BRIEF COMMUNICATION

National Science Review
8: nwab152, 2021
https://doi.org/10.1093/nsr/nwab152
Advance access publication 16 August 2021

MOLECULAR BIOLOGY & GENETICS

Genomic evidence of lung carcinogenesis associated with coal smoke in Xuanwei area, China

Honglei Zhang[2,*,†], Chao Liu[3,†], Li Li[4], Xu Feng[2], Qing Wang[5], Jihua Li[6], Shaobin Xu[7], Shuting Wang[8], Qianlu Yang[1], Zhenghai Shen[1], Jinhua Su[9], Xiaosan Su[2], Ruifen Sun[2], Xuhong Zhou[2], Junliang Wang[2], Yongchun Zhou[1], Baowei Jiao[10,11,12], Wanbao Ding[13], Xianbao Cao[14], Yue Wang[15], Yunchao Huang[1,*] and Lianhua Ye[1,*]

Xuanwei area, in southwestern China, harbors the highest female lung cancer rate in the country (Supplementary Table 1) [1]. Epidemiological studies have shown that lung cancer incidence in Xuanwei area is associated with the use of different types of local coal for household cooking and heating [2]. The genomic landscape of Xuanwei female adenocarcinoma (XWFA), the most distinctive feature of lung cancer in Xuanwei, has yet to be elucidated systematically. Here, we provide the fundamental resource of genomic datasets for further exploration of XWFA molecular mechanisms and development of targetable therapy.

The patients recruited for this study were 117 non-smoker females with untreated primary lung adenocarcinoma (LUAD) from the Xuanwei area, who were receiving surgical treatment at Yunnan Cancer Hospital (Supplementary Table 2). Samples were taken for whole-exome sequencing (WES) (112 pairs of tumor-normal), and 33 normal and 115 tumor samples were sequenced with mRNA-Seq technology. Datasets of 168 TCGA-LUAD female smokers (TLSF) and 102 TCGA-LUAD female non-smokers (TLNF) were adopted from The Cancer Genome Atlas (TCGA) program for genomic comparison between lung cancer associated with cigarettes and that associated with smoky coal [3] (Supplementary Table 3). There were no significant differences in the distribution of samples from different pathologic stages among the XWFA, TLSF and TLNF cohorts (Fig. 1a).

In the XWFA cohort, 35 729 somatic mutations comprising 34 287 single-nucleotide variants (SNVs) and 1442 insertions or deletions (INDELs) were identified (Supplementary Methods). Compared with the TCGA lung cancer samples, XWFA samples possessed higher mutation burdens (median = 2.11) than lung squamous cell carcinomas (LUSC) (median = 1.63) and LUAD (median = 1.41) (Fig. 1b). This suggests substantial differences in the mutational genomic landscape between the XWFA and Western cohorts. Furthermore, MutSig2CV (for details of software, refer to the Supplementary Methods) and oncodriveCLUST algorithms were adopted jointly to identify significantly mutated genes (SMGs) in the XWFA cohort. As demonstrated in Fig. 1c, the most prominent cancer-related variations observed in the XWFA cohort were *EGFR* mutations (found in 52.68% of tumors), followed by mutations in *TP53* (41.07%), *RBM10* (10.71%), *KRAS* (7.14%) and *NKX2-1* (4.46%). Four genes including *EGFR*, *KRAS*, *TPRN* and *SPTLC1*, were identified as driver genes using the oncodriveCLUST algorithm. As SMGs often serve as gatekeepers, which may be targetable or serve as predictive biomarkers for immune checkpoint therapy [4], the relationships among SMGs, mutation load and neoantigens derived from mutations were next examined (Supplementary Fig. 1). The mutation load and the number of neoantigens were significantly correlated (Supplementary Fig. 1a). Higher mutation loads and more neoantigens were observed in *TP53*, *KRAS*, *RBM10* and *POTEC* mutant samples than in the respective WT samples (Supplementary Fig. 1d–k). Taken together, the canonical and novel SMGs identified in the XWFA cohort, which were suspected to be targetable or explored as biomarkers, merit further investigation. LUAD-specific driver gene lists were collected and are listed in Supplementary Table 4.

