Original Article

Fusion gene recurrence in non-small cell lung cancers and its association with cigarette smoke exposure

Neetha Nanoth Vellichirammal[1], Abrar Albahrani[1], Chittibabu Guda[1,2]

[1]Department of Genetics, Cell Biology, and Anatomy, University of Nebraska Medical Center, Omaha, NE, USA; [2]Center for Biomedical Informatics Research and Innovation (CBIRI), University of Nebraska Medical Center, Omaha, NE, USA

Contributions: (I) Conception and design: NN Vellichirammal, C Guda; (II) Administrative support: C Guda; (III) Provision of study materials or patients: C Guda; (IV) Collection and assembly of data: NN Vellichirammal, A Albahrani; (V) Data analysis and interpretation: NN Vellichirammal, A Albahrani; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Chittibabu Guda, PhD. Professor, Department of Genetics, Cell Biology, and Anatomy, University of Nebraska Medical Center, Omaha, NE 68198, USA. Email: babu.guda@unmc.edu; Neetha Nanoth Vellichirammal, PhD. Assistant Professor Department of Genetics, Cell Biology, and Anatomy, University of Nebraska Medical Center, Omaha, NE 68198, USA. Email: n.nanothvellichiram@unmc.edu.

Background: Lung cancer remains the leading cause of cancer-related deaths in the US despite novel treatment protocols, with about 235,000 new cases and 131,000 deaths expected from this cancer in 2021 alone. Lung adenocarcinoma and squamous cell carcinoma, which are both subtypes of non-small cell lung cancer, account for most lung cancer cases, and comparing the molecular signatures in these two cancers can identify novel mechanisms that contribute to non-small cell lung cancer oncogenesis.

Methods: We, in this study, performed a comprehensive gene fusion profiling of these cancers, which is understudied in lung cancers. Using an alignment-free fusion detection tool, 'ChimeRScope', we screened for gene fusions in lung adenocarcinoma and squamous cell carcinoma datasets from The Cancer Gene Atlas database. Fusion profiles in these two cancer subtypes were essentially different with minimal overlap.

Results: Our analysis revealed a positive association of smoking to fusion frequency in lung adenocarcinoma but not in squamous cell carcinoma and identified several fusion genes that could be explored as markers associated with cigarette smoke exposure. We also identified differentially regulated pathways linked to E2F, G2M checkpoint, and MTORC1 signaling upregulated and P53 pathway downregulated in samples containing high fusions in lung adenocarcinoma. Our results indicate that downregulation of the P53 pathway leads to higher gene fusion formation in lung adenocarcinoma.

Conclusions: This manuscript provides a strong rationale for investigating the molecular mechanisms of cigarette smoke-induced gene fusion formation associated with lung cancer. Novel recurrent fusions associated with cigarette smoke were identified in our study, which could further be investigated for patient stratification, personalized therapy, and therapeutic monitoring.

Keywords: Fusion genes; smoke-induced gene fusions; The Cancer Genome Atlas (TCGA); lung adenocarcinoma; ChimeRScope

Submitted Feb 12, 2022. Accepted for publication Aug 22, 2022.
doi: 10.21037/tlcr-22-113
View this article at: https://dx.doi.org/10.21037/tlcr-22-113

Introduction

Technological advancements that enable early detection and implementation of effective treatment protocols have led to the steady decline of cancer-related mortalities across most cancers. Despite this promising progress, lung cancer accounted for the highest percentage of cancer-associated deaths in 2021 (22%) and is still a leading cause of death in men and women (1). Lung cancer is a highly heterogeneous

cancer with multiple histologic and molecular phenotypes, and this plays a significant role in the treatment decisions and development of chemotherapeutic resistance (2). The vast majority of the lung cancer cases belong to non-small cell lung cancer (NSCLC), subtyped into lung adenocarcinoma (LUAD, 40% of lung cancers) and lung squamous cell carcinoma (LUSC, 25–30% of lung cancers), and large-cell carcinoma (5–10% of lung cancers) (3-5). These subtypes of lung cancer originate from different cell types and exhibit differences in molecular characteristics, thus requiring separate treatment protocols (2).

LUADs are more likely to be found at peripheral locations, whereas squamous cell carcinoma is often located centrally and visualized as endobronchial masses (6,7). In addition to these histological differences, several studies strongly link smoking and second-hand smoking exposure to LUSC more than LUAD. A meta-analysis of 17 lung cancer datasets identified that more than half of the smokers developed squamous cell carcinoma, and the majority of non-smokers developed adenocarcinoma, suggesting that smoking is a decisive risk factor for LUSC (8). Several studies have also identified genomic differences across both cancers, including differences in mutational and gene expression profiles. Mutations in *KRAS*, *BRAF*, *ERBB2*, *MET*, and *EGFR* genes are recurrent in LUAD along with recurrent fusion genes (9,10). These fusion genes are reported to partner with other genes like anaplastic lymphoma kinase (*ALK*), the ROS1 receptor tyrosine kinase, and *RET*. On the other hand, the mutational profile was different in LUSC, with mutations in genes such as *DDR2*, *FGFR1*, *FGFR2*, *FGFR3*, and genes involved in the PI3K pathway (11-19). In addition to the mutational differences, gene expression differences in immune-response genes between LUSC and LUAD were also reported. Increased expression of cell proliferation-associated genes and the repression of immune-response genes are suggested to account for the faster disease progression LUSC (20).

Along with molecular differences between LUAD and LUSC, clinical trials have documented differences in therapeutic responses, favoring one subtype over another (21,22). Several targeted therapies for LUAD are available that target genomic variations, including mutations and fusions (23-26). Mutations in epidermal growth factor receptor and ALK fusions are recurrent in LUAD that can be exploited for targeted therapies (27-30).

Even though extensive analyses have identified several genomic changes associated with specific lung cancer subtypes, similarities or differences on how smoking affects these genomic alterations are not well documented in NSCLC. Tobacco smoke contains more than 60 DNA adducts that bind and modify DNA (31), of which the most potent carcinogens are the polycyclic aromatic hydrocarbons (PAHs); benzo[a]pyrene (BAP), and nicotine derived nitrosaminoketone (NNK). Smoking affects an extensive repertoire of genes in airways, alveolar macrophages, and peripheral leukocytes (32-34). The binding of carcinogens in cigarette smoke to the DNA ultimately results in mutations that are characteristic of smoking signatures (31,35). However, the effect of smoking on other gross genomic abnormalities like gene fusions is understudied in lung cancers. Fusion genes, formed from the concatenation of two unrelated genes, could form novel protein products, are often associated with cancer, and serve as effective diagnostic or prognostic markers (36,37). Furthermore, differences in the fusion profile between the lung cancer subtypes and the perturbed molecular pathways facilitating gene fusion formation are unknown. It has been reported that gene fusions contribute to altered gene functions leading to oncogenesis and could be explored as specific and effective diagnostic or prognostic markers (36,37).

We, in this study, have explored the fusion profiles of LUAD and LUSC cancers using the genomic and transcriptomic data from TCGA and correlated these profiles with corresponding clinical characteristics of patients. Several fusions associated with cigarette smoke exposure were identified in our analysis. We also analyzed the effect of smoking on fusion formation in these cancers. Further, the differences in fusion profiles across the two cancer subtypes and the genomic alterations leading to fusion formation were analyzed. We present the following article in accordance with the MDAR reporting checklist (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-22-113/rc).

Methods

Fusion detection from TCGA dataset

Level 1 paired-end RNA sequencing data (aligned BAM files) from The Cancer Genome Atlas (TCGA) were downloaded from the Genomic Data Commons (GDC) data portal for analyses. Fusions were detected from RNA-seq data using our in-house developed ChimeRScope tool described before (38). Briefly, fusion transcripts (also referred to as fusions) were identified from the unmapped RNA-sequencing reads using the Scanner and Sweeper

modules of ChimeRScope and further filtered to remove control fusions (fusions identified from control samples) and sequencing artifacts. The identified fusions were also categorized into canonical and non-canonical depending on the absence or presence of antisense fusion partners, respectively, as described before (38). Fusions were also classified as recurrent if the identified fusion gene pair occurred more than once in the patient cohort. Both NSCLC lung cancers in the TCGA database, LUAD (n=506) and LUSC (n=501), were analyzed in this study. Patient clinical characteristics, including tumor stage, smoking status, and survival data, were also collected from the GDC data portal. After fusion gene identification, LUAD and LUSC samples were further divided into high or low fusion containing groups based on the interquartile range. Samples were also analyzed based on the reported smoking status in each cancer. Patients who were not smoking at the time of interview and had smoked less than 100 cigarettes in their life were categorized as Non-smokers (designated as NS), and those who are currently smoking (daily or non-daily) was categorized as current-smokers (designated as CS). Patients who were not smoking at the time of the interview but had quit smoking were categorized into two groups; current reformed smoker for ≥15 years designated as RF >15 years, and current reformed smoker for ≤15 years designated as RF <15 years. As per the TCGA LUAD clinical data, number of average pack years for the current smokers, current reformed smoker for >15 years and current reformed smoker for ≤15 years were 51.7, 30.1 and 44, respectively. For TCGA LUSC, number of average pack years for the current smokers, current reformed smoker for ≥15 years and current reformed smoker for ≤15 years were 55.8, 38.6 and 55, respectively.

DNA damage repair (DDR) pathway analysis

A list of 276 genes across all significant DNA repair pathways was analyzed for gene fusions in LUAD and LUSC (39) (Table S1). Core DDR pathway genes identified in this gene set consisting of 71 DNA repair pathway-specific and 9 damage response genes (Table S2) were also analyzed in both datasets to understand the mutational and fusion frequencies and gene expression differences across smoking and non-smoking groups. DDR pathways were also compared in samples with high or low fusion profiles in each cancer. Fusions, mutations, and gene expression profiles were compared for each dataset.

Mutation profile comparisons

Mutation profiles for smoking groups and samples with high or low fusions were summarized, annotated, and analyzed for LUAD and LUSC datasets. Somatic mutations for both LUAD and LUSC cohorts were extracted from the MAF file available for download from the TCGA GDC (Genomic Data Commons) website. Somatic variant calling by the GDC pipeline includes four somatic variant callers: *MuTect2*, *VarScan2*, *MuSE*, and *SomaticSniper* (40). These variants are further filtered using several filtering tools to reduce germline variants and other artifacts. The Bioconductor package, *maftools*, was used to analyze and estimate the mutation load among different smoking groups and samples with high or low fusion frequency in both cancers (41). *Maftools* were also used to extract mutational signatures from the datasets analyzed in this study. *Maftools* extract 5' and 3' adjacent bases next to the mutated base and construct a count matrix using the '*trinucleotideMatrix*' function (using three nucleotides). Mutational signatures are extracted using non-negative matrix factorization (NMF) to factorize count matrix M using the '*extractSignatures*' function. Mutational signatures identified are then compared to 30 types of COSMIC signatures (42,43).

Gene expression profile comparison

Read count data from TCGA RNA-seq experiments were downloaded using the R-program, *TCGAbiolinks* (44), and processed further for gene expression comparison between different groups using *limma* (45). TCGA processes the RNA-seq pipeline using *STAR* aligner using a two-pass approach (46). Read counts were normalized and processed using *limma*. Gene expression comparisons between different smoking groups and across samples with high or low fusion frequency were performed in both cancers. A gene was considered differentially expressed if the FDR corrected P value was <0.05 and the absolute foldchange was ≥1.5.

Gene set and pathway enrichment analysis

Gene set and pathway enrichment analyses were performed using *FGSEA* and *IPA*. (I) *FGSEA*: The fast implementation of the Gene Set Enrichment Analysis *GSEA* algorithm (47) using the r-package, *fgsea* (48) was used to identify differentially enriched gene sets for each

comparison. *FGSEA* analysis was conducted pre-ranked mode, using the log2-fold change values from *limma* to rank the gene list. Molecular Signature Database (MSigDB) datasets (Hallmark, immunologic signature dataset, chemical, and genetic perturbations dataset, and Gene Ontology dataset) were used for *FGSEA* analysis. Gene sets with an FDR adjusted P-value <0.1 were considered significantly enriched. (II) *IPA* Analysis: Gene annotation enrichment analysis of differentially regulated genes or genes participating in fusions to identify known functions, pathways, and networks affected was performed using *Ingenuity Pathway Analysis* (IPA) (Ingenuity Systems; CA, USA). The significance was set at a P-value of 0.05.

