The utility of sputum supernatant as an alternative liquid biopsy specimen for next-generation sequencing-based genomic profiling

Ling Qin
Department of Respiratory Medicine, Xiangya Hospital, Central South University, Changsha, 410008, China
Ting Guo
Department of Respiratory Medicine, Xiangya Hospital, Central South University, Changsha, 410008, China
Huaping Yang
Department of Respiratory Medicine, Xiangya Hospital, Central South University, Changsha, 410008, China
Pengbo Deng
Department of Respiratory Medicine, Xiangya Hospital, Central South University, Changsha, 410008, China
Qihua Gu
Department of Respiratory Medicine, Xiangya Hospital, Central South University, Changsha, 410008, China
Chi Liu
Department of Physiology, Xiangya School of Medicine, Central South University, Changsha, 410078, China
Mengping Wu
Department of Physiology, Xiangya School of Medicine, Central South University, Changsha, 410078, China
Analyn Lizaso
Burning Rock Biotech, Guangzhou, 510300, China
Bing Li
Burning Rock Biotech, Guangzhou, 510300, China
Sa Zhang
Burning Rock Biotech, Guangzhou, 510300, China
Zhiqiu Chen
Burning Rock Biotech, Guangzhou, 510300, China
Chengping Hu (huchengp28@csu.edu.cn)

Department of Respiratory Medicine, Xiangya Hospital, Central South University, Changsha, 410008, China https://orcid.org/0000-0002-1285-8579

Research

Keywords: sputum, alternative specimen, non-small-cell lung cancer, molecular testing, tumor DNA

Posted Date: November 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-108555/v1

Abstract

Background

Comprehensive mutation profiling has become a standard clinical practice in the management of advanced lung cancer. In addition to tissue and plasma, other body fluids are also being actively explored as alternative sources of tumor DNA. In this study, we investigated the potential of induced sputum obtained from patients with non-small cell lung cancer (NSCLC) for mutation profiling.

Methods

Capture-based targeted sequencing was performed on matched tumor, plasma, and induced sputum samples of 41 treatment-naïve patients with NSCLC using 168 gene panel.

Results

Comparative analysis on the mutation detection using matched tumor sample as reference revealed detection rates of 76.9% for plasma, 72.4% for sputum-supernatant, and 65.7% for sputum-sediment samples. Plasma, sputum-supernatant, and sputum-sediment achieved positive predictive values of 73.3%, 80.4%, and 55.6% and sensitivities of 50.0%, 36.9%, 31.3%, respectively, relative to tumor samples for 168 genes. Sputum-supernatants had significantly higher concordance rates relative to matched tumor samples (69.2% vs 37.8%; $P = 0.031$) and maximum allelic fraction ($P < 0.001$) than its matched sputum-sediments. Sputum-supernatants had comparable detection rates (71.4% vs. 67.9%; $P = 1$) but with significantly higher maximum allelic fraction than their matched plasma samples ($P = 0.003$). Furthermore, sputum-supernatant from smokers had a significantly higher maximum allelic fraction than sputum-supernatant from non-smokers ($P = 0.021$).

Conclusions

Our study demonstrated that supernatant fraction from induced sputum is a better sampling source than its sediment and has comparable performance as plasma samples. Induced sputum from NSCLC patients could serve as an alternative media for next-generation sequencing-based mutation profiling.

Introduction

In recent years, targeted therapy has transformed the treatment and management of patients with lung cancer into a more personalized approach [1–4]. Personalized medicine mainly relies on the accurate detection of actionable mutations and identification of patients who would benefit from targeted therapy. Molecular testing with either conventional single-gene assays, including amplification refractory mutation system and direct sequencing, or next-generation sequencing has considered tissue biopsy as the gold

standard specimen; however, the biopsy of primary tumor tissues requires invasive sampling procedures [5–8]. The rapid technological advancements in molecular assays have led to the improvement in the diagnostic accuracy in detecting tumor-specific somatic mutations from DNA extracted from samples collected through minimally invasive procedures including liquid biopsy specimens [5–8]. Since blood is easily accessible and permits repeated sampling, plasma-based mutation profiling using next-generation sequencing (NGS) has now become widely used in clinical oncology practice for diagnosis, treatment monitoring, and assessment of mechanisms of drug resistance [8]. Malignant non-blood biological fluids that are in close contact with the tumors, including pleural effusion, ascites, and cerebrospinal fluid, are gaining attention as specimens for molecular testing due to their effectiveness in reflecting tumor genomic profiles [9–12]. Meanwhile, other easily accessible biological fluids including sputum that likely contain tumor-derived DNA are being actively explored for mutation detection [9, 13–17]. Sputum has been explored in the detection of genetic and epigenetic alterations in patients with various stages of lung cancer and in cancer-free chronic smokers who are at higher risk of developing lung cancer [9, 13, 15–27]. These studies consistently demonstrate that induced sputum samples contain circulating cell-free DNA derived from the lungs and lower respiratory tract and are attractive candidate liquid biopsy media for lung cancer diagnosis [13, 15–27]. However, very few studies have explored its use in next-generation sequencing (NGS)-based genomic profiling. In this study, we investigated the potential of induced sputum obtained from treatment-naïve patients with non-small-cell lung cancer (NSCLC) as a medium for comprehensive mutation profiling. To achieve this aim, we performed capture-based targeted NGS on matched tumor, plasma, and sputum from 41 patients. We then performed a comparative analysis to identify the optimal fraction from induced sputum specimens that could yield better mutation detection rate and to establish the utility of induced sputum as an alternative source of tumor DNA for mutation profiling by comparing with matched tumor and plasma samples.

Patients And Methods

Patient selection

Treatment-naïve patients who were diagnosed with locally-advanced to advanced NSCLCs from our hospital between October 2018 and June 2019 were included in this study. The main inclusion criteria for this cohort were as follows: 1. Pathologically confirmed NSCLC; 2. Locally-advanced to advanced disease stage; 3. Have not received prior systemic therapy; 4. Submitted matched tissue, plasma, and sputum samples; and 5. Provided consent for the use of their clinical and molecular data. This study was approved by the Medical Ethics Committee of Xiangya Hospital Central South University (approval number: 201911306) and performed in accordance with the Declaration of Helsinki. Written informed consent was provided by all patients for the use of their biological samples.

Collection and preparation of sputum samples

Before sputum induction, the patients were administered with 400 μg albuterol by inhalation with their lung function measured by spirometry as forced expiratory volume (FEV_1) before and 10 minutes after

albuterol inhalation. Sputum induction was performed with hypertonic saline (4.5%) inhalation for 15 minutes for patients with $FEV_1 \geq 1L$ and isotonic saline (0.9%) was used for patients with $FEV_1 < 1L$. An aliquot of the expectorate was reserved for gene sequencing, and another aliquot was analyzed cytologically for the presence of cancer cells or heterogeneous cell nuclei.

The induced sputum samples (~ 8 mL) collected from each of the patients were treated with 0.25% pancreatin at 37 °C with agitation at 660 rpm for 30 minutes. The digestion condition was adjusted according to the viscosity of the sputum to a maximum of 1:2.5 sputum to pancreatin ratio and/or extension of incubation time until complete liquefaction of the sputum sample. The digest was centrifuged at 3,000 x g for 10 minutes at 4 °C. The sediment fraction was reconstituted in the remaining 1 mL of supernatant and stored at -80 °C until DNA extraction. Meanwhile, the supernatant fractions were transferred to fresh tubes, centrifuged at 16,000 x g for 10 minutes at 4 °C to remove cell debris, aliquoted into fresh tubes, and stored at -80 °C until DNA extraction.

Collection and preparation of tissue and plasma samples

Whole blood samples (approximately 10 mL) and tissue biopsy samples were collected from each of the patients. Lung tumor tissue samples were obtained by biopsy and processed into formalin-fixed, paraffin-embedded (FFPE) cell blocks for storage. Plasma was separated from blood samples collected in EDTA-treated tubes by centrifugation (1,500 x g, 4 °C, 10 minutes). Plasma fractions were transferred into fresh tubes, centrifuged to remove cell debris (16,000 x g, 4 °C, 10 minutes), aliquoted into fresh tubes, and stored at -80 °C until DNA extraction.

DNA isolation and capture-based targeted DNA sequencing

DNA isolation and targeted sequencing were performed at Burning Rock Biotech, a College of American Pathologist (CAP)-accredited/Clinical Laboratory Improvement Amendments (CLIA)-certified commercial clinical laboratory, according to optimized protocols as described previously [10, 11]. DNA was extracted from FFPE tissue biopsy samples and sputum-sediment samples using appropriate QIAamp DNA tissue kits (Qiagen, Hilden, Germany). Circulating cell-free DNA (cfDNA) was extracted from 4–5 ml of plasma samples, and 15 mL sputum-supernatant samples using a QIAamp Circulating Nucleic Acid kit, according to the manufacturer's standard protocol (Qiagen, Hilden, Germany). DNA quality was assessed using Qubit 3.0 fluorimeter with the dsDNA high-sensitivity assay kit (Life Technologies, CA, USA). Tissue DNA was sheared using M220 ultra-focused sonicator (Covaris, MA, USA). Fragments between 200–400 bp from the sheared tissue DNA and cfDNA were purified (Agencourt AMPure XP Kit, Beckman Coulter, CA, USA) and hybridized with capture probes baits using a commercial panel consisting of 168 genes. After purification and amplification of the capture-hybrid library, the quality and the size of the fragments were assessed with high sensitivity DNA kit using the 2100 Bioanalyzer instrument (Agilent Technologies, CA, USA). Indexed samples were sequenced on Nextseq500 (Illumina, Inc., CA, USA) with paired-end reads and average sequencing depth of 1,000· for tumor samples and 10,000· for plasma, sputum supernatant, and sputum sediment samples.