Mutational signatures were further investigated to infer the mutational process during XWFA initiation. Mutation spectrum results showed that the substitution pattern and transversion/transition ratio of the XWFA cohort were similar to those of the TLSF cohort and both showed high C > A substitution (Supplementary Fig. 2). Three mutation signatures including ‘smoking’ (COSMIC_4), ‘APOBEC’ (COSMIC_2) and ‘Aging’ (COSMIC_1) were identified using the nonnegative matrix factorization (NMF) algorithm in the XWFA cohort (Fig. 1d). Polycyclic Aromatic Hydrocarbons (PAHs) and

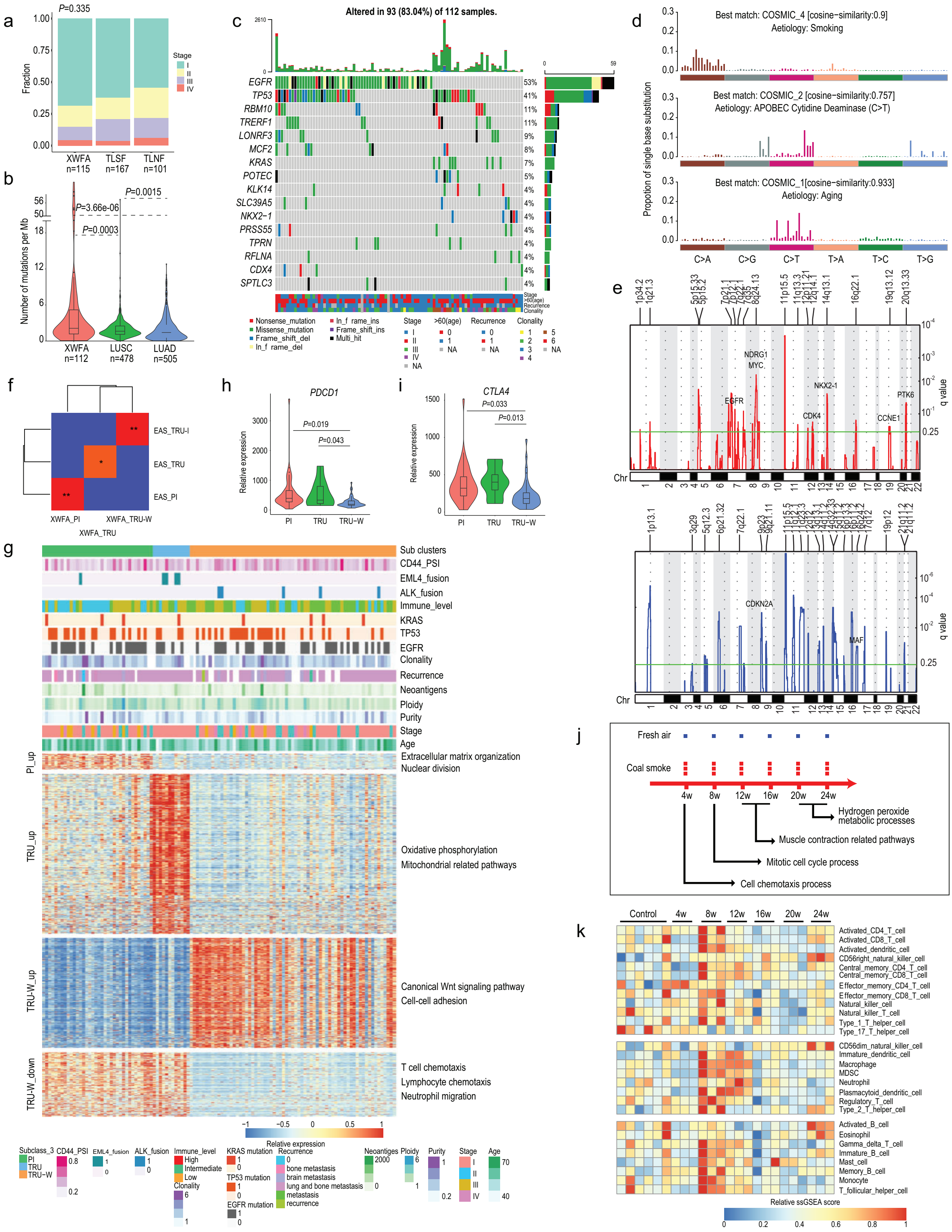
a
P=0.335
1.00
0.75
0.50
0.25
0.00
Fraction
Stage
I
II
III
IV
XWFA n=115
TLSF n=167
TLNF n=101
b
Number of mutations per Mb
P=3.66e-06
P=0.0015
P=0.0003
XWFA n=112
LUSC n=478
LUAD n=505
c
Altered in 93 (83.04%) of 112 samples.
2610
EGFR 53%
TP53 41%
RBM10 11%
TRERF1 11%
LONRF3 9%
MCF2 8%
KRAS 7%
POTEC 5%
KLK14 4%
SLC39A5 4%
NKX2−1 4%
PRSS55 4%
TPRN 4%
RFLNA 4%
CDX4 4%
SPTLC3 4%
Stage
>60(age)
Recurrence
Clonality
Nonsense_mutation
Missense_mutation
Frame_shift_del
In_f rame_del
In_f rame_ins
Frame_shift_ins
Multi_hit
d
Best match: COSMIC_4 [cosine-similarity:0.9]
Aetiology: Smoking
Best match: COSMIC_2 [cosine-similarity:0.757]
Aetiology: APOBEC Cytidine Deaminase (C>T)
Best match: COSMIC_1[cosine-similarity:0.933]
Aetiology: Aging
Propotion of single base substitution
C>A
C>G
C>T
T>A
T>C
T>G
e
NDRG1
MYC
EGFR
NKX2-1
CDK4
PTK6
CCNE1
CDKN2A
MAF
q value
0.25
Chr
f
EAS_TRU-I
EAS_TRU
EAS_PI
XWFA_PI
XWFA_TRU-W
XWFA_TRU
g
Sub clusters
CD44_PSI
EML4_fusion
ALK_fusion
Immune_level
KRAS
TP53
EGFR
Clonality
Recurrence
Neoantigens
Ploidy
Purity
Stage
Age
PI_up
TRU_up
TRU-W_up
TRU-W_down
Extracellular matrix organization
Nuclear division
Oxidative phosphorylation
Mitochondrial related pathways
Canonical Wnt signaling pathway
Cell-cell adhesion
T cell chemotaxis
Lymphocyte chemotaxis
Neutrophil migration
Relative expression
Subclass_3
PI
TRU
TRU−W
High
Intermediate
Low
bone metastasis
brain metastasis
lung and bone metastasis
metastasis
recurrence
h
PDCD1
P=0.019
P=0.043
Relative expression
PI
TRU
TRU−W
i
CTLA4
P=0.033
P=0.013
j
Fresh air
Coal smoke
4w
8w
12w
16w
20w
24w
Hydrogen peroxide metabolic processes
Muscle contraction related pathways
Mitotic cell cycle process
Cell chemotaxis process
k
Control
Activated_CD4_T_cell
Activated_CD8_T_cell
Activated_dendritic_cell
CD56right_natural_killer_cell
Central_memory_CD4_T_cell
Central_memory_CD8_T_cell
Effector_memory_CD4_T_cell
Effector_memory_CD8_T_cell
Natural_killer_cell
Natural_killer_T_cell
Type_1_T_helper_cell
Type_17_T_helper_cell
CD56dim_natural_killer_cell
Immature_dendritic_cell
Macrophage
MDSC
Neutrophil
Plasmacytoid_dendritic_cell
Regulatory_T_cell
Type_2_T_helper_cell
Activated_B_cell
Eosinophil
Gamma_delta_T_cell
Immature_B_cell
Mast_cell
Memory_B_cell
Monocyte
T_follicular_helper_cell
Relative ssGSEA score