Cox proportional hazards model for survival prediction

We used Cox proportional hazards model to investigate how different patient clinicopathological characteristics affect overall survival in LUAD. Univariate Cox regression analysis using gender, age at initial pathologic diagnosis, tumor mutation burden, person neoplasm cancer status, tobacco smoking history, number of pack-years smoked, primary therapy outcome, tumor stage, and fusion status were performed to determine the significance of these features on survival. A multivariate Cox proportional hazards regression model was used to investigate the interactions between multiple covariates. Only covariates significant in univariate analysis (Likelihood ratio test P<0.05) were used in multivariate analysis. The R package, *survival*, was used for performing this analysis (49).

Ethical statement

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Statistical analysis

All analyses were subjected to FDR correction, wherever possible. For gene expression analysis, FDR corrected P value of <0.05 and the absolute foldchange of ≥1.5 was considered as the cut-off threshold. For gene set enrichment analysis, FDR adjusted P value of <0.1 was considered significantly enriched.

Results

We performed a comparative analysis of fusion genes and their mutational and gene expression profiles in three comparison groups: LUAD versus LUSC patients; non-smokers versus smokers (current, reformed >15 years, and reformed <15 years); and low-fusion versus high-fusion groups. In addition, we performed in-depth characterization of genes involved in DNA repair and oncogenic pathways in these cohorts in the context of fusion occurrence and recurrence, gene expression, and mutational burden.

Fusion genes, mutational profiles, and associated clinical characteristics in LUAD and LUSC

Fusions in LUAD

We identified gene fusions using ChimeRscope from LUAD and LUSC datasets and analyzed their fusion profiles. Recurrent gene fusions in LUAD were sparse, but several genes with multiple fusion partners were identified in both canonical and non-canonical fusions (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). Patients reported to be tumor-free (patient was reported to be disease-free at the last contact) had lower fusions than those with tumors (Student's *t*-test P=0.03), suggesting that fusion frequency is correlated to genomic instability associated with cancer. A slight increase in fusion frequency was also found in patients who died of the disease (Student's *t*-test P=0.05), which could also point towards tumor progression and associated genomic instability. A higher frequency of fusions was observed among patients that were CS or had a previous smoking history when compared to NS in LUAD (Figure S1A). CS had the highest fusion frequency among all LUAD patients. RF >15 years had fusion frequency comparable to NS, indicating that fusion frequency decreases over time if the carcinogen is removed from the system. Recurrent fusions in CS and reformed smokers were higher than NS, suggesting that smoking can induce recurrent gene fusions (Figure S1B). We also identified several clinically relevant fusions that were reported previously in lung cancer including *ALK*, *FGFR1*, *FGFR2*, *KRAS*, *MET*, *NTRK2*, *RET*, and *ROS1* (Table S3).

Univariate Cox regression analysis identified that tobacco smoking history (hazard ratio =1.3, 95% CI: 1–1.6), Wald statistic P=0.019) and tumor stage (hazard ratio =1.4, 95% CI: 1.1–1.7, Wald statistic P=0.00079) were associated with the survival of LUAD patients. Fusion status did not influence survival (Wald statistic P=0.7). Multivariate Cox regression hazard analysis model showed that tobacco smoking history and tumor stage significantly predicted survival in LUAD patients (Likelihood ratio test =16.05,

P=3e–04).

EML4-ALK fusions were exclusively identified in NS in LUAD (Figure S2A), whereas recurrent fusions partnered by the *CHRM3* gene were identified in patients with a smoking history (Figure S2B). Several genes participating in fusions were commonly identified in smokers, compared to NS (Figure S2B), indicating that some of these fusions can be explored as biomarkers for cigarette smoke exposure. For example, *ASH1L*, *EYS*, *FOCAD*, *GBP5*, *MIPOL1*, *PTK2*, *ZNF638*, *CHRM3*, and *TDRD5* formed gene fusions LUAD patients with smoking history.

Next, we identified two types of fusions: canonical and non-canonical. Canonical fusions have both fusion partners in the 'sense transcriptional' direction, and non-canonical fusions have at least one gene partner in the 'antisense transcriptional' direction. The canonical and non-canonical status indicate their impact on cellular functions. Canonical fusions in LUAD were more frequent than non-canonical fusions. Of the total canonical fusions identified, only 3.3% were recurrent, while 6% of the non-canonical fusions were recurrent (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). About 20% of the genes participating in canonical fusions were recurrent (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). Top recurrent gene fusions included *CHRM3*&*TDRD5*, *CHRM3*&*EYS*, *ASTN2*&*TMEM212*, and *EML4*&*ALK*. Genes forming canonical fusions were enriched in several signaling pathways, including *E2F*, *VEGF*, *BAG2*, *HIPPO*, *and PTEN* signaling (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx).

Fusions in LUSC

LUSC had a significantly higher frequency of canonical recurrent fusions than LUAD (n=108 compared to n=44, Fisher's exact test P=0.0001). This difference in recurrent fusion frequency was not observed for non-canonical recurrent fusions. Fusion frequency in LUSC was not significantly associated with any of the patients' clinical characteristics, including smoking status. Among the canonical fusions, smokers had a larger frequency of recurrent fusions than the other groups (8.2%, Figure S3A,S3B). Several recurrent fusions partnered by genes *PRKAR1A*, *WDR72*, and *ADGRV1*, were identified in all three smoking groups. Several driver fusions previously associated with lung cancer were also identified from our analysis including *FGFR2*, *FGFR3*, *NRG1*, and *NUTMI* (Table S4).

In LUSC, approximately 6% of the canonical fusions and non-canonical fusions were recurrent (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). *SLCO1A2*&*LOC100996634*, *MRGPRX3*&*SMOX*, *CHRM3*&*TDRD5*, and several canonical fusions containing the *RAB3IP* gene were frequent in LUSC. Genes participating in recurrent canonical fusions in LUSC were enriched in several cancer signaling pathways, including ErBb signaling, BMP signaling, and ERK/MAPK Signaling (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). No differences in tumor mutation burden or mutation frequency were identified across non-smoking or smoking groups in LUSC.

Comparison of fusion profiles between LUAD and LUSC

Several common genes participating in fusions were identified between LUAD and LUSC. Some of these recurring genes had higher fusion frequencies in both cancers. For example, *RAB3IP*, *TDRD5*, and *CHRM3* consistently formed fusions in both cancers (*Figure 1A*,*1B*). A small subset of 74 fusions was shared between the two cancers (*Figure 1C*, available online: available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). Genes participating in fusions identified in LUAD and LUSC cancers were associated with mTOR signaling, Cell Cycle Control of Chromosomal Replication, PXR/RXR Activation, and eNOS Signaling (IPA analysis P<0.05). A set of genes implicated in NSCLC was also identified in this common list using *IPA* (*AP5Z1*, *ARHGAP15*, *CALB1*, *CUL4A*, *EYS*, *FKBP5*, *FOCAD*, *GTPBP1*, *HIF1A*, *KYNU*, *MERTK*, *MMRN1*, *NID1*, *PI4KA*, *PRB3*, *PRIM2*, *QKI*, *SCFD1*, *SEMA4D*, *SH3PXD2A*). We also identified about a 30% overlap of fusions with other reports, indicating that the rest are novel fusions identified by our analysis. Genes participating in fusions that were unique to LUAD were enriched in HOTAIR Regulatory Pathway, Ceramide Degradation, EIF2 Signaling, Regulation of the Epithelial-Mesenchymal Transition Pathway (Figure S4A). Pathways associated with PPARα/RXRα Activation, Aryl Hydrocarbon Receptor Signaling, Apoptosis Signaling, and BMP signaling pathway were enriched in genes forming unique fusion in LUSC (Figure S4B).

Mutations across smoking groups in LUAD

Tumor mutation burden across the four smoking groups varied in LUAD. NS and RF >15 years had a lower mutation burden than CS and RF ≥15 years (*Figure 2A*, available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx).

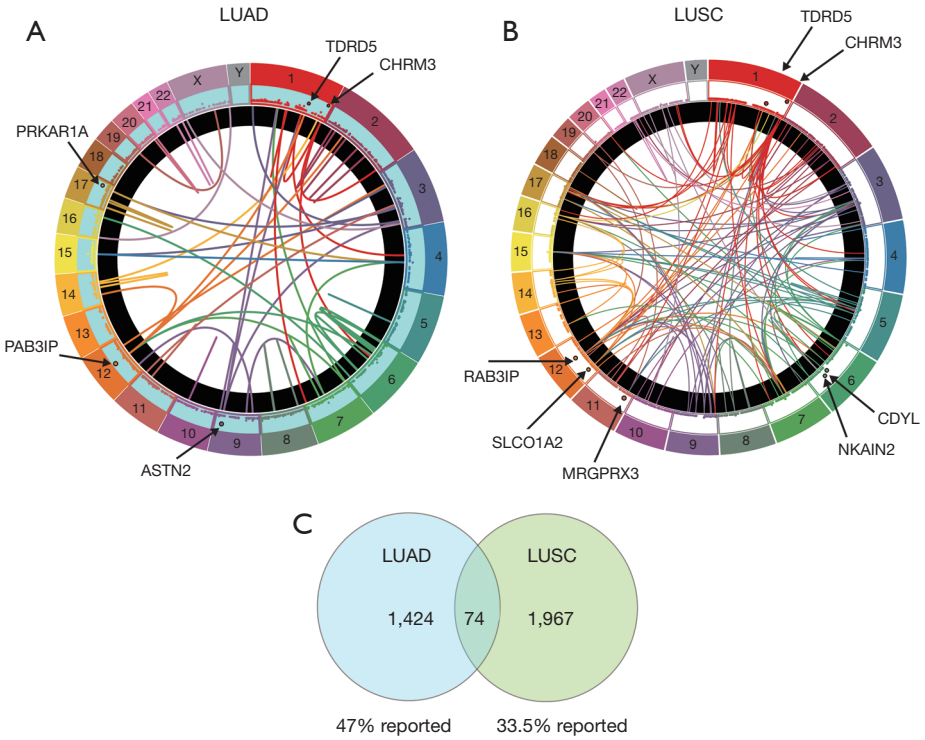

Figure 1 Recurrent fusion profiles in LUAD (A) and LUSC (B) are represented as circos plots. Genes participating in recurrent fusions are highlighted in each cancer. (C) Common fusions identified in LUAD and LUSC are reported, along with the percentage of previously reported fusions. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

The mutation profile of CS is very different from that of non-smokers in LUAD. *TP53* was the most affected gene in both groups, with CS having higher mutations than NS. This was also true for several genes with significantly different mutation profiles between the groups in LUAD (*Figure 2B*, available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). Noteworthy differences in the mutational burden in CS mostly include missense mutations and a significant number of multi-hit, splice site, and non-sense mutations-all leading to disruption of gene function. Overall, *TP53* showed the highest number of in-frame deletions and non-sense mutations; *LRP1B* has the highest number of splice site mutations, while *CSMD3* recorded the most multi-hit mutations in the CS. These highly mutated genes in CS overlapped with the MsigDB chemical and genetic perturbations datasets, 'Ding lung cancer mutated significantly' (Hypergeometric test Padj=7.77e–4) and 'Ding lung cancer by mutation rate' (Hypergeometric test Padj=4.68e–4). This observation is consistent with other reports that suggest smokers accumulate mutations at a higher rate in specific genes (50,51). NS and RF <15 years had fewer differences in mutational profiles. Only *TTN* (Padj=0.022) and *MUC16* (Padj=0.022) were significantly mutated in reformed smokers compared to non-smokers. No difference in mutation was identified between NS and RF >15 years, indicating that the mutations induced by cigarette smoke are rescued after they quit smoking.