NGS data analysis

The NGS sequence data from the patients were mapped to the reference human genome (hg19) using Burrows-Wheeler Aligner version 0.7.10 [28]. Local alignment optimization, duplication marking, and variant calling were performed using the Genome Analysis Tool Kit version 3.2 [29], and VarScan version 2.4.3 [30]. Variants were filtered using the VarScan fpfilter pipeline, loci with depth less than 100 were filtered out. Base-calling in plasma and tissue samples required at least 8 supporting reads for single nucleotide variations (SNV) and 2 and 5 supporting reads for small insertion-deletion variations (Indel), respectively. Variants with population frequency over 0.1% in the ExAC, 1000 Genomes, dbSNP, or ESP6500SI-V2 databases were grouped as single nucleotide polymorphisms and excluded from further analysis. Remaining variants were annotated with ANNOVAR (2016-02-01 release) [31] and SnpEff version 3.6 [32]. Analysis of DNA translocation was performed using Factera version 1.4.3 [33]. Copy number variations (CNV) were analyzed based on the depth of coverage data of capture intervals. Coverage data were corrected against sequencing bias resulting from GC content and probe design. The average coverage of all captured regions was used to normalize the coverage of different samples to comparable scales. Copy number was calculated based on the ratio between the depth of coverage in tumor samples and average coverage of an adequate number ($n > 50$) of samples without copy number variations as references per capture interval. CNV is called if the coverage data of the gene region was quantitatively and statistically significant from its reference control. The cut-off for CNVs is 1.5 for copy number deletion and 2.64 for copy number amplifications.

Statistical analysis

The detection rate was defined as the proportion of samples detected with mutations relative to the total number of samples of the same sample type. Maximum allelic fraction (maxAF) was defined as the maximum fraction of the mutant allele detected from a sample regardless of mutation or gene. The concordance rate was defined as the proportion of the total number of mutations detected from one sample type relative to the reference sample type. Statistical analyses were performed using the Fisher's exact test, paired Student's *t*-test, Wilcoxon signed-rank test, as applicable, in R software. P-value of less than 0.05 was considered statistically significant.

Results

Patient characteristics

A total of 41 treatment-naïve patients with NSCLC consented to participate in the study. The cohort included a majority of males (68.3%, 28/41) with a median age of 65, ranging from 36 to 81 years. A majority (78.0%; 32/41) was diagnosed with lung adenocarcinoma, 7 patients were diagnosed with squamous cell carcinoma, and 2 patients with neuroendocrine tumor. A majority of the patients (92.7%; 38/41) had locally-advanced to advanced disease (stage IIIB to IV), the remaining 3 patients had stage IIIA disease. Table 1 lists the baseline clinicopathologic features of the cohort.

Table 1
Baseline clinicopathologic features of the cohort

Clinicopathologic features	n = 41; n(%)
Age	
Median(range)	65(36 ~ 81)
Gender	
Male	28 (68.3%)
Female	13 (21.7%)
Smoking status	
Smoker	27 (65.9%)
Never smoker	14 (34.1%)
Histology	
Lung adenocarcinoma	32 (78.0%)
Lung squamous cell carcinoma	7 (17.1%)
Other NSCLC	2 (4.9%)
Degree of cellular differentiation of sputum cytology	
Low	20 (48.9%)
Medium	9 (22.0%)
High	7 (17.1%)
NA	5 (12.2%)
Location of primary tumor	
Central	22 (53.7%)
Peripheral	19 (46.3%)
Stage	
≤ IIIA	3 (7.3%)
IIIB-IIIC	7 (17.1%)
IV	31 (75.6%)
Abbreviations: NA, not applicable; NSCLC, non-small-cell lung cancer	

Sample distribution and quality control of the samples

Matched tumor, blood, and sputum were collected from all the patients; however, some samples were unavailable for mutation profiling due to insufficient sample volume (n = 23), and inadequate DNA quality for library construction (n = 2). Of them, tumor samples were available for 38 patients. Blood samples were available for 39 patients. Sputum supernatant and sputum sediment samples were available from 29 and 35 patients, respectively. Table S1 summarizes the distribution of the cohort according to sample type.

DNA was extracted from a total of 141 available samples, with an average DNA yield of 1571.8 ng for tumor, 121.9 ng for plasma, 2766.0 ng for sputum supernatant, and 8144.5 ng for sputum sediment (Figure S1A). The distribution of the library complexity and insert size of all the sequenced samples revealed similar distribution for tumor biopsy samples and sputum sediments, which was distinct from plasma and sputum supernatant samples (Figure S1B). A majority of the tumor and sputum sediment samples had insert sizes between 150 to 250 base pairs; while most of the plasma and sputum supernatant samples had insert sizes between 150 to 175 base pairs (Figure S1B). The sequencing achieved a median depth of 1,275· for tumor samples, 16,326· for plasma samples, 10,549· for sputum supernatant, and 16,660· for sputum sediment (Figure S1C).

We then compared the detection rates and maximum allelic fraction (maxAF) of plasma, sputum supernatant, and sputum sediment samples using matched tumor as reference. In general, relative to tumor samples, the detection rates for the 168-gene panel were 76.9% for plasma (n = 39), 72.4% for sputum supernatant (n = 29), and 65.7% for sputum sediment samples (n = 35) (Fig. 1A). Meanwhile, the detection rates for the 8 oncogenic driver genes and *TP53* (9 genes) were 71.8% for plasma, 62.1% for sputum supernatant, and 51.4% for sputum sediment samples (Fig. 1B). Using tumor tissue samples as reference, sputum supernatant, sputum sediment, and plasma samples achieved a positive predictive value (PPV) of 80.4%, 55.6%, and 73.3% and sensitivity of 36.9%, 31.3%, and 50%, respectively, when considering the 168-genes similarly included in the panels. Meanwhile, all the samples achieved a PPV of 85.7%, 86.7%, and 90.9% and sensitivity of 50.0%, 39.4%, and 51.3%, respectively, when only considering the 8 classic oncogenic genes. The maxAF was significantly higher in tumor samples as compared to plasma ($P < 0.001$), sputum supernatant ($P < 0.001$), and sputum sediment ($P < 0.001$) in either the 168 genes or the 9 genes (Fig. 1C; Fig. 1D). However, maxAF was similar in plasma and sputum supernatant samples in either the 168 genes ($P = 0.81$; Fig. 1C) or the 9 genes ($P = 0.55$; Fig. 1D).

These data indicate that the DNA extracted from sputum supernatant and sputum sediment samples have adequate quality and sufficient quantity for NGS-based genomic profiling.

Mutation detection in sputum supernatant and sediment

Mutation profiling of sputum supernatant samples detected a total of 106 mutations in 52 genes from 21 patients, revealing a detection rate of 72.4% (Fig. 2A). Of these mutations, 81 were missense mutations, 4 were Indels, 6 were frameshift, 4 were splice-site variants, 2 were stop-gained, 3 were CNVs, and 6 were genomic rearrangements. The most frequent mutations detected from sputum supernatants were *TP53* (31.0%), *EGFR* (13.8%), *KRAS* (13.8%), and *ALK* (13.8%). Among the eight classic NSCLC oncogenic driver

genes, actionable mutations were detected in 14 patients, including *EGFR* mutations (p.L858R, n = 1; p. E746_A750del n = 3), *EML4-ALK* fusions (n = 4), *KRAS* G12V/C/D/Q61H mutations (n = 4), *ERBB2* A622S mutation (n = 1), and *CD74-ROS1* fusion (n = 1). No mutations were detected in *BRAF*, *MET*, and *RET* from our cohort.

Meanwhile, a total of 276 mutations in 75 genes were detected from matched sputum sediment samples from 14 patients, revealing a detection rate of 60.9%. Detection rates were comparable between sputum supernatant samples and its matched sputum sediment (78.3% vs 65.2%; $P = 0.51$). By considering the overall number of mutations detected from 23 patients with both sample types using the 168 gene panel, the mutations detected from both sputum supernatant and sediment samples were 47.8% concordant. Figure 2B illustrates the mutations detected in both or either of the sputum supernatant and/or sediment samples from the 23 patients. Based on the distribution of mutation types among the 168 genes, fusions achieved a concordance rate of 80.0%, SNVs and Indels had a concordance rate of 30.4%, while CNVs had very low detection rate, with only 1 sputum supernatant sample and no sputum sediment samples were detected with CNV (Table S2). Actionable mutations were detected from the matched sputum supernatant and sediment from 8 patients, including *EML4-ALK* fusions (n = 3), *EGFR* exon 19 deletion (n = 2), *KRAS* G12C/Q61H mutations (n = 2), and *CD74-ROS1* fusion (n = 1). Moreover, a significantly higher maxAF was observed in sputum supernatant samples than their corresponding sediment samples ($P < 0.001$). The median maxAF in sputum supernatant was 1.26% (range: 0.0%-9.2%); while the median maxAF in sputum sediment was 0.79% (range 0.0%-14.1%).

These data suggest that DNA extracted from both sputum supernatant and sediment could be utilized for mutation profiling; however, the abundance of mutations is significantly higher in sputum supernatant than its corresponding sediment samples.

Concordance of sputum supernatant and sputum sediment with the matched tumor sample

Figure 3A illustrates the mutation profile of tumor tissue samples. Comparing the mutation profile of 26 patients with both the sputum supernatant and tumor samples, 41 mutations were detected from both samples (Fig. 3B), revealing a concordance rate of 69.2%. The detection of fusions from both sputum supernatant and matched tumor samples were highly concordant, achieving 75.0%. SNVs and Indels were only 34.0% concordant, while CNVs were only detected from 2 sputum supernatant samples resulting in a concordance rate of 7.7% (Table S3). It is worth noting that actionable fusions including *EML4-ALK* and *CD74-ROS1* can be detected in sputum supernatant samples with a high concordance of 83.3% (5/6) relative to tumor samples (Table S3). Meanwhile, analysis of the mutation profile of 32 patients with both sputum sediment and tumor samples demonstrated the detection of 45 mutations from both samples (Fig. 3C), revealing a concordance rate of 37.8%. The concordance rates relative to tumor samples of sputum supernatants were significantly higher than sputum sediments (69.2% vs. 37.8%; $P = 0.031$).

