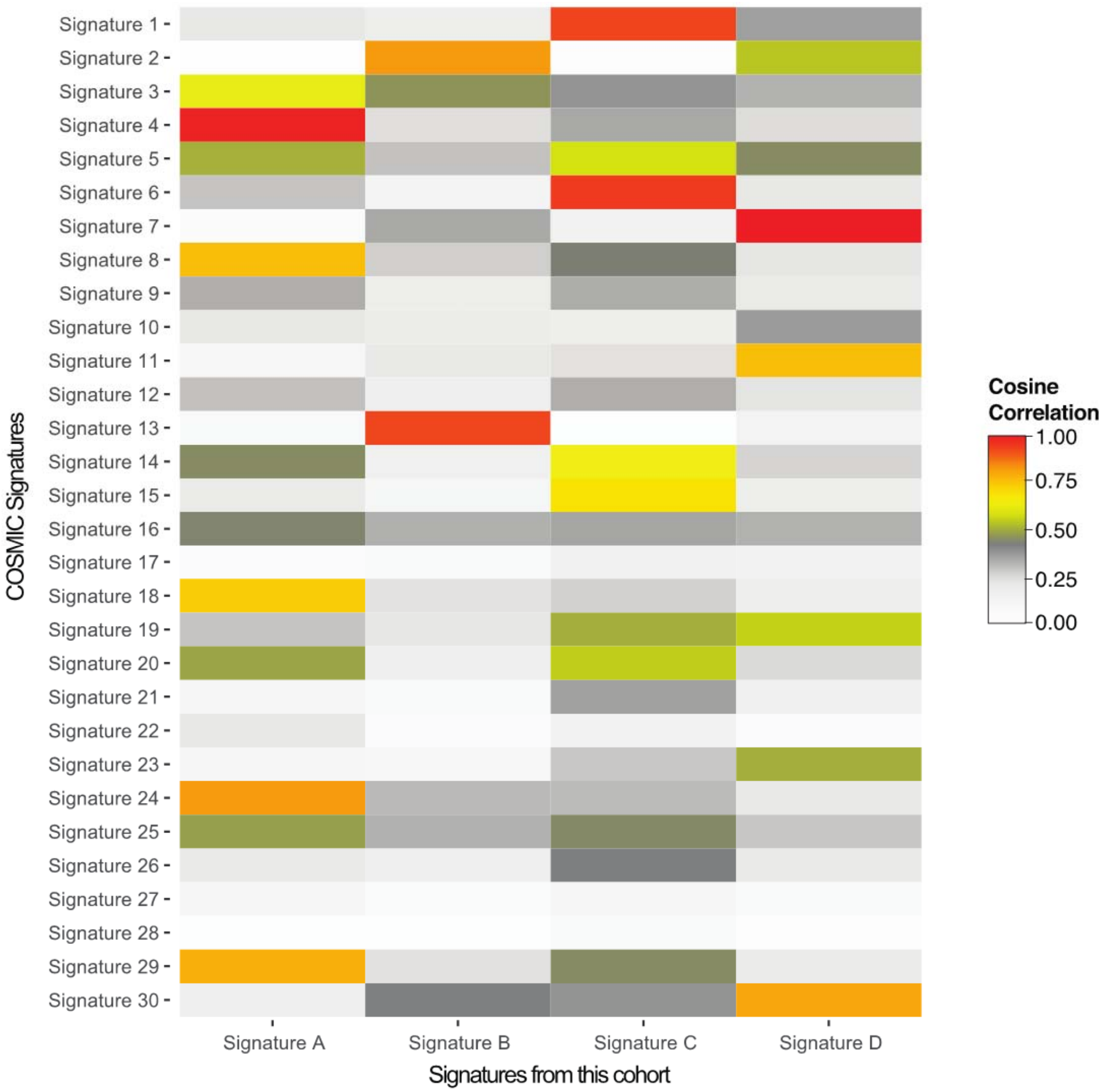
Power calculations for the Fisher's exact test were performed using 50,000 simulations with the "power.fisher.test" function in the "statmod" R package. In this study, the lung adenocarcinoma cohort had 87% power to detect a gene mutated in ≤1% percent in tumors from white patients and ≥10% in tumors from black patients at a significance level of 0.01. The lung squamous cell carcinoma cohort had 32% power using the same parameters.