We also investigated the co-occurrence and mutual exclusivity of mutations in LUAD across groups with different smoking statuses. Mutual exclusivity identifies gene mutations that do not co-exist in the same sample. This mutual exclusivity can be correlated to complementary functions associated with tumor initiation and progression, or the combination could be lethal for tumor cell survival (52). On the other hand, co-occurring gene mutations can activate complementary oncogenic pathways. Interactions between different mutations revealed a co-occurrence of several genes, and the patterns differed across the smoking status (*Figure 3*). *Figure 3* represents the top 25 mutations that are either co-occurring or are mutually exclusive across LUAD patients with different smoking statuses. None of the mutations were found to

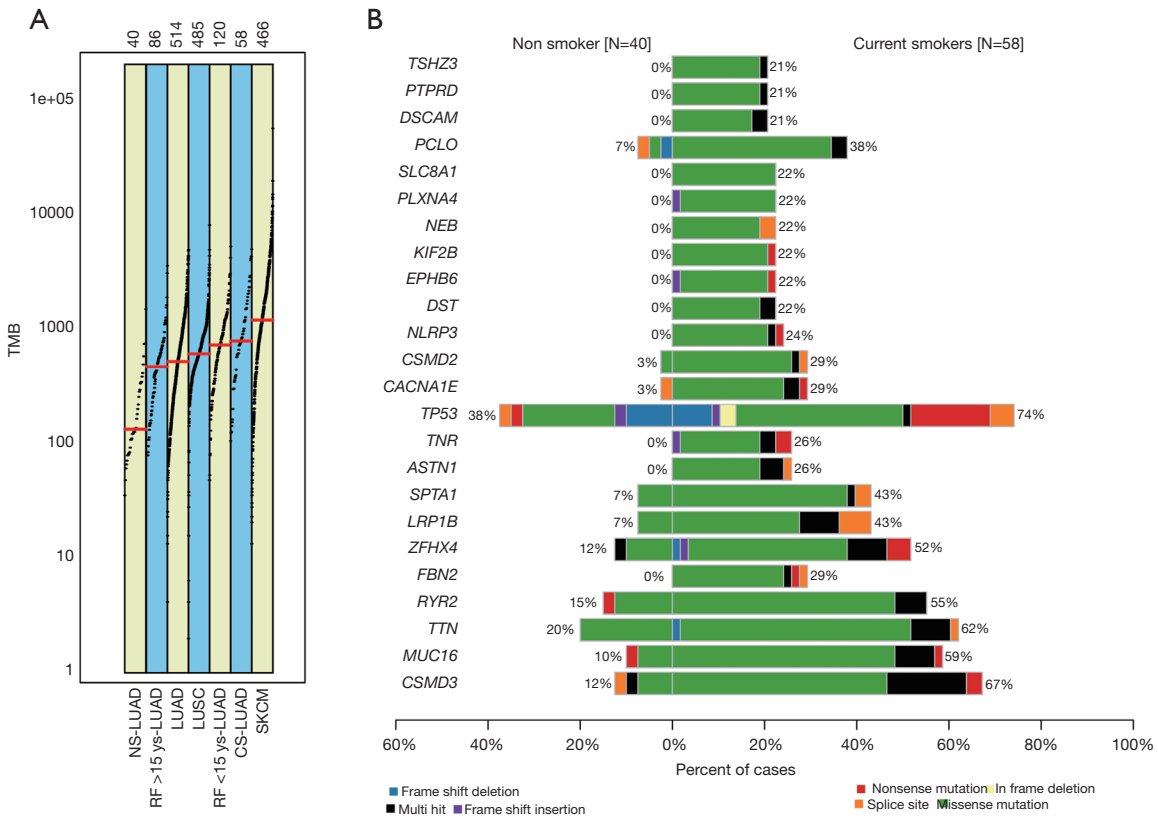

Figure 2 Mutational differences across different smoking groups in LUAD. (A) Comparison of tumor mutation burden among LUAD smoking groups. LUSC and SKCM is shown in the figure for comparison. NS-LUAD: Non-smokers with LUAD, CS-LUAD: Current smokers with LUAD, RF >15 years-LUAD: Reformed smokers for 15 years or more with LUAD, RF <15 years-LUAD: Reformed smokers for less than 15 years with LUAD. (B) Co-Bar plot displaying the mutational profiles between non-smokers and current smokers in LUAD. Most of the genes, including *TP53*, accumulated mutations in smokers. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; SKCM, skin cutaneous melanoma; TMB, tumour mutational burden.

be mutually exclusive among the four groups. Several co-occurring mutations were identified in CS and RF ≤15 years, compared to NS and RF >15 years (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). Though not statistically significant, the percentage of co-occurring mutations among CS and RF ≤15 years were higher (>0.9%) compared to NS and RF >15 years (>0.65%, Chi-Square P=0.06). Eleven common co-occurring mutations were identified among CS and RF ≤15 years. These common co-occurring mutations include *PAPPA2* mutations with *XIRP2*, *TP53*, and *USH2A*; *USH2A* mutations with *MUC17;* and *TTN* mutations with *TP53*, *XIRP2*, *LRP1B*, and *ZFHX4*. The number of co-occurring mutations with *TP53* was comparatively higher in patients with a current or recent smoking history than others.

Fusions, gene expression profiles, and mutations in DDR pathways associated with cigarette smoke exposure

LUAD

We investigated the genes involved in DDR pathways for LUAD and LUSC datasets. Thirty-three of 276 genes involved in the DDR pathway formed fusions in LUAD. Several of these fusions had low fusion frequency, and the majority of them were singletons. *RAD51B* and *PPP4R1* (Homology-dependent recombination pathway; HDR) and *SMARCA4* were found to form recurrent fusions in LUAD (*Tables 1*,*2*). LUAD patients with gene fusions in the DDR pathway were higher among NS (21%) than patients with a smoking history (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). A comparison of DDR pathway genes participating in fusions across the four

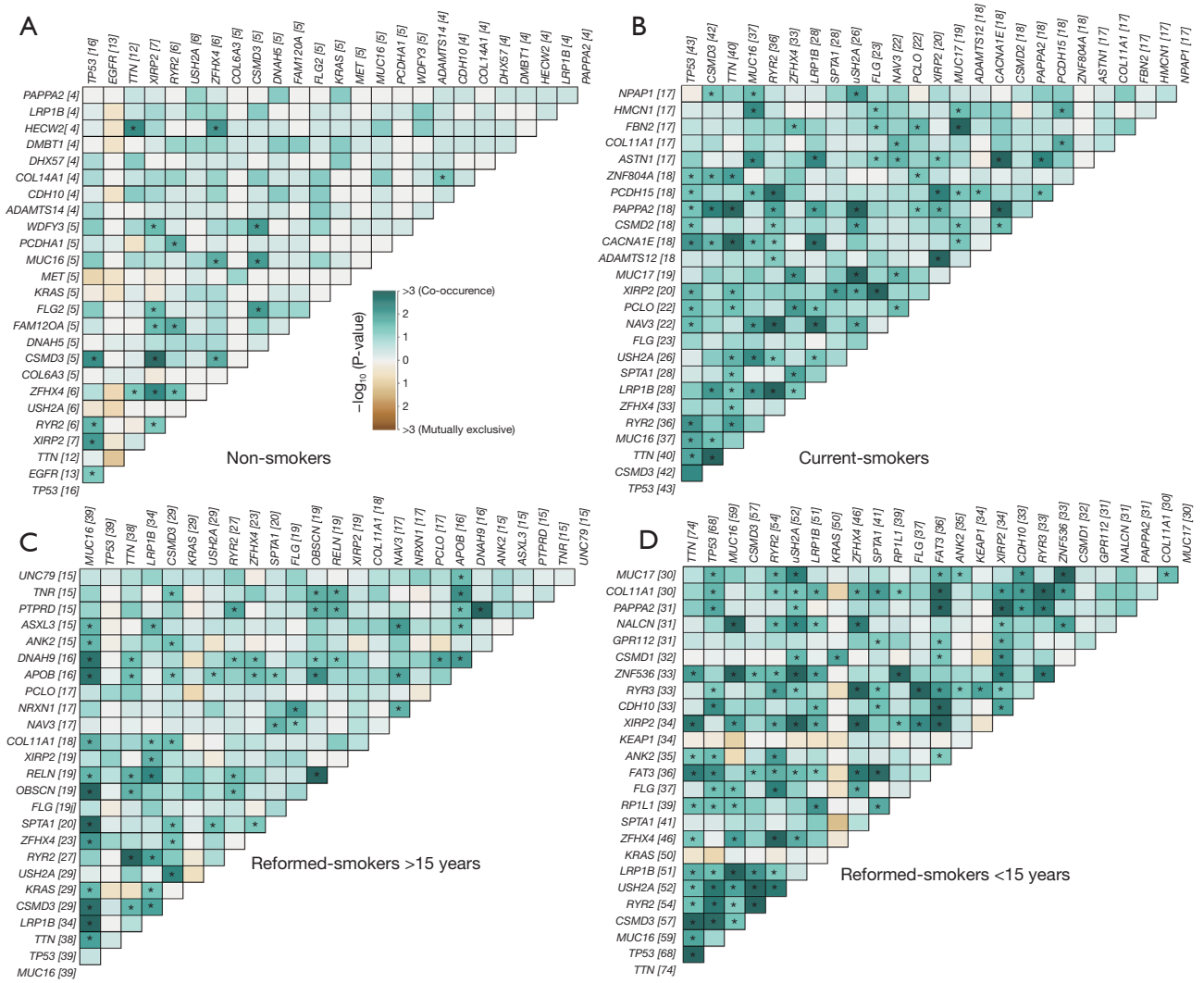

Figure 3 Co-mutation plot for smoking groups in lung adenocarcinoma (LUAD). (A) Non-smokers, (B) current-smokers, (C) reformed smokers >15 years, (D) reformed smokers <15 years. Green in each square represents the co-occurrence of 2 gene mutations, and brown represent mutually exclusive mutations. The frequency of each mutation is indicated adjacent to the gene. *, P<0.05.

smoking groups in LUAD is shown in *Figure 4*. NS had two core genes (*PARP1*, *POLB*) of the 47 genes participating in fusions in the Base Excision repair pathway (BER), and other groups did not present fusions in this pathway. This trend was also observed with the core DDR pathway genes; NS had a higher frequency of fusions involving core DDR pathway genes among all four groups. Fusions involving *PARP1*, *POLB*, *TOP3A*, and *ALKBH3* were identified among NS.

Gene expression of the DDR genes varied across smoking status in LUAD. Thirty-six genes in the DDR pathway were differentially expressed, with most of them having higher expression in CS (Figure S5, available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). Mutation profile differences were rare among the DDR pathway genes in LUAD across the smoking groups. In the DDR pathway, mutations were identified in *APEX1*, *EME1*, *GEN1*, *POLM*, *RBBP8*, and *ERCC5* but were limited to a handful of samples.

LUSC

In LUSC, 48 of the 276 genes in the DDR pathway were found to be forming fusions (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). Genes

Table 1 List of DNA damage repair (DDR) pathway genes involved in fusion formation in lung adenocarcinoma (LUAD)

DDR pathway	Genes
Base excision repair (BER)	*PARP1, POLB, POLD3, MPG, POLD4, APTX*
Nucleotide excision repair (NER)	*POLD3, POLD4, MNAT1, TCEA1, CUL4A, XPC, GTF2H2*
Mismatch repair (MMR)	*POLD3, POLD4*
Fanconi anemia (FA)	*TOP3A, TOP3B, BRCA1, XRCC2, FANCA*
Homology-dependent recombination (HDR)	*PARP1, RAD51B*, TOP3A, POLD3, PPP4R1*, TOP3B, NSMCE2, POLD4, BRCA1, XRCC2, SMARCAD1*
Non-homologous end joining (NHEJ)	*PARP1, POLB*
Direct repair (DR)	*ALKBH3*
Translesion synthesis (TLS)	*POLB, REV1, UBE2N, POLN*
Nucleotide pools (NP)	–
Others	*SMARCA4*, SMARCC1, CLK2, RNF169, SETMAR, MDC1, ATM, HERC2, WEE1, YWHAE, DCLRE1B*

The genes with * indicate recurrent fusions identified in LUAD.

