# Genomic Profiling of Bronchoalveolar Lavage Fluid in Lung Cancer

**Viswam S. Nair**[1,2,3,†,*], **Angela Bik-Yu Hui**[4,†], **Jacob J. Chabon**[4,**], **Mohammad S. Esfahani**[4], **Henning Stehr**[5], **Barzin Y. Nabet**[4,], **Li Zhou**[4,§], **Aadel A. Chaudhuri**[6,¶], **Jalen Benson**[7], **Kelsey Ayers**[7], **Harmeet Bedi**[8], **Meghan Ramsey**[8], **Ryan Van Wert**[8], **Sanja Antic**[9], **Natalie Lui**[7], **Leah Backhus**[7], **Mark Berry**[7], **Arthur W. Sung**[8], **Pierre P. Massion**[9,], **Joseph B. Shrager**[7], **Ash A. Alizadeh**[4,10,11,*], **Maximilian Diehn**[4,6,11,*]

[1]Clinical Research Division, Fred Hutchinson Cancer Research Center

[2]Division of Pulmonary, Critical Care & Sleep Medicine, University of Washington School of Medicine

[3]Department of Radiology, Stanford University School of Medicine

[4]Stanford Cancer Institute, Stanford University School of Medicine

[5]Department of Pathology, Stanford University School of Medicine

[6]Department of Radiation Oncology, Stanford University School of Medicine

[7]Division of Thoracic Surgery, Stanford University School of Medicine

[8]Division of Pulmonary, Allergy & Critical Care Medicine, Stanford University School of Medicine

[9]Division of Allergy, Pulmonary & Critical Care Medicine, Vanderbilt University School of Medicine

[10]Division of Oncology, Department of Medicine, Stanford University School of Medicine

[11]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine

**Corresponding author address:** Viswam S. Nair, 1100 Fairview Ave N, Seattle, WA 98109. phone: (206) 447-4971. vnair@fredhutch.org. [*]Co-corresponding authors.
[†]These authors contributed equally to the manuscript
[**]Current affiliation: Foresight Diagnostics, Aurora, CO
[]Current affiliation: Department of Oncology Biomarker Development, Genentech, South San Francisco, CA
[¶]Current affiliation: Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO
[§]Current affiliation: Grail Inc, Menlo Park, CA
[]Deceased

## Abstract

Genomic profiling of Bronchoalveolar Lavage (BAL) samples may be useful for tumor profiling and diagnosis in the clinic. Here, we compared tumor-derived mutations detected in BAL samples from subjects with non-small cell lung cancer (NSCLC) to those detected in matched plasma samples. CAncer Personalized Profiling by deep Sequencing (CAPP-Seq) was used to genotype DNA purified from BAL, plasma and tumor samples from patients with NSCLC. The characteristics of cell-free DNA (cfDNA) isolated from BAL fluid were first characterized to optimize the technical approach. Somatic mutations identified in tumor were then compared to those identified in BAL and plasma, and the potential of BAL cfDNA analysis to distinguish lung cancer patients from risk-matched controls was explored. In total, 200 biofluid and tumor samples from 38 cases and 21 controls undergoing BAL for lung cancer evaluation were profiled. More tumor variants were identified in BAL cfDNA than plasma cfDNA in all stages (p<0.001) and in stage I-II disease only. Four of 21 controls harbored low levels of cancer-associated driver mutations in BAL cfDNA (mean VAF=0.5%), suggesting the presence of somatic mutations in non-malignant airway cells. Finally, using a Random Forest model with leave-one-out cross validation, an exploratory BAL genomic classifier identified lung cancer with 69% sensitivity and 100% specificity in this cohort and detected more cancers than BAL cytology. Detecting tumor-derived mutations by targeted sequencing of BAL cfDNA is technically feasible and appears to be more sensitive than plasma profiling. Further studies are required to define optimal diagnostic applications and clinical utility.

## Summary Statement

Hybrid-capture, targeted deep sequencing of lung cancer mutational burden in cell-free bronchoalveolar lavage fluid identifies more tumor-derived mutations with increased allele frequencies compared to plasma cell-free DNA.

## Introduction

Lung cancer is the leading cause of cancer deaths worldwide and the majority of patients are diagnosed with metastatic disease that is generally incurable. Survival for lung cancer remains poor in large part due to detection at late stages (1-3). Genotyping is now standard of clinical care for treating lung cancer, but this can be difficult to perform due to issues with the amount and quality of sample acquired during clinical evaluation.

