# Unique spectrum of driver gene mutations in patients with non-small cell lung carcinoma from coal-manufacturing districts in Southwest China

Jun-Ling Wang[1], Chun-Ju Yang[2], Juan Hu[3], Hong-Xia Liu[1], Meng-Xian Li[1], Zhe-Wei Fang[1], Jin-Si Yang[1], Rong Ma[1], Rui Dai[3], Qiang Xie[4], Rui Li[5], Jia-Ling Lv[6], Qiang-Bo Kan[7], Yan-Hong Gao[8], Ying-Yu Yang[9], Kun-Hua He[10], Ce Ci[11], Chao Zhang[6], Hong-Wei Li[1]

[1]Clinical Laboratory, Qujing First People's Hospital, Qujing, China
[2]School of Basic Medical Sciences, Qujing Medical College, Qujing, China
[3]Department of Ophthalmology, Qujing Second People's Hospital, Qujing, China
[4]Department of Respiratory Medicine, Qujing First People's Hospital, Qujing, China
[5]Department of Medical Administration, Qujing First People's Hospital, Qujing, China
[6]Department of Oncology, Qujing First People's Hospital, Qujing, China
[7]Department of Thoracic Surgery, Qujing First People's Hospital, Qujing, China
[8]Department of Traditional Chinese Medicine, Qujing First People's Hospital, Qujing, China
[9]Department of Pathology, Qujing First People's Hospital, Qujing, China
[10]Department of Blood Transfusion, Qujing First People's Hospital, Qujing, China
[11]Bioinformatics Analysis Division, Beijing Life Healthcare Medical Laboratory, Beijing, China

**Corresponding authors:**
Hong-Wei Li
Clinical Laboratory
Qujing First People's Hospital
Qujing, 655000, China
Phone: +86 0874 3311075
E-mail: lihongwei@kmmu.edu.cn

Chao Zhang
Department of Oncology
Qujing First People's Hospital
Qujing, 655000, China
Phone: +86 0874 3311075
E-mail: chesanjin@163.com

**Submitted:** 6 September 2022; **Accepted:** 9 December 2022
**Online publication:** 24 December 2022

Arch Med Sci
DOI: https://doi.org/10.5114/aoms/157288

## Abstract

**Introduction:** The coal-manufacturing districts in the Eastern Yunnan province of Southwest China have the highest rates of occurrence and death from lung tumors. Unique clinical characteristics of non-small cell lung cancer (NSCLC) in patients from these regions were previously reported without a clear understanding of the etiology and molecular characteristics. We aim to identify the unique driver gene mutation spectrum.
**Material and methods:** Samples from 1120 NSCLC patients from Eastern Yunnan were gathered for next-generation sequencing. Seventeen gene targets were sequenced. We compared individuals' medical and genetic features from the coal- and non-coal-manufacturing zones.
**Results:** The mutation rates of *EGFR* (L858R, 19-Del, G719X+L861X, L858R+*EGFR* amplification) and *ERBB2* (20ins) were low in patients from coal-manufacturing regions. Interestingly, *EGFR* (G719X, S768I, G719X+S768I), *KRAS* (G12C), *TP53* (R158L), and *NTRK3* demonstrated a much higher mutation frequency. Furthermore, *EGFR* compound mutations were linked with the patient's job and TNM staging IIIb-IV. The OncodriverCLUST algorithm authenticated 6 genes (*KRAS*, *EGFR*, *ROS1*, *NRAS*, *BRAF*, and *ERBB2*) as driver genes in patients from coal-manufacturing regions. *EGFR* with *KRAS*, *BRAF*, *RET*, and *TP53* with *ALK* and *KRAS* were mutually exclusive. Mutations in the TP53 signaling pathways were the most common in NSCLC patients from the coal-producing districts.
**Conclusions:** Our analyses confirmed the unique spectrum of driver genetic mutations and emphasized the potential of future targeted therapy in NSCLC patients from the coal-manufacturing districts of Eastern Yunnan. Our data broaden the view of NSCLC pathogenesis and its relationship with the environmental conditions in coal-producing regions.

**Key words:** non-small-cell lung cancer, *EGFR*/*KRAS*/*TP53*/*ERBB2*/*NTRK3*, mutation spectrum, coal-manufacturing districts, Eastern Yunnan.

## Introduction

Lung carcinoma is a significant disease that critically endangers people's life and well-being [1]. Its mortality rate ranks first in China [2]. Non-small cell lung (NSCLC) and adenocarcinomas are among the most common types of these tumors, accounting for more than 80% [3]. About 75% of NSCLC patients are diagnosed in the medium and advanced phases of the disease, resulting in high death rates often exceeding 95% for 5 years [4].

People in rural Chinese areas, such as Xuanwei and Fuyuan counties in Qujing city of the East Yunnan Province of Southwest China practically do not smoke. Still, the coal-manufacturing zones are the most severely affected by lung cancer worldwide [5]. Its occurrence is 4–5 times greater than the average Chinese rate, with the incidence of death of about 91.3/per 100,000 estimated for the coal-producing areas of Eastern Yunnan [5]. Local rural residents burn coal for needs over many years [6]. This smoke is the main culprit of inside air contamination [7], in which different chemical substances emitted by coal combustion are the principal malefactors of frequent lung carcinomas in these regions. Data show that a 36- and 99-fold increase in male and female deaths in these regions is associated with coal use [4]. These associations are related to the type of coal produced in the coal-manufacturing districts. For example, coal manufacturing in Yunnan is associated with the production of bituminous coal, especially in the Late Permian C1 coal seam [8], characterized by solid mutagenicity and a high concentration of nano-quartz formed in the Late Permian [9]. Recent evidence also suggested that the abundant nano-quartz and Fe-rich aluminosilicates of interstratified berthierine/chamosite minerals in the C1 coal seam were responsible for activating inflammatory reactions during carcinogenesis [10]. Currently, there is a lack of research to stratify individuals with different lung tumors according to the types of produced coals, which allowed the large-scale whole-exome sequencing of lung cancer biological samples of patients from the eastern parts of Yunnan coal-manufacturing districts.

Previous research on driver genes in NSCLC patients from Xuanwei county showed that the epidermal growth factor receptor (*EGFR*) mutation rate was lower for individuals from this region compared to other parts of China [11]. However, the mutation rates in genes such as *KRAS* [12] and the tumor protein p53 (*TP53*) [13] were higher. Furthermore, the whole-exome sequencing (WES) of lung cancer and distal normal tissues from 112 Xuanwei patients with early lung adenocarcinoma found mutations in *EGFR*, *KRAS*, *TPRN*, and *SPTLC1* genes were identified as driver mutations for individuals with lung cancer from the region of Xuanwei [14]. Other authors detected 10 driver genetic mutations in 526 NSCLC lung cancer patients from Qujing city. They found that Qujing NSCLC patients had a unique driver gene mutation spectrum, in which the *KRAS* gene mutation frequency was higher while the *EGFR* was lower [15]. Moreover, the ratios of *EGFR* G719X + S768I, *EGFR* G719X + L861Q compound double mutation, *KRAS* G12C, and *KRAS* G12D of all molecular mutant subtypes were significantly higher than those of non-Qujing patients. However, ALK and ROS1 fusion gene mutation rates were lower than in non-Qujing patients. The remaining 6 genes (*BRAF*, *RET*, *MET*, *ERBB2/HER2*, *NRAS*, and *PIK3CA*) were unchanged [15]. These data highlighted that afatinib target therapy elongated remission time and amended the health span for most Qujing NSCLC patients [4]. In addition, among some of the oncogenic signaling pathways, receptor tyrosine kinase-rat sarcoma protein (RTK-RAS) [16], phosphoinositide 3-kinase (PI3K) [17], wingless-type mouse mammary tumor virus integration site family (Wnt) [18], tumor protein p53 (TP53) [19], transforming growth factor-β (TGF-β) [20], cell cycle [21], myelocytomatosis viral oncogene homolog (MYC) [22], protein kinase Hippo (Hippo) [23] and the proto-oncogene (*Notch*) pathway [24] appear prominent in NSCLC. Notably, some data link lung carcinoma characterized by *KRAS* and *TP53* mutations in non-smokers with the exposure of these individuals to polycyclic aromatic hydrocarbon (PAH)-rich coal combustion emissions [25]. Chen *et al*. found that Xuanwei lung cancer was mainly linked with perturbations in the PI3K/protein kinase B (Akt), Wnt, and mitogen-activated protein kinase (MAPK) pathways [26]. Further results that demonstrated the success of lung cancer therapy with *KRAS* inhibitors, such as adagrasib [27] and sotorasib [28], proved the necessity of targeted anticancer approaches that complement the genetic population background. Therefore, the search for unique driver mutations and signaling pathways in NSCLC patients from coal-manufacturing districts of East Yunnan is becoming extremely important.

In the current research, we collected tumor tissue or plasma samples from 1120 NSCLC patients. Next-generation sequencing (NGS) technology was used for comparative analysis of the exome mutation profiles of 17 genes (Table I) in individuals diagnosed with NSCLC in coal- and non-coal-manufacturing districts in Eastern Yunnan. We further investigated the correlation between patients' clinical characteristics and the types of the detected unique driver mutations' frequency and subtypes, which allowed us to map a unique spectrum of driver mutations in NSCLC

**Table I.** NCBI ID, symbol, gene name, UniProt KB and PubMed PMID for the 17 genes

| NCBI Gene ID | Symbol | Gene name | UniProt KB AC | PubMed PMID |
|---|---|---|---|---|
| 1956 | EGFR | Epidermal growth factor receptor | P00533 | 20887192 |
| 238 | ALK | ALK receptor tyrosine kinase | Q9UM73 | 17625570 |
| 6098 | ROS1 | ROS proto-oncogene 1, receptor tyrosine kinase | P08922 | 35200557 |
| 3845 | KRAS | KRAS proto-oncogene, GTPase | P01116 | 34089836 |
| 673 | BRAF | B-Raf proto-oncogene, serine/threonine kinase | P15056 | 29729495 |
| 5979 | RET | Ret proto-oncogene | P07949 | 34503226 |
| 4233 | MET | MET proto-oncogene, receptor tyrosine kinase | P08581 | 32615820 |
| 2064 | ERBB2 | Erb-b2 receptor tyrosine kinase 2 | P04626 | 15457249 |
| 4893 | NRAS | NRAS proto-oncogene, GTPase | P01111 | 32979462 |
| 5290 | PIK3CA | Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha | P42336 | 32908885 |
| 4914 | NTRK1 | Neurotrophic receptor tyrosine kinase 1 | P04629 | 26565381 |
| 7157 | TP53 | Tumor protein p53 | P04637 | 34088750 |
| 2260 | FGFR1 | Fibroblast growth factor receptor 1 | P11362 | 33984662 |
| 3815 | KIT | KIT proto-oncogene, receptor tyrosine kinase | P10721 | 34107476 |
| 5156 | PDGFRA | Platelet-derived growth factor receptor alpha | P16234 | 30867736 |
| 4915 | NTRK2 | Neurotrophic receptor tyrosine kinase 2 | Q16620 | 32540558 |
| 4916 | NTRK3 | Neurotrophic receptor tyrosine kinase 3 | Q16288 | 30215037 |

individuals from the coal-producing areas of Eastern Yunnan.

## Material and methods

### Patients' medical records

One thousand one hundred twenty patients diagnosed with NSCLC who were above 18 years old, living in Qujing City, Yunnan Province, and its border area with Western Guizhou, were recruited for the retrospective study. Inclusion criteria were: (1) the patients visited Qujing First People's Hospital and Xuanwei County People's Hospital between September 2016 and March 2022, (2) histologically or cytologically confirmed with NSCLC, (3) adults (> 18 years) who were dwelling in Qujing city of Eastern Yunnan province, 9 counties (Qilin, Fuyuan, Xuanwei, Huize, Zhanyi, Malong, Luliang, Shizong, Luoping), and its border area with Western Guizhou province, 2 counties (Panzhou and Shuicheng), (4) previous *EGFR* or NGS exome gene mutation testing was performed, (5) each patient underwent a complete medical evaluation and staging. There were no exclusion criteria for individuals meeting the inclusion criteria. All who participated in this study provided written informed consent. Raw data from all the NSCLC patients are presented in Supplementary Table SI. Among them were individuals from 9 counties in Qujing City, Yunnan Province, and 2 counties in Liupanshui City, Guizhou Province. These locations are provided in previous articles from our laboratory [29]. In addition, the electronic medical records of all NSCLC patients were retrieved from the Zhiye Medical Record Workstation of the 2 medical institutions. They included patients' basic demographic, behavioral and clinical information such as gender, age, origin, histopathology, specimen type, lesion site, tumor node metastasis classification (TNM) staging, smoking history, brain metastasis, family history of malignant tumors, ethnicity group, and occupation.