Table 2 List of core DNA damage repair (DDR) pathway genes involved in fusion formation in lung adenocarcinoma (LUAD)

DDR pathway	Genes
Base excision repair (BER)	*PARP1, POLB*
Nucleotide excision repair (NER)	*XPC*
Mismatch repair (MMR)	–
Fanconi anemia (FA)	*FANCA*
Homology-dependent recombination (HDR)	*TOP3A, BRCA1*
Non-homologous end joining (NHEJ)	–
Direct repair (DR)	*ALKBH3*
Translesion synthesis (TLS)	*REV1, POLN*
Damage sensor, etc.	*MDC1, ATM*

associated with Nucleotide Excision Repair (NER) and Homology-dependent recombination (HDR) had a higher frequency of fusions in LUSC (Figure S6). NS had higher genes with fusions in the DDR pathways, followed by smokers (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). CS and RF >15 years had the lowest representation of fusions in the DDR pathway genes (5.5% for non-smokers compared to 3.9% for smokers and reformed smokers). Among the core DDR pathway genes, smokers in LUSC had a higher percentage of fusions than non-smokers (0.6% *vs.* 0%). *CUL4*, *RBX1*, *PMS2*, *BRE*, *POLB*, *PTEN*, *CHEK2*, *RBBP8*, *YWHAE*, and *SETMAR* were recurrently forming fusions in LUAD (*Tables 3*,*4*). Among the core DDR pathway genes, *PMS2*, *RBBP8*, and *CHEK2* formed recurrent fusions. Gene expression and mutation differences in DDR pathway genes across smoking groups were not detected in LUSC (data not shown).

Fusions, gene expression profiles, and mutational profiles of patients with high or low fusion frequency

LUAD

To investigate the genomic profile of the LUAD samples with diverse fusion profiles, we grouped the samples into high or low fusion groups and compared the gene expression, mutation, and fusion profiles between them. Non-smokers had a lower percentage of fusions compared to other smoking groups (Fisher exact test P=0.0009). Next, we compared the gene expression difference between the high and low fusion groups independently in both cancers. Differential gene expression analysis revealed a total of 1,358 genes differentially expressed between the high and low fusion groups, with 685 genes upregulated in the high fusion group (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). Pathway enrichment analysis using *FGSEA* with different MSigDB gene sets identified several differentially expressed pathways. Enrichment analysis with Hallmark gene sets (53) identified pathways associated with G2M checkpoint, E2F

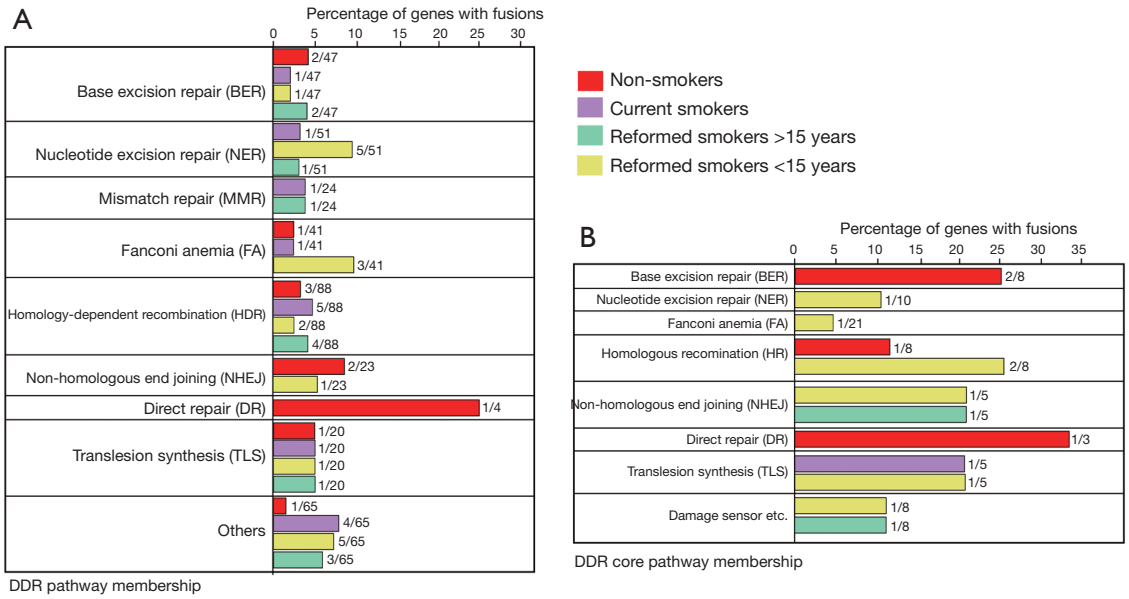

Figure 4 Genes that participate in fusions in the DDR pathways in smoking groups in LUAD. (A) Genes in DDR pathway. (B) Genes in DDR core pathway. All relevant DDR pathway genes and core pathway genes are represented separately for each group. The percentage of genes participating in fusions in the DDR pathway is represented as percentages along with the number of genes. All genes associated with the DDR pathways were compiled from Knijnenburg *et al.*, 2018 (39). DDR, DNA damage repair; LUAD, lung adenocarcinoma.

Table 3 List of DNA damage repair (DDR) pathway genes involved in fusion formation in lung squamous cell carcinoma (LUSC)

DDR pathway	Genes
Base excision repair (BER)	*PARG, POLB*
Nucleotide excision repair (NER)	*CUL4A*, RBX1*, MNAT1, CUL3, GTF2H1, GTF2H5, TCEB2, DDB1, XPA*
Mismatch repair (MMR)	*PMS2**
Fanconi anemia (FA)	*BRE*, FANCA*
Homology-dependent recombination (HDR)	*PARG, NSMCE1, PAXIP1, SHFM1, HFM1, NSMCE2, POLH, RAD51D, RAD51B, NBN, POLQ, RBBP8**
Non-homologous end joining (NHEJ)	*POLB, NBN, PRKDC, PARG*
Direct repair (DR)	–
Translesion synthesis (TLS)	*POLB*, POLQ, POLN, HLTF*
Nucleotide pools (NP)	–
Others	*PTEN*, CHEK2*, YWHAE*, HERC2, CDC25A, POLA1, RRM2B, SOX4, TYMS, YWHAG, SMARCA4, SETMAR*, ATRX*

The genes with * indicate recurrent fusions identified in LUSC.

targets, MTOC1 signaling upregulated, and P53 pathway downregulated in samples with high fusions (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx, *Figure 5*). Among the gene sets associated with chemical and genetic perturbations, the 'Shedden lung cancer poor survival A6' dataset was enriched in the high fusion group (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). This dataset is associated with poor survival in NSCLC. FGSEA, with the canonical dataset from MSigDB identified several pathways related

Table 4 List of core DNA damage repair (DDR) pathway genes involved in fusion formation in lung squamous cell carcinoma (LUSC)

DDR pathway	Genes
Base excision repair (BER)	*POLB*
Nucleotide excision repair	*XPA*
Mismatch repair (MMR)	*PMS2**
Fanconi anemia (FA)	*FANCA*
Homologous recombination (HR)	*SHFM1, NBN, RBBP8**
Non-homologous end joining (NHEJ)	*PRKDC*
Direct repair (DR)	–
Translesion Synthesis (TLS)	–
Damage Sensor, etc.	*CHEK2**

The genes with * indicate recurrent fusions identified in LUSC.

to cell cycle, DNA replication, and kinesins enriched in samples with high fusion. Among the immunologic signature gene set from MSigDB, the up-regulation of 'GSE21063_WT_VS_NFATC1_KO_8H_ANTI_IGM_STIM_BCELL_UP' (Padj=0.02; available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx) and 'GSE13411_PLASMA_CELL_VS_MEMORY_BCELL_UP' (Padj=0.03; available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx) indicates the activation of B cells in samples with high fusion groups. The enrichment of the 'GSE18893_TCONV_VS_TREG_24H_TNF_STIM_UP' gene set in samples with high fusion also suggested tumor necrosis factor (TNF) induced transcription program that led to upregulation of the cytokines, various anti-apoptotic genes; and immune-response genes in these samples. GO pathway analysis with ontology gene sets

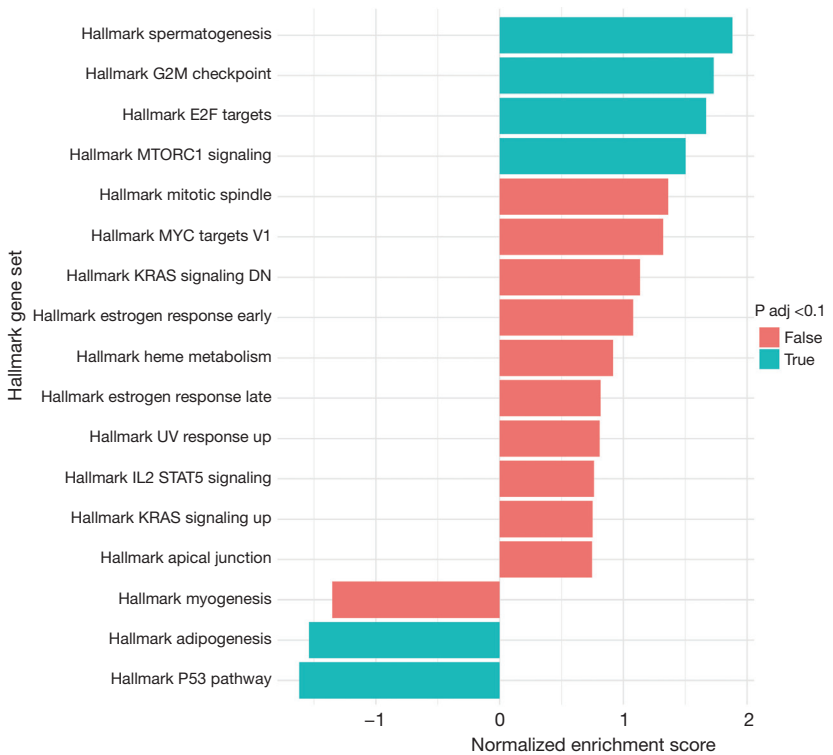

Figure 5 Fast Gene Set Enrichment (*FGSEA*) analysis of differentially expressed genes with high or low fusion frequency in LUAD using Hallmark gene set from The Molecular Signatures Database (MSigDB). Pathways enriched in samples with high fusion frequency are represented here with a higher Normalized Enrichment Score. True indicates pathways enriched with adjusted P≤0.1. LUAD, lung adenocarcinoma.

from MSigDB identified several GO terms linked to cell cycle, chromosome segregation, and signal transduction in response to DNA damage enriched in samples with high fusion frequency (Padj<0.01, available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx).

Of the 275 DDR pathway genes, 47 were differentially expressed between high and low fusion samples in LUAD. The majority of these genes were upregulated in the high fusion group, indicating that the DDR pathway is relatively active in samples with high fusion compared to low fusion (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). Though not statistically significant, trends in fusion frequency of genes in DDR pathways among samples with high or low fusion frequency were noted. The frequency of fusions in the DDR pathway genes in samples with high fusion in LUAD was relatively higher (4.1%) than samples with low fusion (0.1%, Fisher's exact test P=0.08, available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). But this observation was reversed when only the core DDR pathway genes (N=80) were analyzed. The low-fusion group had a higher percentage of fusions within the core DDR pathway genes than the high-fusion group (2.7% compared to 0.8%, Fisher's exact test P=0.07, available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx. Fanconi Anemia (FA) and Homology-dependent recombination (HDR) pathways were most affected by fusions in samples with high fusion. In contrast, the Direct repair pathway was the most affected in samples with low fusion, indicating that fusion patterns differed between the two groups. Genes participating in fusions among the DDR pathway in samples with high fusion included *BRCA1*, *XRCC2*, *PARP1*, *RAD51B*, and *SMARCA4* (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). Few genes, including *SMARCA4*, *POLB*, and *XPC*, formed recurrent fusions among LUAD samples with low fusion frequency (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx).

Tumor mutation burden differed across high and low fusion groups in LUAD (Figure S7). Samples with a low fusion rate also had a lower tumor mutation burden than those with high fusions in LUAD (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). Further comparison of mutation frequency between LUAD samples with high versus low frequency identified a set of genes with significantly different mutation profiles (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). These genes included *ADAMTS2*, *TP53*, *HEPHL1*, *NTNG1*, *NRXN1*, *DNAH1*, *LRRTM3*, *DCC*, and *SCN1A* (Padj<0.05). Top mutated genes in high (*Figure 6A*) and low fusion (*Figure 6B*) samples are represented as oncoplots. *TP53* gene mutation was higher in samples with high fusion (Padj=0.01), which correlated with the lower activity of the P53 pathway in this group. We also compared mutation profiles within each smoking group in LUAD to identify genes differentially mutated between high and low fusion groups. Mutations in *DNAH3*, *NRXN1*, *MUC17*, *TSHZ3*, *TP53*, *LRP2*, *SI*, and *TTN* had higher mutations (Padj=0.05) in samples with high fusion frequency among CS compared to samples with low fusions within this group (Figure S8A). Among RF ≤15 years, the high fusion group had a higher frequency of mutations in *TP53*, *TTN*, *ADAMTS12*, *CSMD2*, *USH2A*, and *COL6A3* (Figure S8B, Padj<0.05). We checked the overlap among the MSigDB gene sets for significantly mutated genes in samples with high fusions. Significant overlap for the 'Ding Lung Cancer Mutated Significantly' gene set (part of MSiGDB chemical and genetic perturbations gene set) was identified (Hypergeometric test Padj=8.2e–6), indicating that these mutated genes are involved in lung cancer.