These data indicate that sputum supernatant samples are better than their sediment fraction in reflecting tumor-related mutations.

Concordance of sputum supernatant with their matched plasma sample

Furthermore, Fig. 4A illustrates the mutation profile of plasma samples. Comparing the mutation profile of 28 patients with both the sputum supernatant and plasma samples, 32 mutations were detected from both samples, revealing a concordance rate of 53.6% (Fig. 4B). Sputum supernatant and plasma samples had comparable detection rates (71.4% vs. 67.9%; $P = 1$) but significantly higher median allelic fraction than its matched plasma sample ($P = 0.034$). Sputum supernatant and matched plasma samples were highly concordant in detecting fusions, achieving 83.5%. SNVs and Indels were 31.0% concordant, while CNVs were only detected from 3 sputum supernatant samples resulting in a concordance rate of 7.7% (Table S4).

These data indicate that sputum supernatant samples have comparable detection rates as plasma samples, suggesting its utility as an alternative sample for comprehensive genomic profiling, particularly for non-CNV mutations.

Sputum supernatant from smokers and non-smokers

Next, we investigated the clinical factors that are associated with a better detection rate for induced sputum samples. All the clinical features analyzed including age, gender, disease stage, smoking history, and histology, were not statistically correlated with mutation detection rate in either sputum supernatant or sediment samples (Table S5). However, significantly higher maxAF ($P = 0.018$; Table S5) and AF ($P = 0.021$; Figure S2) were observed in the sputum supernatant samples from smokers than from non-smokers.

These data suggest that sputum supernatant samples, particularly from smokers, could provide valuable genetic information.

Case Vignette

Of the 13 patients evaluable for sputum cytology, 38.5% (5/13) were identified with malignant cells. Figure 5 illustrates the apparent heterogeneous cell nuclei in the sputum cytology of samples from three patients. The three patients had stage IVA-IVB lung cancer of various histologies.

Figure 5A was the sputum cytology findings for Patient P23, who was a 56-year-old female non-smoker diagnosed with stage IVA pulmonary sarcomatoid carcinoma. *TP53* c.993 + 1G > C and *PIK3CA* p.H1047R were detected from both the matched tissue and sputum supernatant samples but were undetected from the sputum sediment sample. With no actionable mutations detected, she received pemetrexed, carboplatin, and bevacizumab as front-line therapy.

Figure 5B was the sputum cytology findings for Patient P27 was a 55-year-old male smoker diagnosed with stage IVB well-differentiated squamous cell lung carcinoma. *EGFR* exon 19 deletion E746_A750 (19del) was detected from his matched tissue, plasma, and sputum supernatant samples, but was undetected from the sputum sediment sample. His disease achieved partial response with cisplatin, paclitaxel, and pembrolizumab as the front-line regimen. Upon detection of *EGFR* 19del, he received icotinib as the second-line regimen.

Figure 5C was the sputum cytology findings for Patient P41was a 66-year-old female non-smoker diagnosed with stage IVB poorly-differentiated lung adenocarcinoma. *EGFR* 19del was detected from all her matched samples, including tissue, plasma, sputum supernatant, and sputum sediment samples. In addition to the *EGFR* 19del, *EGFR* copy number amplification and *TP53* c.783-1G > T were detected from her tissue samples, which were undetected in the other sample types. She received icotinib as the front-line regimen and achieved complete response.

Discussion

Exfoliative cytology, which involves the microscopic study of the cells exfoliated from tumors in various samples including saliva, sputum, and bronchial secretions, has been well-established non-invasive procedure in providing diagnostic information [34, 35]. The diagnostic accuracy of sputum cytology for lung cancer diagnosis has been demonstrated to achieve a specificity of 90%, sensitivity of 87%, and positive predictive value of 79%, which in the absence of necrotizing pneumonia could exceed 95% [34]. As early as 1994, the use of sputum samples had been explored in the detection of gene mutations using polymerase chain reaction (PCR)-based single gene assays [13, 14, 17, 18, 25]; however, only one study thus far has explored its use in NGS-based mutation profiling [16]. In our study, we have demonstrated the feasibility of using sputum supernatant as an alternative liquid biopsy specimen for comprehensive molecular profiling. The quality and quantity of circulating cell-free DNA isolated from the supernatant and sediment fractions from the induced sputum samples of our cohort were adequate for molecular profiling. Based on its significantly higher positive predictive value (80.4% vs. 55.6%) and concordance rate with tumor tissue samples as gold standard (69.2% vs. 37.8%; $P = 0.031$), sputum supernatant is the optimal fraction for the accurate detection of tumor-related non-CNV mutations than the sediment fraction. The higher concordance rate with tumor samples also suggests two important points: first, the concentration of circulating tumor DNA found in sputum supernatant is higher as compared to sputum sediment fraction; and second, the molecular profile derived from sputum supernatant samples more accurately reflects the non-CNV mutations found in the primary lung tumor tissues. The overall concordance rate in mutations detected from sputum samples relative to the matched tumor tissue samples we have observed from our cohort (69.2%) was consistent with the recent study by Wu and colleagues, which demonstrates a 74% overall concordance rate [16]. The comparable mutation detection rates, particularly in actionable non-CNV mutations, between sputum supernatant and plasma samples further suggest the feasibility of using sputum as an alternative liquid biopsy specimen for molecular testing. Similar to the observations by Wu and colleagues [16], our study also demonstrated differences and similarities in mutation profile in matched sputum fractions, plasma, and tumor tissues, which might

be related to spatial genetic heterogeneity inherent in small volume needle biopsy samples from tumor tissues. However, since sputum supernatant and sediment fractions were derived from 1 collection tube and were only physically separated *in vitro*, it is safe to conclude that sputum supernatant is the optimal fraction for molecular profiling.

In clinical practice, the alternative use of induced sputum can minimize the need for obtaining tumor samples using tissue biopsy in molecular testing applications. Due to convenience and easy accessibility, sputum specimens can also be used when blood samples are difficult to obtain. Our observations on the significantly higher maxAF in sputum supernatant than plasma samples suggest that sputum supernatant might be useful in patients with early-stage lung cancer and can also be explored for monitoring of treatment response. Previous studies have demonstrated the feasibility of detecting genetic and epigenetic alterations in sputum samples from cancer-free chronic smokers as a strategy for early detection of lung cancer [13, 18, 19, 22, 25]. NGS-based molecular profiling of sputum supernatant samples was adequately sensitive in detecting actionable mutations, particularly fusions, which has clinical value in guiding the use of appropriate targeted therapies. However, based on the data from our cohort, the detection rate for CNVs was low for sputum samples, which could be due to the limit of detection of CNVs in circulating cell-free DNA. We speculate that the use of unique modifier identification-based capture probes could improve the detection rate of CNVs and other mutations with ultralow allele frequencies.

Our study is limited by the small cohort size which included patients who participated in only a single-center. A study with a larger cohort is warranted to establish the utility of induced sputum specimens as a liquid biopsy specimen for NGS-based mutation testing and explore its application in treatment monitoring in advanced lung cancer patients or early detection of disease recurrence in both early and advanced lung cancer patients.

Conclusions

Our study demonstrates that the supernatant fraction of induced sputum specimens has comparable performance as plasma samples. Induced sputum from NSCLC patients with advanced-stage NSCLC could serve as an alternative source of tumor DNA for comprehensive mutation profiling. Our study contributes to the growing knowledge of alternative media for NGS-based mutation profiling.

Abbreviations

CNV, copy number variations; FEV_1, forced expiratory volume; FFPE, formalin-fixed, paraffin-embedded; Indel, insertion-deletion variations; maxAF, maximum allelic fraction; NGS, next-generation sequencing; NSCLC, non-small-cell lung cancer; PPV positive predictive value; SNV, single nucleotide variations

Declarations

Ethical approval: All procedures performed in studies involving human participants were performed in accordance with the Declaration of Helsinki. This study was approved by the Medical Ethics Committee of Xiangya Hospital Central South University (approval number: 201911306). Written informed consent was obtained from all participants included in the study.

Consent for publication: Consent for publication has been obtained.

Data Sharing and Data Accessibility: All authors confirm adherence to the policy. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing interest: A. Lizaso, B. Li, S. Zhang, and Z. Chen are employed by Burning Rock Biotech. The other authors declare no potential conflicts of interest.

Authors' contributions: L. Qin, T. Guo, H. Yang, P. Deng, Q. Gu, C. Liu, M. Wu, C. Hu collected the clinical data and participated in the data interpretation. L. Qin, T. Guo, and C. Hu conceived of the study and drafted the manuscript. A. Lizaso analyzed and interpreted the data and wrote the manuscript. B. Li performed the statistical analysis. S. Zhang was involved in the optimization of NGS-based detection from sputum. Z. Chen analyzed and interpreted the genomic data. All the authors contributed in the revision and approved the final manuscript.

Funding: This study did not receive any financial support in the form of grants.

Acknowledgments

The authors would like to thank all the patients and their family for their cooperation and support. We also thank the investigators, study coordinators, operation staff, and the whole project team who worked on this project.