eFigures


222,463 variants detected in 509 tumors
NFE **or** AA frequency > 0.0005
NFE **and** AA frequency < 0.0005
133,116 Variants
89,347 Variants
NFE **or** AA frequency > 0.01
Within coding exon
104,961 Variants
17,146 Variants
SNVs only
SNVs with AF <0.40 or >0.60 and valid trascript strand
Missense, splice, and silent SNVs in COSMIC
98,619 Variants (3,517 unique sites)
12,174 Variants
3,874 Variants
PCA Analysis
Mutational Signatures
Association with clinical phenotypes

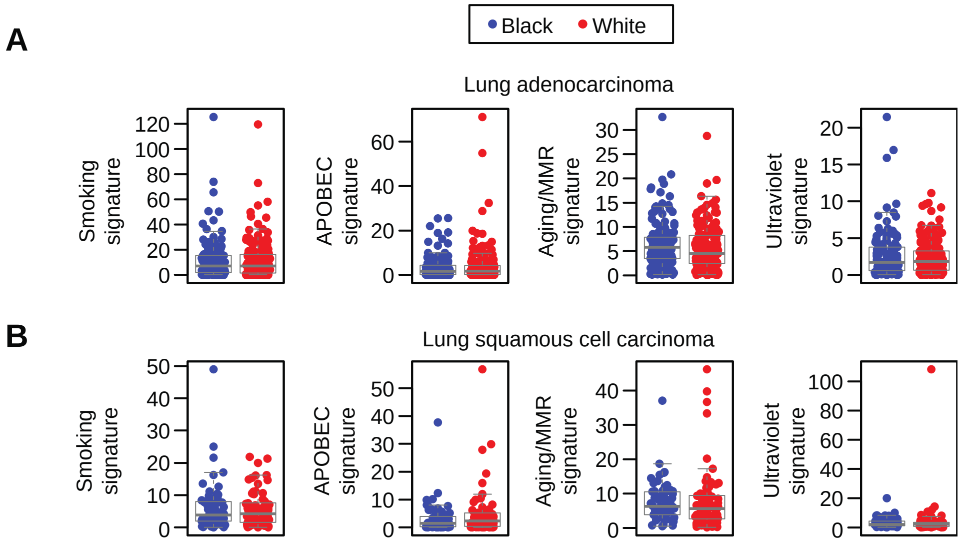
eFigure 1. Flow chart of variant filtering procedure.

As matched normal DNA was not available for these patients, we applied several criteria to limit the influence of inherited (i.e. germline) variants in the analysis. First, we excluded variants found in the Exome Aggregation Consortium (ExAC, version 3, http://exac.broadinstitute.org/) in a frequency greater than 0.0005 in the African American (AA) or Non-Finnish European (NFE) populations. Multi-allelic variants in ExAC were decomposed and normalized with vt (https://github.com/atks/vt) before matching with variants from this cohort. Second, we excluded variants outside of coding exons such as those in flanking regions, UTRs, and introns. For mutational signature inference, variants with alternate allele fractions between 0.4 and 0.6 or that did not have an annotated transcript strand were excluded. For analyses examining the association between clinical phenotypes and mutation status, missense, splice site, and nonsense SNVs were limited to variants previously observed in tumors as described in the Catalogue of Somatic Mutations in Cancer (COSMIC) database v74. Additional variants excluded after manual review included those in *MUC2* (chr11:1092891C>A and chr11:1093204C>A) and in *MSN* (chrX:64956743A>G and chrX:64956699_G>A).



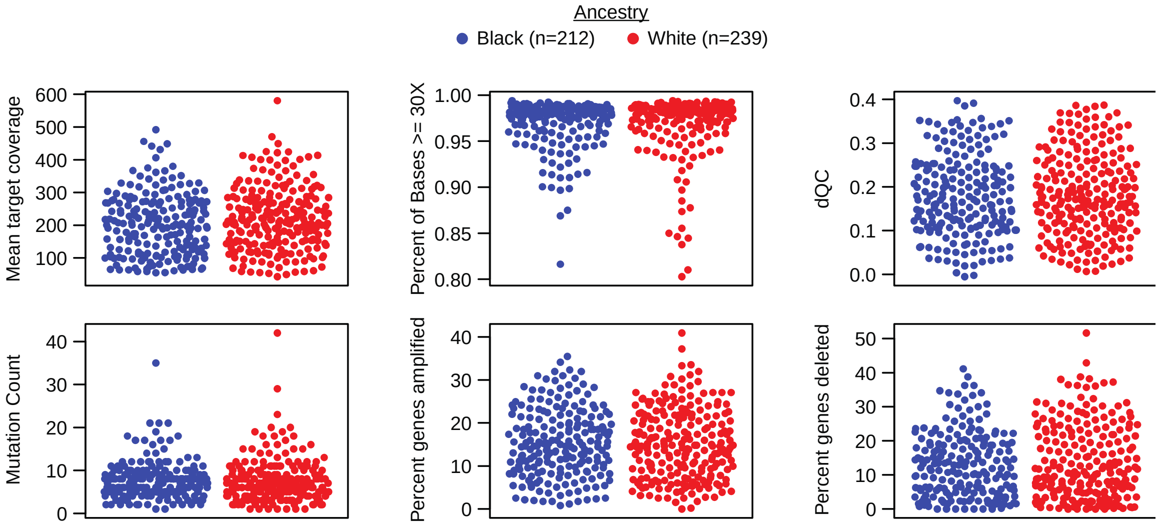
eFigure 2. Correlation between mutational signatures derived in this cohort and previously defined signatures from COSMIC.