Therapies that target specific oncogenic alterations in lung cancers have been approved for clinical use (4, 5). Appropriate use of these therapies requires molecular profiling of lung tumor tissue. The diagnosis and genomic profiling of lung cancer is achieved by biopsy or surgical resection for patients with early-stage disease (localized or regional disease) or biopsy alone for those with advanced disease (widespread). Unfortunately, biopsy samples are difficult to obtain in a significant subset of patients and/or contain insufficient material for mutation profiling. As an example, for bronchoscopic biopsies of peripheral lung lesions, diagnostic yield is 50% (6, 7) and adequate molecular profiling of diagnostic biopsies ranges from 60-100% (8, 9). Patients are therefore frequently subjected to repeat sampling via multiple procedures, or clinicians are left to render treatment decisions with incomplete data.

Chest practitioners have routinely performed bronchoalveolar lavage (BAL) for decades at academic and community medical centers across the world with little change in practice and few new assays developed to increase utility (10). BAL is simple to perform, easily obtained, relatively inexpensive to process, and safely retrieved during bronchoscopy for lung cancer evaluation, but it has a poor yield for cancer diagnosis by cytology alone (11). Currently, there is no role for BAL fluid to molecularly profile lung cancer, despite evidence that it might provide complementary information to tissue sampling (7).

Genomic profiling from a routine blood draw is a convenient and promising approach to non-invasively profile lung cancer (12), yet its utility for detecting and profiling early tumors is limited by the low quantities of circulating tumor DNA (ctDNA) shed in patients with low tumor burden (13).

Studying proximal fluids such as pleural fluid, urine, ascites or cerebrospinal fluid can enhance tumor detection since they are often enriched for molecular markers of the cancer of interest. Molecular analyses of biofluids to detect tumors and their associated mutation profiles is therefore an important and emerging approach for solid cancer evaluation (14-16). For lung cancer, new paradigms of biofluid analysis for genomic analysis could reduce patient morbidity, decrease costs of care, shorten time to treatment, and increase treatment efficacy.

We hypothesized that BAL fluid from early-stage lung cancer patients is enriched for lung tumor derived DNA when compared to blood. To explore this question, we utilized CAncer Personalized Profiling by deep Sequencing (CAPP-Seq) to identify mutations in regions of the genome that are commonly altered in lung cancer (12, 17-19). We identified putative tumor mutations in BAL fluid and then compared these results to tumor tissue and plasma profiling. We demonstrate that BAL fluid obtained during routine bronchoscopy frequently contains lung tumor-derived mutant DNA and could add value to routine BAL cytology.

## Materials and Methods

### Patient enrollment

Our goal was to explore whether molecular profiling of BAL fluid might have utility for genotyping and detection of lung cancer by utilizing the CAncer Personalized Profiling by deep Sequencing (CAPP-Seq) platform (Supplemental Figure 1). We performed routine bronchoscopy and phlebotomy in patients undergoing clinical evaluation for lung cancer to collect BAL and plasma samples from 2015-17 at Stanford Health Care. We enrolled a separate cohort of patients without cancer who underwent lung cancer screening from 2016-17 at Vanderbilt Health to identify field cancerization and to then develop a BAL genome classifier. Additional details on sample collection, DNA extraction, library preparation and statistical analyses of biofluids are provided in the Supplementary Methods.

Written informed consent was obtained from the patients in this study in accordance with ethical guidelines put forth by the Declaration of Helsinki at each participating center. Protocols were approved by the institutional review board at each center prior to initiation of the study.

## CAPP-Seq analysis

Targeted capture and sequencing analysis of all samples was performed using CAPP-Seq (17). We employed a 302 kb CAPP-Seq selector targeting 771 non-contiguous regions of the human genome, spanning 276 genes (18). A maximum of 32 ng DNA was input into sequencing library preparation. For plasma and BAL fluid samples with less than 32 ng of isolated cfDNA, all the extracted cfDNA was used for library preparation, down to a minimum of 16 ng. Samples were sequenced using 2x100 or 2x150 reads on an Illumina HiSeq 2500 or 4000. Sequencing data were processed using a previously described bioinformatics pipeline (12, 17). SNVs and indels were genotyped in all samples (17).