Figure 1. Molecular characterization of XWFA and rat model recapitulation of lung cancer initiation. (a) Comparison of distributions of samples from different pathology stages (I, II, III and IV) across the XWFA, TLSF and TLNF cohorts. Chi-square test was used to calculate *P* value. (b) Violin plot showing the tumor mutation burden (TMB) across the XWFA, TCGA-LUAD and TCGA-LUSC cohorts. The box indicates the interquartile range (IQR), the middle line indicates median, whiskers indicate the highest and lowest values within 1.5 × IQR away from the box, and dots plot values $>1.5 \times$ IQR away from the box. (c) Co-mutation plot of tumor samples from the XWFA cohort. Significantly mutated genes were identified with MutSig2CV algorithm (FDR corrected $P < 0.25$) and oncodriveCLUST (FDR corrected $P < 0.1$) and were ranked in order of decreasing prevalence. Clinical features such as pathology stages, ages, recurrence status and clonality were indicated. (d) Top three mutation signatures derived from single nucleotide variants were identified with Cosmic Mutational Signatures (version 2.0). (e) Focal-level CNV across chromosomes 1–22 in XWFA cohort, with GISTIC FDR q values on the y axis. Amplifications were labeled red (top) and deletions were labeled blue (bottom). Selected genes in the Cancer Gene Census (CGC) were labeled in the significant peak regions. (f) Association analysis of subgroups identified in the XWFA and EAS cohorts with Submap. Significant correspondence between subgroups highlighted in red with Bonferroni adjusted *P* values. *: FDR < 0.1; **: FDR < 0.01. (g) Phenotypes of RNA-based subgroups in XWFA. The annotation rows showing the genomic and clinicopathologic features for each patient. The heat map showing the normalized mean expression of subgroup differentially expressed genes. Representative GO biological processes were labeled right. (h and i) Comparison of *PDCD1* (h) and *CTLA4* (i) expression among PI, TRU and TRU-W clusters in XWFA cohort. (j) Schematic showing the overall experimental design for the rat_coal model (top). Red dot: rats treated with smoke from local smoky coal; blue dots: rats treated with fresh air. The arrow below the timeline indicates the most enriched disturbed biological processes during smoke treatment. (k) Heatmap showing the relative infiltration of immune cells derived from RNA-based ssGSEA scores in lung tissue across six time points. Two-sided Mann-Whitney U test was used for *P* value calculation in (b, h and i).