The trinucleotide probabilities for 30 signatures generated from 10,952 exomes and 1,048 whole-genomes across 40 were obtained from COSMIC (http://cancer.sanger.ac.uk/cancergenome/assets/signatures_probabilities.txt). A pair-wise cosine correlation was performed between all COSMIC signatures and signatures from this study. The top correlated COSMIC signatures were used determine the identity of each signature in this cohort and each signature was renamed accordingly.



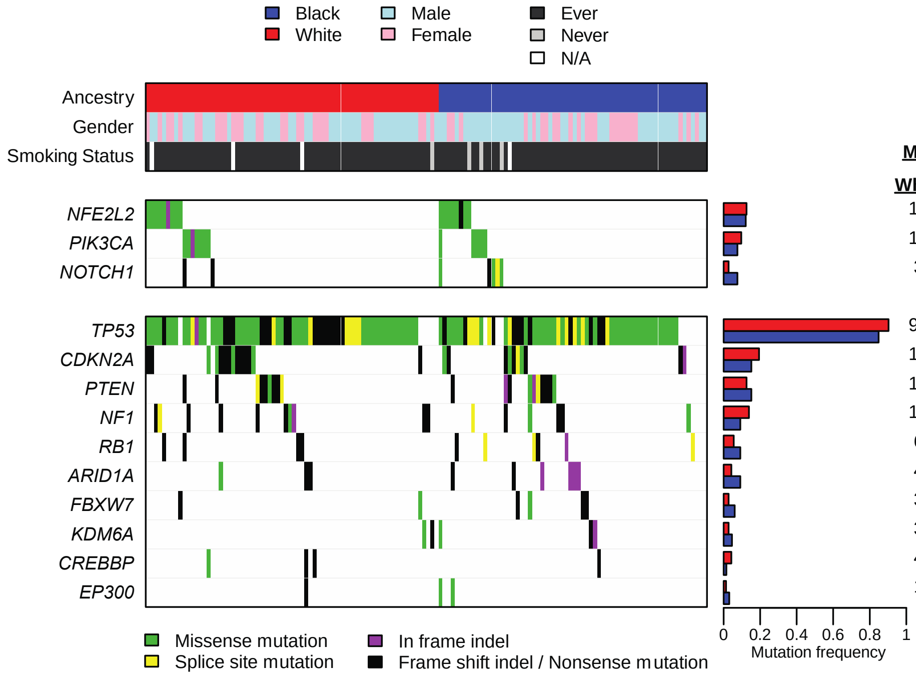
eFigure 3. Association of mutational signatures with ancestry.

(A) Mutational signatures were not significantly different between tumors from black patients (n=146) or white patients (n=167) in lung adenocarcinoma. (p > 0.05, Wilcoxon rank-sum test). **(B)** Mutational signatures were not significantly different between tumors from black patients (n=66) or white patients (n=72) in lung squamous cell carcinoma (p > 0.05, Wilcoxon rank-sum test).



eFigure 4. No association between ancestry and quality control statistics or overall numbers of mutations.

Exome sequencing quality control statistics including mean target coverage and the percentage of bases that had at least 30x sequencing depth were not associated with ancestry ($p > 0.05$). The ReCapSeq statistic “dQC” which measures goodness of fit for the segmentation was not associated with ancestry ($p > 0.05$). Overall numbers of putative somatic mutations or percentages of genes that were called amplified or deleted by ReCapSeg were not associated with ancestry ($p > 0.05$). This analysis was limited to lung adenocarcinomas and squamous cell carcinomas with confirmed ancestry (n=451).

