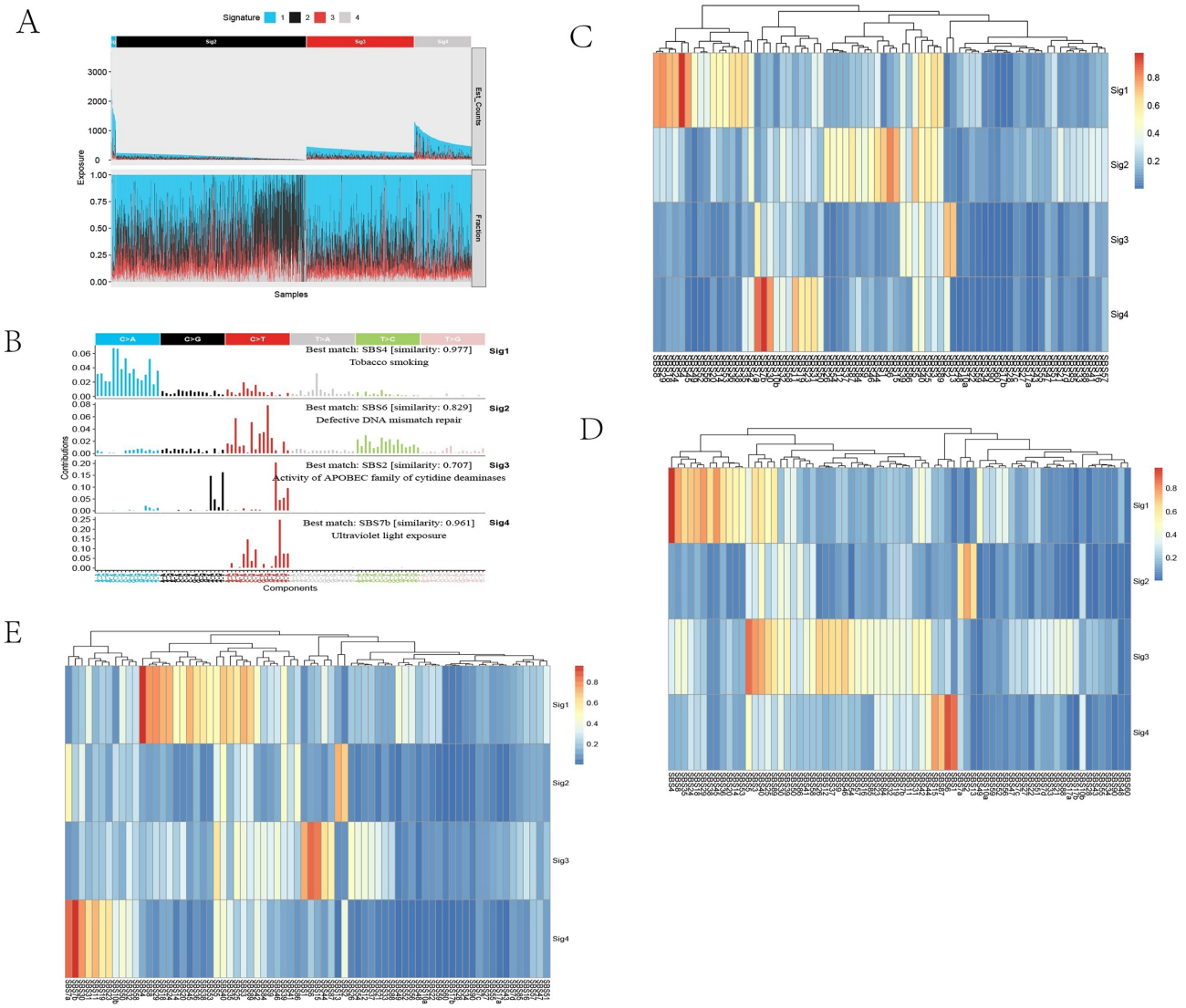

Figure S1 Identification of the mutational landscape in NSCLC. (A) Each NSCLC patient's relative contribution of each feature code (bottom panel) and the estimated number of copy number segments (top panel) are shown as histograms. The lung adenocarcinoma samples were divided into 2 groups based on the consensus matrix of multiple NMF runs, with each group specified by an enriched feature code. (B) Map of *de novo* extracted mutation features identified from NSCLC mutation data. Each feature is shown as the percentage (y-axis) of mutations attributed to the 96 SBS categories (x-axis) defined by color-coded substitution categories and sequence contexts. (C-E) A heat map of the cosine similarity of mutation features extracted from the NSCLC, LUAD, and LUSC mutation datasets with published mutation features from COSMIC-2. NSCLC, nonsmall cell lung carcinoma; NMF, the nonnegative matrix decomposition; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