Mutational signature identified in samples with high fusion contained a signature similar to SBS1 (cosine-similarity: 0.905), associated with spontaneous or enzymatic deamination 5-methylcytosine to thymine which generates G:T mismatches in double-stranded DNA (Figure S9). This signature was not identified in samples with low fusions in LUAD. Mutational signatures identified in both high and low fusion containing groups in LUAD were similar to SBS2 (APOBEC Cytidine Deaminase (C>T), SBS6 (defective DNA mismatch repair), SBS4 (exposure to tobacco (smoking) mutagens).

Next, to investigate the impact of *TP53* mutation on gene fusion formation, a subset of lung adenocarcinoma cell lines were analyzed. Fusion profiles in cell lines containing mutant *TP53* (NCI-H1299, NCI-H441, NCI-H1437, NCI-H727, NCI-H23, NCI-H358) were compared against cell lines containing wild type *TP53* gene (NCI-H460, A549, HCC827, NCI-H1395, NCI-H226, NCI-H1666, NCI-H1563, NCI-H292) using the same protocol mentioned earlier. Average fusion frequency in *TP53* mutant cell lines was higher than fusions observed in *TP53* wild-type cell lines (113 *vs.* 70, *t*-test P=0.03, available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). This link between *TP53* and fusion gene frequency warrants further investigations into the mechanisms of genomic and transcriptomic gene fusion formation.

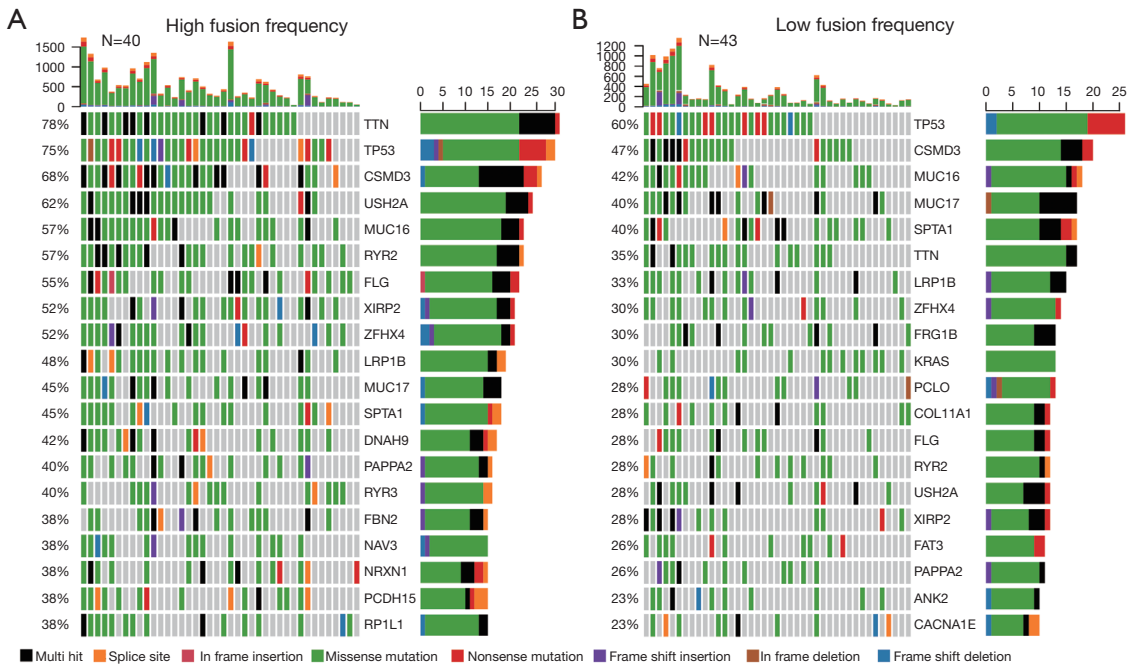

Figure 6 Oncoplot representing the mutational profile and frequency of mutation among samples with high (A) or low (B) fusion frequency in LUAD. Each column represents a patient sample. Barplot corresponding to each column represents the frequency of mutations in each patient. The horizontal barplot represents the frequency of mutations in each gene. Only genes with high frequency of mutations are represented in this plot. The number of mutated samples, frequency, and the type of mutations in each gene is represented. LUAD, lung adenocarcinoma.

LUSC

The percentage of DDR fusions did not differ among the high and low fusion groups in LUSC (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx, Fisher Exact test P>0.05). Among the different DDR pathways, genes participating in Fanconi Anemia (FA), Translesion Synthesis (TLS), Non-homologous End Joining (NHEJ), and Homologous Recombination (HR) frequently formed fusions in samples with high fusion. Among the core genes involved in DDR pathways, only genes associated with Mismatch Repair were affected in samples with low fusion frequency (available online: https://cdn.amegroups.cn/static/public/tlcr-22-113-1.xlsx). We did not find differences in gene expression or mutation across samples with high and low fusions in LUSC. In addition, no differences in TMB were identified across samples with high or low fusion frequency in LUSC.

Discussion

This study analyzed the fusion profile of two TCGA lung cancer cohorts, lung adenocarcinoma, and squamous cell carcinoma, to compare the genomic alterations and transcription profiles between the cancers. Fusion gene profiling was performed at the transcript level using a non-alignment-based fusion detection algorithm, ChimeRScope, which performed superior to other popular tools for cancer datasets in our previous studies (38,54,55). Our analysis revealed several recurrent canonical and non-canonical fusions in LUAD and LUSC and only a minor overlap between the two datasets. LUAD had fewer recurrent fusions than LUSC, and its fusion frequency was also linked to several patient clinical characteristics. Specifically, lower fusion frequency was linked to the tumor-free status, suggesting the link between tumor progression and associated genomic instability. Genomic instability is one of

cancer's hallmarks (56) and is identified in precancerous lesions (57,58) and advanced tumors (59,60). Both chromosomal instability (CIN) and non-CIN forms of genomic instability can directly measure tumor aggressiveness (61) and result in fusion genes (62,63). However, tumor status was not associated with fusion frequency in LUSC, suggesting that both cancers differ genetically.

LUAD had a higher frequency of fusions among patients that were either current or past smokers than non-smokers. However, this positive association of smoking and fusion frequency was not identified in LUSC. Even though LUSC is strongly linked to a history of smoking (8), the association of fusion frequency and tumor mutation burden to smoking status exclusive to LUAD could be attributed to the differences in cancer originating cell types (64,65). We also observed that an extended period of smoking cessation lowered fusion frequency, which indicated that quitting enhanced the replacement of DNA-damaged cells by normal cells. This phenomenon was also reported for mutations in the lungs and suggested that quitting smoking promotes bronchial epithelium replenishment from mitotically quiescent cells (66).

Analysis of gene expression, fusion patterns, and mutational profiles among different smoking groups in LUAD and LUSC revealed interesting patterns. In both cancers, non-smokers had a higher frequency of fusions in the DDR pathway genes than smokers. This observation needs to be investigated further to identify the underlying molecular mechanism leading to higher fusions in DDR genes associated with non-smokers. On the other hand, differentially expressed genes in the DDR pathway in LUAD had a relatively higher expression in CS than NS, indicating that DNA damage induced by cigarette smoke can activate DDR pathway genes. An earlier report also identified a correlation between increased expression of DNA repair pathway genes with smoking in LUAD (67). Several fusions characteristic of smokers were identified; for example, fusions involving *CHRM3* (cholinergic receptor muscarinic 3) were strongly associated with smokers. This gene has been associated with cigarette smoke-induced proinflammatory role for *CHRM3* in the bronchial epithelium of a mice model (68). Cigarette smoke also increases acetylcholine production, increasing *CHRM3* expression, thereby activating proinflammatory signaling in bronchial epithelial cells (69). Fusions involving *CHRM3* could be further explored as a biomarker for smoking-associated molecular changes. These fusions associated with smoking can be further investigated for patient stratification to identify high-risk patient subgroups or those responding to targeted therapy. In addition, novel gene inhibitors can be developed and tested against specific target genes involved in these fusions. For example, *RET* fusions identified in NSCLS can now be targeted by *RET*-specific inhibitors, selpercatinib and pralsetinib, which have been successfully tested in the clinics and are now FDA approved (70).

Comparison of fusion high and low fusion groups in LUAD identified differences in gene expression and mutation patterns among the two groups. Gene Set Enrichment Analysis with Hallmark gene set identified pathways associated with G2M checkpoint, E2F targets, and MTOC1 signaling upregulated in samples with high fusions, and P53 pathway downregulated in samples with high fusion frequency. This observation indicates that alterations in DNA damage checkpoints (G2/M) and P53 pathway could be related to higher fusion frequency in LUAD. Cell cycle checkpoints act as a regulator to ensure the integrity of chromosomes before they proceed through these vital replication stages (71,72). G1M and G2M checkpoints regulate the progression of a cell into the S phase and the mitosis phase, respectively, and are controlled by the damage response signaling pathways that will halt the progression of replication mitosis to accelerate DNA repair activity, if necessary (73,74). The expression of genes in the DDR pathways was also elevated in samples with high fusion. Alterations in the expression of the G2M and DDR pathway indicate that a link between perturbed DNA repair mechanism and G2M regulation could lead to increased fusion frequency.

We also report a decreased activity of the P53 pathway in LUAD samples with high fusion, which could be associated with higher *TP53* mutations. This observation was also validated in LUAD cell lines with *TP53* functional mutation. *TP53* is a tumor suppressor gene known to be mutated in several cancers and a critical player in the anti-cancer defense mechanism (75,76). TP53 activation orchestrates cell cycle arrest and apoptosis and regulates numerous cellular processes (77). TP53 activation through the p53-p21-DREAM/RB pathway indirectly leads to repression of cell cycle genes in addition to affecting the expression of several transcription factors (78,79). Decreased activity of *TP53* ultimately leads to increased genomic instability manifested as mutations, fusions, and other gross chromosomal abnormalities (80-82). Consistent with these reports, samples with higher fusion frequency also had a higher mutation burden signifying increased genomic instability, probably contributed by decreased activity of the *TP53* pathway. Further investigations are warranted to

understand the exact molecular mechanism that connects the *TP53* pathway to fusion gene formation in LUAD.

Conclusions

This report investigated the fusion profiles of two lung cancer datasets, LUAD and LUSC, from the TCGA database in the context of the smoking status of the four smoking cohorts. Fusion profiles of LUAD and LUSC were different, with increased recurrent fusions in LUSC and minimal overlap between the two cancers. We identified a significant influence of smoking on fusion formation in LUAD. Several smoking-associated fusions were identified that could be explored for biomarker development for translational use in patient classification. Further examination of the molecular differences between samples with high or low fusion frequency in LUAD revealed altered cell cycle regulation, TP53 activation, and DNA damage response pathways associated with different fusion profiles. These pathways should be further scrutinized for understanding the exact mechanism of fusion formation in LUAD with or without the influence of smoking. Novel recurrent fusions identified in this study can be further explored as biomarkers for therapeutic response and to stratify high-risk patient subgroups or patients responding to a targeted therapy. Exploring how cigarette smoke further impacts these pathways and leads to gene fusion formation could unravel previously unknown molecular mechanisms of carcinogenesis and can impact early biomarker development.

Acknowledgments

The authors would like to thank the Bioinformatics and Systems Biology Core (BSBC) facility at UNMC for providing the computational infrastructure and support. The authors also acknowledge the Holland Computing Center of the University of Nebraska-Lincoln for computational resources, which receives support from the Nebraska Research Initiative.

Funding: This work was supported by the Nebraska Research Initiative and multiple National Institute of Health awards (Nos. 5P20GM103427, 5P30CA036727, 5P30MH062261, 5U54GM115458) to the Bioinformatics and Systems Biology Core led by CG.

