## Supplementary Figures

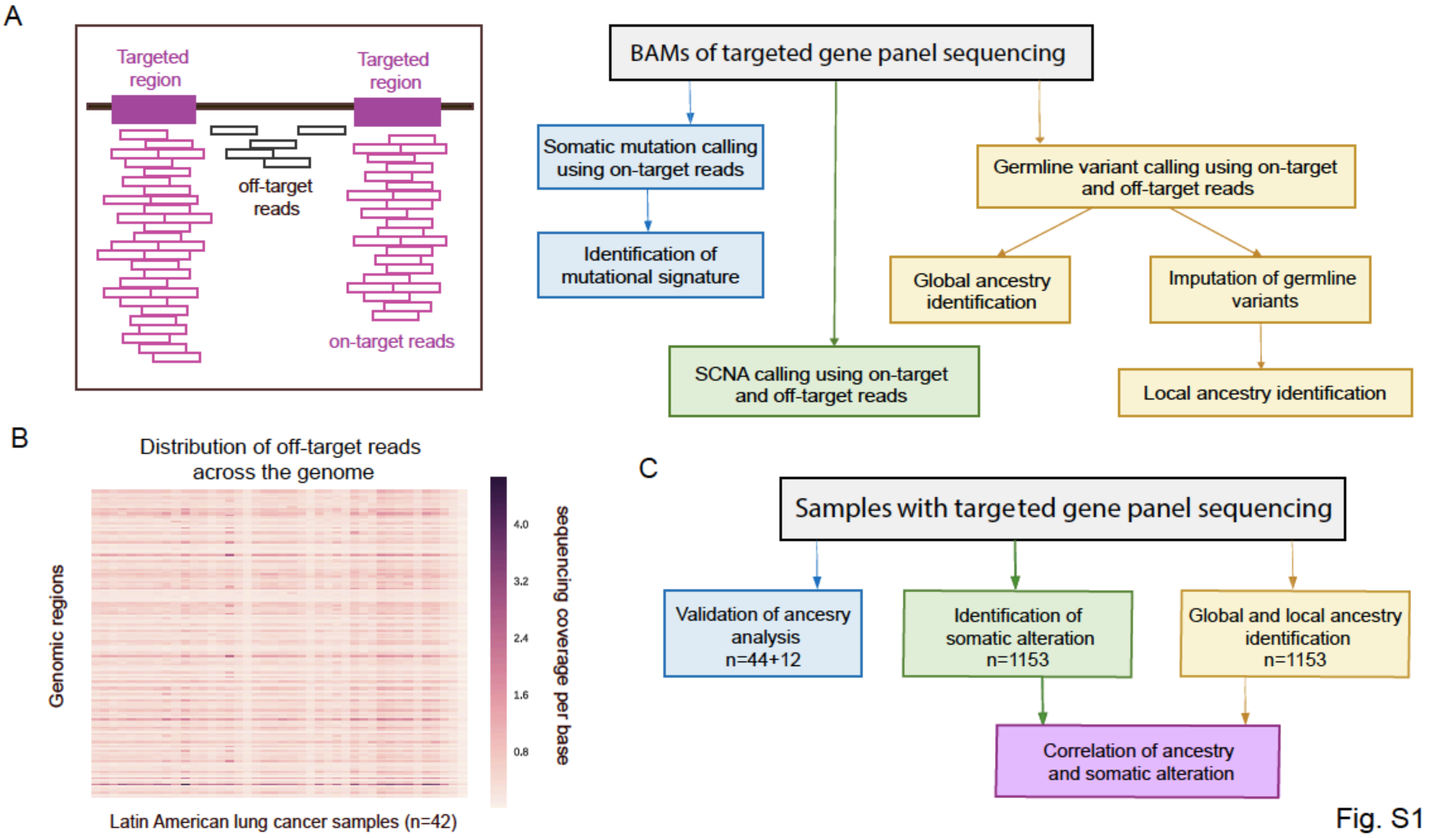

A
Targeted region
Targeted region
off-target reads
on-target reads
BAMs of targeted gene panel sequencing
Somatic mutation calling using on-target reads
Identification of mutational signature
SCNA calling using on-target and off-target reads
Germline variant calling using on-target and off-target reads
Global ancestry identification
Imputation of germline variants
Local ancestry identification
B
Distribution of off-target reads across the genome
Genomic regions
Latin American lung cancer samples (n=42)
sequencing coverage per base
4.0
3.2
2.4
1.6
0.8
C
Samples with targeted gene panel sequencing
Validation of ancesry analysis n=44+12
Identification of somatic alteration n=1153
Global and local ancestry identification n=1153
Correlation of ancestry and somatic alteration