### Samples, gene mutation detection methods and NGS gene panels

Tumor tissues from formalin-fixed paraffin-embedded (FFPE) samples and biopsies, blood plasma, surgically resected fresh tumor tissues, biopsy fresh tumor tissues, or malignant pleural effusion cell specimens from the enrolled NSCLC patients were analyzed for genetic mutations in 17 genes, displayed in Table I. The analyzed cancerous tissues were collected by surgical resection, namely open-chest lung biopsy or through the lung, ultrasound-assisted transdermal core needle lung, or pleural biopsy. The cytological specimens were mainly from malignant pleural effusions. All specimens were dyed with the H&E Staining Kit (Hematoxylin and Eosin) (Solarbio Life Science, China) to detect cancer cell contents. Tissue specimens containing more than 20% tumor cells were qualified for amplification refractory mutation system – polymerase chain reaction (ARMS-PCR) and NGS. Whole blood samples (10 ml) were collected in BD $K_2$ EDTA anticoagulation tubes (BD Biosciences, USA). Centrifugation followed (1500 g/20 min, 4°C) (Centrifuge 5804 R, Eppendorf, Germany). The blood plasma was gathered within 2 h and spun down (13000 g/10 min, 4°C) to collect 3 ml of it to extract circulating tumor DNA. Genomic DNA of FFPE and fresh tumor tissues, malignant pleural effusion cells, and plasma was isolated using QIAamp DNA FFPE, QIAamp Fast DNA Tissue, Blood & Cell Culture DNA and miRNeasy Serum/Plasma Advanced genomic DNA extraction kits (Qiagen, Netherlands), respectively. The DNA concentration was quantified using the NanoDrop 1000 spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA), a Qubit fluorometer 3.0, and a Qubit dsDNA High Sensitivity (HS) Assay Kit (Invitrogen, Carlsbad, CA, USA). The DNA fragment length was measured using an Agilent 2100 Bioanalyzer and DNA HS Kit (Agilent Technologies, Santa Clara, CA, USA). The sample quality fulfilled the certain as follows: total amount ≥ 20 ng for cell-free DNA samples with ~ 170 bp fragments, and 100 ng for tissue DNA with > 1000 bp fragments, 100 ng for tissue DNA > 1000 bp fragments. Samples were then stored immediately at –80°C for the following experiments. Real-time ARMS-PCR studied 73 specimens of NSCLC patients with the AmoyDx *EGFR* 29 Mutations Detection Kit (Amoy Diagnostics, Xiamen, China). The other 1047 specimens were sequenced by NGS using 2 commercial human exome capture platforms (Life Healthcare Clinical Laboratories and KingMed Diagnostics, Beijing and Guangzhou, China) (Figure 1, Supplementary Table SII).

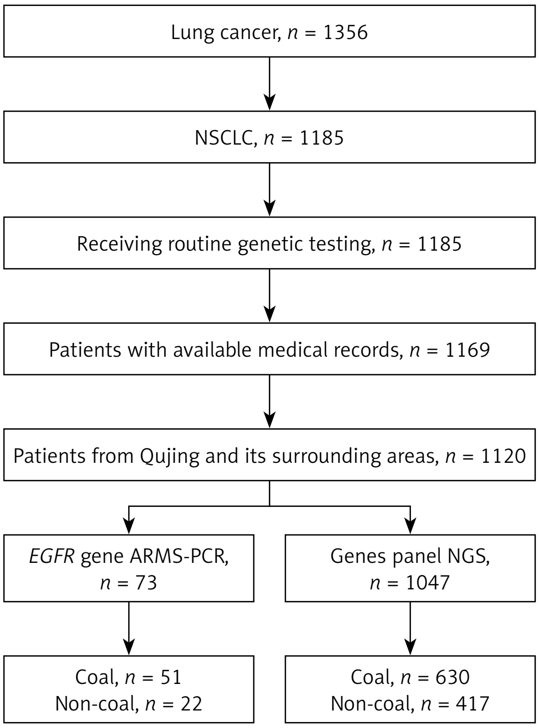

Lung cancer, n = 1356
NSCLC, n = 1185
Receiving routine genetic testing, n = 1185
Patients with available medical records, n = 1169
Patients from Qujing and its surrounding areas, n = 1120
EGFR gene ARMS-PCR, n = 73
Genes panel NGS, n = 1047
Coal, n = 51
Non-coal, n = 22
Coal, n = 630
Non-coal, n = 417

**Figure 1.** Flowchart of the selection process of study participants for the analyses

### NGS library construction

The extracted 100 ng of DNA was sheared with a Covaris E210 DNA Sonicator (Covaris, Woburn, MA, USA) into fragments approximately 200 bp in length. All DNA samples underwent library preparation using an Accel-NGS 2S DNA Library Kit (Swift Biosciences) and xGen Lockdown Probes kit (IDT) (Integrated DNA Technologies, USA). The custom xGen Lockdown probe was synthesized by IDT, Inc. (Integrated DNA Technologies, USA) for the exons and selected intronic regions of 8, 17, 55, 58, 499, 618, and 876 genes, respectively. The panels were designed to detect mutations and small insertions and deletions. Library PCR was performed with a KAPA HiFi HotStart ReadyMix PCR Kit (Kapa Biosystems, Boston, MA). The prepared library was quantified using the Qubit 3.0 Fluorometer, and quality and fragment size were further measured using an Agilent 2100 Bioanalyzer (reference fragment size: 280–350 bp; DNA quality: 0.5–50 ng/µl; DNA Integrity Number (DIN) > 3).

### Exome hybrid capture and sequencing

Each library replicate was amplified using unique combinations of dual-indexed PCR primers, and pooling libraries were simultaneously enriched using xGen Lockdown probes during two rounds of capture. Hybridization capture was performed in a customized xGen Lockdown probe panel (Integrated DNA Technologies, USA) at 65°C overnight, followed by post-hybridization washes. The NimbleGenSeqCap EZ Hybridization and Wash Kit (Nimblegen, Roche Diagnostics, Mannheim, Germany) was applied. Exome identification was performed with the Nextera Rapid Capture Exome Kit (Illumina, California, USA). Each hybrid-selected library was performed by qPCR using the KAPA SYBR FAST qPCR Master Mix (Kapa Biosystems; Roche Diagnostics Corporation, Indianapolis, IN, USA) for Illumina sequencing platforms on a Rotor-Gene Q thermocycler (Qiagen, Hilden, Germany), and normalized in relation to its size. Individual libraries were normalized to 5 nM in preparation for sequence analysis. Sequencing libraries were chemically denatured and applied to an Illumina NovaSeq flow cell using the NovaSeq XP workflow (Illumina). Following a transfer of the flow cell to the Illumina NovaSeq 6000 instrument, the catalog number of the sequencing kit was NovaSeq 6000 S4 Reagent kit v1.5 (300 cycles; cat. no. 20028312; Illumina Inc.). Samples underwent paired-end sequencing on an Illumina NovaSeq 6000 platform (Illumina) with paired-end 2 × 150-bp read length. Median coverage of 2593 × (range: 201–10489) and 4241 × (range: 1523–9762) was achieved for tumor tissue DNA and cell-free plasma DNA (cfDNA), respectively. To confirm the concordance between different panels, 27 circulating tumor DNA (ctDNA) samples were duplicated and tested with the 63-gene, 128-gene, and 1460-gene panels.

### Bioinformatics analysis of sequencing data

For bioinformatics analysis, we additionally analyzed the data using a custom pipeline. FASTQ files were trimmed with Cutadapt (1.15) to remove adapter sequences and sample barcode identifiers, which included demultiplexing the raw data to FASTQ files using bcl2fastq (v2.20) followed by a quality assessment of the FASTQ files using Trimmomatic (v0.39) or Illumina NGS data analyses [30]. The Fastq file was converted into the unmapped BAM (uBAM) format using the FastqToSam (Picard, 2.19.2) tool, and the sequence paired-end molecular tag information was extracted using the ExtractUmisFromBam (Fgbio, 0.8.0) tool and stored in the RX tag of the uBAM file to be analyzed later. The sequencing data were compared to the human genome by BWA (0.7.12-r1039) software (http://bio-bwa.sourceforge.net/), and ANNOVAR (date: 2015-06-17) was used to annotate the mutation sites based on dbSNP [31], Clinvar, and 1000 genomes [32]. Mapping was performed on Ensembl hg19 (February 2009 [GRCh37]) using the Burrows-Wheeler Aligner. Common single nucleotide polymorphism (SNP) and small indel detection was performed with GATK HaplotypeCaller (version 3.8-0) and were annotated by ANNOVAR (http://annovar.openbioinformatics.org/). We set a series of filtering cutoff values for reliable SNP calling by GATK: (1) Sites with Phred-scaled strand bias at this position (SB) > 60, homopolymer length to the right of report indel position (Hrun) ≥ 8, 20 bp better reads before and after mutation (neighbor_20) < 10, alt-forward bases (DP2) < 5, alt-reverse bases (DP3) < 5, and better reads supporting the mutation (Alt_reads) < 8 were removed. (2) Sites with sequencing depth DP < 1000, 1000 person frequency AF > 0.01, and mutation frequency AF < 0.01 were removed. Copy number variations were assessed using the OncoCNV (v6.8) package (Paris, France, https://oncocnv.curie.fr). Finally, the Maftools R package was applied in oncoplot drawing, driver gene identification, signaling pathway, and association analyses.

### Ethics approval

This retrospective study was approved by the institutional review boards of Qujing First People's Hospital (approval number: 2016-023-01). All procedures performed in studies that involved human participants were in accordance with the ethical standards of the institutional and/or national research committees and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. All samples were obtained from participants who signed an informed consent form.

### Statistical analysis

All patients were allocated to 11 counties in Eastern Yunnan and Western Guizhou, divided into coal-manufacturing areas and non-coal districts [29]. The links between the detected genetic variations and clinical characteristics of the studied patients were statistically evaluated by SPSS 26.0 (IBM Inc., Illinois, USA). $\chi^2$ or Fisher's exact test was used to compare two categorical variables. Result interpretation: (1) The total number of cases ≥ 40, all theoretical frequencies ≥ 5. See Pearson $\chi^2$ results. (2) The total number of cases is ≥ 40, and there is one theoretical frequency of ≥ 1 and < 5. The $\chi^2$ test must be corrected for continuity, and the continuity correction result shall prevail. (3) The total number of cases ≥ 40, at least 2 theoretical frequencies ≥ 1 and < 5; see the results of Fisher's exact test (exact significance, 2-sided). (4) The total number of cases is less than 40, or the theoretical frequency is less than 1; see the results of Fisher's exact test (exact significance, 2-sided). In addition, a multivariate binary logistic regression model was used to evaluate risk factors for *EGFR* complex mutation. *P*-values less than 0.05, 0.01, and 0.001 indicated statistically significant differences.

## Results

### NSCLC patients' medical data from the coal-manufacturing and non-coal eastern Yunnan districts

It is widely known that the coal-manufacturing Eastern Yunnan districts have a high incidence of lung cancer in China [33]. Compared with previous studies, we increased the sample size and the number of detected genes and further divided the population in Eastern Yunnan into coal-manufacturing districts and non-coal-manufacturing to explore the unique molecular characteristics of NSCLC patients in these places. The tumor tissue and plasma samples of the 1120 studied NSCLC patients were screened by ARMS PCR and DNA sequencing for the 17 genes (Table I). There were 681 (60.80%) NSCLC patients from the coal-manufacturing Eastern Yunnan districts and 439 (39.20%) patients in non-coal ones. Clinicopathological features, including gender, age, histopathology, smoking history, family history, TNM staging, lesion site, and occupation, are summarized in Table II. 29 sites of *EGFR* gene exons 18, 19, 20, and 21 in 73 patients were detected. The entire exons of the genes *EGFR*, *ALK*, *ROS1*, *KRAS*, *BRAF*,