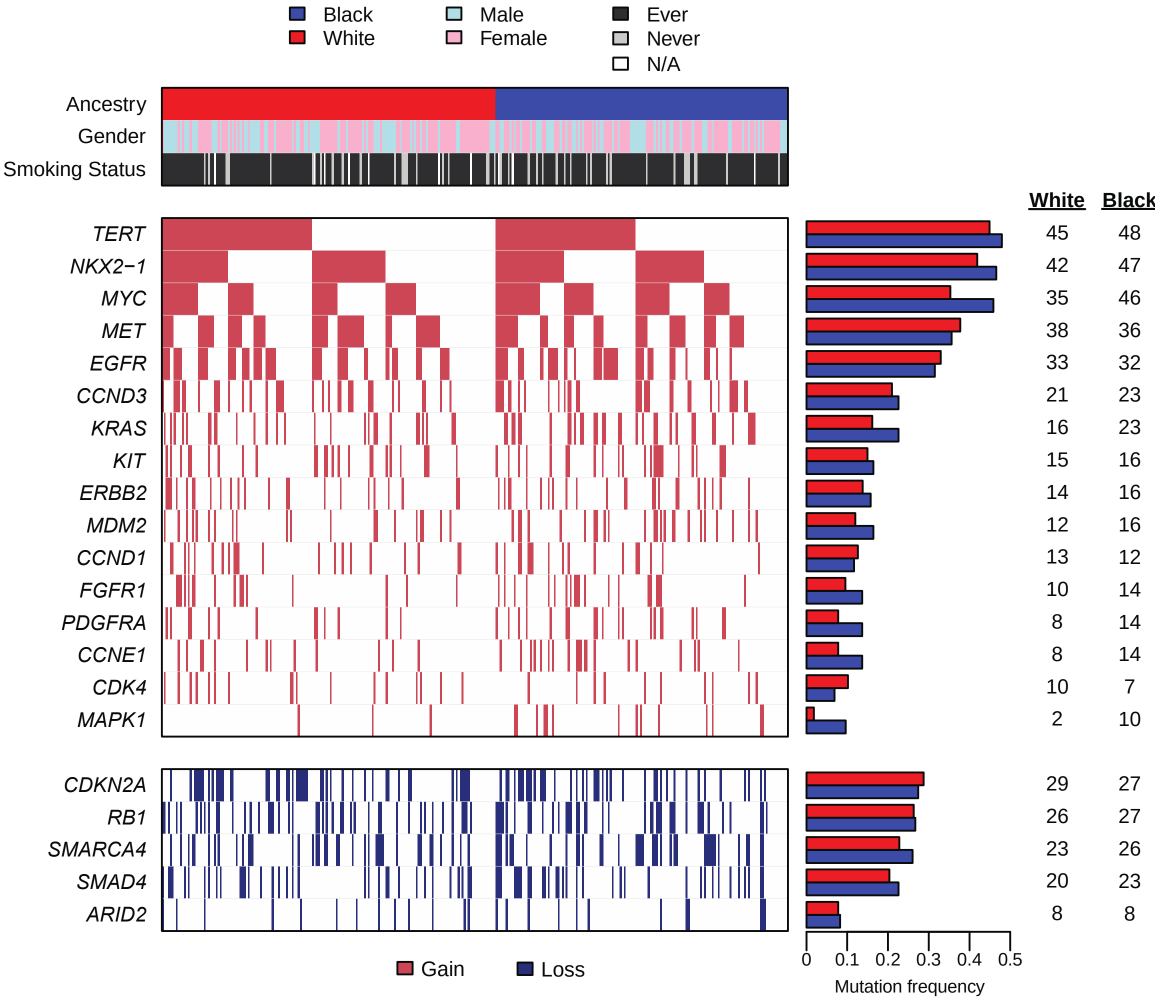


Mutational Frequency

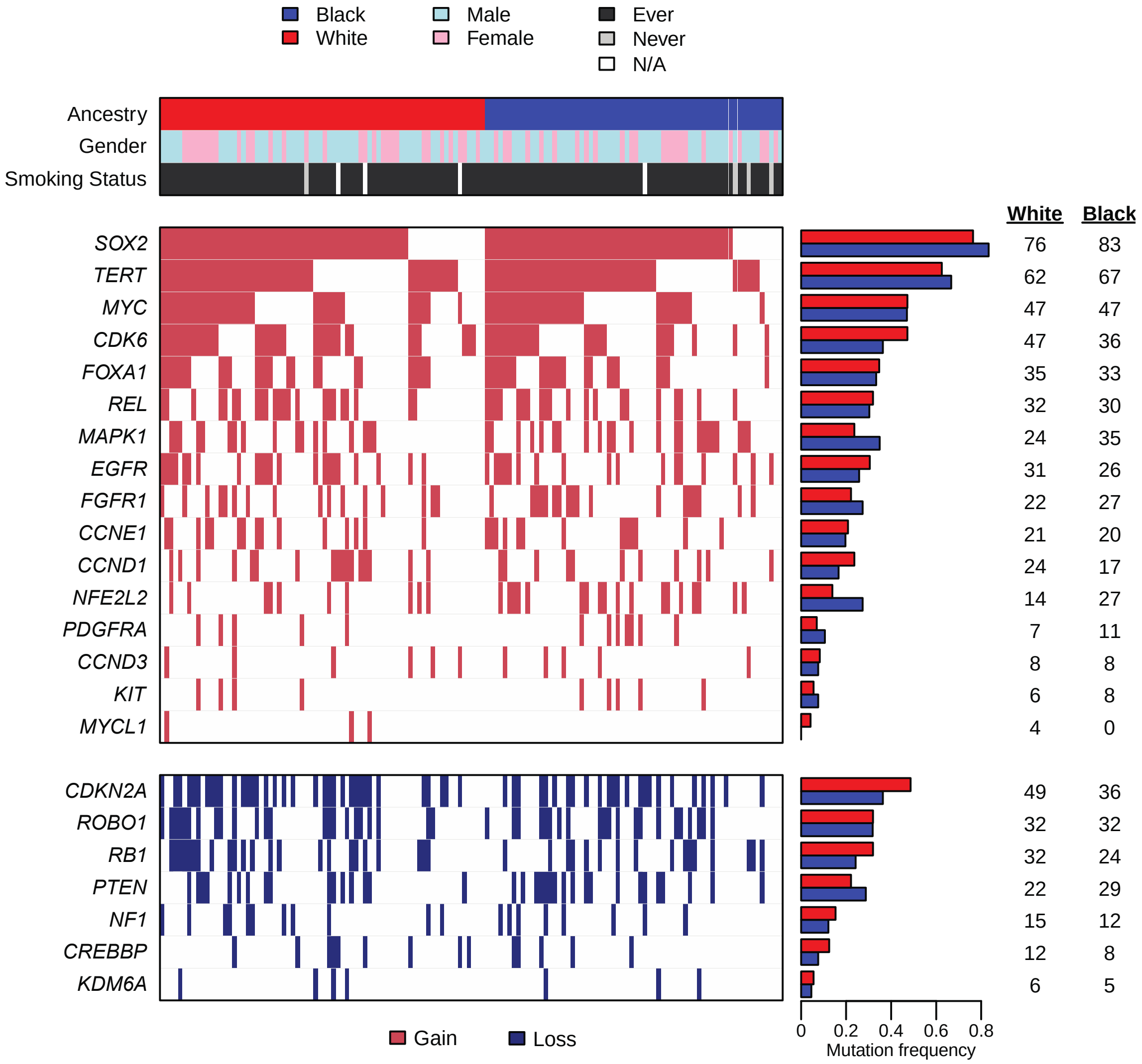
White	Black	TCGA LUSC
12	12	14
10	8	11
3	8	8
90	85	86
19	15	16
12	15	12
14	9	11
6	9	7
4	9	7
3	6	5
3	5	5
4	2	7
1	3	5

eFigure 5. Frequencies of mutations and fusions for known lung squamous cell carcinoma driver genes.

The frequencies of putative acquired alterations in each gene were compared between tumors from black and white ancestry using the Fisher's exact test. No genes reached statistical significance after correction for multiple hypothesis testing (FDR q-value > 0.05). Frequencies of alterations are shown for previously characterized oncogenes and tumor suppressors in lung squamous cell carcinoma. These frequencies are also compared to those found in a cohort of 484 lung squamous cell carcinomas (LUSC) from TCGA. No significant differences were observed between the mutational frequencies in tumors from African-Americans in this cohort and the mutational frequencies from TCGA using the Fisher's exact test (p > 0.05). Columns correspond to tumors and rows correspond to genes or clinical annotation. The individual boxes are colored according to the type of alteration for that gene in that tumor.