References

1. Chan BA, Hughes BG: **Targeted therapy for non-small cell lung cancer: current standards and the promise of the future.***Transl Lung Cancer Res* 2015, **4:**36-54.
2. Mascaux C, Tomasini P, Greillier L, Barlesi F: **Personalised medicine for nonsmall cell lung cancer.***Eur Respir Rev* 2017, **26:**170066.
3. Yuan M, Huang LL, Chen JH, Wu J, Xu Q: **The emerging treatment landscape of targeted therapy in non-small-cell lung cancer.***Signal Transduct Target Ther* 2019, **4:**61.
4. Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL: **Molecular profiling for precision cancer therapies.***Genome Med* 2020, **12:**8.
5. Kim L, Tsao MS: **Tumour tissue sampling for lung cancer management in the era of personalised therapy: what is good enough for molecular testing?***Eur Respir J* 2014, **44:**1011-1022.
6. Balla A, Hampel KJ, Sharma MK, Cottrell CE, Sidiropoulos N: **Comprehensive Validation of Cytology Specimens for Next-Generation Sequencing and Clinical Practice Experience.***J Mol Diagn* 2018,

20:812-821.

7. Heitzer E, Haque IS, Roberts CES, Speicher MR: **Current and future perspectives of liquid biopsies in genomics-driven oncology.***Nat Rev Genet* 2019, **20:**71-88.
8. Rossi G, Ignatiadis M: **Promises and Pitfalls of Using Liquid Biopsy for Precision Medicine.***Cancer Res* 2019, **79:**2798-2804.
9. Takano T, Ohe Y, Tsuta K, Fukui T, Sakamoto H, Yoshida T, Tateishi U, Nokihara H, Yamamoto N, Sekine I, et al: **Epidermal growth factor receptor mutation detection using high-resolution melting analysis predicts outcomes in patients with advanced non small cell lung cancer treated with gefitinib.***Clin Cancer Res* 2007, **13:**5385-5390.
10. Mao X, Zhang Z, Zheng X, Xie F, Duan F, Jiang L, Chuai S, Han-Zhang H, Han B, Sun J: **Capture-Based Targeted Ultradeep Sequencing in Paired Tissue and Plasma Samples Demonstrates Differential Subclonal ctDNA-Releasing Capability in Advanced Lung Cancer.***J Thorac Oncol* 2017, **12:**663-672.
11. Li YS, Jiang BY, Yang JJ, Zhang XC, Zhang Z, Ye JY, Zhong WZ, Tu HY, Chen HJ, Wang Z, et al: **Unique genetic profiles from cerebrospinal fluid cell-free DNA in leptomeningeal metastases of EGFR-mutant non-small-cell lung cancer: a new medium of liquid biopsy.***Ann Oncol* 2018, **29:**945-952.
12. Guo Z, Xie Z, Shi H, Du W, Peng L, Han W, Duan F, Zhang X, Chen M, Duan J, et al: **Malignant pleural effusion supernatant is an alternative liquid biopsy specimen for comprehensive mutational profiling.***Thorac Cancer* 2019, **10:**823-831.
13. Mao L, Hruban RH, Boyle JO, Tockman M, Sidransky D: **Detection of oncogene mutations in sputum precedes diagnosis of lung cancer.***Cancer Res* 1994, **54:**1634-1637.
14. Boldrini L, Gisfredi S, Ursino S, Camacci T, Baldini E, Melfi F, Fontanini G: **Mutational analysis in cytological specimens of advanced lung adenocarcinoma: a sensitive method for molecular diagnosis.***J Thorac Oncol* 2007, **2:**1086-1090.
15. Scott SN, Ostrovnaya I, Lin CM, Bouvier N, Bochner BH, Iyer G, Solit D, Berger MF, Lin O: **Next-generation sequencing of urine specimens: A novel platform for genomic analysis in patients with non-muscle-invasive urothelial carcinoma treated with bacille Calmette-Guerin.***Cancer Cytopathol* 2017, **125:**416-426.
16. Wu Z, Yang Z, Li CS, Zhao W, Liang ZX, Dai Y, Zhu Q, Miao KL, Cui DH, Chen LA: **Differences in the genomic profiles of cell-free DNA between plasma, sputum, urine, and tumor tissue in advanced NSCLC.***Cancer Med* 2019, **8:**910-919.
17. Wang Z, Zhang L, Li L, Li X, Xu Y, Wang M, Liang L, Jiao P, Li Y, He S, et al: **Sputum Cell-Free DNA: Valued Surrogate Sample for Detection of EGFR Mutation in Patients with Advanced Lung Adenocarcinoma.***J Mol Diagn* 2020, **22:**934-942.
18. Kersting M, Friedl C, Kraus A, Behn M, Pankow W, Schuermann M: **Differential frequencies of p16(INK4a) promoter hypermethylation, p53 mutation, and K-ras mutation in exfoliative material mark the development of lung cancer in symptomatic chronic smokers.***J Clin Oncol* 2000, **18:**3221-3229.

19. Belinsky SA, Klinge DM, Dekker JD, Smith MW, Bocklage TJ, Gilliland FD, Crowell RE, Karp DD, Stidley CA, Picchi MA: **Gene promoter methylation in plasma and sputum increases with lung cancer risk.***Clin Cancer Res* 2005, **11:**6505-6511.
20. Belinsky SA, Liechty KC, Gentry FD, Wolf HJ, Rogers J, Vu K, Haney J, Kennedy TC, Hirsch FR, Miller Y, et al: **Promoter hypermethylation of multiple genes in sputum precedes lung cancer incidence in a high-risk cohort.***Cancer Res* 2006, **66:**3338-3344.
21. Machida EO, Brock MV, Hooker CM, Nakayama J, Ishida A, Amano J, Picchi MA, Belinsky SA, Herman JG, Taniguchi S, Baylin SB: **Hypermethylation of ASC/TMS1 is a sputum marker for late-stage lung cancer.***Cancer Res* 2006, **66:**6210-6218.
22. Castagnaro A, Marangio E, Verduri A, Chetta A, D'Ippolito R, Del Donno M, Olivieri D, Di Cola G: **Microsatellite analysis of induced sputum DNA in patients with lung cancer in heavy smokers and in healthy subjects.***Exp Lung Res* 2007, **33:**289-301.
23. Li R, Todd NW, Qiu Q, Fan T, Zhao RY, Rodgers WH, Fang HB, Katz RL, Stass SA, Jiang F: **Genetic deletions in sputum as diagnostic markers for early detection of stage I non-small cell lung cancer.***Clin Cancer Res* 2007, **13:**482-487.
24. van der Drift MA, Prinsen CF, Hol BE, Bolijn AS, Jeunink MA, Dekhuijzen PN, Thunnissen FB: **Can free DNA be detected in sputum of lung cancer patients?***Lung Cancer* 2008, **61:**385-390.
25. Baryshnikova E, Destro A, Infante MV, Cavuto S, Cariboni U, Alloisio M, Ceresoli GL, Lutman R, Brambilla G, Chiesa G, et al: **Molecular alterations in spontaneous sputum of cancer-free heavy smokers: results from a large screening program.***Clin Cancer Res* 2008, **14:**1913-1919.
26. Katz RL, Zaidi TM, Fernandez RL, Zhang J, He W, Acosta C, Daniely M, Madi L, Vargas MA, Dong Q, et al: **Automated detection of genetic abnormalities combined with cytology in sputum is a sensitive predictor of lung cancer.***Mod Pathol* 2008, **21:**950-960.
27. Hulbert A, Jusue-Torres I, Stark A, Chen C, Rodgers K, Lee B, Griffin C, Yang A, Huang P, Wrangle J, et al: **Early Detection of Lung Cancer Using DNA Promoter Hypermethylation in Plasma and Sputum.***Clin Cancer Res* 2017, **23:**1998-2005.
28. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.***Bioinformatics* 2009, **25:**1754-1760.
29. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.***Genome Res* 2010, **20:**1297-1303.
30. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.***Genome Res* 2012, **22:**568-576.
31. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.***Nucleic Acids Res* 2010, **38:**e164.
32. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the**

genome of Drosophila melanogaster strain w1118; iso-2; iso-3.*Fly (Austin)* 2012, **6:**80-92.

33. Newman AM, Bratman SV, Stehr H, Lee LJ, Liu CL, Diehn M, Alizadeh AA: **FACTERA: a practical method for the discovery of genomic rearrangements at breakpoint resolution.***Bioinformatics* 2014, **30:**3390-3393.
34. Jay SJ, Wehr K, Nicholson DP, Smith AL: **Diagnostic sensitivity and specificity of pulmonary cytology: comparison of techniques used in conjunction with flexible fiber optic bronchoscopy.***Acta Cytol* 1980, **24:**304-312.
35. Hubers AJ, Prinsen CF, Sozzi G, Witte BI, Thunnissen E: **Molecular sputum analysis for the diagnosis of lung cancer.***Br J Cancer* 2013, **109:**530-537.

List Of Supplementary

Figure S1. Quality control metrics including the DNA yield expressed in nanograms (A), the correlation between library complexity and insert size of the fragments (in base pairs) for sequencing (B), and the median sequencing depth (C) of all the samples analyzed in the study.

Figure S2. Smokers had significantly higher allelic fraction of sputum supernatant. Box plot illustrating the distribution of allelic frequencies in sputum supernatant (SPU) from smokers and never-smokers.