## Mutation identification pipeline

**Tumor informed:** Mutational profiling of primary tumor biopsy samples in 34 patients was performed in the Stanford pathology department using the Solid Tumor Actionable Mutation Panel (STAMP), a CLIA certified tumor genotyping assay (20). Variants that were identified in tumor tissue were then compared to the corresponding biofluids from these patients (18). We limited our analysis to genomic positions targeted by both STAMP and the lung cancer-focused CAPP-seq selector. Sixty-two genes (53.5 kb total) overlapped between the STAMP (150 total genes) and CAPP-Seq (276 total genes) panels. Overlapping positions were identified to generate patient-specific tumor variant lists after removing germline variants that were identified by sequencing germline DNA from plasma depleted whole blood. Sequencing results from corresponding plasma, BAL cfDNA and BAL cell pellets were then queried for the presence of single nucleotide variants (SNVs) and small structural variants (indels) that were identified in matched tumor tissue using our previously described Monte-Carlo based ctDNA detection index (12, 14, 21, 22). We refer to variants detected in tumor and biofluids as tumor-derived.

**Tumor-naive:** BAL is usually performed during a biopsy procedure or prior to surgical resection, so determining whether BAL mutation detection is useful without *a priori* knowledge of a tumor's molecular profile is relevant for clinical utility assessment. To study this question, we applied an adaptation of a previously described tumor naïve genotyping strategy developed by our group to analyze BAL cfDNA (19). SNVs for cancer and non-cancer patients for all sequenced plasma cfDNA and BAL cfDNA were identified without using primary biopsy data after comparing to matched germline DNA as previously described (12, 17, 19).

## BAL genomic classifier model development

We also performed an exploratory analysis to determine whether BAL fluid profiling might be useful for diagnosis of lung cancer. In order to classify benign controls vs. lung cancer cases, the following set of features was defined to summarize the mutations identified in a BAL sample: (a) the mean variant allele frequency across all the mutations identified (mAF), (b) the total number of mutations (n), (c) the maximum allele frequency (mxAF), (d) the number of mutations identified that were observed in 1 cancer cases observed in the COSMIC database (CosmicGenomeScreens v85, [nCOSMIC1]), (e) the number of mutations identified that were observed in 1 lung cancer cases in

the COSMIC database (CosmicGenomeScreens v85, [nCOSMICL1]), (f) the number of mutations identified that were observed in 10 lung cancer cases observed in the COSMIC database (CosmicGenomeScreens v85, [nCOSMICL10]), (g) the number of mutations in canonical lung cancer driver genes [lung_driver] (23), (h) the mean allele frequency of non-synonymous mutations [ns-mAF], (i) the number of nonsynonymous mutations [nns], (j) the fraction of mutations present in matched leukocyte DNA [nGp], and (k) fraction of mutations with a control-compared empirical p-value 0.05 or nCOSMICL1>0 [nEp].

Using these 11 features a Random Forest model with n=1000 trees was then trained, and a classifier was generated through a leave-one-out cross-validation framework. A receiver operating characteristic (ROC) curve and its area under the curve (AUC) was used to summarize model performance. Diagnostic sensitivity and specificity were calculated per standard methods based on the derived ROC curves. R packages randomForest and pROC were used for the analysis. To evaluate individual feature importance for the proposed classifier, we used the 'mean decrease in accuracy' metric and summarized this metric across the models generated in the leave-one-out cross-validation. A comparison to BAL cytology was performed by AUC analysis, where BAL cytology was dichotomized as not diagnostic for cancer for atypical, suspicious or no malignancy detected and diagnostic for cancer if definitive malignant cells were reported.

## Data Availability

The data generated in this study are available within the article and its supplementary data files. Additional data generated not reported herein are available upon request from the corresponding author. Raw data for this study were generated at Stanford University and are available from the corresponding author upon request if permitted by the local institutional review board.

## Results

### Cohort

In total, we analyzed 200 samples from 59 participants that included 38 subjects with lung cancer and 21 high risk controls without cancer (Supplemental Table 1). Controls were patients who had nodules detected on CT scan that ultimately were diagnosed as not having cancer (n=11) or were undergoing lung cancer screening based on age and tobacco history and did not have cancer (n=10). Thirty-four patients with lung cancer had their primary tumor sequenced by STAMP (Supplementary methods). For each subject we profiled matching BAL fluid, plasma and plasma depleted whole blood (PDWB) by CAPP-Seq (17, 19). Most BAL specimens from cancer patients were adequate for library preparation and sequencing after extraction (35 of 38, 92%). All 21 control subjects at risk for lung cancer had adequate cfDNA isolated from BAL fluid and plasma to carry forward for sequencing.