nicotine-derived nitrosamines, two smoking carcinogens reported to be strongly associated with C > A transversion hotspots [5], were also found in high levels in local smoky coal from Xuanwei [2]. This explained the similarities in mutation spectra and mutation signatures between the XWFA and TLSF cohorts, which further supports the hypothesis that the high lung cancer rate in Xuanwei area results, at least partially, from domestic use of local smoky coal.

Copy number variations (CNVs) play pivotal roles in tumor initiation. We identified significantly altered CNVs with Sequenza and the Genome Identification of Significant Targets in Cancer (GISTIC) 2.0 algorithm in the XWFA cohort (Fig. 1e, Supplementary Fig. 3 and Supplementary Table 5). Generally, numbers of CNV-affected genes (both amplification and deletion) in the XWFA cohort were higher than in the TLNF and TLSF cohorts (Supplementary Fig. 3e and 3f). However, many focal amplification CNVs around driver genes such as *MYC*, *PVT1* and EGFR, and deletion CNVs such as *CDKN2A* and *CDKN2B*, were identified in all three cohorts (Fig. 1e, Supplementary Fig. 3a–f), which suggests pivotal roles for those genes in initiation of lung cancer. Both amplification and deletion genes detected by ABSOLUTE and Sequenza in the XWFA cohort were comparable, with only a small proportion of genes identified software-specifically (Supplementary Fig. 3g), indicating stable detection of the CNVs. Overall, our results suggest that the XWFA cohort had more genomic CNVs than the TLSF and TLNF cohorts, which further suggests substantial differences in the genomic landscape between XWFA and Western cohorts and that the significantly CNV-affected genes need further investigation.

Unsupervised clustering of RNA-seq data from XWFA tumor samples revealed three subgroups and the SubMap module was applied to compare the subgroups between the XWFA (N = 115) and ESA [6] (N = 230) cohorts. Although we found subgroups from the XWFA cohort that were significantly correlated to the PI, TRU and TRU-I subgroups from the ESA cohort (Fig. 1f), we also found a subgroup highly expressing Wnt signaling pathway genes and designated this the TRU-W subgroup, corresponding to the TRU-I subgroup in the ESA cohort. To further explore the TRU-W subgroup, we identified its up- and down-regulated genes and found that low immune infiltration was the most remarkable feature (Fig. 1g). This result indicates that WNT/β-catenin pathway activation correlates with immune exclusion across human cancers. Furthermore, we found that expression of *PDCD1* and *CTLA4* were significantly lower in the TRU-W subgroup compared with PI and TRU samples (Fig. 1h and 1i). All these results suggest that the TRU-W subgroup from the XWFA cohort formed a specific cluster with low immune infiltration and high Wnt signaling, which should be considered in further immunotherapy. Clinical and molecular features (including immune cell infiltration status, Supplementary Fig. 8 and Supplementary Methods) among the subgroups in the XWFA cohort were further compared. The TRU-W subgroup was enriched with EML4 fusion events and low and intermediate immune infiltration samples (Supplementary Fig. 5k and 5q); other features showed no significant differences (Supplementary Fig. 5a–j and m–p).