Table S1. Distribution of samples according to available sample type

Table S2. By variant comparison between the mutations detected from sputum supernatant and corresponding sediment samples of the 23 patients

Table S3. By variant comparison between the mutations detected from sputum supernatant and matched tissue samples of the 26 patients

Table S4. By variant comparison between the mutations detected from sputum supernatant and matched plasma samples of the 28 patients

Figures

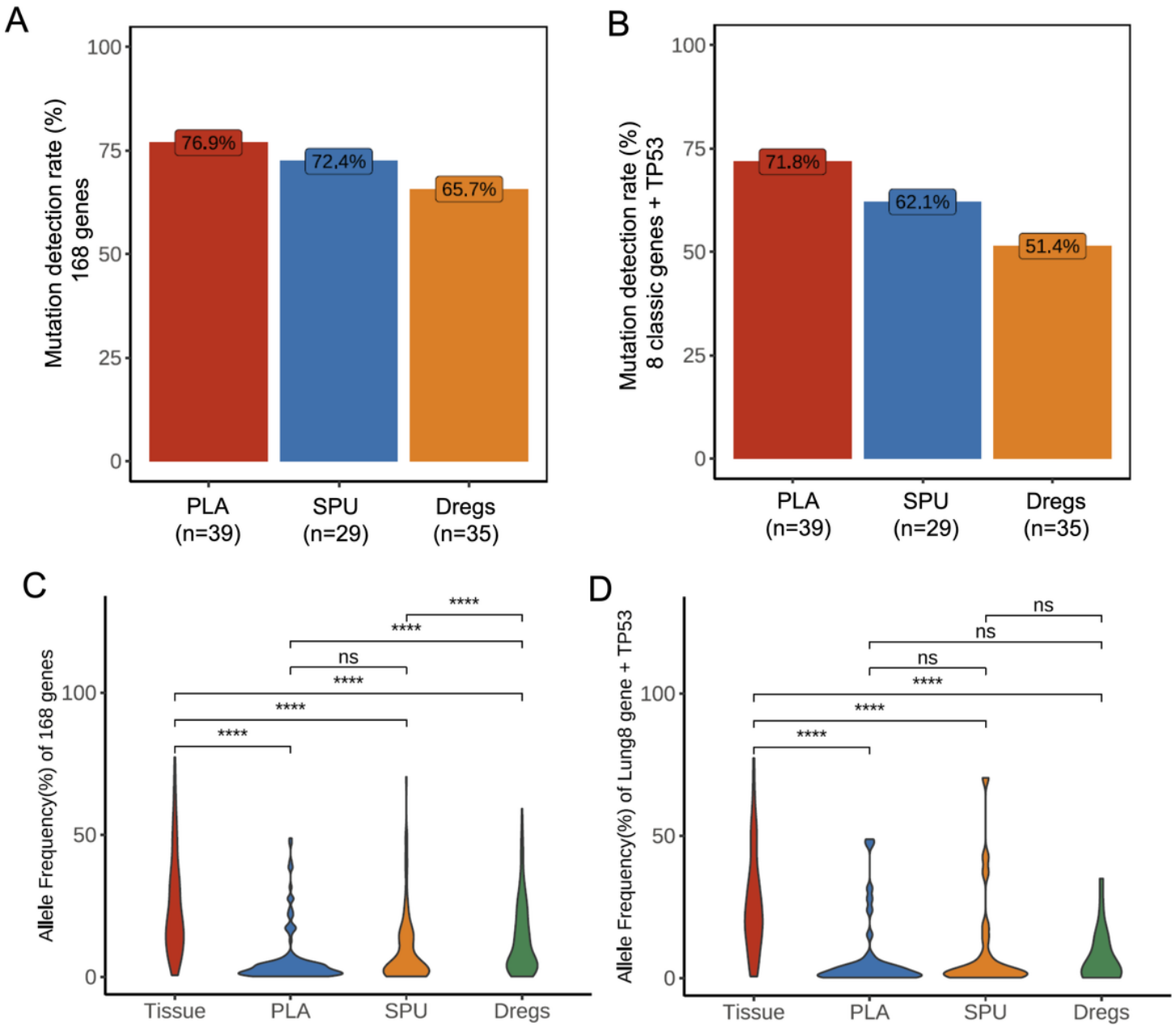

Figure 1

The mutation detection rate and allelic fraction from plasma and sputum supernatant were comparable. A-B. Bar plots summarizing the mutation detection rates from 168 genes (A) and 9 genes (8 classic NSCLC oncogenic driver genes and TP53; B) with tumor samples as reference. C-D. Violin plots summarizing the maximum allele frequency from 168 genes (C) and 9 genes (D).

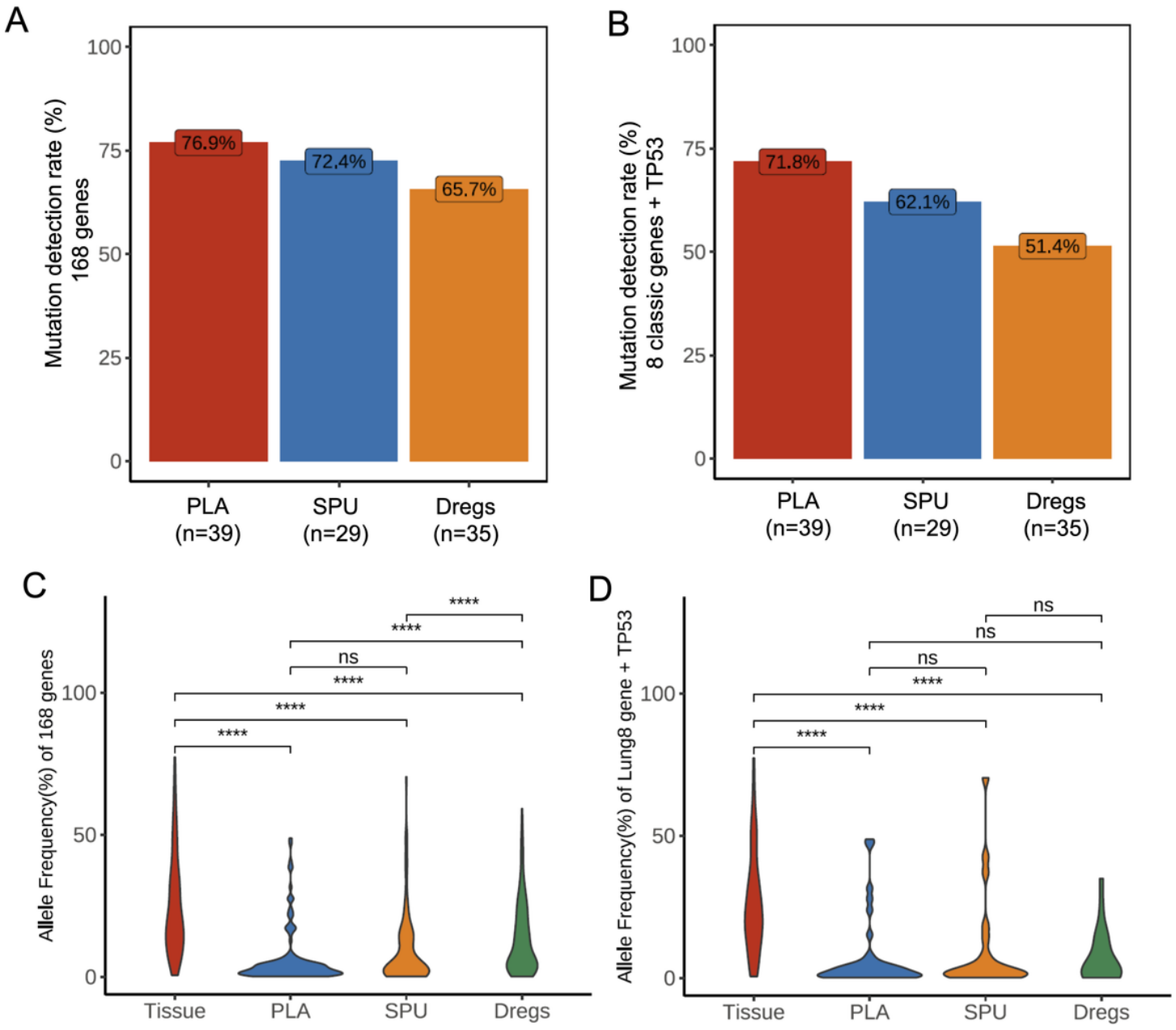

Figure 1

The mutation detection rate and allelic fraction from plasma and sputum supernatant were comparable. A-B. Bar plots summarizing the mutation detection rates from 168 genes (A) and 9 genes (8 classic NSCLC oncogenic driver genes and TP53; B) with tumor samples as reference. C-D. Violin plots summarizing the maximum allele frequency from 168 genes (C) and 9 genes (D).