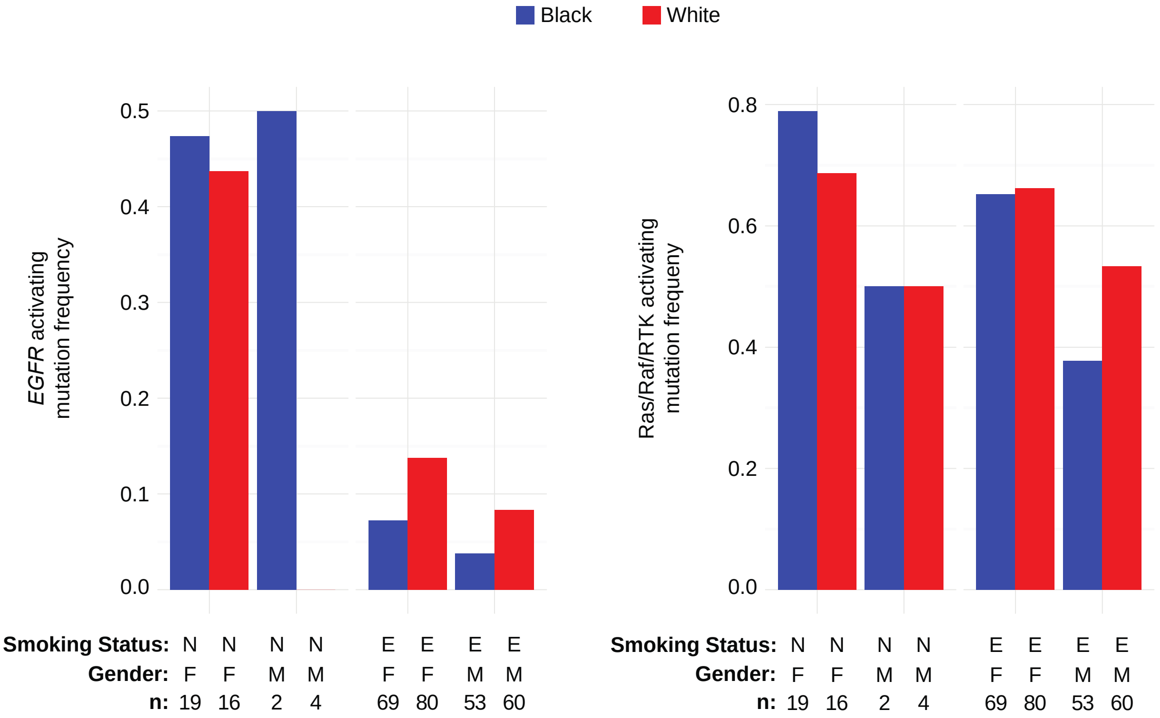


eFigure 6. Copy number alterations for known lung adenocarcinoma driver genes. Copy number gains and losses for each gene were compared between tumors from black and white ancestry using the Fisher's exact test. No genes reached statistical significance after correction for multiple hypothesis testing (FDR q-value > 0.05). The frequencies of the copy number gains and losses are shown for previously characterized oncogenes and tumor suppressors in lung adenocarcinoma. Columns correspond to tumors and rows correspond to genes or clinical annotation.



eFigure 7. Copy number alterations for known lung squamous cell carcinoma driver genes.

Copy number gains and losses for each gene were compared between tumors from black and white ancestry using the Fisher's exact test. No genes reached statistical significance after correction for multiple hypothesis testing (FDR q-value > 0.05). The frequencies of the copy number gains and losses are shown for previously characterized oncogenes and tumor suppressors in lung squamous cell carcinoma. Columns correspond to tumors and rows correspond to genes.



eFigure 8. Frequencies of EGFR or Ras/Raf/RTK alterations by ancestry, gender, and smoking status.

Using a logistic regression model and examining lung adenocarcinomas with sufficient clinical data (n=303), we found that *EGFR* mutation status was significantly increased in never smokers (N) compared to ever smokers (E) (p = 2.85 x 10^{-6}) and modestly increasing in females (F) compared to males (M) (p=0.08; left panel). No significant association was observed with ancestry (p=0.35). The frequency of Ras/Raf/RTK alterations was higher in females compared to males (p = 0.0007), but not associated with smoking status (p = 0.3296) or ancestry (p = 0.3591) in a logistic regression model (right panel).