To explore experimentally the role of emissions from indoor combustion of C1 bituminous coal in the initiation stage of XWFA, a lung cancer model (rat_coal model) derived from female F344 rat was established (Supplementary Methods, Supplementary Fig. 6a, Fig. 1j and Supplementary Fig. 7). We firstly investigated the biological process alterations in both rat_coal and mouse_cigarette models [7]. Our results showed clear step-wise alterations of biological processes in both models during smoke treatment. Cell chemotaxis processes, mitotic cell cycle process, muscle contraction-related pathways and hydrogen peroxide metabolic processes were mostly altered after 4 weeks, 8 weeks, 12–16 weeks, 20–24 weeks treatment in the rat_coal model (Fig. 1j and Supplementary Fig. 8). Parallelly, immune response, regulation of cell cycle, immune response, smooth muscle construction and oxidation-related process after 1 week, 1 month, 3 months, 6 months and 9 months treatment were identified in the mouse_cigarette model (Supplementary Fig. 6b and Supplementary Fig. 9). These parallel step-wise alterations of biological process from both the rat_coal model and mouse_cigarette models reflected progressive tumor initiation starting from inflammation.

We further investigated the trend of tumor-infiltrating lymphocytes (TILs) during tumor initiation with a single-sample gene set enrichment analysis (ssGSEA) method based on RNA-Seq profiling. The remarkable feature was the wave of TIL profiles in both models. Specifically, TILs rose notably after 8 weeks treatment, decreased gradually to 20 weeks and increased again at 24 weeks in the rat_coal model (Fig. 1k). A parallel trend was observed in the mouse_cigarette model (Supplementary Fig. 6c). This trend of TILs was correlated with alteration of biological pathways in lung tissues. Specifically, activated CD4/CD8 T cells, CD56bright/dim natural killer cells and activated B cells were more enriched at 24 weeks, which was accompanied by up-regulation of hydrogen peroxide metabolic processes. It has been proved that hydrogen peroxide-induced oxidative stress further triggered an innate immune response in A549 cells [8]. These results reflect the cross-talk between tissue cells and TILs during tumor initiation. Further exploration of expression of immune-related genes revealed several potential therapeutic targets in lung cancer initiation (Supplementary Fig. 6d). For example, B and T lymphocyte attenuator (BTLA), belonging to the CD28 superfamily and similar to programmed cell death-1 (PD-1) and cytotoxic T lymphocyte associated antigen-4 (CTLA-4) in terms of its structure and function, induces immunosuppression by inhibiting B and T cell activation and proliferation [9]. Another promising target is KDR (VEGFR-2), which is the main mediator of VEGF-induced tumor cell proliferation, migration, survival and increased permeability [10]. All the above mouse homologous genes showed remarkable up-regulation after 24 weeks in the rat_coal model (Supplementary Fig. 6d), indicating a suppressive state of both adaptive immune and innate immunity by these suppressors, which also served as promising therapeutic targets during the initiation stage of lung cancer.

Taking all the results together, our study establishes a valuable resource for XWFA, provides insight into the initiation process and indicates that therapies targeting the SMGs, early-stage pathway alterations or blocking immune-cancer cross-talk, show potential and merit further investigation.

DATA AVAILABILITY

Clinical data (deidentified) are listed in Supplementary Table 2. Raw sequencing data including WES and RNA-seq datasets from human samples have been deposited in the Genome Sequence Archive under the accession code HRA000124. The RNA-seq dataset from rat model samples has been deposited in the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) under the accession code GSE162001 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162001]. RNA-seq dataset for mouse˙cigarette model analyses were downloaded from GEO accession: GSE76205 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76205]. RNA-seq and somatic mutation datasets of TCGA-LUAD and TCGA-LUSC were downloaded from XENA (https://xena.ucsc.edu/). All other relevant data are available within the article, Supplementary data or available from the authors upon request.

SUPPLEMENTARY DATA

Supplementary data are available at *NSR* online.

ACKNOWLEDGEMENTS

We would like to thank Kaiyang Ma, Li yang, Rufu Dong from the Institute of Information, Yunnan University of Chinese Medicine for the network assistance. We thank Prof. Wenming Zhao and Prof. Zhang from the Beijing Institute of Genomics, Chinese Academy of Sciences for data deposition assistance. We also thank Prof. Bing Su from Kunming Institute of Zoology for the professional manuscript review.