Fig. S1

**Fig. S1: Overview of analytical pipeline.** A) On-target and off-target reads are obtained from BAM files. On-target reads are used to identify somatic SNVs/indels/fusions. Both on-target and off-target reads are used to identify SCNAs, and to infer overall genetic ancestry (global ancestry) and germline SNP genotypes. Missing SNPs are imputed and then ancestry is assigned to each genomic region (local ancestry identification). B) Distribution of off-target reads across the genome from panel sequencing of 42 samples. Average coverage per base within 10MB are shown for each sample. C) Somatic alteration, global ancestry and local ancestry are identified using panel sequencing data of 1153 tumor only DNA. A subset of samples is selected for validation of the ancestry analysis using low-pass whole-genome sequencing (n=44) and SNP array (n=12). Correlation analysis of NAT ancestry and global ancestry is performed on the full 1153 samples.

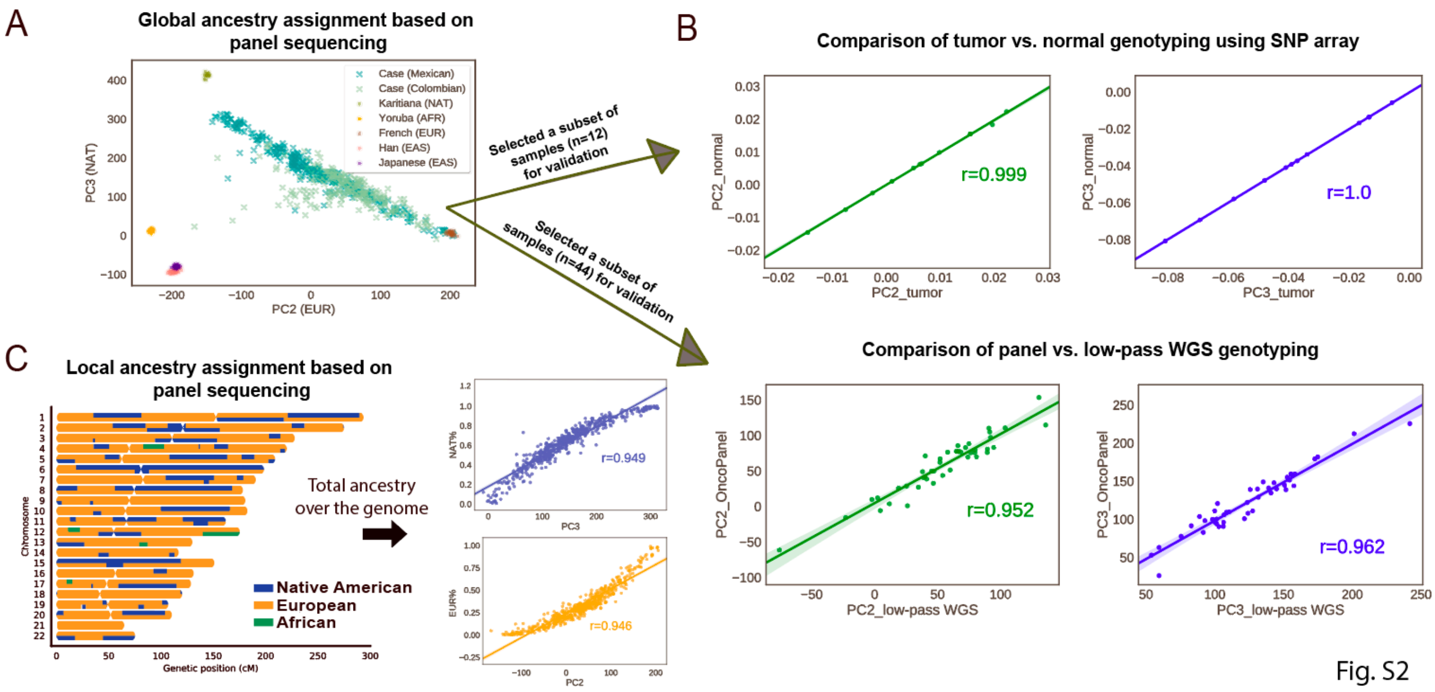


Fig. S2

**Fig. S2: Validation of analytical pipeline.** A) PC2 vs. PC3 from PCA of the Latin American cases in this study analyzed together with 939 reference samples from HGDP. The Latin American cases stretch out along PC2 and PC3, indicating the admixture of ancestries. B) High correlation of PC2 and PC3 between genotyping data of paired tumor vs. normal DNA from 12 patients (upper panel); and between the two sequencing approaches, panel vs. low-pass whole genome sequencing (lower panel) in 44 cases. The r values are obtained from Pearson's correlation. C) An example of local ancestry identification based on panel sequencing. Each genomic region is assigned with NAT, EUR or AFR ancestry. The proportion of NAT and EUR ancestry per sample obtained from local ancestry highly correlates with PC3 and PC2, respectively. The r values are obtained from Spearman's correlation.