### Characteristics of BAL cellular DNA

Because sequencing of BAL fluid is not well characterized in the literature (24-26), we assessed the quality of DNA and sequencing between the BAL cell pellet (BAL cellular

DNA) and supernatant (BAL cell free DNA, BAL cfDNA). DNA fragments from BAL cfDNA were consistently larger in size when compared to plasma fragments that have a stereotyped size distribution with a mode fragment size of ~167 bp (Figure 1A, B) (27). We therefore sequenced BAL cfDNA before and after DNA shearing to understand the impact of fragment size on our ability to genotype tumor-derived mutations in BAL cfDNA. Sheared BAL cfDNA samples yielded more unique, "deduplicated" DNA reads and detected more mutations than unsheared BAL cfDNA (Figure 1C, D). Based on these results, all subsequent BAL cfDNA samples analyzed were sheared before library preparation and sequencing.

In 20 lung cancer patients, BAL cellular DNA concentrations ranged from 38.5-44,213 ng/ml (median, 1,007 ng/ml), which was 34 times higher compared to BAL cfDNA (range 1.3-7,795 ng/ml, median 29.3 ng/ml) (p< 0.001, Figure 1E). Despite more DNA being present in cellular BAL, the median variant allele frequency (VAF) of tumor derived mutations was significantly lower in BAL cellular DNA (0.11%, range, 0–34%) compared to BAL cfDNA (0.99%, range, 0–51%; n=15; p=0.048, Figure 1F & Supplemental Table 2). Furthermore, 87% of BAL cfDNA samples contained tumor-derived DNA as determined by tumor-informed monitoring using a cutoff of p<0.05 to identify variants compared to only 60% in cellular BAL. Finally, the mean VAF of all mutations detected in BAL cfDNA samples was higher in 11 of the 15 BAL samples that analyzed both BAL cfDNA and BAL cellular DNA. Taken together, these results suggest that BAL cfDNA best captures tumor derived fragments, and we therefore focused on sequencing BAL cfDNA for our subsequent analyses.

### Cell free BAL fluid contains a higher concentration of tumor DNA than plasma

DNA was extracted from a median of 4.7 ml of cell free BAL (IQR 4.1-7.0 ml) and 4.0 ml plasma (IQR 3.8-4.9 ml) from 38 cancer subjects to input for library preparation and sequencing (Supplemental Table 3). DNA quantity extracted from BAL cfDNA (p=0.71) and plasma (p=0.29) did not differ stratified by stage I/II vs. III/IV lung cancer. Median deduplicated read depth per sample for BAL cfDNA and plasma cfDNA samples was (2,228 [IQR 1,366-3,240) and (3,612 [IQR 2,666-4,903]) respectively (p<0.001, Supplemental Figure 2). Read depth did not differ by tumor stage for BAL cfDNA (p=0.46) or plasma cfDNA (p=0.30).

We identified a median of 4 mutations per tumor. Mutations affecting *TP53* and *KRAS* were the most frequent alterations detected (41% and 35% of tumors respectively, Supplemental Table 4). Patient smoking status (ever vs. never) was associated with more tumor mutations (p=0.015), but age (>65 years, p=0.66), gender (p=0.30) and stage (I/II vs. III/IV, p=0.33) were not.

Using tumor-informed analysis (which leverages prior knowledge of somatic mutations from tumor tissue sequencing), we identified tumor-derived variants (SNVs and indels) in 81% of BAL cfDNA samples and 47% of plasma cfDNA samples (p=0.016, Table 1). There was no significant association between stage and the number of tumor derived mutations identified in BAL cfDNA (I-II vs. III-IV, p=0.96, Supplemental Table 5). Single nucleotide variants were identified in 21 of 27 BAL cfDNA samples (78%) compared to 14 of 27 plasma

cfDNA samples (52%, p=0.12, Figure 2A & Supplemental Table 5). Furthermore, the tumor derived mean VAF% of SNVs was higher in BAL cfDNA than for plasma cfDNA in 22 of 27 patients (p=0.001, Figure 2B).