**Table II.** Clinical features of NSCLC patients in coal- and non-coal-manufacturing Yunnan regions

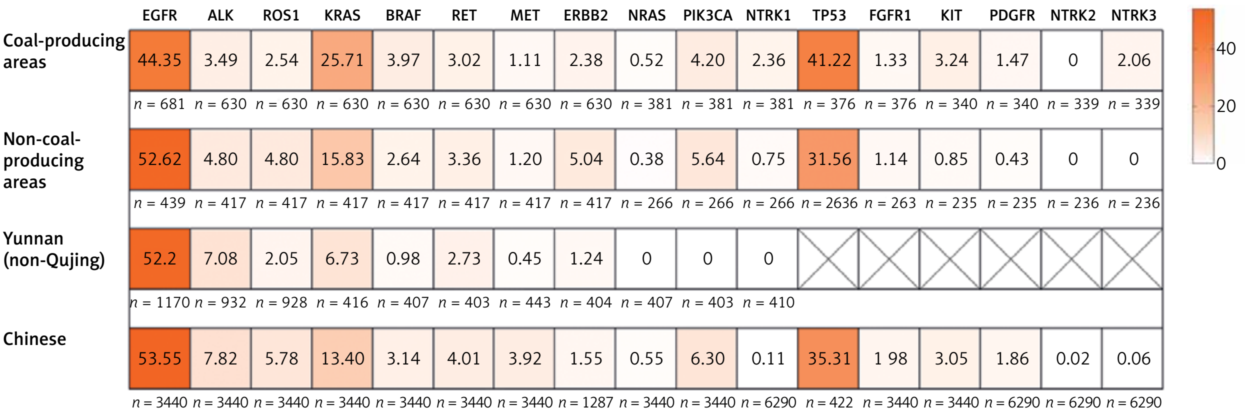
| Characteristic | All patients (*n* = 1120) | Region | | *P*-value |
|---|---|---|---|---|
| | | Coal-producing areas (*n* = 681) | Non-coal-producing areas (*n* = 439) | |
| Gender: | | | | **0.035** |
| Male | 508 (45.36%) | 326 (47.87%) | 182 (41.46%) | |
| Female | 612 (54.64%) | 355 (52.13%) | 257 (58.54%) | |
| Age: | | | | |
| Median (range) | | 56 (19-88) | 58 (26-91) | 0.758 |
| ≤ 40 | 36 (3.21%) | 21 (3.08%) | 15 (3.42%) | |
| > 40 | 1084 (96.79%) | 660 (96.92%) | 424 (96.58%) | |
| Histopathology: | | | | **0.027** |
| Adenocarcinoma | 1023 (91.34%) | 624 (91.63%) | 399 (90.89%) | |
| Squamous carcinoma | 36 (3.21%) | 15 (2.20%) | 21 (4.78%) | |
| Unknown (NSCLC) | 61 (5.45%) | 42 (6.17%) | 19 (4.33%) | |
| Smoking history: | | | | **0.025** |
| Yes | 291 (25.98%) | 193 (28.34%) | 98 (22.32%) | |
| No | 829 (74.02%) | 488 (71.66%) | 341 (77.68%) | |
| Family history: | | | | **< 0.001** |
| Yes | 184 (16.43%) | 138 (20.26%) | 46 (10.48%) | |
| No | 936 (83.57%) | 543 (79.74%) | 393 (89.52%) | |
| TNM staging: | | | | **0.009** |
| I–IIIa | 827 (73.84%) | 484 (71.07%) | 343 (78.13%) | |
| IIIb–IV | 293 (26.16%) | 197 (28.93%) | 96 (21.87%) | |
| Lesion site: | | | | 0.461 |
| Left | 427 (38.13%) | 252 (37.00%) | 175 (39.86%) | |
| Right | 647 (57.77%) | 398 (58.44%) | 249 (56.72%) | |
| Bilateral | 46 (4.11%) | 31 (4.55%) | 15 (3.42%) | |
| Occupation: | | | | **< 0.001** |
| Farmer | 923 (82.41%) | 612 (89.87%) | 311 (70.84%) | |
| Non-farmer/unknown | 197 (17.59%) | 69 (10.13%) | 128 (29.16%) | |

*RET*, *MET*, and *ERBB2* were examined by DNA sequencing in 1047 other NSCLC patients (Supplementary Table SI). Six hundred forty-seven tumor samples were profiled for mutations in the entire exons of the genes *NRAS*, *PIK3CA*, and *NTRK1*. Six hundred thirty-nine tumor samples were tested for mutations in the whole exons of *TP53* and *FGFR1*. 575 were tested for *KIT*, *PDGFRA*, *NTRK2*, and *NTRK3* (Supplementary Table SI).

We have compared patients' demographic and clinicopathological characteristics between the 2 types of districts. This comparison highlighted the following main demographic factors and baseline clinical characteristics of NSCLC individuals from the coal regions. These factors were female gender ($p$ = 0.035), adenocarcinoma as histopathology ($p$ = 0.027), non-smoking as a smoking history ($p$ = 0.025), no family history ($p$ < 0.001), TNM staging I-IIIa ($p$ = 0.009) and farmers as an occupation ($p$ < 0.001) (Table II).

### Mutation frequencies of *EGFR*, *KRAS*, *TP53*, *ALK*, *ROS1*, *BRAF*, *RET*, *MET*, *ERBB2*, *NRAS*, *KIT*, *PIK3CA*, *FGFR1*, *PDGFRA*, *NTRK1*, *NTRK2* and *NTRK3* genes in NSCLC patients in coal-manufacturing Eastern Yunnan

Among all studied NSCLC individuals in Eastern Yunnan, 1023 (91.34%) were diagnosed with lung adenocarcinoma, 36 (3.21%) with lung squamous cell carcinoma, and 61 (5.45%) with unspecified NSCLC (Table II). The comparative examination of the mutation rates of the 17 studied genes (Table I) showed a unique mutation spectrum in NSCLC patients from the coal-manufacturing Yunnan zones (Figure 2). *EGFR* had a mutation rate of 44.35% in coal districts vs. 52.62% in non-coal ones, $p$ = 0.006 (Figure 3 A), while *ERBB2* had 2.38% vs. 5.04%, $p$ = 0.021, in coal and non-coal regions, respectively (Figure 4 C). These data showed that the *EGFR* and *ERBB2* gene mutation rates in NSCLC



| | EGFR | ALK | ROS1 | KRAS | BRAF | RET | MET | ERBB2 | NRAS | PIK3CA | NTRK1 | TP53 | FGFR1 | KIT | PDGFR | NTRK2 | NTRK3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Coal-producing areas | 44.35 | 3.49 | 2.54 | 25.71 | 3.97 | 3.02 | 1.11 | 2.38 | 0.52 | 4.20 | 2.36 | 41.22 | 1.33 | 3.24 | 1.47 | 0 | 2.06 |
| | *n* = 681 | *n* = 630 | *n* = 630 | *n* = 630 | *n* = 630 | *n* = 630 | *n* = 630 | *n* = 630 | *n* = 381 | *n* = 381 | *n* = 381 | *n* = 376 | *n* = 376 | *n* = 340 | *n* = 340 | *n* = 339 | *n* = 339 |
| Non-coal-producing areas | 52.62 | 4.80 | 4.80 | 15.83 | 2.64 | 3.36 | 1.20 | 5.04 | 0.38 | 5.64 | 0.75 | 31.56 | 1.14 | 0.85 | 0.43 | 0 | 0 |
| | *n* = 439 | *n* = 417 | *n* = 417 | *n* = 417 | *n* = 417 | *n* = 417 | *n* = 417 | *n* = 417 | *n* = 266 | *n* = 266 | *n* = 266 | *n* = 2636 | *n* = 263 | *n* = 235 | *n* = 235 | *n* = 236 | *n* = 236 |
| Yunnan (non-Qujing) | 52.2 | 7.08 | 2.05 | 6.73 | 0.98 | 2.73 | 0.45 | 1.24 | 0 | 0 | 0 | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ |
| | *n* = 1170 | *n* = 932 | *n* = 928 | *n* = 416 | *n* = 407 | *n* = 403 | *n* = 443 | *n* = 404 | *n* = 407 | *n* = 403 | *n* = 410 | | | | | | |
| Chinese | 53.55 | 7.82 | 5.78 | 13.40 | 3.14 | 4.01 | 3.92 | 1.55 | 0.55 | 6.30 | 0.11 | 35.31 | 1 98 | 3.05 | 1.86 | 0.02 | 0.06 |
| | *n* = 3440 | *n* = 3440 | *n* = 3440 | *n* = 3440 | *n* = 3440 | *n* = 3440 | *n* = 3440 | *n* = 1287 | *n* = 3440 | *n* = 3440 | *n* = 6290 | *n* = 422 | *n* = 3440 | *n* = 3440 | *n* = 6290 | *n* = 6290 | *n* = 6290 |

**Figure 2.** Variation rates of 17 genes associated with NSCLC in individuals from coal and non-coal regions of Eastern Yunnan, Yunnan (non-Qujing), and the Chinese population

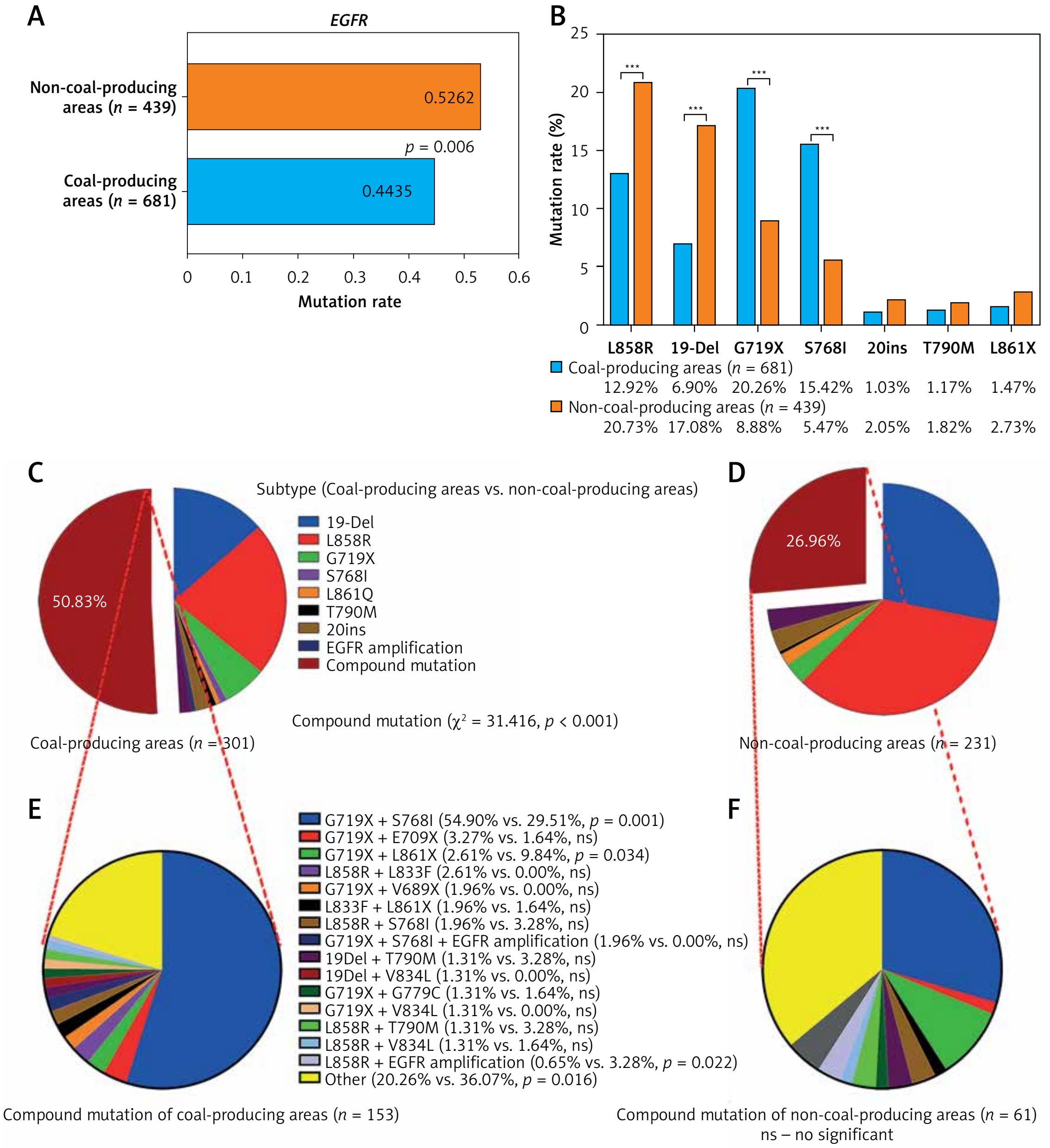


**Figure 3.** Frequency of *EGFR* variants, subtypes, and compound mutations in NSCLC individuals. **A** – Prevalence of *EGFR* mutations in coal-producing regions compared with patients from the non-coal-manufacturing Yunnan provinces. **B** – G719X and S768I point mutation rates were higher, and the frequency of L858R and 19-Del variants was lower in NSCLC patients from coal-producing areas of Eastern Yunnan. **C**, **D** – Dispersal of *EGFR* subtypes in coal and non-coal Yunnan provinces. (**E**, **F**) Distribution of *EGFR* compound