Footnote

Reporting Checklist: The authors have completed the MDAR reporting checklist. Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-22-113/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-22-113/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Siegel RL, Miller KD, Fuchs HE, et al. Cancer Statistics, 2021. CA Cancer J Clin 2021;71:7-33.
2. Wang DC, Wang W, Zhu B, et al. Lung Cancer Heterogeneity and New Strategies for Drug Therapy. Annu Rev Pharmacol Toxicol 2018;58:531-46.
3. Herbst RS, Heymach JV, Lippman SM. Lung cancer. N Engl J Med 2008;359:1367-80.
4. Available online: https: //www.cancer.gov/types/lung/hp/non-small-cell-lung-treatment-pdq
5. Bender E. Epidemiology: The dominant malignancy. Nature 2014;513:S2-3.
6. Buccheri G, Barberis P, Delfino MS. Diagnostic, morphologic, and histopathologic correlates in bronchogenic carcinoma. A review of 1,045 bronchoscopic examinations. Chest 1991;99:809-14.
7. Chasin WD. Atlas of Flexible Bronchofiberscopy. JAMA 1975;231:662.
8. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers--a different disease. Nat Rev Cancer 2007;7:778-90.
9. Davies H, Bignell GR, Cox C, et al. Mutations of the BRAF gene in human cancer. Nature 2002;417:949-54.

10. Santos E, Martin-Zanca D, Reddy EP, et al. Malignant activation of a K-ras oncogene in lung carcinoma but not in normal tissue of the same patient. Science 1984;223:661-4.
11. Engelman JA, Zejnullahu K, Mitsudomi T, et al. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science 2007;316:1039-43.
12. Fernandez-Cuesta L, Plenker D, Osada H, et al. CD74-NRG1 fusions in lung adenocarcinoma. Cancer Discov 2014;4:415-22.
13. Kohno T, Ichikawa H, Totoki Y, et al. KIF5B-RET fusions in lung adenocarcinoma. Nat Med 2012;18:375-7.
14. Rikova K, Guo A, Zeng Q, et al. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell 2007;131:1190-203.
15. Soda M, Choi YL, Enomoto M, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448:561-6.
16. Stephens P, Hunter C, Bignell G, et al. Lung cancer: intragenic ERBB2 kinase mutations in tumours. Nature 2004;431:525-6.
17. Vaishnavi A, Capelletti M, Le AT, et al. Oncogenic and drug-sensitive NTRK1 rearrangements in lung cancer. Nat Med 2013;19:1469-72.
18. Weiss J, Sos ML, Seidel D, et al. Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. Sci Transl Med 2010;2:62ra93.
19. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489:519-25.
20. Chen M, Liu X, Du J, et al. Differentiated regulation of immune-response related genes between LUAD and LUSC subtypes of lung cancers. Oncotarget 2017;8:133-44.
21. Scagliotti GV, Parikh P, von Pawel J, et al. Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. J Clin Oncol 2008;26:3543-51.
22. Johnson DH, Fehrenbacher L, Novotny WF, et al. Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer. J Clin Oncol 2004;22:2184-91.
23. Lynch TJ, Bell DW, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 2004;350:2129-39.
24. Paez JG, Jänne PA, Lee JC, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 2004;304:1497-500.
25. Shepherd FA, Rodrigues Pereira J, Ciuleanu T, et al. Erlotinib in previously treated non-small-cell lung cancer. N Engl J Med 2005;353:123-32.
26. Zugazagoitia J, Molina-Pinelo S, Lopez-Rios F, et al. Biological therapies in nonsmall cell lung cancer. Eur Respir J 2017;49:1601520.
27. Huang L, Jiang XL, Liang HB, et al. Genetic profiling of primary and secondary tumors from patients with lung adenocarcinoma and bone metastases reveals targeted therapy options. Mol Med 2020;26:88.
28. Pan Y, Zhang Y, Ye T, et al. Detection of Novel NRG1, EGFR, and MET Fusions in Lung Adenocarcinomas in the Chinese Population. J Thorac Oncol 2019;14:2003-8.
29. Zhang X, Jiang W, Yang N, et al. Afatinib response in a lung adenocarcinoma with novel compound S720F+L861R mutation in EGFR. Lung Cancer 2020;148:170-2.
30. Taus Á, Camacho L, Rocha P, et al. Dynamics of EGFR Mutation Load in Plasma for Prediction of Treatment Response and Disease Progression in Patients With EGFR-Mutant Lung Adenocarcinoma. Clin Lung Cancer 2018;19:387-394.e2.
31. Hecht SS. Progress and challenges in selected areas of tobacco carcinogenesis. Chem Res Toxicol 2008;21:160-71.
32. Beane J, Sebastiani P, Liu G, et al. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. Genome Biol 2007;8:R201.
33. Heguy A, O'Connor TP, Luettich K, et al. Gene expression profiling of human alveolar macrophages of phenotypically normal smokers and nonsmokers reveals a previously unrecognized subset of genes modulated by cigarette smoking. J Mol Med (Berl) 2006;84:318-28.
34. Büttner P, Mosig S, Funke H. Gene expression profiles of T lymphocytes are sensitive to the influence of heavy smoking: A pilot study. Immunogenetics 2007;59:37-43.
35. Pfeifer GP, Denissenko MF, Olivier M, et al. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene 2002;21:7435-51.
36. Mertens F, Johansson B, Fioretos T, et al. The emerging complexity of gene fusions in cancer. Nat Rev Cancer 2015;15:371-81.
37. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 2007;7:233-45.
38. Vellichirammal NN, Albahrani A, Banwait JK, et al. Pan-

Cancer Analysis Reveals the Diverse Landscape of Novel Sense and Antisense Fusion Transcripts. Mol Ther Nucleic Acids 2020;19:1379-98.
39. Knijnenburg TA, Wang L, Zimmermann MT, et al. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell Rep 2018;23:239-254.e6.
40. Zhang Z, Hernandez K, Savage J, et al. Uniform genomic data analysis in the NCI Genomic Data Commons. Nat Commun 2021;12:1226.
41. Mayakonda A, Lin DC, Assenov Y, et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res 2018;28:1747-56.
42. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. Nature 2020;578:94-101.
43. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Deciphering signatures of mutational processes operative in human cancer. Cell Rep 2013;3:246-59.
44. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 2016;44:e71.
45. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
46. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15-21.
47. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.
48. Korotkevich G, Sukhov V, Nikolay Budin N, et al. Fast gene set enrichment analysis. bioRxiv 060012.
49. Therneau T. A Package for Survival Analysis in R. R package version 3.2-10, Available online: https://CRAN.R-project.org/package=survival. 2021.
50. Ding L, Getz G, Wheeler DA, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature 2008;455:1069-75.
51. Govindan R, Ding L, Griffith M, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell 2012;150:1121-34.
52. Deng L, Kimmel M, Foy M, et al. Estimation of the effects of smoking and DNA repair capacity on coefficients of a carcinogenesis model for lung cancer. Int J Cancer 2009;124:2152-8.
53. Liberzon A, Birger C, Thorvaldsdóttir H, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015;1:417-25.
54. Li Y, Heavican TB, Vellichirammal NN, et al. ChimeRScope: a novel alignment-free algorithm for fusion transcript prediction using paired-end RNA-Seq data. Nucleic Acids Res 2017;45:e120.
55. Vellichirammal NN, Albahrani A, Li Y, et al. Identification of Fusion Transcripts from Unaligned RNA-Seq Reads Using ChimeRScope. Methods Mol Biol 2020;2079:13-25.
56. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57-70.
57. Gorgoulis VG, Vassiliou LV, Karakaidos P, et al. Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. Nature 2005;434:907-13.
58. Bartkova J, Horejsí Z, Koed K, et al. DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. Nature 2005;434:864-70.
59. Nowell PC. The clonal evolution of tumor cell populations. Science 1976;194:23-8.
60. Lengauer C, Kinzler KW, Vogelstein B. Genetic instability in colorectal cancers. Nature 1997;386:623-7.
61. Correa AF, Ruth KJ, Al-Saleem T, et al. Overall tumor genomic instability: an important predictor of recurrence-free survival in patients with localized clear cell renal cell carcinoma. Cancer Biol Ther 2020;21:424-31.
62. Penserga ET, Skorski T. Fusion tyrosine kinases: a result and cause of genomic instability. Oncogene 2007;26:11-20.
63. Venkatesan S, Natarajan AT, Hande MP. Chromosomal instability--mechanisms and consequences. Mutat Res Genet Toxicol Environ Mutagen 2015;793:176-84.
64. Liu H, Hu X, Zhu Y, et al. Up-regulation of SRPK1 in non-small cell lung cancer promotes the growth and migration of cancer cells. Tumour Biol 2016;37:7287-93.
65. Okabe N, Ezaki J, Yamaura T, et al. FAM83B is a novel biomarker for diagnosis and prognosis of lung squamous cell carcinoma. Int J Oncol 2015;46:999-1006.
66. Yoshida K, Gowers KHC, Lee-Six H, et al. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature 2020;578:266-72.
67. Hammouz RY, Kostanek JK, Dudzisz A, et al. Differential expression of lung adenocarcinoma transcriptome with signature of tobacco exposure. J Appl Genet 2020;61:421-37.
68. Kistemaker LE, Bos IS, Hylkema MN, et al. Muscarinic receptor subtype-specific effects on cigarette smoke-induced inflammation in mice. Eur Respir J 2013;42:1677-88.
69. Profita M, Bonanno A, Montalbano AM, et al. Cigarette smoke extract activates human bronchial epithelial cells

affecting non-neuronal cholinergic system signalling in vitro. Life Sci 2011;89:36-43.
70. Thein KZ, Velcheti V, Mooers BHM, et al. Precision therapy for RET-altered cancers with RET inhibitors. Trends Cancer 2021;7:1074-88.
71. Hartwell LH, Kastan MB. Cell cycle control and cancer. Science 1994;266:1821-8.
72. Zhou BB, Elledge SJ. The DNA damage response: putting checkpoints in perspective. Nature 2000;408:433-9.
73. Canman CE, Lim DS, Cimprich KA, et al. Activation of the ATM kinase by ionizing radiation and phosphorylation of p53. Science 1998;281:1677-9.
74. Terzoudi GI, Manola KN, Pantelias GE, et al. Checkpoint abrogation in G2 compromises repair of chromosomal breaks in ataxia telangiectasia cells. Cancer Res 2005;65:11292-6.
75. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature 2020;578:82-93.
76. Kastenhuber ER, Lowe SW. Putting p53 in Context. Cell 2017;170:1062-78.
77. Chen J. The Cell-Cycle Arrest and Apoptotic Functions of p53 in Tumor Initiation and Progression. Cold Spring Harb Perspect Med 2016;6:a026104.
78. Uxa S, Bernhart SH, Mages CFS, et al. DREAM and RB cooperate to induce gene repression and cell-cycle arrest in response to p53 activation. Nucleic Acids Res 2019;47:9087-103.
79. Fischer M, Quaas M, Steiner L, et al. The p53-p21-DREAM-CDE/CHR pathway regulates G2/M cell cycle genes. Nucleic Acids Res 2016;44:164-74.
80. Eischen CM. Genome Stability Requires p53. Cold Spring Harb Perspect Med 2016;6:a026096.
81. Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. Nature 1998;396:643-9.
82. Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability--an evolving hallmark of cancer. Nat Rev Mol Cell Biol 2010;11:220-8.