We then applied a tumor naïve calling approach that would approximate a clinical scenario where tumor mutation data was not available, such as in the diagnostic setting (Supplemental Tables 6, 7). We again observed that analysis of BAL cfDNA identified more tumor-associated variants than analysis of plasma (SNVs and indels, p=0.004). Using the tumor naïve approach, we detected at least one tumor-derived mutation in 20 of 27 (74%) patients using BAL cfDNA but only 4 of 27 (15%) patients using plasma cfDNA (Figure 2C & Supplemental Figure 3A). Additionally, among early-stage patients BAL cfDNA harbored more variants (12/18, 67% vs. 2/18, 11%) with a higher median VAF% (0.74% vs. 0.0%; p=0.002) than plasma (Figure 2D). Nineteen of the 27 BAL samples with variants identified by tumor naïve calling had higher mean VAF% in BAL cfDNA compared to plasma cfDNA (p=0.003, Figure 2E).

As expected, fewer tumor-derived mutations were detected in both BAL cfDNA and plasma cfDNA using a naïve calling strategy (n=27 subjects) when compared to the informed approach. This is due to the fact that tumor informed CAPP-Seq analysis decreases multiple hypothesis testing by only interrogating positions known to be mutant in the matching tumor and achieves an 10-fold lower detection limit (~0.01% vs ~0.1-0.5%) (12, 17). However, due to the higher concentrations of tumor-derived DNA, analysis of BAL cfDNA identified more mutations than plasma cfDNA (p<0.001, Supplemental Figure 3B). Importantly, statistically significant decreases between the two approaches were noted for plasma cfDNA (p<0.001) but not for BAL cfDNA. When focusing on cancer driver genes only, the proportion of mutations in these genes detected in BAL cfDNA was again higher compared to plasma cfDNA (p<0.001) and not significantly different between the two approaches (Supplemental Figure 3C).

### Development of a diagnostic classifier from tumor-derived mutations in bronchoalveolar lavage fluid

A common indication for bronchoscopy and BAL is for diagnosis of lung cancer in patients with lung nodules. Therefore, we wished to explore if BAL cfDNA analysis might aid in distinguishing lung cancer patients from at risk controls. One potential complication for such an approach is field cancerization (FC), which refers to the acquisition of somatic mutations in morphologically normal appearing tissues (28). Because the lung is susceptible to FC, (29-31) we aimed to compare results of tumor-naïve BAL cfDNA analysis in 35 cancer patients and 21 non-cancer controls at risk for lung cancer who underwent bronchoscopy as part of research studies at two medical centers (Supplementary Methods).

The 21 non-lung cancer controls consisted of 7 subjects with benign nodules or masses who underwent bronchoscopy for lung cancer evaluation and 14 subjects that were current or ex-smokers from a lung cancer screening program, 4 of whom had lung nodules detected on their screening CT. Gender (p=0.72), and smoking status (p=0.53) were not significantly different between cases and controls with age showing a trend towards difference (p=0.07). The concentration of cfDNA and sequencing depth did not differ in the two groups

(p=0.79 and p=0.21 respectively; Supplemental Table 3). While the mean VAF% of detected mutations (Figures 3A) and the frequency of lung cancer driver mutations (Figure 3B) was significantly lower in both biofluids for at-risk controls compared to cancer patients, mutations in lung cancer driver genes were identified in 4/21 (19%) of BAL cfDNA controls (Figure 3C; Supplemental Table 8). A total of 38 driver mutations were detected in BAL cfDNA among the lung cancer patients, 25 of which were also present in matched tumor tissue (concordance 66%) This indicates that the presence of mutations in cancer driver genes alone is insufficiently specific for distinguishing between controls and patients with lung cancer using BAL cfDNA.

We therefore performed an exploratory analysis to test if it is possible to develop a machine learning-based classifier to distinguish between the two groups of patients. Specifically, we trained a multivariable BAL genomic classifier using a random forest model on 56 samples (35 cases and 21 controls) and eleven gene features (Supplemental Table 9). Performance was evaluated using leave-one-out cross validation (Figure 4A). The three features with the largest impact on model performance were mean VAF% of detected mutations, number of cancer driver mutations detected, and number of total single nucleotide variants detected (Figure 4B, Supplemental Figure 4A). The genomic classifier achieved an AUC of 0.84 with all eleven features incorporated (Figure 4C), with 69% sensitivity at 100% specificity. Notably, the classifier outperformed BAL cytology (p=0.001) for both early- and late-stage patients (Figure 4D), and it was not associated with patient age, gender, and smoking status (Supplemental Figure 4B). Of the 17 lung cancer cases profiled with BAL cytology, 2 (12%) were diagnosed with lung cancer by cytology, compared to 11 (65%) using the BAL genomic classifier (Figure 4E).