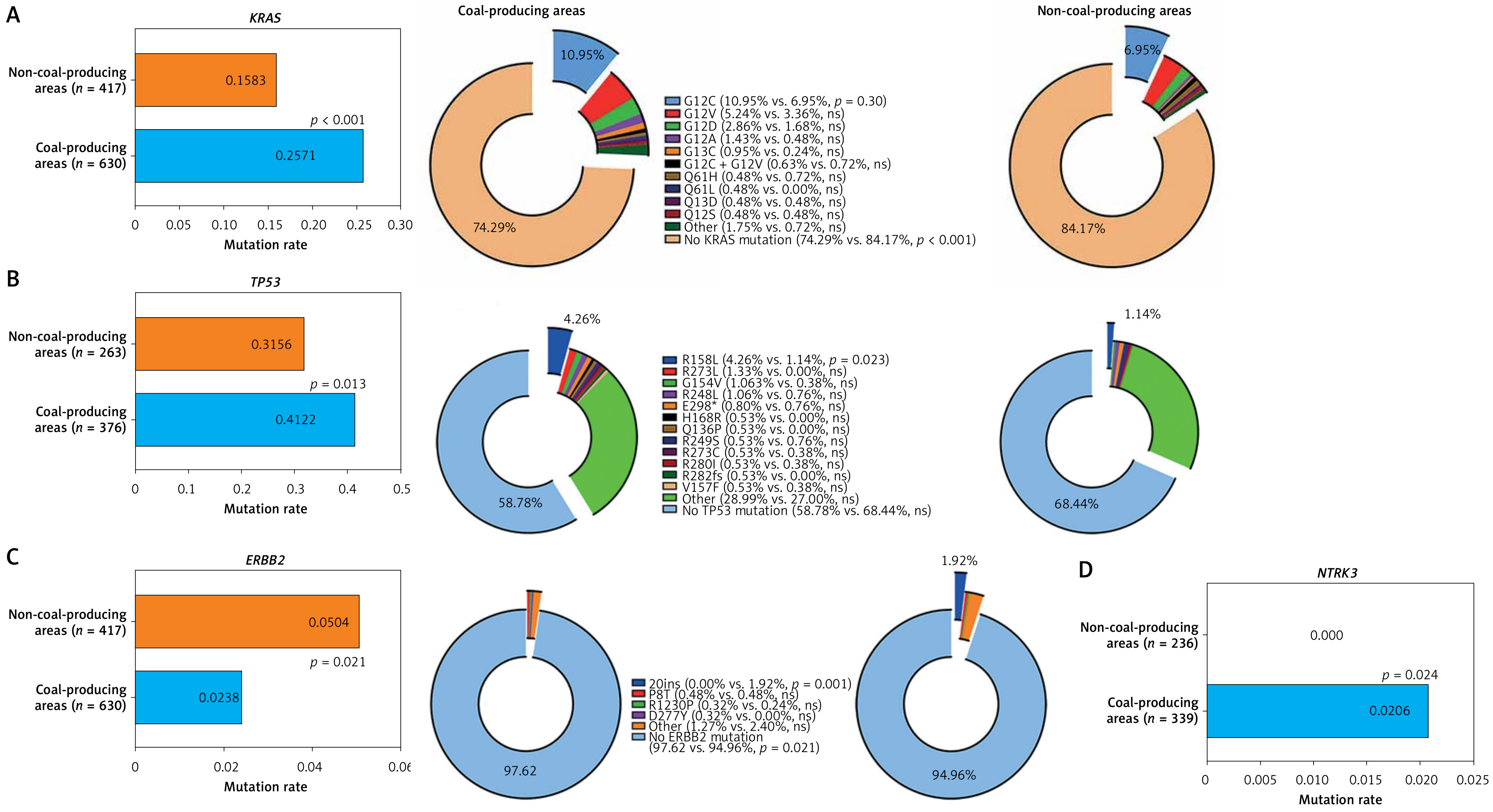


**Figure 4.** Profiles of *KRAS/TP53/ERBB2/NTRK3* variants and subtypes

**Table III.** The association between patients' medical features and *EGFR*, *KRAS*, *TP53*, *ERBB2* and *NTRK3* in NSCLC individuals

| Characteristics | EGFR | | | | KRAS | | | | TP53 | | | | ERBB2 | | | | NTRK3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mut | WT | $\chi^2$ | *p* | Mut | WT | $\chi^2$ | *p* | Mut | WT | $\chi^2$ | *p* | Mut | WT | $\chi^2$ | *p* | Mut | WT | $\chi^2$ | *p* |
| Gender: | | | | | | | | | | | | | | | | | | | | |
| Male | 118 | 208 | | | 95 | 210 | | | 94 | 88 | | | 10 | 295 | | | 4 | 159 | | |
| Female | 183 | 172 | 16.242 | < 0.001 | 67 | 258 | 9.137 | 0.003 | 61 | 133 | 15.822 | < 0.001 | 5 | 320 | 2.050 | 0.152 | 3 | 173 | | 0.715 |
| Age: | | | | | | | | | | | | | | | | | | | | |
| ≤ 40 | 10 | 11 | | | 3 | 15 | | | 3 | 8 | | | 0 | 18 | | | 0 | 11 | | |
| > 40 | 291 | 369 | 0.103 | 0.749 | 159 | 453 | 0.381 | 0.537 | 152 | 213 | 0.414 | 0.520 | 15 | 597 | | 1.000 | 7 | 321 | | 1.000* |
| Histopathology: | | | | | | | | | | | | | | | | | | | | |
| AD | 286 | 341 | | | 152 | 425 | | | 144 | 211 | | | 14 | 563 | | | 7 | 312 | | |
| Non-AD | 15 | 39 | 6.413 | 0.011 | 10 | 43 | 1.42 | 0.233 | 11 | 10 | 1.143 | 0.285 | 1 | 52 | | 1.000 | 0 | 20 | | 1.000 |
| Smoking history: | | | | | | | | | | | | | | | | | | | | |
| Yes | 61 | 132 | | | 63 | 118 | | | 48 | 50 | | | 6 | 175 | | | 3 | 87 | | |
| No | 240 | 248 | 17.319 | < 0.001 | 99 | 350 | 10.991 | 0.001 | 107 | 171 | 3.291 | 0.070 | 9 | 440 | 0.473 | 0.492 | 4 | 245 | 0.308 | 0.579 |
| Family history: | | | | | | | | | | | | | | | | | | | | |
| Yes | 53 | 85 | | | 41 | 90 | | | 37 | 44 | | | 3 | 128 | | | 1 | 79 | | |
| No | 248 | 295 | 2.356 | 0.125 | 121 | 378 | 2.699 | 0.100 | 118 | 177 | 0.846 | 0.358 | 12 | 487 | | 1.000* | 6 | 253 | | |
| Ethnicity: | | | | | | | | | | | | | | | | | | | | |
| Han | 298 | 376 | | | 160 | 463 | | | 155 | 218 | | | 15 | 608 | | | 7 | 329 | | |
| Non-Han | 3 | 4 | | 1.000 | 2 | 5 | | 1.000 | 0 | 3 | | 0.271 | 0 | 7 | | 1.000* | 0 | 3 | 0.019 | 0.891 |
| Staging: | | | | | | | | | | | | | | | | | | | | |
| I–IIIa | 217 | 267 | | | 140 | 315 | | | 115 | 157 | | | 11 | 444 | | | 6 | 265 | | |
| IIIb–IV | 84 | 113 | 0.274 | 0.601 | 22 | 153 | 21.911 | < 0.001 | 40 | 64 | 0.453 | 0.501 | 4 | 171 | | 1.000* | 1 | 67 | | 1.000* |
| Lesion site: | | | | | | | | | | | | | | | | | | | | |
| Left | 116 | 136 | | | 56 | 174 | | | 63 | 76 | | | 6 | 253 | | | 3 | 122 | | |
| Right | 175 | 223 | 0.265 | 0.606 | 97 | 274 | 0.242 | 0.623 | 83 | 136 | 1.941 | 0.164 | 9 | 362 | 0.008 | 0.929 | 4 | 203 | | 1.000* |
| Occupation: | | | | | | | | | | | | | | | | | | | | |
| Farmer | 269 | 343 | | | 145 | 419 | | | 137 | 197 | | | 12 | 552 | | | 6 | 295 | | |
| Non-farmer/unknown | 32 | 37 | 0.148 | 0.701 | 17 | 49 | | 0.993 | 18 | 24 | 0.052 | 0.819 | 3 | 63 | 0.628 | 0.428 | 1 | 37 | | 1.000* |

*AD – adenocarcinoma, SCC – squamous cell carcinoma, *Fisher's exact test.*

patients in coal-manufacturing Yunnan zones were considerably lower than those in non-coal ones. In contrast, the *KRAS* mutation rates in the 2 regions were 25.71% (coal-manufacturing) vs. 15.83% (non-coal ones) ($p$ = 0.000) (Figure 4 A), *TP53* (41.22% vs. 31.56%, $p$ = 0.013) (Figure 4 B) and *NTRK3* (2.06% vs. 0.00%, $p$ = 0.024) (Figure 4 D). These data proved that these particular gene mutation frequencies were higher than those in non-coal Yunnan districts. The same results also appeared in lung adenocarcinoma patients in Eastern Yunnan. Also, there were also significant differences in the mutation rates of the genes *ROS1* (2.43% (coal-manufacturing) vs. 4.95% (non-coal ones), $p$ = 0.036) and *NTRK1* (2.22% vs. 0.00%, $p$ = 0.024, respectively) (Supplementary Figure S1). Further analysis revealed that the mutation rates of *EGFR* (52.29% vs. 65.46%, $p$ = 0.002), *KRAS* (21.33% vs. 14.52%, $p$ = 0.042), *BRAF* (4.33% vs. 1.24%, $p$ = 0.035) and *ROS1* (2.33% vs. 6.22%, $p$ = 0.023) in females, non-smokers and individuals with lung adenocarcinoma between coal- and non-coal-producing Eastern Yunnan areas were also significantly different (Supplementary Figure S2). However, no statistically significant differences were detected among the mutation rates of the gene *TP53* in female, non-smoking, and lung adenocarcinoma patients. Compared with NSCLC individuals in non-coal Eastern Yunnan zones, mutation of only one of the 17 genes (*EGFR*, *KRAS*, *TP53*, *ALK*, *ROS1*, *BRAF*, *RET*, *MET*, *ERBB2*, *NRAS*, *KIT*, *PIK3CA*, *FGFR1*, *PDGFRA*, *NTRK1*, *NTRK2*, and *NTRK3*) was less frequent in coal-manufacturing ones (47.42% (coal-manufacturing) vs. 57.69% (non-coal ones), $p$ = 0.024). They had higher mutation rates of complex mutations (52.58% vs. 42.31%, $p$ = 0.024) (Supplementary Figure S3).

### Relationship between clinical characteristics of NSCLC patients in coal-manufacturing Eastern Yunnan districts and mutation rates in the genes *EGFR*, *KRAS*, *TP53*, *ERBB2,* and *NTRK3*

Our subsequent analyses included investigating the link between the mutation rates in the genes *EGFR*, *KRAS*, *TP53*, *ERBB2,* and *NTRK3* and NSCLC individuals' medical data. The results proved that among those patients from the observed coal-manufacturing districts, the *EGFR* gene was predominantly expressed in females ($p$ < 0.001), non-smokers ($p$ < 0.001), and individuals with lung adenocarcinoma ($p$ = 0.011) (Table III). The *KRAS* variations were commonly detected in males ($p$ = 0.003), non-smokers ($p$ = 0.001), and patients in TNM stage I-IIIa ($p$ = 0.000) (Table III). Among smokers, *KRAS* gene variations were more frequent in NSCLC patients ($p$ = 0.007) from non-coal-manufacturing zones (Supplementary Table SIV). In contrast, among non-smoking ones, the gene variations in *KRAS* were predominant in individuals from non-coal zones ($p$ = 0.013) (Supplementary Table SIV). Overall, the *KRAS* gene variations were predominant in smokers from non-coal-producing areas ($p$ = 0.010) (Supplementary Table SIV). The analysis of *TP53* mutation rates showed that they were predominant in males ($p$ < 0.001), and there were no differences between smokers and non-smokers (Table III). *ERBB2* and *NTRK3* mutations were not associated with clinical characteristics such as gender, age, histopathology, smoking history, family history, ethnicity, TNM staging, lesion site, and occupation in NSCLC patients from Eastern Yunnan coal-manufacturing places (Table III).

### Mutation subtypes of *EGFR* among the studied patients

NSCLC individuals from Eastern Yunnan coal-producing areas had a significantly higher frequency of point mutations G719X (20.26% vs. 8.88%, $p$ < 0.001) and S768I (15.42% vs. 5.47%, $p$ < 0.001). On the other hand, the 19Del (6.90% vs. 17.08%, $p$ < 0.001) and L858R SNPs (single-nucleotide polymorphisms) (12.92% vs. 20.73%, $p$ = 0.001) were significantly less frequent. Results are displayed in Figure 3 B. Furthermore, highly significant variation of *EGFR* compound mutations (50.83% vs. 26.96%, $p$ < 0.001) as well as *EGFR* G719X + S768I (54.90% vs. 29.51 %, $p$ = 0.001), *EGFR* G719X + L861X (2.61% vs. 9.84%, $p$ = 0.034) and *EGFR* L858R + EGFR amplification (0.65% vs. 3.28%, $p$ = 0.022) was detected in NSCLC patients in the eastern Yunnan coal-producing area (Figures 3 C–F). We performed logistic multivariate regression evaluations of these data and detected that patients' jobs (occupied in rural areas, such as agrarians) had OR = 2.430 and 95% (95% CI: 1.031–5.727), and the disease TNM staging (IIIb–IV) displayed OR = 6.820 and 95% (95% CI: 3.639–12.782), and these characteristics were unconventionally linked with higher *EGFR* compound mutation rates (Table IV).