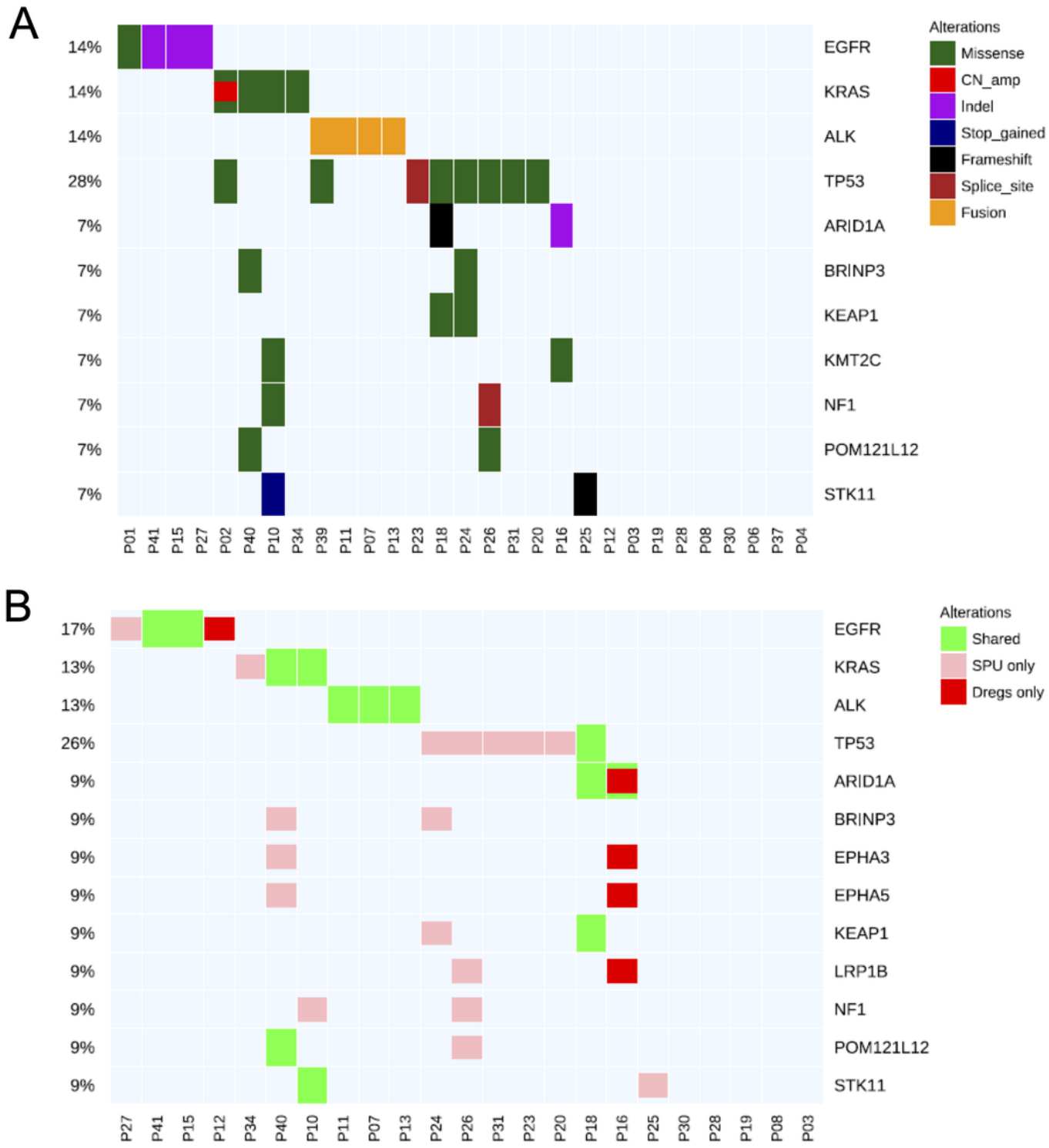

Figure 2

Sputum supernatant is a more optimal sputum fraction for molecular profiling than its corresponding sediment. A-B. Oncoprints illustrating the mutation landscape derived from the sputum supernatant of 29 patients (A) and the comparison between the mutation profiles derived from the sputum supernatant and its corresponding sediment in 23 patients. The colors indicate either the mutation types (A) or the status of each mutation whether detected in both samples (Shared) or detected in only the supernatant (SPU

only) or the sediment (Dregs only). Each column represents one patient. Each row represents a gene. Side bar represents the mutation rate of a certain gene. The oncoprint was summarized to only reflect the genes with 2 or more mutations detected.

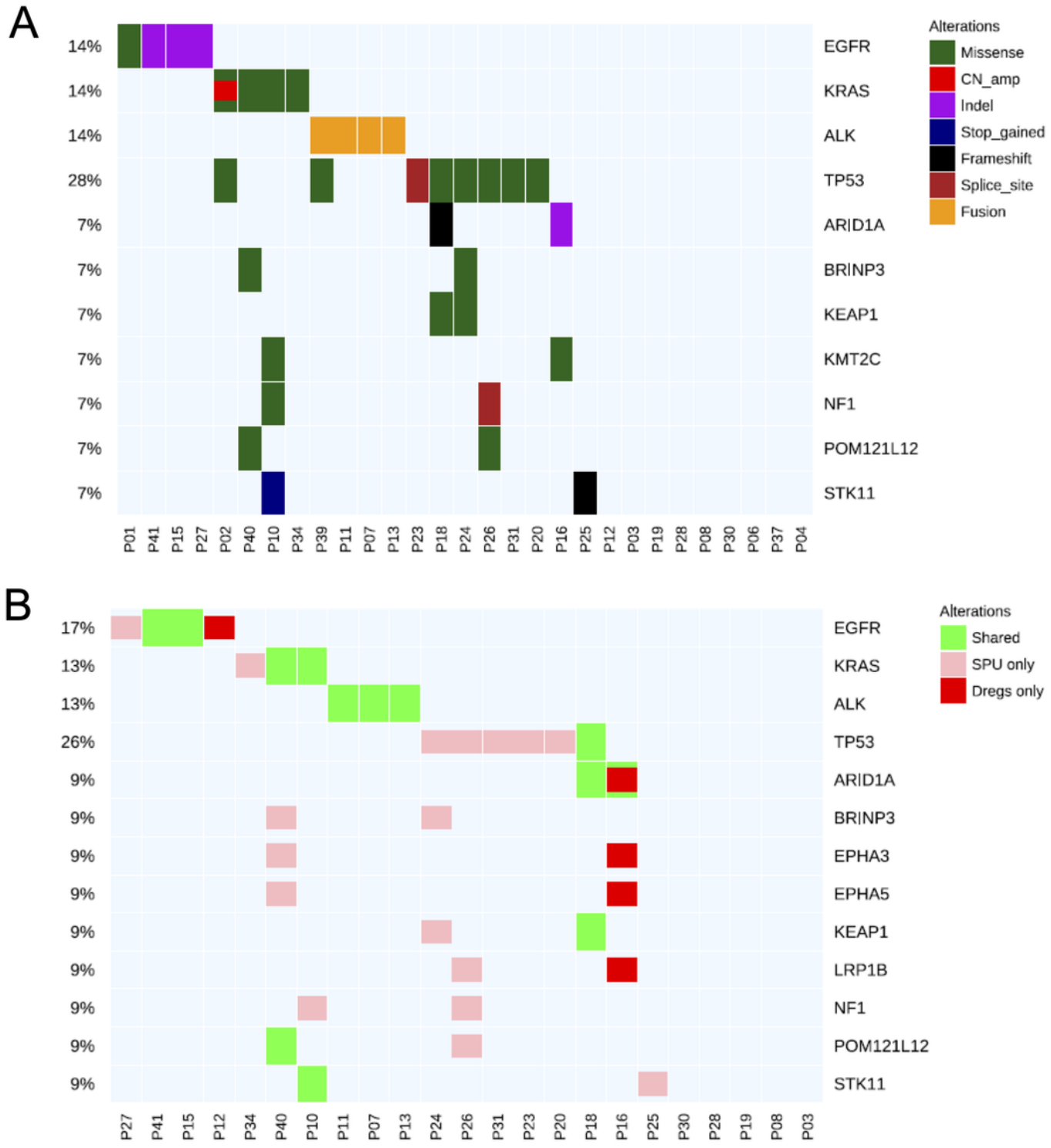

Figure 2

Sputum supernatant is a more optimal sputum fraction for molecular profiling than its corresponding sediment. A-B. Oncoprints illustrating the mutation landscape derived from the sputum supernatant of 29

patients (A) and the comparison between the mutation profiles derived from the sputum supernatant and its corresponding sediment in 23 patients. The colors indicate either the mutation types (A) or the status of each mutation whether detected in both samples (Shared) or detected in only the supernatant (SPU only) or the sediment (Dregs only). Each column represents one patient. Each row represents a gene. Side bar represents the mutation rate of a certain gene. The oncoprint was summarized to only reflect the genes with 2 or more mutations detected.