## Discussion

Here, we compared BAL cfDNA and plasma cfDNA as two different sources of tumor derived DNA and found that BAL cfDNA analysis is more sensitive than plasma for identifying lung cancer-derived mutations. We also explored the potential of tumor naïve BAL cfDNA analysis for detection of lung cancer. Our results suggest that BAL cfDNA analysis could have clinical utility for identification of mutations in lung cancer patients and, potentially, for the diagnosis of lung cancer.

Although previous studies have demonstrated the feasibility of tumor genotyping via BAL (25, 26, 32), we directly compared BAL to plasma cfDNA. We observed that tumor DNA concentrations were significantly higher in BAL than in plasma and more likely to be above our assay's detection limit. Using tumor-informed analysis, we found higher concentrations of tumor derived DNA and an increased sensitivity for identifying tumor mutations in BAL cfDNA samples (17/22, 77%) compared to plasma cfDNA samples 10/22 (45%) in early stage I-II disease. Tumor naïve analysis, which is clinically relevant in the diagnostic setting when tumor profiling is not available, also demonstrated higher ctDNA concentrations and better performance in BAL cfDNA (12/18, 67%) than plasma cfDNA in stage I-II disease (2/18, 11%). Only a subset of stage I-II lung cancers can be detected using ultra-sensitive plasma cfDNA profiling methods (19, 33, 34). Additionally, our study adds to an emerging literature that has demonstrated successful high throughput genome sequencing of proximal

fluids in several cancer types including CSF for gliomas (35-37), urine for genito-renal cancers (38, 39), and lavage or ascites for gynecologic cancers (40-42).

Several groups have demonstrated that gene expression profiling of histologically normal bronchial mucosal tissue can identify patients at high risk for developing lung cancer, due to a process called field cancerization (FC) (29, 30, 43). Genomic analysis of brushing specimens and single cell analysis has recently confirmed this effect at the DNA level (31, 44). Our data are consistent with these findings. Specifically, our observation of mutations in cancer driver genes in BAL cfDNA that were not found in the matching tumor specimens and were also detected in at risk patients without lung cancer (Fig. 3C) supports the existence of FC and less likely tumor heterogeneity. We speculate that these mutations reflect clonal mutagenesis of airway epithelial cells, analogous to the clonal mutations observed in tissues such as blood, esophagus, skin, and uterus (45-48). Finally, our exploratory analysis developing a BAL cfDNA-based classifier suggests that it may be possible to leverage FC to diagnose lung cancer in patients with non-diagnostic bronchoscopic biopsies.

Strengths of our work include benchmarking of alterations identified in plasma cfDNA and BAL cfDNA to those present in matched tumor samples, sequencing of matched leukocytes to remove germline alterations and alterations arising due to CH, use of risk-matched controls to account for FC, and utilization of a validated and ultra-sensitive sequencing method to detect tumor derived mutations from cfDNA.

Our study has a number of limitations. First, we performed this work on patients enrolled in an observational study of BAL biomarkers without pre-defined assay criteria, since this was a proof-of-concept and method development study. Second, the majority of patients had adenocarcinoma and we therefore did not have sufficient power to examine the impact of tumor histology on the ability to detect tumor-derived variants in BAL. Third, we developed our BAL genomics classifier using a cross-validation framework, and therefore validation in an independent cohort will be required. Lastly, collection protocols were not standardized between the two centers.

Before we can realistically consider BAL genomics for mutational profiling or detection of lung cancer in the clinic, further studies are required to enable a better understanding of how pre-analytical variables influence tumor DNA identification in BAL. Standardization of collection methods will facilitate robust genomic profiling of BAL in larger cohorts of patients. Additionally, it will be important to investigate alternative tumor DNA detection strategies such as whole genome sequencing or DNA methylation analysis (34, 49, 50) and to compare these to our mutation-based approach. Lastly, it will be informative to explore if features such as histology, grade and location in the lung are associated with shedding of tumor DNA in BAL. These studies will more fully elucidate the clinical utility of BAL genomic profiling during lung cancer evaluation as a complementary liquid biopsy approach to blood-based analyses (51).