### *KRAS*, *TP53*, *ERBB2,* and *NRTK3* gene mutation subtypes in NSCLC patients in Eastern Yunnan coal-producing areas

We studied the mutation frequency and mutation subtypes in 1047 patients by NGS and found that 228 patients (21.78%) had mutations in the *KRAS* gene (Figure 4 A), including 162 individuals from the studied coal-manufacturing districts and 66 from the non-coal ones. *KRAS* G12C mutation had an increased rate in the individuals from the coal regions compared to the others (10.95% vs. 6.95%, $p$ = 0.030) (Figure 4 A). In addition, 639

**Table IV.** Multivariate regression analysis of the association between *EGFR* compound mutations and demographic factors in NSCLC patients from Eastern Yunnan coal-producing areas

| Characteristics | Exp(B) | EXP(B) 95% CI | | *P*-value |
|---|---|---|---|---|
| | | Lower | Upper | |
| Sex (Male vs. Female) | 1.022 | 0.530 | 1.967 | 0.949 |
| Age | 1.342 | 0.313 | 5.758 | 0.692 |
| Occupation (Farmer vs. non-Farmer) | 2.430 | 1.031 | 5.727 | **0.042** |
| Race (Han vs. non-Han) | 0.362 | 0.028 | 4.657 | 0.436 |
| Smoking (Yes vs. No) | 0.657 | 0.294 | 1.466 | 0.305 |
| Family history (Yes vs. No) | 0.740 | 0.383 | 1.430 | 0.370 |
| Lesion site (Left vs Right) | 1.061 | 0.674 | 1.669 | 0.799 |
| Histopathology (AD vs. SCC) | 1.207 | 0.409 | 3.562 | 0.733 |
| TNM staging (I–IIIa vs. IIIb–IV) | 6.820 | 3.639 | 12.782 | **< 0.001** |

patients underwent NGS analysis of the mutation rates and subtypes in the gene *TP53*, and the results showed that 238 (37.25%) had *TP53* gene mutations (Figure 4 B), among which 155 were from the coal regions, while 83 were from the non-coal ones. The frequency of *TP53* R158L mutation was predominant and with a high rate in the coal regions in contrast to the others (4.26% vs. 1.14%, respectively) ($p$ = 0.023) (Figure 4 B). In addition, 1047 patients underwent *ERBB2* gene analysis. Thirty-six patients (3.44%) had *ERBB2* gene mutations (Figure 4 C), including 15 from the coal-manufacturing regions and 21 from the non-coal ones. *ERBB2* 20ins mutation in the NSCLC patients from the non-coal regions was more predominant than in the coal ones (0.00% vs. 1.92%, $p$ = 0.001) (Figure 4 C). Furthermore, 575 individuals underwent screening for *NRTK3* gene variations, and the results demonstrated that 7 (2.06%) patients from the coal zones had *NRTK3* variations (Figure 4 D).

### Prevalence and subtype distribution of driver gene mutations in patients with early and advanced-stage NSCLC

Early-stage NSCLC encompasses stages I and II, while advanced-stage NSCLC includes stages III and IV. A cohort of 765 early-stage and 355 advanced-stage NSCLC patients was used separately for *EGFR*, *KRAS*, *TP53*, *ERBB2*, and *NTRK3* gene mutation analysis. Among the early-stage NSCLC cases, 436 patients were from coal-producing areas, and 329 were from non-coal-producing areas. Two hundred forty-five advanced NSCLC patients were from coal-producing areas, and 110 were from non-coal-producing areas. The results showed that the frequencies of *EGFR* and *ERBB2* mutation in patients with early and advanced-stage NSCLC from coal-producing areas were significantly lower than those of non-coal-producing areas (Figures 5 A, C, E). In contrast, the *KRAS* mutation frequency in early-stage NSCLC patients from coal-producing areas was significantly higher than that from non-coal-producing areas (Figure 5 B). The frequency of G719X + S768I compound double mutation in early and advanced-stage NSCLC patients from coal-producing areas was also significantly higher than that from the non-coal-producing areas (Figures 5 D, F). However, the frequency of 19-Del mutation in the early-stage and L858R in advanced-stage NSCLC patients from coal-producing areas was significantly lower than that from non-coal-producing areas (Figures 5 D, F). There were no differences in the mutation frequencies of the remaining genes and subtypes between coal-producing and non-coal-producing patients. Then, compared with the driver gene mutation rates of early and advanced-stage patients, the mutation rates of *EGFR* gene 19-Del and L858R and the *KRAS* gene in early-stage patients were significantly higher than those in advanced-stage patients. At the same time, G719X + S768I was significantly lower (Figure 5 G).

### Mutation frequencies of base transitions and transversions in the studied genes in NSCLC patients from the studied regions in Yunnan

Nucleotide base G>T point mutation in the *TP53* gene was related to the pollution of benzopyrene from smoke particles [25]. We then studied the frequencies of appearance of the detected point mutations in all studied genes among those 1120 NSCLC patients. Our results showed that 681 patients from the coal-producing areas had the following frequencies of gene point mutations: G>T (51.04%), T>G (10.30%), and C>T (3.94%) (Figure 6 A). On the other hand, 439 NSCLC patients from non-coal-producing areas had the point mutation G>T with 30.05% frequency, the mutation T>G with 24.31%, and C>T polymorphism with 8.26%

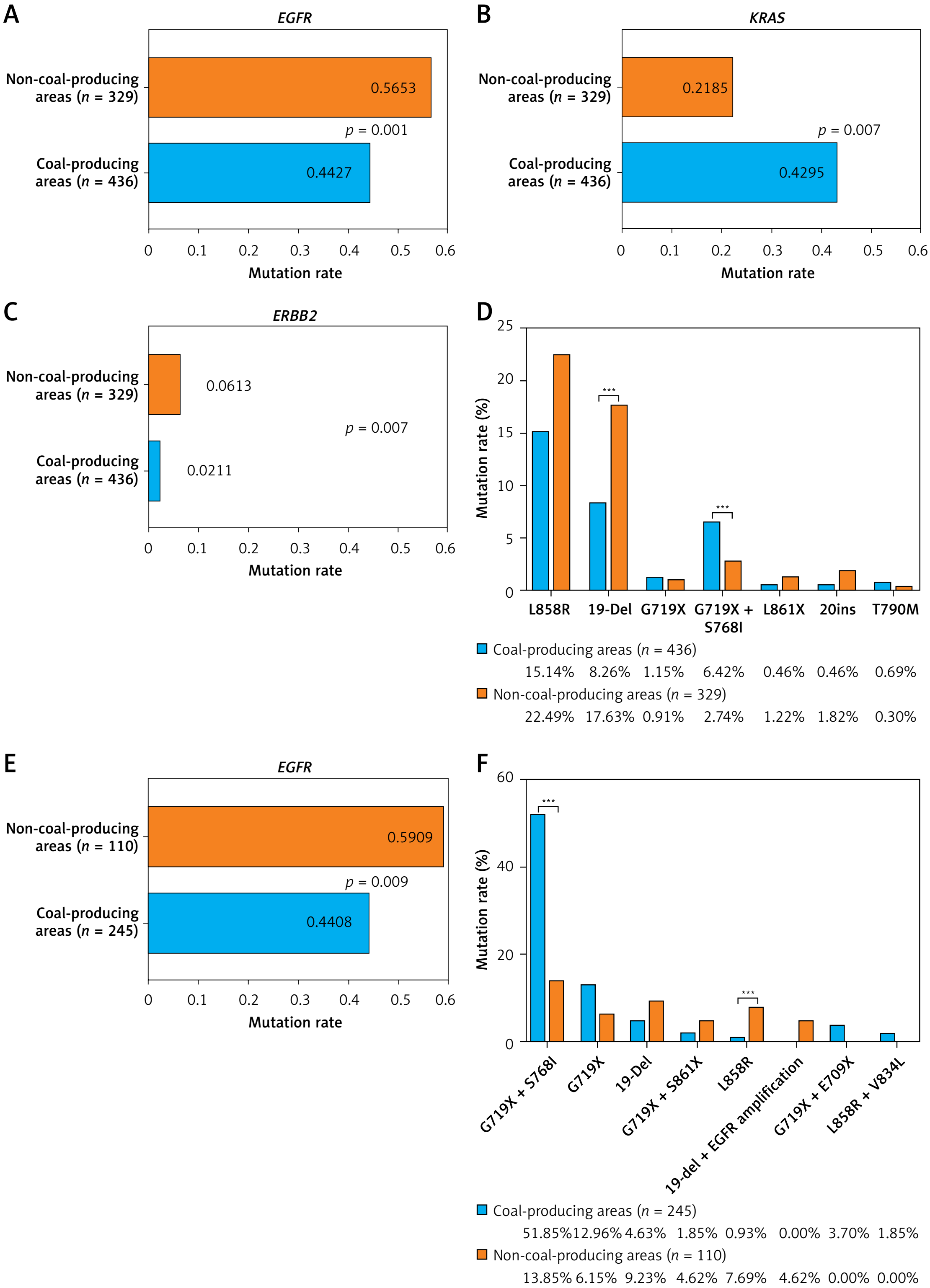
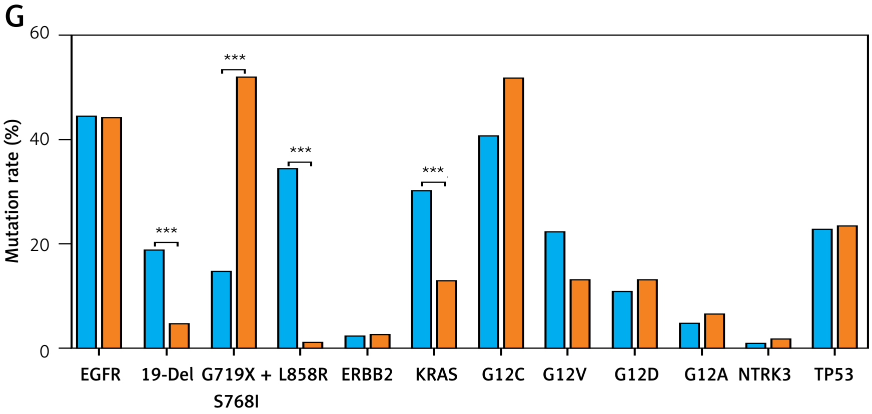


**Figure 5. A** – *EGFR* gene mutation frequency in early-stage NSCLC patients from coal-producing and non-coal-producing areas. **B** – *KRAS* gene mutation frequency in early-stage NSCLC patients from coal-producing and non-coal-producing areas. **C** – *ERBB2* gene mutation frequency in early-stage NSCLC patients from coal-producing and non-coal-producing areas. ****p* < 0.001, ***p* < 0.01, **p* < 0.05. **D** – Distribution and frequency of driver gene mutation subtypes in early-stage NSCLC patients from coal-producing and non-coal-producing regions. **E** – *EGFR* gene mutation frequency in advanced-stage NSCLC patients from coal-producing and non-coal-producing areas. **F** – Distribution and frequency of driver gene mutation subtypes in advanced-stage NSCLC patients from coal-producing and non-coal-producing regions. ****p* < 0.001, ***p* < 0.01, **p* < 0.05. ****p* < 0.001, ***p* < 0.01, **p* < 0.05



| | EGFR | 19-Del | G719X + S768I | L858R | ERBB2 | KRAS | G12C | G12V | G12D | G12A | NTRK3 | TP53 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 🟦 I and II staging | 44.27% | 18.65% | 14.51% | 34.20% | 2.06% | 30.05% | 40.46% | 22.14% | 10.69% | 4.58% | 0.69% | 22.48% |
| 🟧 II and IV staging | 44.08% | 4.63% | 51.85% | 0.93% | 2.45% | 12.65% | 51.61% | 12.90% | 12.90% | 6.45% | 1.63% | 23.27% |

**Figure 5.** Cont. **G** – Driver gene mutation frequencies and subtypes in early and advanced NSCLC in coal-producing regions. ***$p$ < 0.001, **$p$ < 0.01, *$p$ < 0.05

frequency (Figure 6 A). The overall G>T ($p$ < 0.001) point mutation frequency in all NSCLC patients from the coal-manufacturing eastern Yunnan districts was considerably more significant than that in the non-coal zones. In contrast, the two-point mutations T>G ($p$ < 0.001) and C>T ($p$ = 0.001) did not have a high rate of occurrence when compared to the non-coal districts of Yunnan (Figure 6 A). Six hundred thirty-nine individuals were screened for *TP53* gene variations by exosome NGS, among which 376 were from coal-producing areas. Among those patients, 155 possessed *TP53* gene mutation (41.22%), containing the point mutation G>T at the 52.87% occurrence rate and A>T transversion mutation at 11.46%. On the other hand, 263 individuals were from non-coal zones; among them, 83 had *TP53* gene mutation (31.56%), with G>T at 37.50% frequency and A>T at 3.75%. The mutation frequency of G>T ($p$ = 0.025) and A>T ($p$ = 0.048) point mutations in the gene *TP53* in the NSCLC patients from the coal regions was greater than that in the non-coal ones (Figure 6 B).