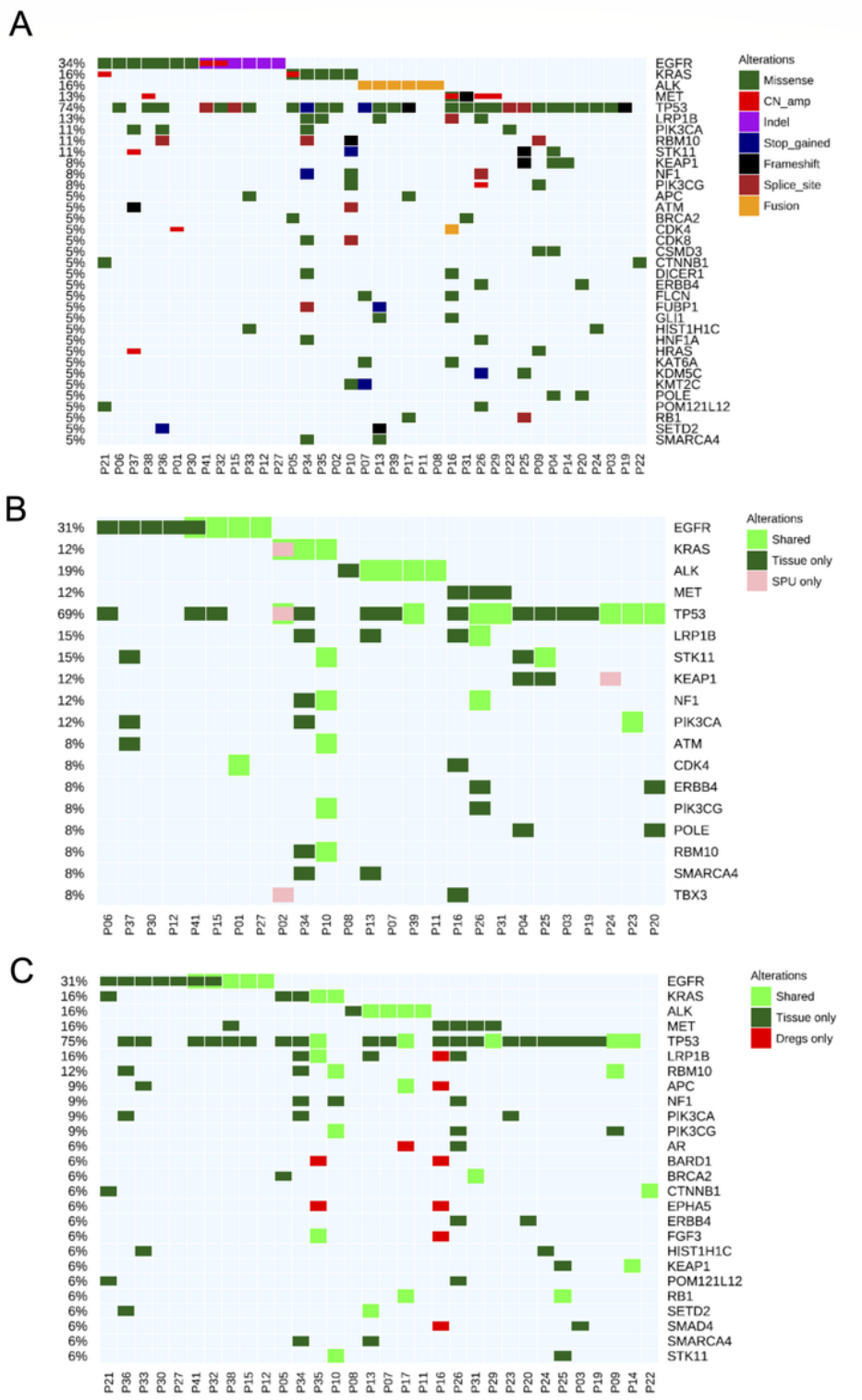

Figure 3

Sputum supernatant reflects more tumor-related mutations than its corresponding sediment. A-C. Oncoprints illustrating the mutation landscape derived from the tumor samples of 38 patients (A), the comparison of mutation profiles between the tumor samples and sputum supernatant samples in 26 patients (B), and between tumor sample and sputum sediment in 32 patients (C). The colors indicate either the mutation types (A) or the status of each mutation whether detected in both samples (Shared) or detected in only the tumor (Tissue only), supernatant (SPU only), or sediment (Dregs only). Each column represents one patient. Each row represents a gene. Side bar represents the mutation rate of a certain gene. The oncoprint was summarized to only reflect the genes with 2 or more mutations detected.

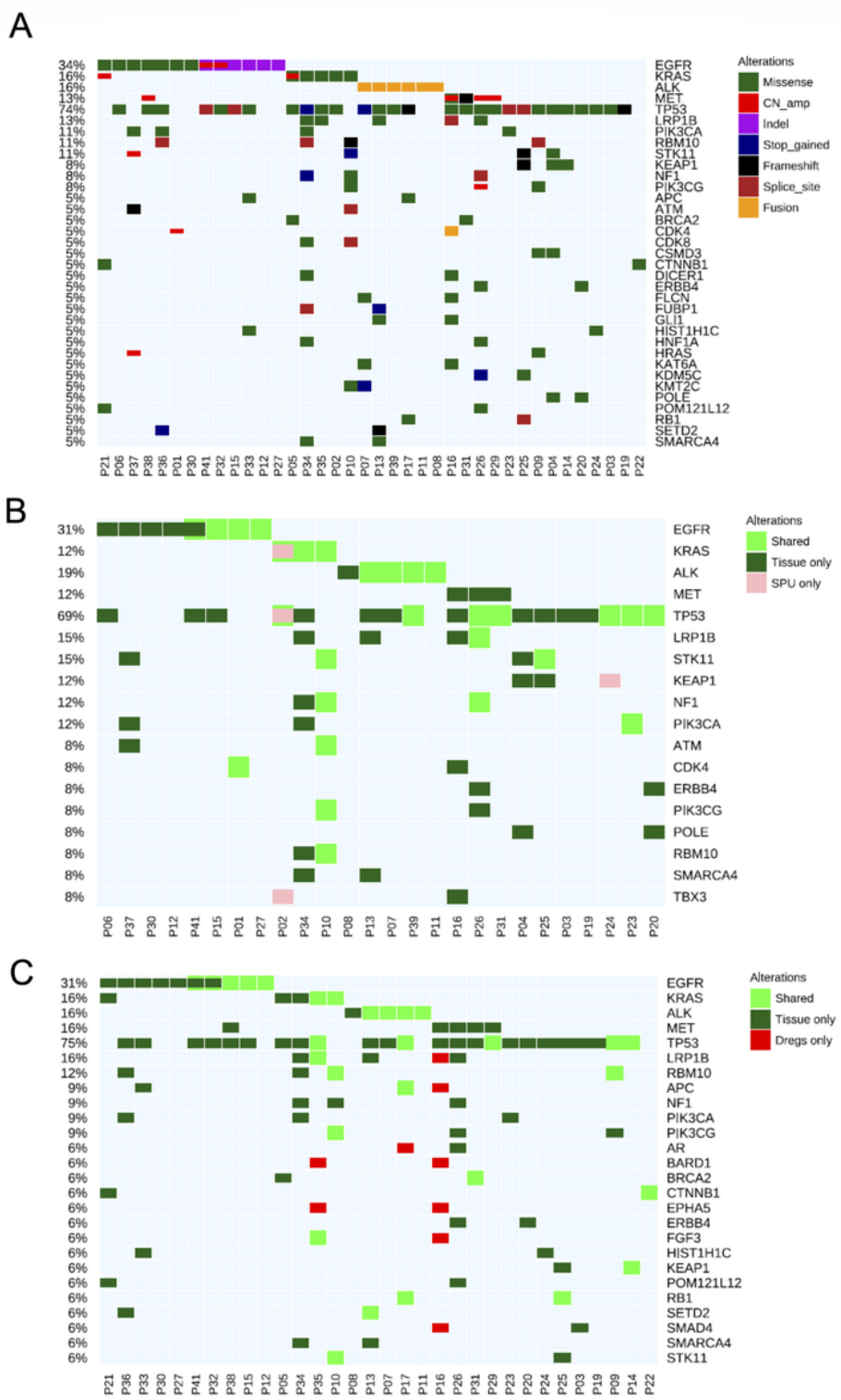

Figure 3

Sputum supernatant reflects more tumor-related mutations than its corresponding sediment. A-C. Oncoprints illustrating the mutation landscape derived from the tumor samples of 38 patients (A), the comparison of mutation profiles between the tumor samples and sputum supernatant samples in 26 patients (B), and between tumor sample and sputum sediment in 32 patients (C). The colors indicate either the mutation types (A) or the status of each mutation whether detected in both samples (Shared) or

detected in only the tumor (Tissue only), supernatant (SPU only), or sediment (Dregs only). Each column represents one patient. Each row represents a gene. Side bar represents the mutation rate of a certain gene. The oncoprint was summarized to only reflect the genes with 2 or more mutations detected.

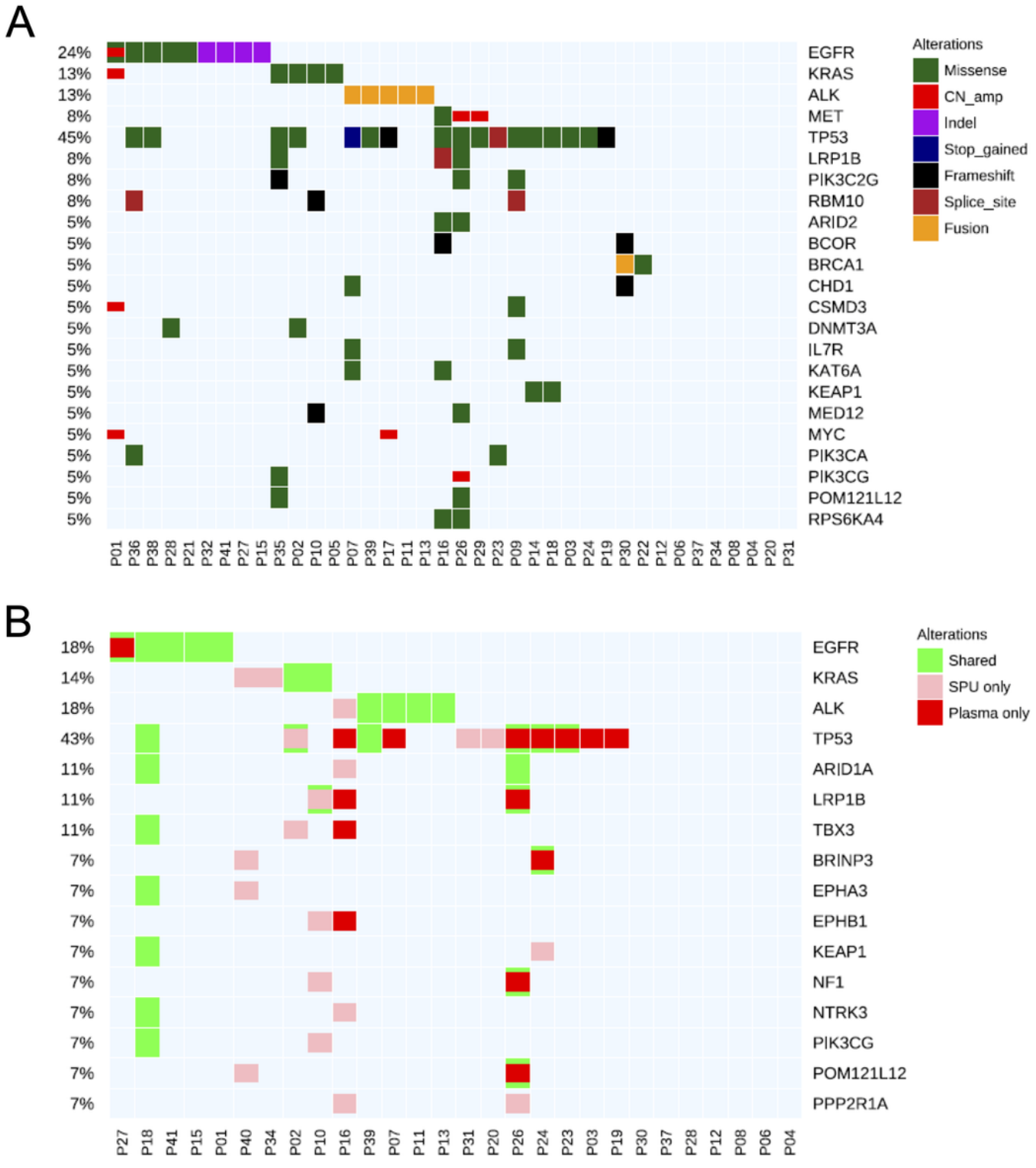

Figure 4

Mutation profile derived from sputum supernatant is comparable with plasma samples. A-B. Oncoprints illustrating the mutation landscape derived from the plasma samples of 38 patients (A), and the

comparison of mutation profiles between the plasma samples and sputum supernatant samples in 28 patients (B). The colors indicate either the mutation types (A) or the status of each mutation whether detected in both samples (Shared) or detected in only the plasma samples (Plasma only) or supernatant (SPU only) (B). Each column represents one patient. Each row represents a gene. Side bar represents the mutation rate of a certain gene. The oncoprint was summarized to only reflect the genes with 2 or more mutations detected.