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements:

## References

1. Chansky K, Sculier JP, Crowley JJ, Giroux D, Van Meerbeeck J, Goldstraw P, et al. The International Association for the Study of Lung Cancer Staging Project: prognostic factors and pathologic TNM stage in surgically managed non-small cell lung cancer. J Thorac Oncol. 2009;4(7):792–801. [PubMed: 19458556]

2. Pignon JP, Tribodet H, Scagliotti GV, Douillard JY, Shepherd FA, Stephens RJ, et al. Lung adjuvant cisplatin evaluation: a pooled analysis by the LACE Collaborative Group. J Clin Oncol. 2008;26(21):3552–9. [PubMed: 18506026]

3. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502(7471):333–9. [PubMed: 24132290]

4. Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, Maki RG, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med. 2010;363(18):1693–703. [PubMed: 20979469]

5. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 2004;350(21):2129–39. [PubMed: 15118073]

6. Folch EE, Pritchett MA, Nead MA, Bowling MR, Murgu SD, Krimsky WS, et al. Electromagnetic Navigation Bronchoscopy for Peripheral Pulmonary Lesions: One-Year Results of the Prospective, Multicenter NAVIGATE Study. J Thorac Oncol. 2019;14(3):445–58. [PubMed: 30476574]

7. Ost DE, Ernst A, Lei X, Kovitz KL, Benzaquen S, Diaz-Mendoza J, et al. Diagnostic Yield and Complications of Bronchoscopy for Peripheral Lung Lesions. Results of the AQuIRE Registry. Am J Respir Crit Care Med. 2016;193(1):68–77. [PubMed: 26367186]

8. Lim JH, Kim MJ, Jeon SH, Park MH, Kim WY, Lee M, et al. The optimal sequence of bronchial brushing and washing for diagnosing peripheral lung cancer using non-guided flexible bronchoscopy. Sci Rep. 2020;10(1):1036. [PubMed: 31974454]

9. Tajarernmuang P, Ofiara L, Beaudoin S, Gonzalez AV. Bronchoscopic tissue yield for advanced molecular testing: are we getting enough? J Thorac Dis. 2020;12(6):3287–95. [PubMed: 32642252]

10. Baughman RP. Technical aspects of bronchoalveolar lavage: recommendations for a standard procedure. Semin Respir Crit Care Med. 2007;28(5):475–85. [PubMed: 17975775]

11. Rivera MP, Mehta AC, Wahidi MM. Establishing the diagnosis of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest. 2013;143(5 Suppl):e142S–e65S. [PubMed: 23649436]

12. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nature medicine. 2014;20(5):548–54.

13. Hori SS, Gambhir SS. Mathematical model identifies blood biomarker-based early cancer detection strategies and limitations. Sci Transl Med. 2011;3(109):109ra16.

14. Dudley JC, Schroers-Martin J, Lazzareschi DV, Shi WY, Chen SB, Esfahani MS, et al. Detection and Surveillance of Bladder Cancer Using Urine Tumor DNA. Cancer Discov. 2019;9(4):500–9. Epub 2018/12/24. [PubMed: 30578357]

15. Wang Y, Li L, Douville C, Cohen JD, Yen TT, Kinde I, et al. Evaluation of liquid from the Papanicolaou test and other liquid biopsies for the detection of endometrial and ovarian cancers. Sci Transl Med. 2018;10(433). Epub 2018/03/23.

16. Springer SU, Chen CH, Rodriguez Pena MDC, Li L, Douville C, Wang Y, et al. Non-invasive detection of urothelial cancer through the analysis of driver gene mutations and aneuploidy. Elife. 2018;7. Epub 2018/03/21.

17. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nature biotechnology. 2016;34(5):547–55. Epub 2016/03/29.

18. Chaudhuri AA, Chabon JJ, Lovejoy AF, Newman AM, Stehr H, Azad TD, et al. Early Detection of Molecular Residual Disease in Localized Lung Cancer by Circulating Tumor DNA Profiling. Cancer Discov. 2017;7(12):1394–403. [PubMed: 28899864]

19. Chabon JJ, Hamilton EG, Kurtz DM, Esfahani MS, Moding EJ, Stehr H, et al. Integrating genomic features for non-invasive early lung cancer detection. Nature. 2020;580(7802):245–51. [PubMed: 32269342]

20. Yang SR, Lin CY, Stehr H, Long SR, Kong CS, Berry GJ, et al. Comprehensive Genomic Profiling of Malignant Effusions in Patients with Metastatic Lung Adenocarcinoma. J Mol Diagn. 2018;20(2):184–94. [PubMed: 29269277]

21. Chabon JJ, Simmons AD, Lovejoy AF, Esfahani MS, Newman AM, Haringsma HJ, et al. Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. Nature communications. 2016;7:11815.