FUNDING

This research was funded by the National Natural Science Foundation of China (31801249, 81860325, 81860513, 81760519 and 81760818, 82060519, 81960322), Scientific Research Fund Project of Yunnan Education Department (2019J1313, 2017zzx197 and 2019J1061), The National Key Research and Development Program of China (2017YFC0907902 and 2016YFE0103400-01), High-level Health Technology Personnel Training Program of Yunnan Province (L-2017006), Joint Program of Applied Basic Research of Yunnan Provincial Department of Science and Technology-Kunming Medical University (2019FE001 (-066), 2018FE001 (-247), 2018FE001(-247)), Health research project of Kunming Health Committee (2019-03-10-006), Reserve talent project of young and middle-aged academic and technical leaders in Yunnan province (202005AC160048), Medical edible flowers technology innovation team of universities in Yunnan province (2020YGC01), Applied Basic Research Project of Yunnan (202001AU070096) and Medical Reserve Personnel Training Program of Yunnan Provincial Health Commission (H-2018097).

AUTHOR CONTRIBUTIONS

Honglei Zhang designed the study, analyzed the data, prepared tables/figures and wrote the manuscript. Lianhua Ye and Yunchao Huang provided funding and samples for the project and helped design the study. Chao Liu collected samples and were involved in the experiment design with the assistance of Li Li. Xu Feng conducted raw data processes of WES and RNA-Seq datasets. Qianlu Yang carried out establishment of the rat_coal model. Qing Wang, Shuting Wang, Zhenghai Shen, Jinhua Su, Xiaosan Su, Ruifen Sun, Xuhong Zhou, Junliang Wang, Wanbao Ding and Xianbao Cao performed a medical record review and performed DNA and RNA extraction. Shaobin Xu provided help and facilities with the supercomputer operations. Yongchun Zhou and Baowei Jiao reviewed the manuscript and made suggestions.

Conflict of interest statement. None declared.

Honglei Zhang [2,*,†], Chao Liu[3,†], Li Li[4], Xu Feng[2], Qing Wang[5], Jihua Li[6], Shaobin Xu[7], Shuting Wang[8], Qianlu Yang[1], Zhenghai Shen[1], Jinhua Su[9], Xiaosan Su[2], Ruifen Sun[2], Xuhong Zhou[2], Junliang Wang[2], Yongchun Zhou[1], Baowei Jiao [10,11,12], Wanbao Ding[13], Xianbao Cao[14], Yue Wang[15], Yunchao Huang[1,*] and Lianhua Ye[1,*]

[1]Department of Thoracic Surgery I, Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, China; [2]Center for Scientific Research, Yunnan University of Chinese Medicine, China; [3]Department of Nuclear Medicine, Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, China; [4]Biotherapy Center, Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, China; [5]Department of Oncology, Qujing First People's Hospital, China; [6]Qujing Center for Disease Control and Prevention, China; [7]Supercomputing Platform of Kunming Institute of Zoology, Kunming Biological Diversity Center of Instruments, Kunming Institute of Zoology, Chinese Academy of Sciences, China; [8]Department of Thoracic Surgery II, Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, China; [9]Department of Thoracic Surgery, Xuanwei People's Hospital, China; [10]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, China; [11]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, China; [12]KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming Institute of Zoology, Chinese Academy of Sciences, China; [13]Department of Oncology, Yan'an Affiliated Hospital of Kunming Medical University, China; [14]Department of Otolaryngology, First People's Hospital of Yunnan Province, the Affiliated Hospital of Kunming University of Science and Technology, China and [15]The Affiliated Hospital of Guizhou Medical University, China

*__Corresponding authors.__ E-mails: Lhye1204@aliyun.com; huangych2001@aliyun.com; hlzhang2014@163.com.

[†]Equally contributed to this work.

REFERENCES

1. Li J, He J and Zhang Y *et al. Zhongguo Fei Ai Za Zhi* 2019; **22**: 477–87.
2. Lan Q, Chapman RS and Schreinemachers DM *et al. J Natl Cancer Inst* 2002; **94**: 826–35.
3. Goldman MJ, Craft B and Hastie M *et al. Nat Biotechnol* 2020; **38**: 675–8.
4. Ghorani E, Reading JL and Henry JY *et al. Nat Cancer* 2020; **1**: 546–61.
5. Hollstein M, Sidransky D and Vogelstein B *et al. Science* 1991; **253**: 49–53.
6. Chen J, Yang H and Teo ASM *et al. Nat Genet* 2020; **52**: 177–86.
7. Miller MA, Danhorn T and Cruickshank-Quinn CI *et al. PLoS One* 2017; **12**: e0178281.
8. Upadhyay S, Vaish S and Dhiman M. *Mol Cell Biochem* 2019; **450**: 135–47.
9. Ning Z, Liu K and Xiong H. *Front Immunol* 2021; **12**: 654960.
10. Holmes K, Roberts OL and Thomas AM *et al. Cell Signal* 2007; **19**: 2003–12.