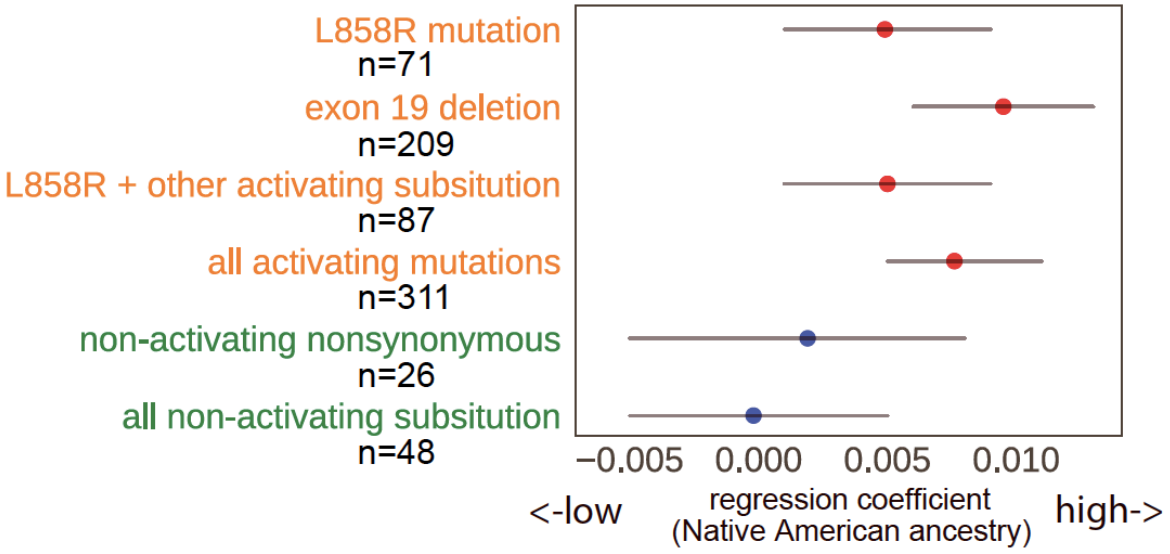


Fig. S3

**Fig. S3: *EGFR* mutations in association with NAT ancestry.** Logistic regression is used, coding *EGFR*-mutant cases as 1 and other oncogene-mutant cases as 0, and accounting for country of sample collection. Logistic regression coefficients are denoted by dots and 95% confidence intervals are denoted by lines. Red dots represent correlations with P values less than 0.05. Sample size (n) for each mutation group of *EGFR* is indicated.