Cite this article as: Vellichirammal NN, Albahrani A, Guda C. Fusion gene recurrence in non-small cell lung cancers and its association with cigarette smoke exposure. Transl Lung Cancer Res 2022;11(10):2022-2039. doi: 10.21037/tlcr-22-113

Supplementary

Table S1 List of DNA Damage Repair Pathway Genes

Base Excision Repair (BER)	Nucleotide Excision Repair (NER)-includes TC-NER and GC-NER	Mismatch Repair (MMR)	Fanconi Anemia (FA)	Homology-dependent recombination (HDR)	Non-homologous End Joining (NHEJ)	Direct Repair (DR)	Translation Synthesis (TLS)	Nucleotide pools (NP)	Others
ALKBH1	CCNH	EXO1	APITD1	BARD1	DCLRE1C	ALKBH2	HLTF	NUDT1	AEN
APEX1	CDK7	HMGB1	BARD1	BLM	DNTT	ALKBH3	MAD2L2	NUDT15	ATM
APEX2	CETN2	LIG1	BLM	BRCA1	FAM175A	ASCC3	PCNA	NUDT18	ATR
APLF	CUL3	MLH1	BRCA1	BRCA2	LIG4	MGMT	POLB	RRM1	ATRIP
APTX	CUL4A	MLH3	BRCA2	BRIP1	MRE11A		POLH	RRM2	ATRX
FEN1	CUL5	MSH2	BRE	DMC1	NBN		POLI		BABAM1
HMGB1	DDB1	MSH3	BRIP1	DNA2	NHEJ1		POLK		BCAS2
HMGB2	DDB2	MSH6	ERCC1	EID3	PARG		POLM		BRCC3
LIG1	ERCC1	PCNA	ERCC4	EME1	PARP1		POLN		CDC25A
LIG3	ERCC2	PMS1	FAAP100	EME2	PARP3		POLQ		CDC25B
MBD4	ERCC3	PMS2	FAAP20	ERCC1	PNKP		RAD18		CDC25C
MPG	ERCC4	POLD1	FAAP24	EXO1	POLB		REV1		CDC5L
MUTYH	ERCC5	POLD2	FAN1	FANCM	POLL		REV3L		CHAF1A
NEIL1	ERCC6	POLD3	FANCA	FEN1	POLM		SHPRH		CHEK1
NEIL2	ERCC8	POLD4	FANCB	GEN1	PRKDC		UBE2A		CHEK2
NEIL3	GTF2H1	RFC1	FANCC	H2AFX	RAD50		UBE2B		CLK2
NTHL1	GTF2H2	RFC2	FANCD2	HELQ	RIF1		UBE2N		DCLRE1A
OGG1	GTF2H3	RFC3	FANCE	HFM1	RNF168		UBE2V2		DCLRE1B
PARG	GTF2H4	RFC4	FANCF	INO80	RNF8		USP1		DUT
PARP1	GTF2H5	RFC5	FANCG	KAT5	TP53BP1		WDR48		ENDOV
PARP2	LIG1	RPA1	FANCI	LIG1	XRCC4				EXO5
PARP3	MMS19	RPA2	FANCL	MRE11A	XRCC5				GADD45A
PARP4	MNAT1	RPA3	FANCM	MUS81	XRCC6				GADD45G
PCNA	PCNA	RPA4	HELQ	NBN					HERC2
PNKP	POLD1		HES1	NFATC2IP					HUS1
POLB	POLD2		MAD2L2	NSMCE1					IDH1
POLD1	POLD3		PALB2	NSMCE2					MDC1
POLD2	POLD4		RAD51	NSMCE3					MORF4L1
POLD3	POLE		RAD51C	NSMCE4A					MPLKIP
POLD4	POLE2		RMI1	PALB2					MRPL40
POLE	POLE3		RMI2	PARG					NABP2
POLE2	POLE4		SLX1A	PARP1					PER1
POLE3	RAD23A		SLX4	PARP2					PLK3
POLE4	RAD23B		STRA13	PARPBP					PLRG1
POLK	RBX1		TELO2	PAXIP1					POLA1
POLL	RFC1		TOP3A	PCNA					POLG
RFC1	RFC2		TOP3B	POLD1					PRPF19
RFC2	RFC3		UBE2T	POLD2					PTEN
RFC3	RFC4		USP1	POLD3					RAD1
RFC4	RFC5		WDR48	POLD4					RAD17
RFC5	RPA1		XRCC2	POLH					RAD9A
SMUG1	RPA2			POLQ					RAD9B
TDG	RPA3			PPP4C					RIF1
TDP1	RPA4			PPP4R1					RNF169
UNG	TCEA1			PPP4R2					RNF4

Table S1 (*continued*)

https://dx.doi.org/10.21037/tlcr-22-113

Table S1 (*continued*)

Base Excision Repair (BER)	Nucleotide Excision Repair (NER)-includes TC-NER and GC-NER	Mismatch Repair (MMR)	Fanconi Anemia (FA)	Homology-dependent recombination (HDR)	Non-homologous End Joining (NHEJ)	Direct Repair (DR)	Translation Synthesis (TLS)	Nucleotide pools (NP)	Others
WRN	TCEB1			PPP4R4					RNMT
XRCC1	TCEB2			RAD50					RRM2B
	UVSSA			RAD51					SETMAR
	XAB2			RAD51B					SLX4
	XPA			RAD51C					SMARCA4
	XPC			RAD51D					SMARCC1
				RAD52					SOX4
				RAD54B					SPRTN
				RAD54L					SWI5
				RBBP8					TDP2
				RDM1					TOPBP1
				RECQL					TP53
				RECQL4					TREX1
				RECQL5					TREX2
				RFC1					TTK
				RFC2					TYMS
				RFC3					WEE1
				RFC4					YWHAB
				RFC5					YWHAE
				RMI1					YWHAG
				RMI2					
				RPA1					
				RPA2					
				RPA3					
				RPA4					
				RTEL1					
				SHFM1					
				SLX1A					
				SLX1B					
				SLX4					
				SMARCAD1					
				SMC5					
				SMC6					
				SPO11					
				SWSAP1					
				TOP3A					
				TOP3B					
				TP53BP1					
				UIMC1					
				WRN					
				XRCC2					
				XRCC3					
				ZSWIM7					

https://dx.doi.org/10.21037/tlcr-22-113

Table S2 Core DNA Damage Repair Genes

Base Excision Repair (BER)	Nucleotide Excision Repair (NER, including TC-NER and GC-NER))	Mismatch Repair (MMR)	Fanconi Anemia (FA)	Homologous Recombination (HR)	Non-homologous End Joining (NHEJ)	Direct Repair (DR)	Translation Synthesis (TLS)	Damage Sensor etc.
APEX1	CUL5	EXO1	FANCA	BARD1	LIG4	ALKBH2	POLN	ATM
APEX2	ERCC1	MLH1	FANCB	BLM	NHEJ1	ALKBH3	POLQ	ATR
FEN1	ERCC2	MLH3	FANCC	BRCA1	POLL	MGMT	REV1	ATRIP
PARP1	ERCC4	MSH2	FANCD2	BRCA2	POLM		REV3L	CHEK1
POLB	ERCC5	MSH3	FANCI	BRIP1	PRKDC		SHPRH	CHEK2
TDG	ERCC6	MSH6	FANCL	EME1	XRCC4			MDC1
TDP1	POLE	PMS1	FANCM	GEN1	XRCC5			RNMT
UNG	POLE3	PMS2	UBE2T	MRE11A	XRCC6			TOPBP1
	XPA			MUS81				TREX1
	XPC			NBN				
				PALB2				
				RAD50				
				RAD51				
				RAD52				
				RBBP8				
				SHFM1				
				SLX1A				
				TOP3A				
				TP53BP1				
				XRCC2				
				XRCC3				

Table S3 Clinically relevant fusions identified in LUAD

Fusion	No of Fusions	Frequency (%)
EML4&ALK	4	0.79
FGFR2&ATE1	1	0.20
WHSC1L1&FGFR1	1	0.20
CAPZA2&MET	1	0.20
TRIM24&NTRK2	1	0.20
TRIM33&RET	1	0.20
CLTC&ROS1	1	0.20
ROS1&FBXO9	1	0.20
CCDC6&RET	1	0.20
KRAS&SLC2A14	1	0.20
WHSC1L1&FGFR1	1	0.20

Table S4 Clinically relevant fusions identified in LUSC

Fusion	No of Fusions	Frequency (%)
FGFR2&CCAR2	1	0.20
FGFR3&TACC3	2	0.40
SMAD4&NRG1	1	0.20
THAP7&NRG1	1	0.20
WHSC1L1&NUTM1	1	0.20

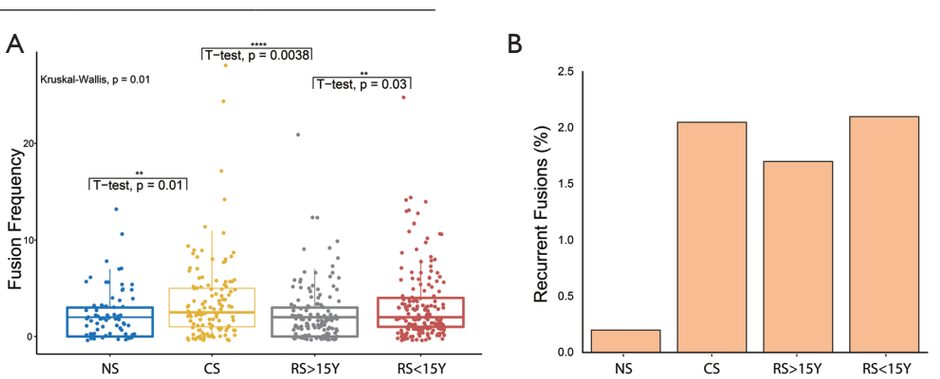

Figure S1 (A) Fusion frequencies identified across different LUAD smoking groups. (B) Percentage of recurrent fusions identified across different LUAD groups. CS, Current Smokers; NS, Nonsmokers, RF>15Ys, Reformed smokers for more than 15Ys; RF<15Ys, Reformed smokers for less than 15Ys.

https://dx.doi.org/10.21037/tlcr-22-113

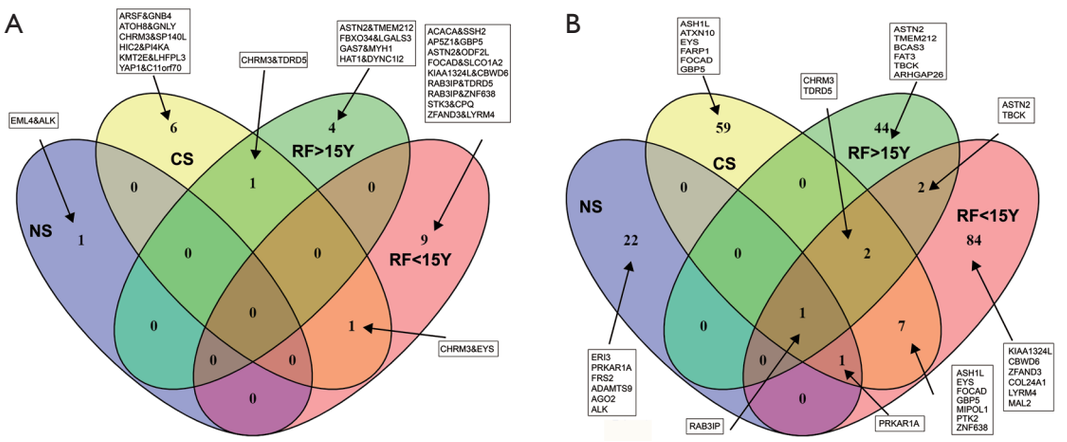

Figure S2 Unique and shared fusions across different LUAD groups. (A) Shared genes participating in fusions. (B) Shared gene fusion pairs. CS, Current Smokers; NS, Nonsmokers; RF>15Ys, Reformed smokers for more than 15Ys; RF<15Ys, Reformed smokers for less than 15Ys.

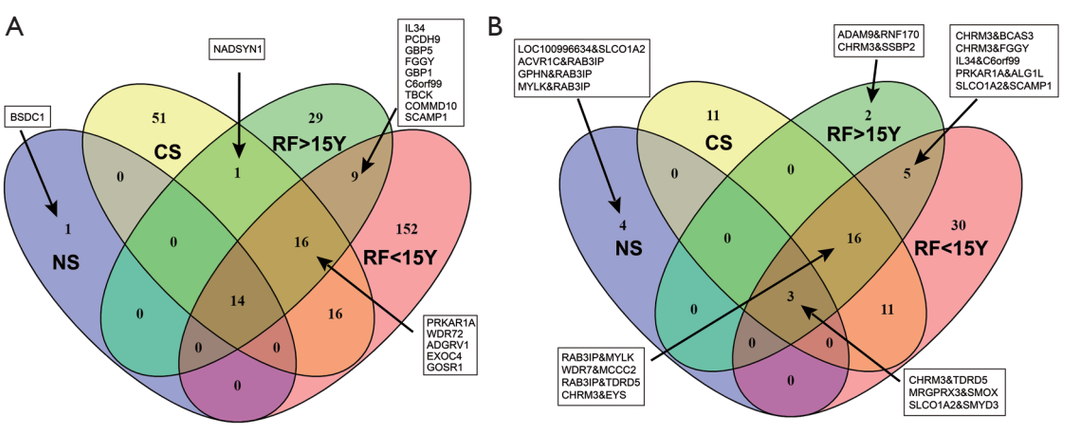

Figure S3 Unique and shared fusions across different LUSC groups. (A) Shared genes participating in fusions. (B) Shared gene fusion pairs. CS, Current Smokers; NS, Nonsmokers; RF>15Ys, Reformed smokers for more than 15Ys; RF<15Ys, Reformed smokers for less than 15Ys.