https://dx.doi.org/10.21037/tlcr-23-150

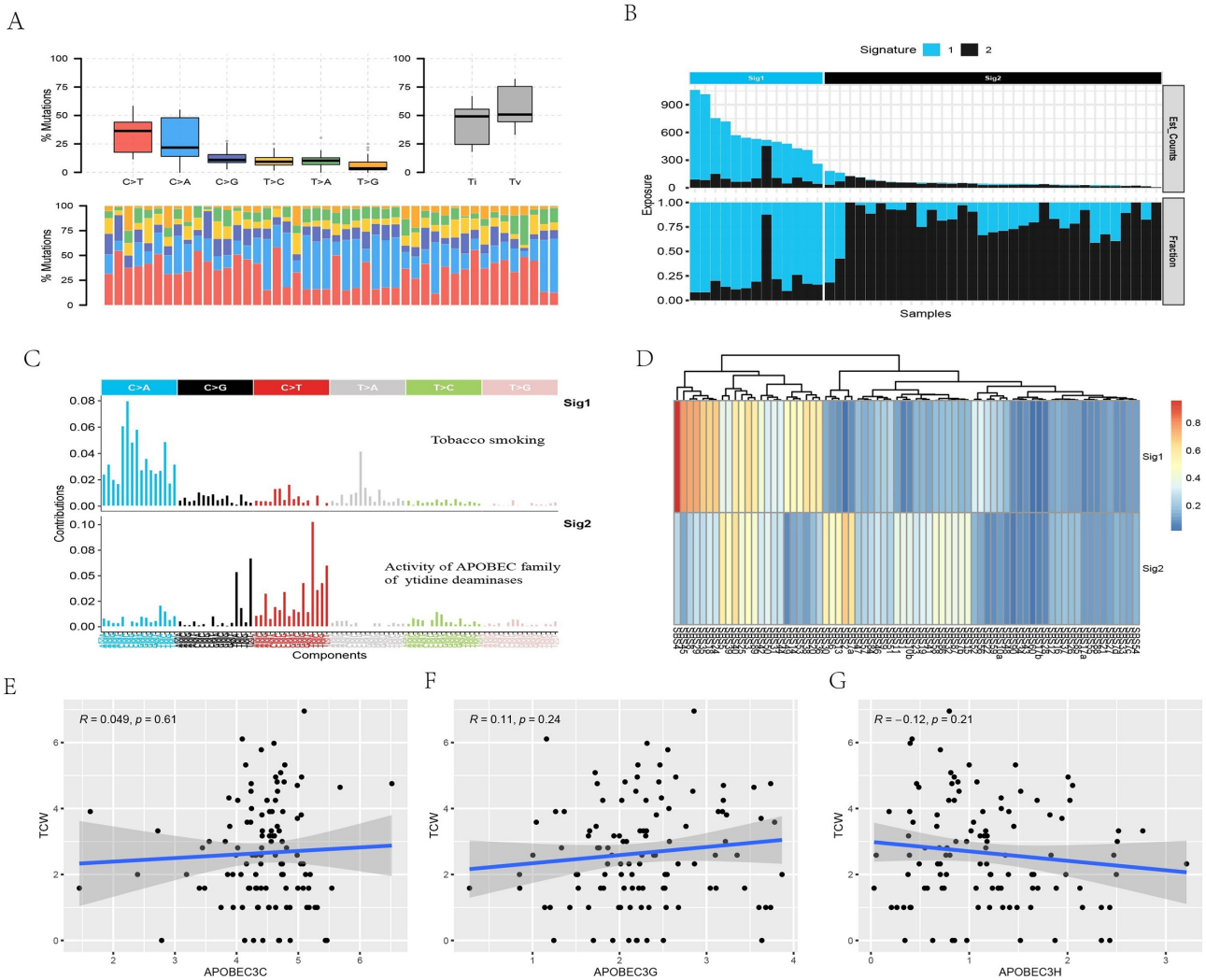

Figure S2 Validating the mutation landscape of nonsmoking patients in the validation cohort. (A) Mutations in somatic cells of luad_cptac_2020 nonsmoking patients. Top left: statistics of base mutations, top right: proportion of base reversals and conversions, Ti refers to conversions (purines replaced by purines and pyrimidines replaced by pyrimidines), and Tv refers to reversals (substitution between purines and pyrimidines). Lower panel: the proportion of base mutations in all samples, with colors corresponding to the box plot. The box plot (top left) presents the total mutations in all samples, and the percentage bar (bottom) presents the mutations in each sample. (B) For each luad_cptac_2020 patient, the relative contribution of each feature code (bottom panel) and the estimated number of copy number segments (top panel) are shown as histograms. The lung adenocarcinoma samples were divided into 2 groups based on the consensus matrix of multiple NMF runs, with each group specified by an enriched feature code. (C) Map of the *de novo* extracted mutation features identified from the luad-cptac_2020 mutation data. Each feature is shown as the percentage (y-axis) of mutations attributed to the 96 SBS categories (x-axis) defined by color-coded substitution categories and sequence context. (D) Heat map of the cosine similarity of mutation features extracted from the luad_cptac_2020 mutation dataset to those published in COSMIC-2. (E-G) APOBEC3C, APOBEC3H, and APOBEC3G were not significantly correlated with TCW counts. Ti, Transition; Tv, Transversion; NMF, the nonnegative matrix decomposition; SBS, single base substitutions.