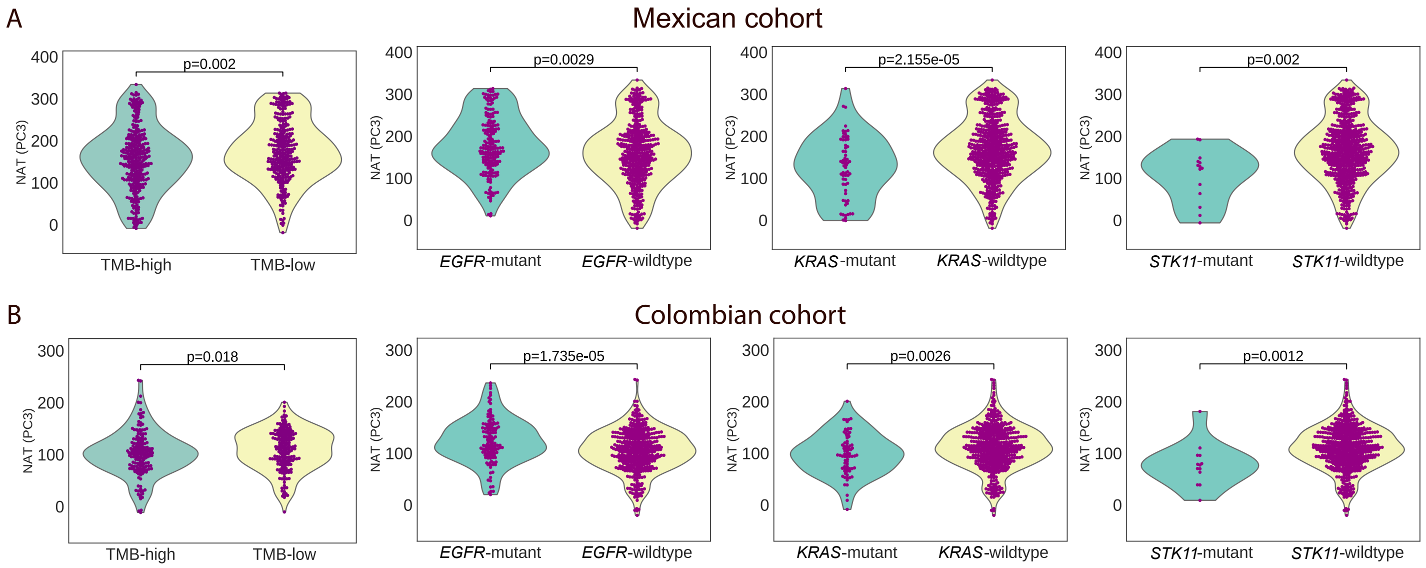


Fig. S4

**Fig. S4: Correlation of genetic ancestry and somatic mutations in Mexican and Colombian samples separately.** Comparison of PC3 indicating the NAT ancestry in cases with TMB-high and TMB low, and in cases with or without *EGFR*, *KRAS* or *STK11* mutations in A) the Mexican cohort and B) the Colombian cohort. TMB greater than the median TMB for each cohort is defined as TMB-high. PC3 is obtained from PCA of all Mexican and Colombian cases with 939 HGDP samples as reference. P values are obtained from Mann-Whitney U test.

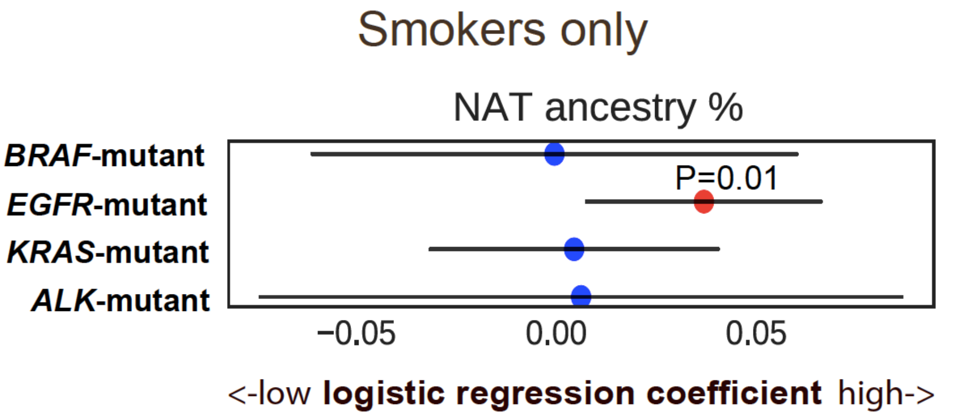


Fig. S5

**Fig. S5: Targetable LUAD driver genes associated with genetic ancestry in reported smokers.** Association of targetable LUAD driver genes with NAT ancestry, mutational signature and gender in patients reported themselves as smokers (n=77). Multivariable logistic regression P values are shown, with NAT ancestry percentage, gender, smoking and APOBEC signature as covariates. Red dots represent P value <0.05. Lines represent 95% confidence intervals.