### Analysis of the driver genes and signaling pathways in the studied NSCLC patients from the eastern Yunnan coal-producing areas

To identify the unique "driver genes" in the studied NSCLC patients from the Eastern Yunnan coal-producing areas, we used the OncodriverCLUST algorithm. It allowed us to analyze the significantly mutated genes in the 519 NSCLC patients screened genetically by NGS at the College of American Pathologists (CAP)-Certified Laboratory (Life Healthcare, Beijing, China). As shown in Figure 7 A, the most predominant cancer-related genetic variants observed in these patients were mutations in the *EGFR* gene (found in 53.56% of NSCLC patients from the studied coal-producing area), followed by the genes *TP53* (47.21%), *KRAS*

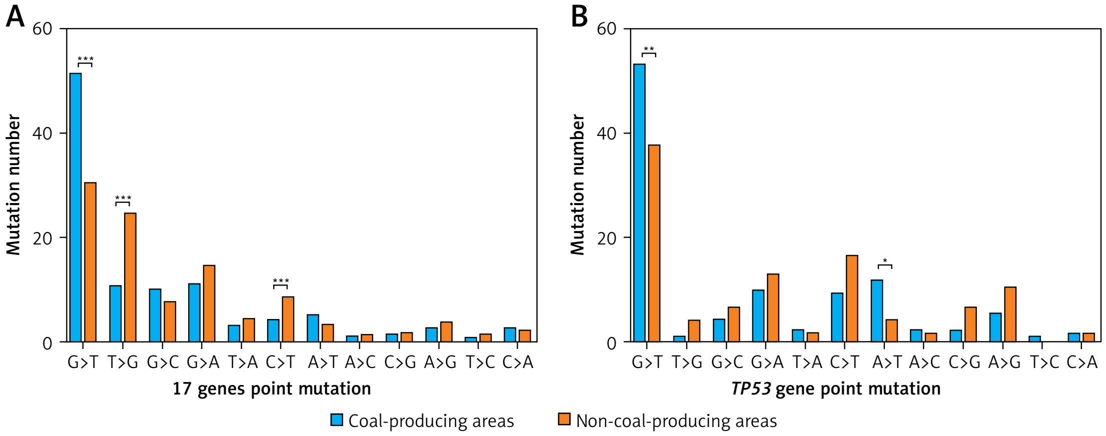
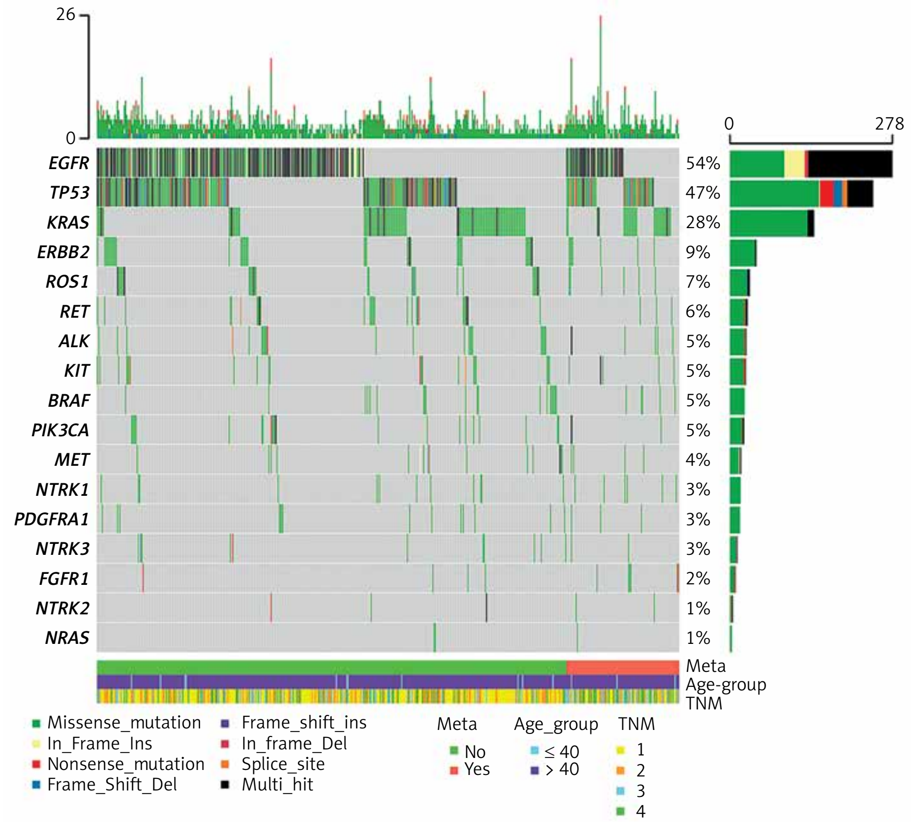


**Figure 6. A** – Frequency of base transition and transversion mutations in the 17 studied genes in NSCLC patients. **B** – Frequency of base transition and transversion variants in *TP53*. ***$p$ < 0.001, **$p$ < 0.01, *$p$ < 0.05

(27.75%), *ERBB2* (8.67%), *ROS1* (6.55%), *RET* (5.78%), *ALK* (5.39%), *KIT* (5.20%), *BRAF* (4.82%), *PIK3CA* (4.62%), *MET* (3.66%), *NTRK1* (3.47%), *PDGFRA* (3.28%), *NTRK3* (2.50%), *FGFR1* (1.93%), *NTRK2* (0.96%) and *NRAS* (0.59%) (Figure 7 A). Six genes, i.e. *KRAS*, *EGFR*, *ROS1*, *NRAS*, *BRAF*, and *ERBB2* studied here, were identified as NSCLC driver mutation genes for patients from the coal-manufacturing regions (Figure 7 C; $p < 0.05$). Our data further confirmed that the *EGFR* somatic mutations were mutually exclusive in NSCLC individuals from the coal-fabricating districts with mutations in the genes *KRAS*, *BRAF*, and *RET* ($p < 0.05$). Furthermore, the same results were obtained for mutations in *TP53*, *ALK*, and *KRAS* ($p < 0.05$). However, *KRAS* gene mutation co-occurred with mutations in the genes *FGFR1* and *RET*, e.g. *RET* and *BRAF* mutations co-occurred. Furthermore, the same results were obtained for the co-occurrence of mutations in the genes *ALK*

A



B

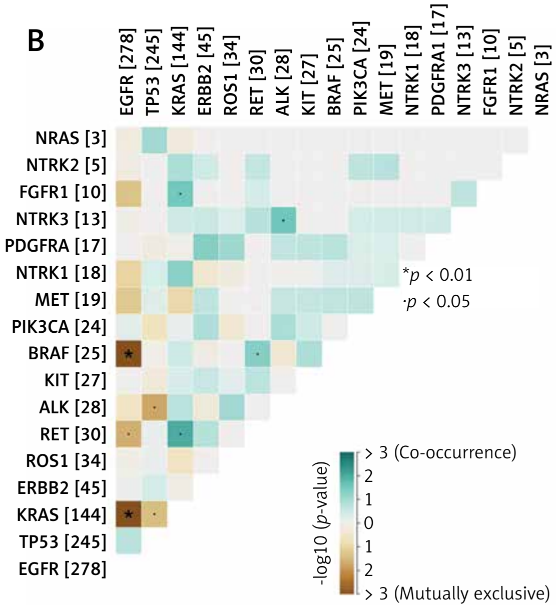
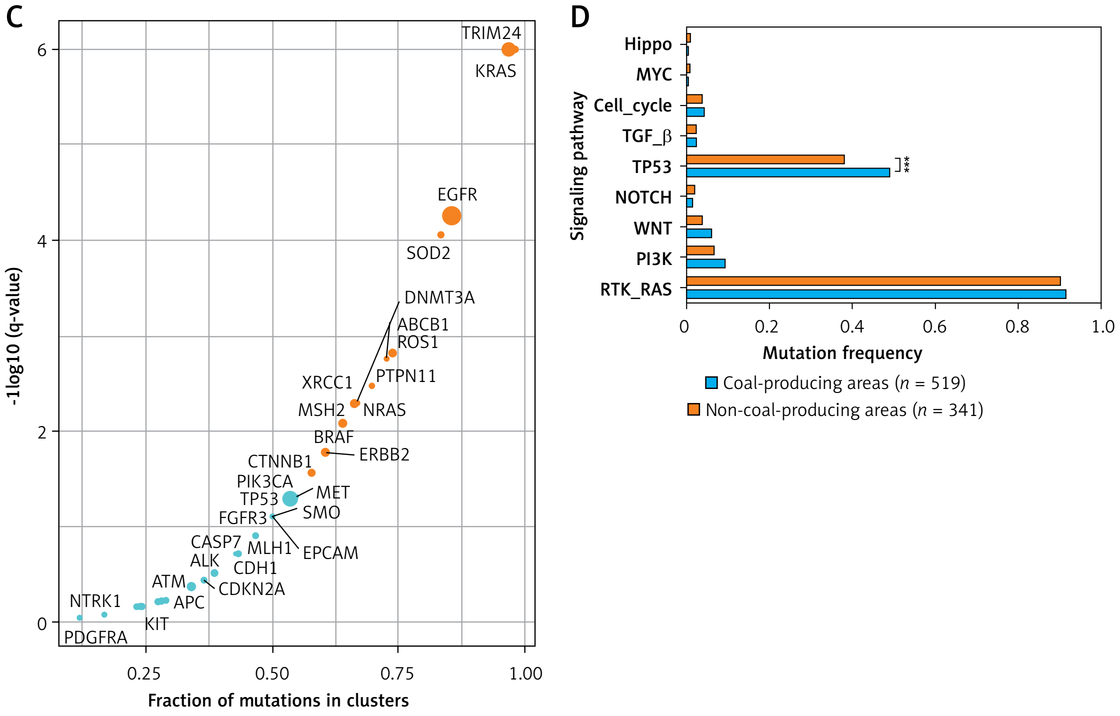


**Figure 7. A** – Outlines of 17 gene mutations in 519 NSCLC patients from Eastern Yunnan coal-producing areas. Clinical features such as tumor metastasis, age, and TNM stage are indicated. **B** – Somatic cell interactions in NSCLC patients from Eastern Yunnan coal-producing area. 17 mutually exclusive or co-occurring genes in NSCLC patients in the Eastern Yunnan coal-producing area are shown. The number after the gene name represents the number of mutations. Fisher's test was used to distinguish major gene couples. $^*p < 0.01$, $p < 0.05$. **C** – NSCLC candidate driver genes in coal-producing regions of Eastern Yunnan. Tumor driver functions were analyzed in the R package Maftools 3. The x-axis shows the proportion of mutations observed in each cluster. Red dots represent significance ($p < 0.05$). Blue dots mean no significance. The size of these points was proportional to the number of clusters found in the gene. **D** – Comparison of gene mutation rates in signaling pathways between coal-manufacturing and non-coal regions in Eastern Yunnan. $^{***}p < 0.001$,$^{**}p < 0.01$, $^*p < 0.05$



and *NTRK3* among individuals from the coal regions ($p$ < 0.05) (Figure 7 B).

Next, we analyzed the Hippo, MYC, cell cycle, TGF-β, TP53, Notch, Wnt, PI3K, and RTK-RAS oncogenic signaling pathways to understand the biological functions of the detected mutated genes and link them with the development of NSCLC in individuals from regions that produce coal. We further compared the genetic mutations identified in patients from non-coal-producing regions. We found that mutations in the TP53 and ATM (ataxia telangiectasia mutated) genes were the most predominant in NSCLC patients from the coal regions and were associated with the TP53 signaling pathway. In contrast, others did not demonstrate any statistically significant differences (Figure 7 D).

## Discussion

### Comparison of clinical characteristics of patients from coal-manufacturing districts in Eastern Yunnan and other parts of China diagnosed with lung tumors

90% of coal mines in China were in Shanxi, Shaanxi, Inner Mongolia, Xinjiang, Gansu, Guizhou, and Ningxia [34]. Moreover, though data showed that the death rates of lung tumors in the coal-producing regions of these provinces were not higher than those in Eastern Yunnan, the epidemiological characteristics of the patients were markedly different [35]. Epidemiological features of NSCLC patients from Shanxi, Shaanxi, Inner Mongolia, Xinjiang, and Ningxia were mainly represented by smoking males with squamous cell carcinomas, aged over 60 years, and workers as occupation, while lung cancer patients of Gansu and Guizhou were primarily smoking males with adenocarcinoma and farmers as an occupation [36].

### Distinctive mutation rates in genes associated with NSCLC in individuals from coal-producing areas of Eastern Yunnan

To address the problem, we performed large-scale exome sequencing of 17 lung cancer genes in 1120 NSCLC patients. Our research indicated that the mutation frequencies of *EGFR*, *ERBB2*, *KRAS*, *TP53*, and *NTRK3* genes in NSCLC patients from the Yunnan coal districts were considerably dissimilar from those in the non-coal regions (Figures 3 and 4). In contrast, the mutation rates in the remaining 12 genes had no significant difference. *KRAS*, *TP53,* and *NTRK3* gene mutation frequencies were more significant than those in non-coal zones (Figure 4). In comparison, the mutation frequencies of *EGFR* and *ERBB2* were substantially lower (Figures 3 A, 4 C). The findings of *EGFR* and *KRAS* gene mutation frequencies in this study were consistent with Zhou *et al.*'s study, which investigated NSCLC patients from Qujing in Yunnan province. In contrast, the *ALK* and *ROS1* mutation rates were inconsistent [15]. The probable reason for these results was that most patients (2146) in the study of Zhou *et al.* were only tested for *ALK* and *ROS1* gene fusions using the ARMS-PCR method, while here we applied NGS exome sequencing for detection of all mutation sites in these genes. Other authors such as Zhang *et al.* detected *EGFR* mutations in 52.68% of tumor samples from Xuanwei lung cancer patients, followed by *TP53* (41.07%) and *KRAS* (7.14%) [14], which was consistent with our results. The screening of the gene *KRAS* in our studies was quite different from Zhang *et al.* [14] as in their study the authors included only 117 never-smoking women with lung adenocarcinoma. Compared with the Yunnan population (except the Qujing population) [15], NSCLC patients in eastern Yunnan coal-producing areas had lower mutation rates of *EGFR* and *ALK* genes and higher mutation rates of *KRAS*, *BRAF*, *PIK3CA*, and *NTRK1* genes (Figure 2). Compared with the Chinese population [37], NSCLC patients in Eastern Yunnan coal-producing areas had lower mutation rates of *EGFR*, *ALK*, *ROS1*, *MET*, and *PIK3CA* genes and higher mutation rates of *KRAS*, *NTRK1*, *TP53*, and *NTRK3* (Figure 2). In general, due to the complex molecular mechanisms of the occurrence and development of NSCLC in the Eastern Yunnan coal-manufacturing regions and due to the lack of large-scale research cohorts, the mutation frequency in the lung tumor-associated genes in the coal-manufacturing Eastern Yunnan was controversial (Table V), and there was no evidence to directly differentiate the abnormality driver gene mutation spectrum for this disease in the studied regions. These data reveal a need for detailed screening of those acceptable mechanisms that underlie the high incidence of lung cancer in these coal Yunnan regions.