https://dx.doi.org/10.21037/tlcr-23-150

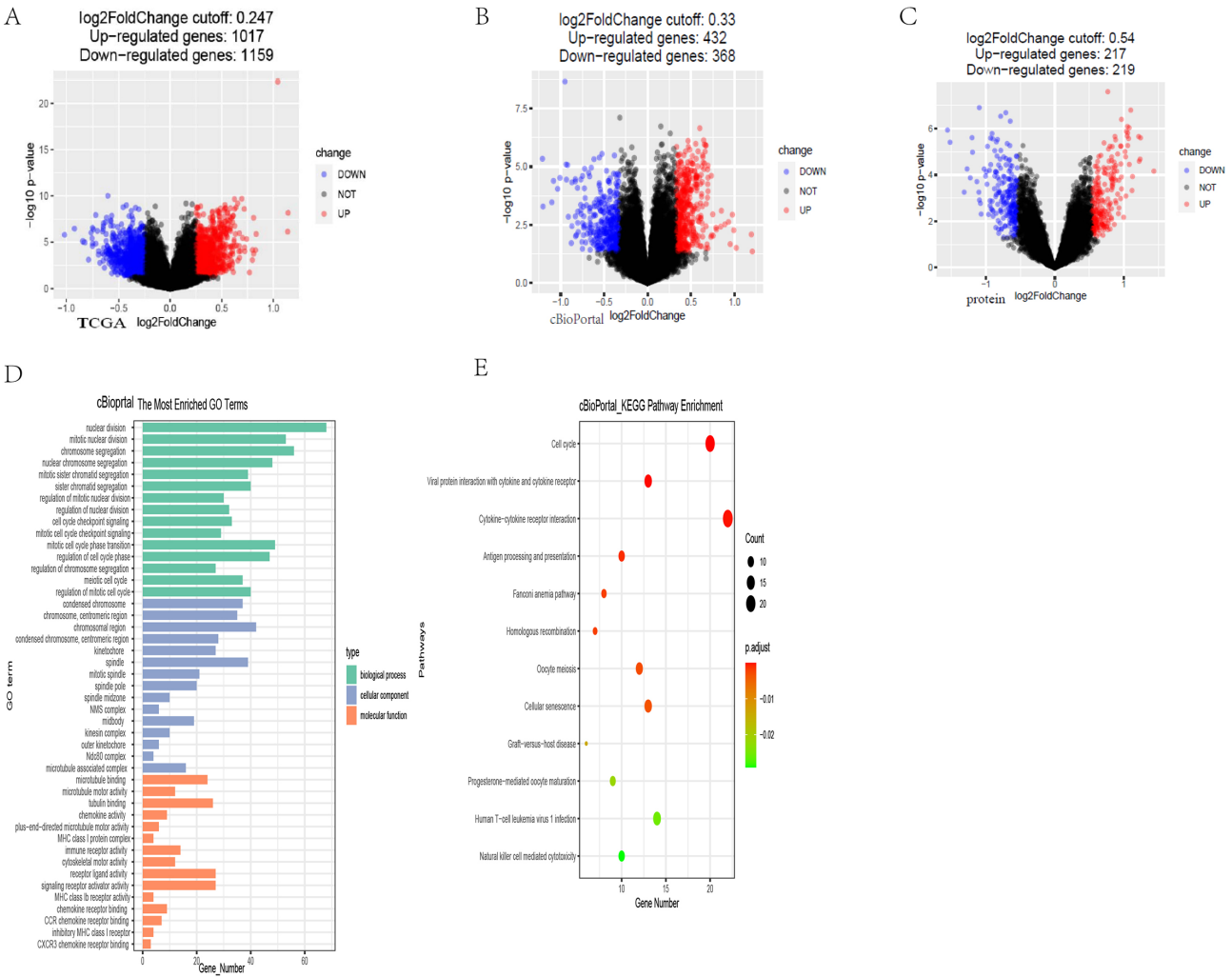

Figure S3 Validating the immune landscape of the APOBEC3 score in the validation cohort. (A-C) Volcano plots of mRNA differences and volcano plots of proteomic differences between the 2 fractions of TCGA, luad_CPTAC_2020 nonsmoking patients, respectively. (D) Gene ontology enrichment analysis of differential genes in high and low subgroups in luad_CPTAC_202 samples. (E) KEGG enrichment analysis of differential genes between high and low scoring groups in the luad_CPTAC_2020 sample. TCGA, the cancer genome atlas; KEGG, Kyoto Encyclopedia of genes and genomes.

https://dx.doi.org/10.21037/tlcr-23-150