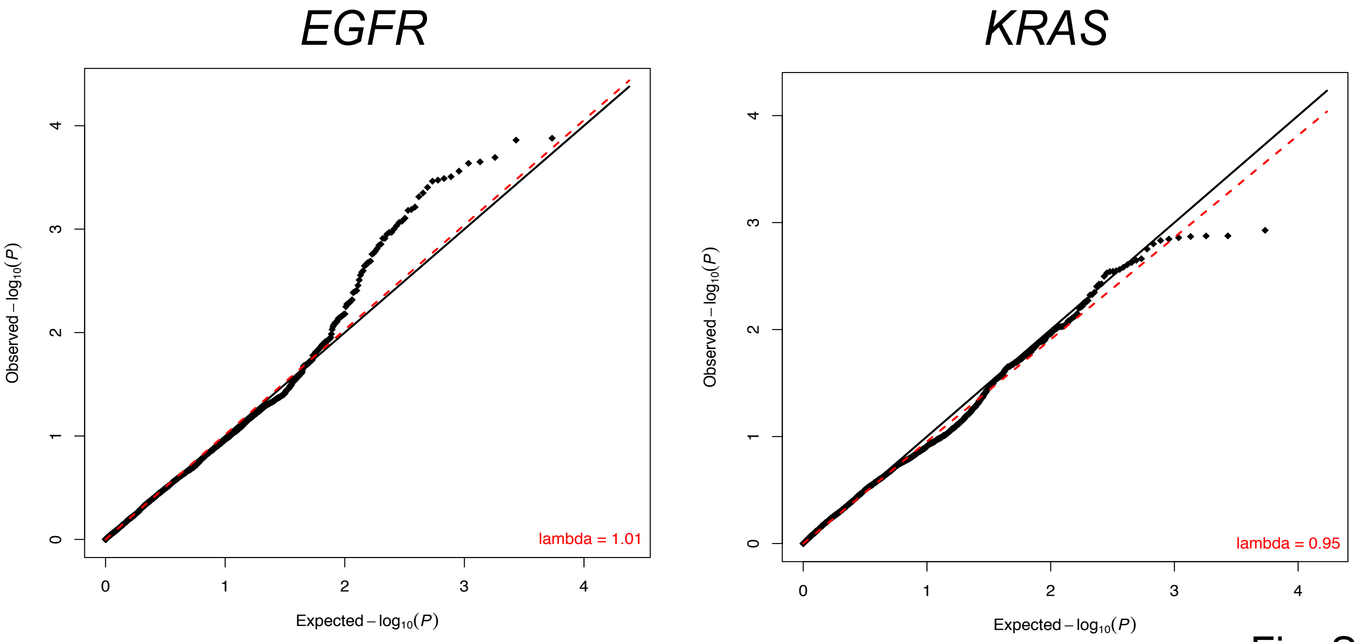
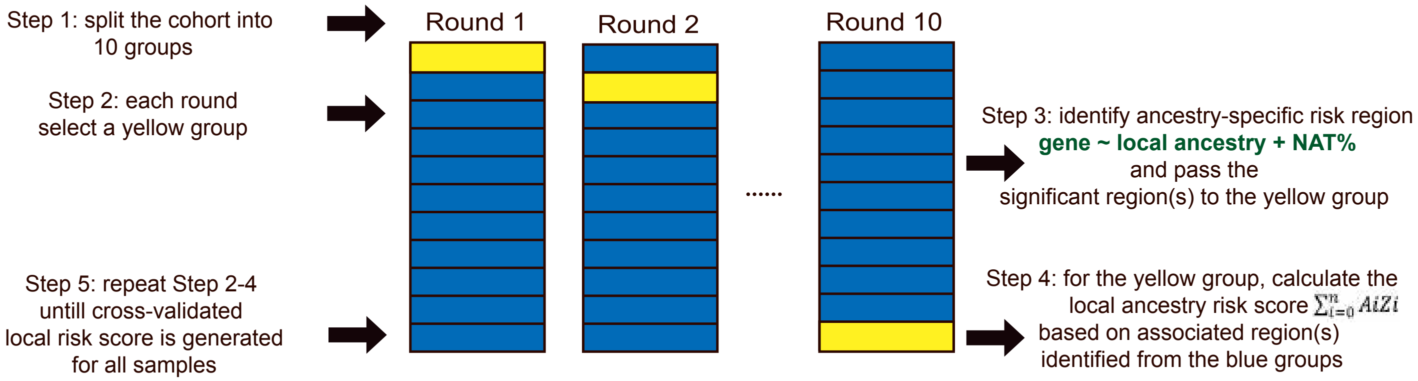


Fig. S6

**Fig. S6: Genome-wide association of local NAT ancestry with somatic mutations.** A) QQ plots demonstrating the distribution of P values of association with *EGFR*, and B) *KRAS*.



Step 1: split the cohort into 10 groups

Step 2: each round select a yellow group

Step 3: identify ancestry-specific risk region **gene ~ local ancestry + NAT%** and pass the significant region(s) to the yellow group

Step 4: for the yellow group, calculate the local ancestry risk score $\sum_{i=0}^{n} AiZi$ based on associated region(s) identified from the blue groups

Step 5: repeat Step 2-4 untill cross-validated local risk score is generated for all samples

Fig. S7

**Fig. S7: Schematic diagram for calculations of local ancestry risk scores** (weighted sum of the NAT ancestry in associated genomic loci).