### Predominant *EGFR* G719X+S768I subtypes of NSCLC in patients from coal-manufacturing regions of Eastern Yunnan

Our results from this study and other authors' data suggested that the most common mutations in the gene *EGFR* in NSCLC patients from the studied coal districts were G719X+S768I [38]. Interestingly, studies by John *et al.* (2022), including patients from South Korea, Singapore, France, Japan, Greece, Taiwan, and Brazil, reported G719X as rare, accounting for less than 4.8% of all *EGFR* gene variations, followed by S768I with 0.5–2.5% [39]. Our data proved that the *EGFR* gene compound mutation differed from those found in individuals from the non-coal regions (Figures 3 E, 3 F), which suggested that the effectiveness of inhibitors of this particular gene in NSCLC individuals from Yunnan areas could vary considerably [15]. Likewise, tumor samples expressing the combination of the following variants, G719X and S768I had a favorable reaction to afatinib – a blocker of pan-ERBB [40]. Our data demonstrated that afatinib therapy was more promising than EGFR- specific inhibitors in NSCLC individuals from coal regions. However, more clinical data are needed to confirm these cell-based findings. Moreover, in the literature, the therapeutic results of EGFR TKI application in individuals with the same *EGFR* variants varied widely without any detected reason. One hypothesis is that the *EGFR* amplification (*EGFR* amp) mutations have an essential role [41]. Some data confirm that in 72 Hispanic individuals diagnosed with lung carcinoma and with *EGFR* amp mutation, this variant influenced the erlotinib treatment with an OS time of 27.5 months (95% CI: 12.4–42.5) for individuals with the specific mutation *EGFR* amp+L8585R ($p$ < 0.001) [41]. The compound mutation *EGFR* amp+L858R in NSCLC patients from the non-coal regions was predominant, suggesting that individuals from these regions with *EGFR* amp may be more suitable for

**Table V.** Mutation characteristics of the 17 studied genes in Xuanwei/Fuyuan/Qujing lung cancer patients presented in previous studies

| Study | Patients | *n* | Gene | Mutation rate |
|---|---|---|---|---|
| Zhou *et al.* 2021 | NSCLC | 752 | EGFR | *EGFR*: 46.01% |
| | | | | G719X: 23.01% |
| | | | | S768I: 10.24% |
| | | | | G719X + S768I: 19.65% |
| | | | | G719X + L861Q: 21.10% |
| | NSCLC | 265 | KRAS | *KRAS*: 23.02% |
| | | | | G12C: 51.11% |
| | | | | G12D: 6.67% |
| | NSCLC | 600 | ALK | 3.17% |
| | NSCLC | 598 | ROS1 | 0.50% |
| | NSCLC | 259 | BRAF | 1.16% |
| | NSCLC | 257 | RET | 0.78% |
| | NSCLC | 290 | MET | 0.34% |
| | NSCLC | 258 | ERBB2 | 0.39% |

**Table V.** Cont.

| Study | Patients | *n* | Gene | Mutation rate |
|---|---|---|---|---|
| | NSCLC | 259 | NRAS | 0.00% |
| | NSCLC | 257 | PIK3CA | 0.00% |
| Guo 2021 | NSCLC | 146 | EGFR | *EGFR*: 46.60% |
| | | | | G719X: 47.60% |
| | | | | S768I: 24.60% |
| | | | | G719X + S768I: 15.4% |
| | | | | G719X + L861Q: 2.50% |
| | NSCLC | 146 | KRAS | *KRAS*: 26.6% |
| | | | | G12C: 53.8% |
| | NSCLC | 146 | ALK | 1.40% |
| | NSCLC | 146 | BRAF | 3.40% |
| | NSCLC | 146 | RET | 0.70% |
| | NSCLC | 146 | MET | 0.70% |
| | NSCLC | 146 | ERBB2 | 3.40% |
| | NSCLC | 146 | PIK3CA | 8.00% |
| | NSCLC | 146 | PDGFRA | 5.00% |
| | NSCLC | 146 | NTRK3 | 5.00% |
| Ma 2020 | NSCLC | 119 | EGFR | *EGFR*: 39.5% |
| | | | | G719X: 8.51% |
| | | | | S768I: 2.13% |
| | | | | G719X + S768I: 57.45% |
| | | | | G719X + L861Q: 2.13% |
| | NSCLC | 119 | KRAS | 23.53% |
| | NSCLC | 119 | ALK | 1.68% |
| | NSCLC | 119 | ROS1 | 0.84% |
| | NSCLC | 119 | BRAF | 4.20% |
| Yang *et al.* 2016 | NSCLC | 63 | EGFR | *EGFR*: 55.60% |
| | | | | G719X: 14.30% |
| | | | | S768I: 17.10% |
| | | | | G719X + S768I: 17.10% |
| | NSCLC | 63 | KRAS | *KRAS*: 6.30% |
| | | | | G12C: 50.00% |
| Hosgoog III 2013 | NSCLC, never smoking female | 40 | EGFR | *EGFR*: 35.00% |
| | | | | G719X: 46.00% |
| | NSCLC, never smoking female | 40 | KRAS | 15.00% |
| Keohavong 2003 | Lung cancer, never smoking female | 41 | KRAS | 21.90% |
| Chen 2015 | NSCLC | 90 | EGFR | *EGFR*: 56.67% |
| | | | | G719X: 7.84% |
| | | | | S768I: 3.92% |
| | | | | G719X + S768I: 45.10% |
| | | | | G719X + L861Q: 1.96% |
| Chen 2019 | NSCLC | 205 | ALK | 8.80% |
| | | 205 | ROS1 | 2.50% |

erlotinib applications for improving patients' lifespan and for precise selection of individuals who are more suitable for EGFR TKI management.

### *KRAS* G12C gene mutation subtypes in NSCLC patients from the coal-producing Eastern Yunnan regions

In this study, we found that the mutations in the *KRAS* gene mainly occurred in stage I-IIIa NSCLC patients (Table III). The possible reason was that the Eastern Yunnan coal-producing area government provided residents with free CT screening for early lung cancer, which enriched the cohort of patients at these particular stages of the tumors. *KRAS* G12C was diagnosed as the main variant for NSCLC from Eastern Yunnan coal-producing areas, which was in unison with the results obtained in NSCLC individuals from Xuanwei [42]. It was also found that *KRAS* G12C mutations were less abundant in Asian individuals and more common in Asian male patients than in females [43]. The results of our study were consistent with these data. The *KRAS* mutation rate (25.71%) in NSCLC patients in the Eastern Yunnan coal-producing area exceeded the overall mutation rate in whites (13%), ranking first in Asia. Although in the Eastern Yunnan coal-producing areas, most NSCLC patients were females, *KRAS* mutations were predominantly distributed among males (Table III). In addition, *KRAS* G12C was more common in smokers than in non-smokers [43], suggesting that coal burning in these regions led to particle-induced lung cancer, the same as in the case with tobacco smoke-induced lung cancer [25].

Chemotherapy was usually less effective for *KRAS*-mutated NSCLC patients [44]. However, the research on *KRAS*-targeted drugs has made significant progress in the past two years [27]. Sotorasib [28] and adagrasib [27], for example, were introduced as new targeted drugs that effectively treated individuals with *KRAS* G12C. We have also suggested that NSCLC patients from coal-producing areas of Eastern Yunnan could have available target drugs and may profit from directed therapies with sotorasib, adagrasib, and AMG510 [45].

### A typical *TP53* R158L mutation in patients from the coal-producing Eastern Yunnan regions

Our study showed that the variant *TP53* R158L was predominant in patients with NSCLC from coal-producing areas in eastern Yunnan. Furthermore, there is proof that benzo(a)pyrene was the leading cause of the V157F and R158L mutations [46]. Furthermore, studies with yeast functional tests proved that the defective transactivation ability of the V157F and R158L mutants resulted in the loss of *TP53* target gene expression [46]. It was further concluded that *TP53* R158L mutation regulated a new transcriptome of the lungs, which endowed cancer cells with neonatal functions [46]. Regardless of these data, there is no targeted therapy based on *TP53* mutations. Instead, they are commonly applied to identify individuals with dismal prognoses and poor responses to EGFR TKIs [47]. Therefore, we propose *TP53* variants as biomarkers to guide stratified targeted therapy for NSCLC patients and to broaden the understanding of the pathogenesis of NSCLC in the studied regions [48].

### Comparison of detection rates of NSCLC driver gene mutations in patients' plasma and tissues

We compared mutation detection rates for driver genes in plasma and tissue samples. The results showed that the detection rates of *EGFR*, *KRAS*, and *TP53* in plasma were significantly lower than those in tissues (Supplementary Table SV), and *ERBB2* and *NTRK3* detection rates were not significantly different, which indicated that the use of circulating tumor DNA for tumor mutation gene detection had a certain probability of false negatives. It suggested the need for increasing the sequencing depth and precise selection of tissue samples for gene detection. For genes with lower mutation frequencies, there was no difference in the detection rates of plasma and tissue samples. It would be beneficial to expand the sample size further to analyze these genes' mutation frequency and subtype distribution.

### A large number of G>T transition mutations in coal-producing regions

Alexandrov *et al.* (2016) found that tobacco smoking could lead to DNA damage, which was characterized by a G>T transversion mutation in human cancer [49]. Smoky coal may be associated with increased multiple distinct mutational signatures. Our results indicate that lung cancer patients in coal mining areas were also accompanied by a large number of G>T mutations (Figure 6). Smoky coal may be associated with increased multiple distinct mutational signatures. The carcinogenic mechanisms of smoke from smoking and coal combustion might be similar. Polycyclic aromatic hydrocarbons (PAHs) are the main carcinogens found in the emissions from coal burning, which can interact with DNA to form polycyclic aromatic hydrocarbon dihydrodiol epoxide (PAH-DNA adducts). These adducts can combine with the nucleophilic group of the exocyclic amino group in guanine (G), which then pairs with thymine (T) instead of cytosine (C) during the DNA replication process.

### Analyses of NSCLC driver genes and signaling pathways

Hosgood III *et al*. found that *EGFR* and *KRAS* were driver genes in individuals from Xuanwei diagnosed with lung tumors. Both *EGFR* and *KRAS* genes' variations were mutually exclusive [50]. Zhou *et al*. also found that *EGFR*, *KRAS*, and *ALK* were the driver genes in lung cancer patients from Qujing [15]. Zhang *et al*. found that *EGFR*, *TP53*, *RBM10,* and *KRAS* were the driver genes in 117 lung cancer patients from Xuanwei [14]. Another study analyzed 84 lung cancer patients from Xuanwei based on genome-wide sequencing and RNA expression profiles. It concluded that *CREB3L4*, *TRIP13*, and *CCNE2* were potential oncogenes in lung cancer patients from Xuanwei, while *MYC* did not have any effect [51]. A comprehensive study involving molecular profiling of lung adenocarcinoma suggested that *NF1*, *MET*, *ERBB2* and *RIT1* played a driving role in tumorigenesis [52]. Nevertheless, *TP53*, *NFE2L2*, *KEAP1*, *CDKN2A*, and *RB1* mutations were the principal causes of squamous cell carcinoma [53]. Our study identified *KRAS*, *EGFR*, *ROS1*, *NRAS*, *BRAF*, and *ERBB2* as driver genes, whereas other results from our team also highlighted *TRIM24*, *SOD2*, *DNMT3A*, *ABCB1*, *PTPN11*, *XRCC1*, *MSH2* and *CTNNB1* as NSCLC driver genes ($p$ < 0.05) in the studied Eastern Yunnan coal-producing areas. The last cohort of genes was not included in this work (Figure 7 C). In general, we had inconclusive results for the driver genes of NSCLC in the coal-manufacturing Yunnan districts. The possible reasons for this were: (1) the obtained results were from a limited sample size [51]; (2) they were inconsistent when detecting part of the exome and the whole exome [14]; (3) the regional distribution of the included patients was inconsistent in the coal-producing provinces of Xuanwei, Qujing and Eastern Yunnan [15].