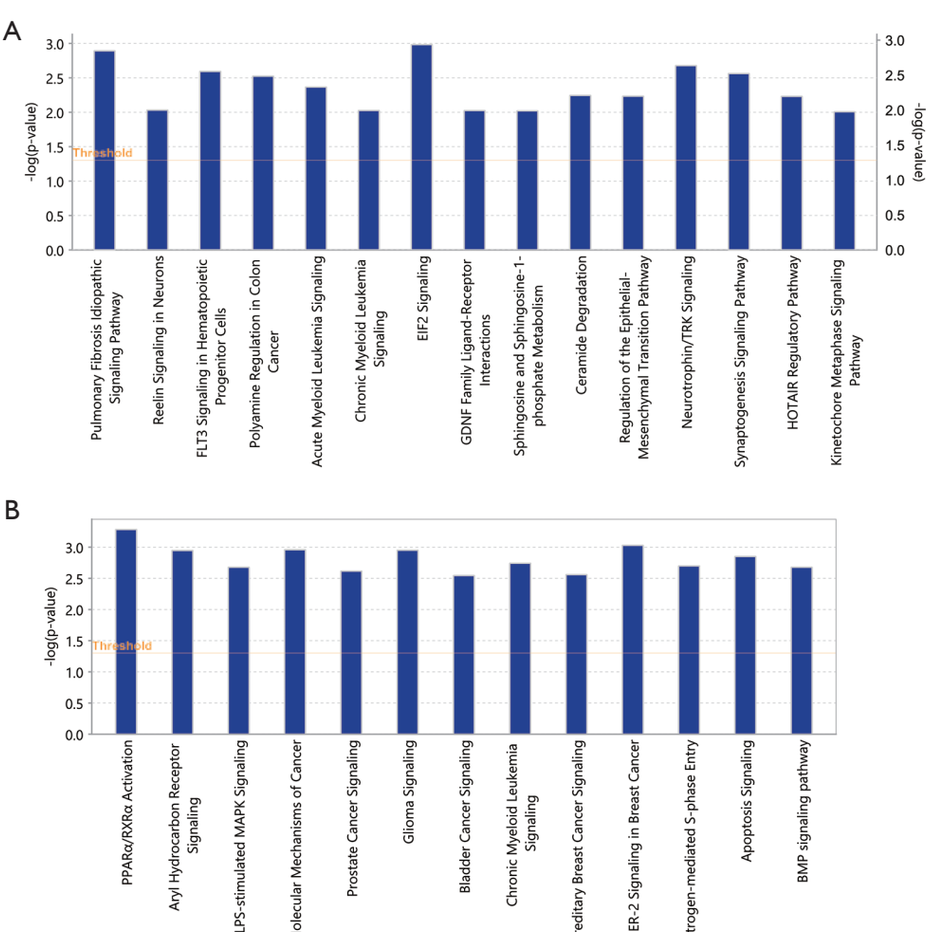

Figure S4 IPA enrichment pathways identified in LUAD (A) and LUSC (B).

https://dx.doi.org/10.21037/tlcr-22-113

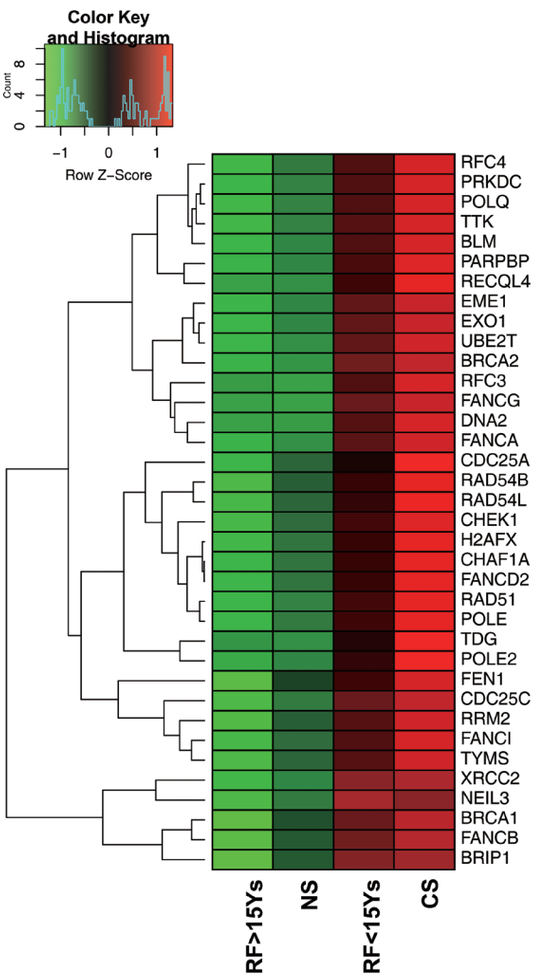

Figure S5 Genes in the DDR pathway differentially expressed across the different groups in LUAD. CS, Current Smokers; NS, Nonsmokers; RF>15Ys, Reformed smokers for more than 15Ys; RF<15Ys, Reformed smokers for less than 15Ys.

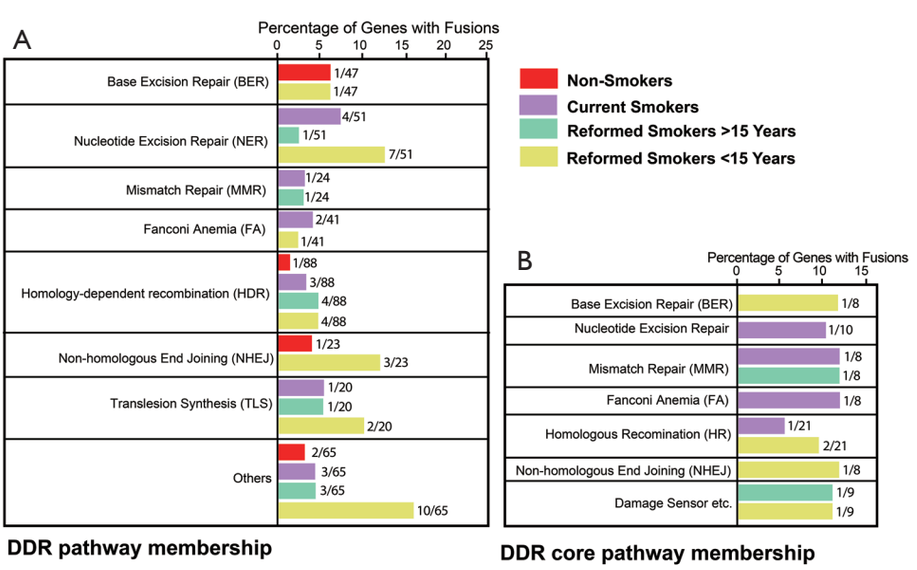

Figure S6 Percentage of genes in the DDR pathway that participated in fusions among different smoking groups in LUAD.

https://dx.doi.org/10.21037/tlcr-22-113

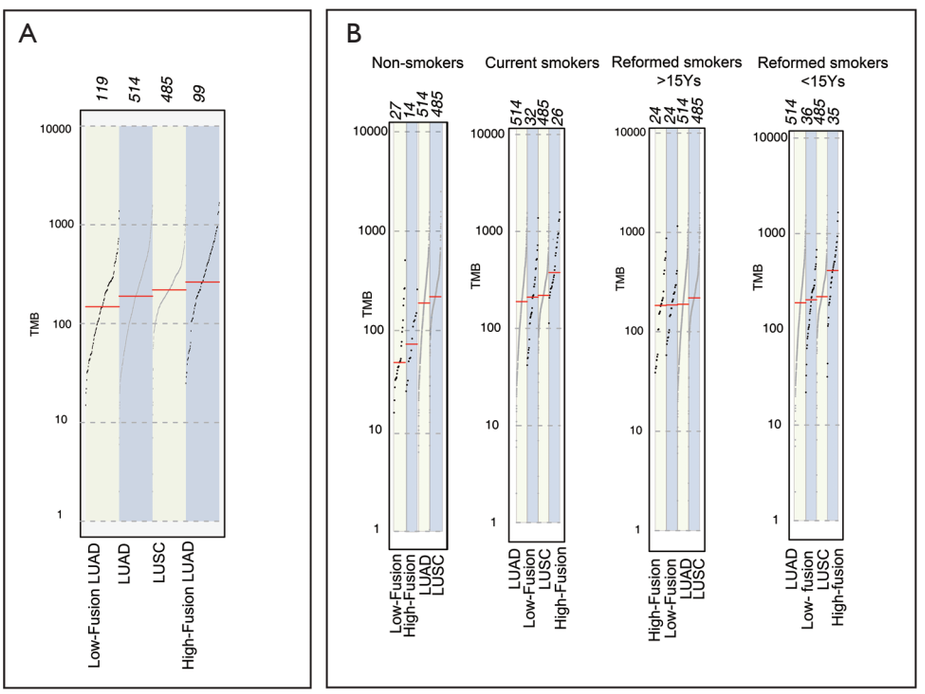

Figure S7 Tumor mutation burden across LUAD and LUSC (A). Tumor mutation burden across different smoking groups in LUAD is shown in B.

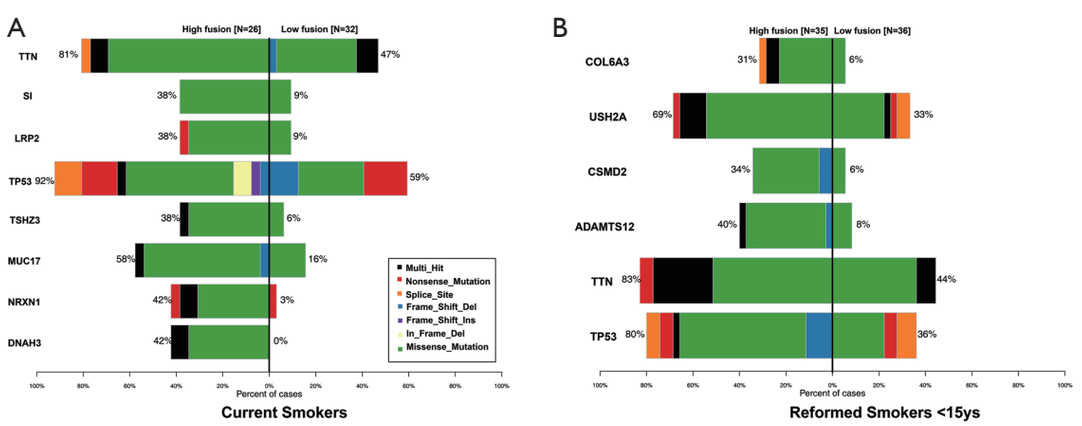

Figure S8 Comparison of mutations across LUAD samples with high or low fusions in Current Smokers (A) and Reformed Smokers <15ys. The percentage of mutated samples in each group is represented along with the type of mutations.

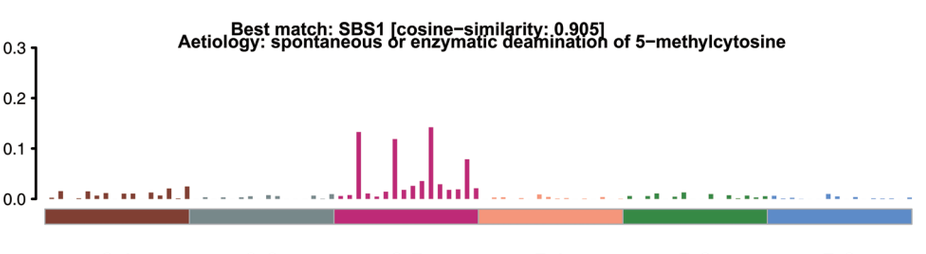

Figure S9 Mutational signatures identified in LUAD samples with high fusions.

https://dx.doi.org/10.21037/tlcr-22-113