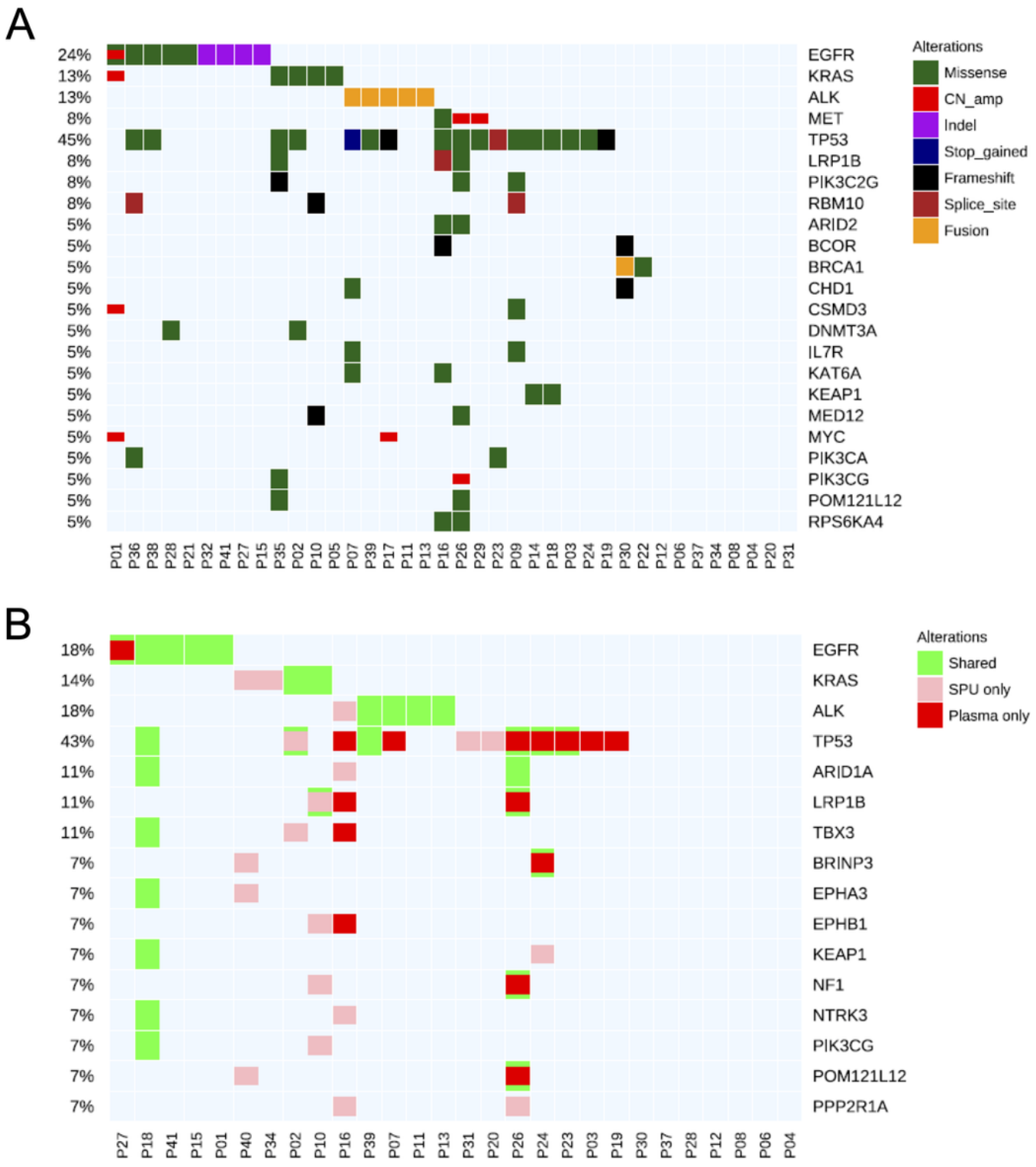

Figure 4

Mutation profile derived from sputum supernatant is comparable with plasma samples. A-B. Oncoprints illustrating the mutation landscape derived from the plasma samples of 38 patients (A), and the comparison of mutation profiles between the plasma samples and sputum supernatant samples in 28 patients (B). The colors indicate either the mutation types (A) or the status of each mutation whether detected in both samples (Shared) or detected in only the plasma samples (Plasma only) or supernatant (SPU only) (B). Each column represents one patient. Each row represents a gene. Side bar represents the mutation rate of a certain gene. The oncoprint was summarized to only reflect the genes with 2 or more mutations detected.

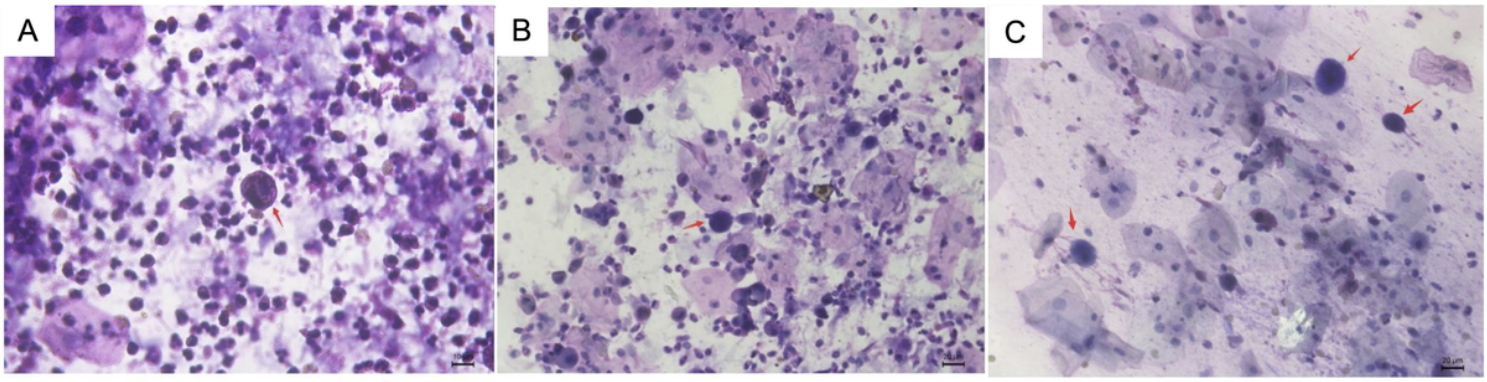

Figure 5

Malignancy in sputum cytology characterized by heterogeneous cell nuclei. Hematoxylin-eosin staining of sputum samples from three patients, P23 (A), P27 (B), and P41 (C), with various advanced non-small-cell lung cancer histology. The heterogeneous cell nucleus is indicated by the red arrow.

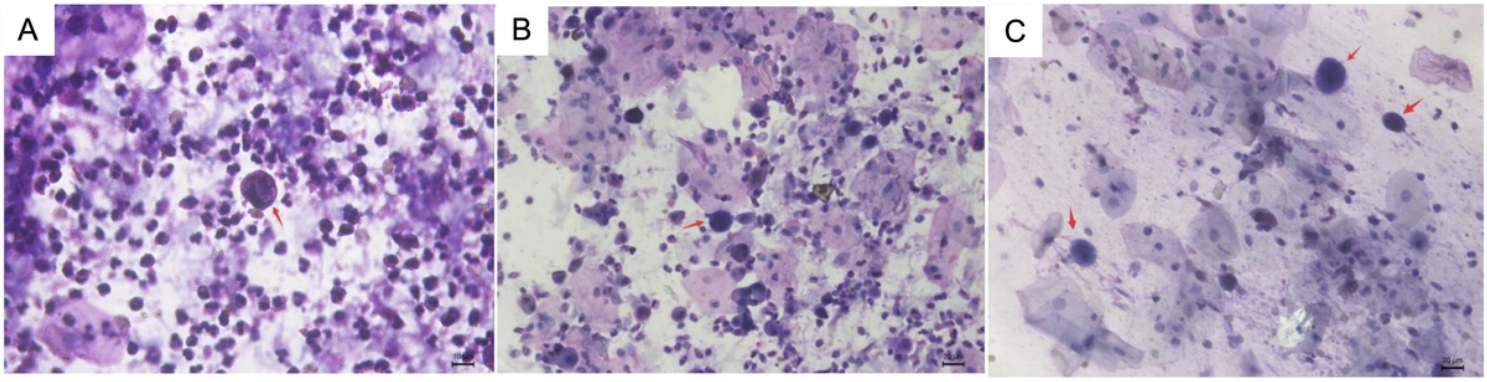

Figure 5

Malignancy in sputum cytology characterized by heterogeneous cell nuclei. Hematoxylin-eosin staining of sputum samples from three patients, P23 (A), P27 (B), and P41 (C), with various advanced non-small-cell lung cancer histology. The heterogeneous cell nucleus is indicated by the red arrow.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfiguresandtables.docx
- Supplementaryfiguresandtables.docx