Another study by Zhang *et al*. confirmed that RTK-RAS, Wnt, and Notch signaling pathways were significantly affected in individuals with lung tumors from Xuanwei with specific genomic features in driver composition, gene mutation frequency, and oncogenic signaling pathways [14]. Chen *et al*. confirmed that overexpression of miR34a decreased *CDK6* expression by increasing *PTEN*, thereby partially inhibiting the growth and metastasis of YTMLC-90 and XWLC-05 cells, resulting in the subsequent inhibition of the PI3K/AKT pathway [54]. Guo *et al*. found that genetic variants in the Wnt/MAPK/ERBB signaling pathway were predominant in Xuanwei NSCLC individuals [4]. Our data highlight the TP53 pathway as the leading player in disease pathology.

Our study has a significant advantage, which is the big cohort of NSCLC individuals screened with NGS exome sequencing. The obtained data are an excellent reference for future investigations. However, the work has some limitations too. First, this was a retrospective investigation, and although the included data were from 2 prominent medical institutions, the data from Qujing First People's Hospital were the main ones. Second, not all patients had molecular testing with the 17-gene panel. In addition, we did not collect treatment and predictive data for these patients, so it was unclear how the directed therapy would affect them.

In conclusion, our results highlight the unique spectrum of driver genetic variant characteristics in NSCLC individuals from Eastern Yunnan coal-producing areas. Data showed that these individuals had more *EGFR* variants besides the common 19Del and L858R, and they included combinations such as *EGFR* G719X + S768I, G719X + L861X, and L858R + *EGFR* amplification compound mutations. The frequency rate of *KRAS* G12C, *TP53* R158L, and *NTRK3* was more significant, while the ratio of *ERBB2* 20ins was lower. *KRAS*, *EGFR*, *ROS1*, *NRAS*, *BRAF*, and *ERBB2* were also NSCLC driver genes. Activation of the TP53 signaling pathway was the leading carcinogenic cause of NSCLC. These results proved the pathogenic mechanism of NSCLC in patients from Eastern Yunnan coal-producing areas, suggesting that local patients should adopt different treatment strategies.

## Acknowledgments

We are sincerely grateful for the cooperation and unfailing support of the medical, nursing, and paramedical staff of the Department of Thoracic Surgery, Oncology, and Pathology of Qujing First People's Hospital. Service rendered by the Department of Medical Administration of Qujing First People's Hospital for checking the completeness of patients' medical data is also gratefully acknowledged. We would also like to express our sincere thanks to Beijing Life Healthcare Medical Laboratory for the mutant bioinformatic analysis.

This study was supported by the Scientific Research Fund of Yunnan Provincial Department of Science and Technology and Kunming Medical University Joint Project 202201AY070001-214, Hong-Wei Li, 2019FE001(-105), Kun-Hua He, Scientific Research Fund of Yunnan Provincial Department of Education, 2020J0800, Chun-Ju Yang, 2022J0226, Jun-Ling Wang, 2022J0230, Hong-Wei Li, Scientific Research Fund of Qujing First People's Hospital, 2019YJKT03, Jun-Ling Wang.

Jun-Ling Wang and Chun-Ju Yang contributed equally to this work.

## Conflict of interest

The authors declare no conflict of interest.

## References

1. Ferlay J, Colombet M, Soerjomataram I, et al. Cancer statistics for the year 2020: an overview. Int J Cancer 2021; 149: 778-89.
2. Wu F, Wang L, Zhou C. Lung cancer in China: current and prospect. Curr Opin Oncol 2021; 33: 40-6.
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019; 69: 7-34.
4. Guo G, Li G, Liu Y, et al. Next-generation sequencing reveals high uncommon EGFR mutations and tumor mutation burden in a subgroup of lung cancer patients. Front Oncol 2021; 11: 628.
5. Li J, He J, Zhang YS, et al. Survival in lung cancer among female never-smokers in rural Xuanwei and Fuyuan counties in Eastern Yunnan Province, China. Chinese J Lung Cancer 2019; 22: 477-87.
6. Zhang Y, Meliefste K, Hu W, et al. Household air pollution from, and fuel efficiency of, different coal types following local cooking practices in Xuanwei, China. Environ Pollut 2021; 290: 117949.
7. Mumford J, He X, Chapman R, et al. Lung cancer and indoor air pollution in Xuan Wei, China. Science 1987; 235: 217-20.
8. Xiao Y, Shao Y, Yu X, Zhou G. The epidemic status and risk factors of lung cancer in Xuanwei City, Yunnan Province, China. Front Med 2012; 6: 388-94.
9. Zhang M, Shao L, Jones T, Hu Y, Adams R, BéruBé K. Hemolysis of PM10 on RBCs in vitro: an indoor air study in a coal-burning lung cancer epidemic area. Geoscience Frontiers 2022; 13: 101176.
10. Li X, Dai S, Nechaev VP, et al. Mineral matter in the late permian C1 coal from Yunnan Province, China, with emphasis on its origins and modes of occurrence. Minerals 2020; 11: 19.
11. Hosgood III HD, Sapkota AR, Rothman N, et al. The potential role of lung microbiota in lung cancer attributed to household coal burning exposures. Environ Mol Mutagen 2014; 55: 643-51.
12. Keohavong P, Lan Q, Gao WM, et al. K-ras mutations in lung carcinomas from nonsmoking women exposed to unvented coal smoke in China. Lung Cancer 2003; 41: 21-7.
13. Keohavong P, Lan Q, Gao WM, et al. Detection of p53 and K-ras mutations in sputum of individuals exposed to smoky coal emissions in Xuan Wei County, China. Carcinogenesis 2005; 26: 303-8.
14. Zhang H, Liu C, Li L, et al. Genomic evidence of lung carcinogenesis associated with coal smoke in Xuanwei area, China. Natl Sci Rev 2021; 8: nwab152.
15. Zhou Y, Ge F, Du Y, et al. Unique profile of driver gene mutations in patients with non-small-cell lung cancer in Qujing city, Yunnan province, southwest China. Front Oncol 2021; 11: 644895.
16. Imperial R, Toor OM, Hussain A, Subramanian J, Masood A. Comprehensive pancancer genomic analysis reveals (RTK)-RAS-RAF-MEK as a key dysregulated pathway in cancer: its clinical implications. Semin Cancer Biol 2019; 54: 14-28.
17. Tan AC. Targeting the PI3K/Akt/mTOR pathway in non-small cell lung cancer (NSCLC). Thorac Cancer 2020; 11: 511-8.
18. Stewart DJ. Wnt signaling pathway in non–small cell lung cancer. J Natl Cancer Inst 2014; 106: djt356.
19. Mogi A, Kuwano H. TP53 mutations in nonsmall cell lung cancer. J Biomed Biotechnol 2011; 2011: 583929.
20. Li J, Shen C, Wang X, et al. Prognostic value of TGF-β in lung cancer: systematic review and meta-analysis. BMC Cancer 2019; 19: 691.
21. Shcherba M, Liang Y, Fernandes D, Perez-Soler R, Cheng H. Cell cycle inhibitors for the treatment of NSCLC. Expert Opin Pharmacother 2014; 15: 991-1004.
22. Massó-Vallés D, Beaulieu ME, Soucek L. MYC, MYCL, and MYCN as therapeutic targets in lung cancer. Expert Opin Ther Targets 2020; 24: 101-14.
23. Wang Y, Ding W, Chen C, Niu Z, Pan M, Zhang H. Roles of Hippo signaling in lung cancer. Indian J Cancer 2015; 52 (suppl 1): p1-5.
24. Galluzzo P, Bocchetta M. Notch signaling in lung cancer. Expert Rev Anticancer Ther 2011; 11: 533-40.
25. DeMarini DM, Landi S, Tian D, et al. Lung tumor KRAS and TP53 mutations in nonsmokers reflect exposure to PAH-rich coal combustion emissions. Cancer Res 2001; 61: 6679-81.
26. Chen S, Zhou Y, Chen Y, et al. Specific microRNA expression profiles of lung adenocarcinoma in Xuanwei region and bioinformatic analysis for predicting their target genes and related signaling pathways. J Southern Med Univ 2017; 37: 238-44.
27. Jänne PA, Riely GJ, Gadgeel SM, et al. Adagrasib in non-small-cell lung cancer harboring a KRASG12C mutation. N Engl J Med 2022; 387: 120-31.
28. Skoulidis F, Li BT, Dy GK, et al. Sotorasib for lung cancers with KRAS p. G12C mutation. N Engl J Med 2021; 384: 2371-81.
29. Wang JL, Fu YD, Gao YH, et al. Unique characteristics of G719X and S768I compound double mutations of epidermal growth factor receptor (EGFR) gene in lung cancer of coal-producing areas of East Yunnan in Southwestern China. Genes Environment 2022; 44: 17.
30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30: 2114-20.
31. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001; 29: 308-11.
32. Consortium GP. A global reference for human genetic variation. Nature 2015; 526: 68-74.
33. Cao Y, Gao H. Prevalence and causes of air pollution and lung cancer in Xuanwei City and Fuyuan County, Yunnan Province, China. Front Med 2012; 6: 217-20.
34. Hussain R, Luo K, Chao Z, Xiaofeng Z. Trace elements concentration and distributions in coal and coal mining wastes and their environmental and health impacts in Shaanxi, China. Environ Sci Pollut Res Int 2018; 25: 19566-84.
35. Li J, Ran J, Chen LC, et al. Bituminous coal combustion and Xuan Wei Lung cancer: a review of the epidemiology, intervention, carcinogens, and carcinogenesis. Arch Toxicol 2019; 93: 573-83.
36. Cao M, Chen W. Epidemiology of lung cancer in China. Thoracic Cancer 2019; 10: 3-7.
37. Sun S, Du W, Sun Q, et al. Driver gene alterations profiling of Chinese non-small cell lung cancer and the effects of co-occurring alterations on immunotherapy. Cancer Med 2021; 10: 7360-72.
38. Lv L, Liu Z, Liu Y, et al. Distinct EGFR mutation pattern in patients with non-small cell lung cancer in Xuanwei region of China: a systematic review and meta-analysis. Front Oncol 2020; 10: 519073.

39. John T, Taylor A, Wang H, Eichinger C, Freeman C, Ahn MJ. Uncommon EGFR mutations in non-small-cell lung cancer: a systematic literature review of prevalence and clinical outcomes. Cancer Epidemiol 2022; 76: 102080.
40. Qian J, ye X, huang A, et al. Afatinib 30 mg in the treatment of common and uncommon EGFR-mutated advanced lung adenocarcinomas: a retrospective, single-center, longitudinal study. J Thorac Dis 2022; 14: 2169-77.
41. Nicoś M, Wojas-Krawczyk K, Krawczyk P, et al. Assessment of EGFR gene mutations in circulating free DNA in monitoring of response to EGFR tyrosine kinase inhibitors in patients with lung adenocarcinoma. Arch Med Sci 2020; 16: 1496-500.
42. Liu Y, Liang J, Zhao J, et al. The impact of genomic mutational status and correlation with tumor mutation burden in non-small cell lung cancer of Xuanwei, Yunnan Province, China. Cancer Res 2020; 80: 5885.
43. Reita D, Pabst L, Pencreach E, et al. Direct targeting KRAS mutation in non-small cell lung cancer: focus on resistance. Cancers 2022; 14: 1321.
44. Wood K, Hensing T, Malik R, Salgia R. Prognostic and predictive value in KRAS in non-small-cell lung cancer: a review. JAMA Oncol 2016; 2: 805-12.
45. Liu SY, Sun H, Zhou JY, et al. Clinical characteristics and prognostic value of the KRAS G12C mutation in Chinese non-small cell lung cancer patients. Biomark Res 2020; 8: 22.
46. Barta JA, Pauley K, Kossenkov AV, McMahon SB. The lung-enriched p53 mutants V157F and R158L/P regulate a gain of function transcriptome in lung cancer. Carcinogenesis 2020; 41: 67-77.
47. Hou H, Qin K, Liang Y, et al. Concurrent TP53 mutations predict poor outcomes of EGFR-TKI treatments in Chinese patients with advanced NSCLC. Cancer Manag Res 2019; 11: 5665.
48. Canale M, Andrikou K, Priano I, et al. The role of TP53 mutations in EGFR-mutated non-small-cell lung cancer: clinical significance and implications for therapy. Cancers 2022; 14: 1143.
49. Alexandrov LB, Ju YS, Haase K, et al. Mutational signatures associated with tobacco smoking in human cancer. Science 2016; 354: 618-22.
50. Hosgood III HD, Pao W, Rothman N, et al. Driver mutations among never smoking female lung cancer tissues in China identify unique EGFR and KRAS mutation pattern associated with household coal burning. Respir Med 2013; 107: 1755-62.
51. Network CGAR. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014; 511: 543-50.
52. Network CGAR. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012; 489: 519-25.
53. Zhang Y, Xue Q, Pan G, et al. Integrated analysis of genome-wide copy number alterations and gene expression profiling of lung cancer in Xuanwei, China. PLoS One 2017; 12: e0169098.
54. Chen Y, Hou C, Zhao LX, et al. The association of microRNA-34a with high incidence and metastasis of lung cancer in Gejiu and Xuanwei Yunnan. Front Oncol 2021; 11: 619346.