22. Scherer F, Kurtz DM, Newman AM, Stehr H, Craig AF, Esfahani MS, et al. Distinct biological subtypes and patterns of genome evolution in lymphoma revealed by circulating tumor DNA. Sci Transl Med. 2016;8(364):364ra155.

23. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;173(2):371–85 e18. [PubMed: 29625053]

24. Buttitta F, Felicioni L, Del Grammastro M, Filice G, Di Lorito A, Malatesta S, et al. Effective assessment of egfr mutation status in bronchoalveolar lavage and pleural fluids by next-generation sequencing. Clinical cancer research : an official journal of the American Association for Cancer Research. 2013;19(3):691–8. [PubMed: 23243218]

25. Ryu JS, Lim JH, Lee MK, Lee SJ, Kim HJ, Kim MJ, et al. Feasibility of Bronchial Washing Fluid-Based Approach to Early-Stage Lung Cancer Diagnosis. Oncologist. 2019;24(7):e603–e6. [PubMed: 31036768]

26. Roncarati R, Lupini L, Miotto E, Saccenti E, Mascetti S, Morandi L, et al. Molecular testing on bronchial washings for the diagnosis and predictive assessment of lung cancer. Mol Oncol. 2020;14(9):2163–75. [PubMed: 32441866]

27. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer. 2017;17(4):223–38. [PubMed: 28233803]

28. Curtius K, Wright NA, Graham TA. An evolutionary perspective on field cancerization. Nat Rev Cancer. 2018;18(1):19–32. Epub 2017/12/09. doi: 10.1038/nrc.2017.102. [PubMed: 29217838]

29. Kadara H, Fujimoto J, Yoo SY, Maki Y, Gower AC, Kabbout M, et al. Transcriptomic architecture of the adjacent airway field cancerization in non-small cell lung cancer. Journal of the National Cancer Institute. 2014;106(3):dju004. [PubMed: 24563515]

30. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nature medicine. 2007;13(3):361–6.

31. Kadara H, Sivakumar S, Jakubek Y, San Lucas FA, Lang W, McDowell T, et al. Driver Mutations in Normal Airway Epithelium Elucidate Spatiotemporal Resolution of Lung Cancer. Am J Respir Crit Care Med. 2019.

32. Zeng D, Wang C, Mu C, Su M, Mao J, Huang J, et al. Cell-free DNA from bronchoalveolar lavage fluid (BALF): a new liquid biopsy medium for identifying lung cancer. Ann Transl Med. 2021;9(13):1080. [PubMed: 34422992]

33. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature. 2017;545(7655):446–51. [PubMed: 28445469]

34. Klein EA, Richards D, Cohn A, Tummala M, Lapham R, Cosgrove D, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. Ann Oncol. 2021;32(9):1167–77. [PubMed: 34176681]

35. De Mattos-Arruda L, Mayor R, Ng CK, Weigelt B, Martinez-Ricarte F, Torrejon D, et al. Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. Nature communications. 2015;6:8839.

36. Pentsova EI, Shah RH, Tang J, Boire A, You D, Briggs S, et al. Evaluating Cancer of the Central Nervous System Through Next-Generation Sequencing of Cerebrospinal Fluid. J Clin Oncol. 2016;34(20):2404–15. [PubMed: 27161972]

37. Panditharatna E, Kilburn LB, Aboian MS, Kambhampati M, Gordish-Dressman H, Magge SN, et al. Clinically Relevant and Minimally Invasive Tumor Surveillance of Pediatric Diffuse Midline Gliomas Using Patient-Derived Liquid Biopsy. Clinical cancer research : an official journal of the American Association for Cancer Research. 2018;24(23):5850–9. [PubMed: 30322880]

38. Sin MLY, Mach KE, Sinha R, Wu F, Trivedi DR, Altobelli E, et al. Deep Sequencing of Urinary RNAs for Bladder Cancer Molecular Diagnostics. Clinical cancer research : an official journal of the American Association for Cancer Research. 2017;23(14):3700–10. [PubMed: 28193625]

39. Dudley JC, Schroers-Martin J, Lazzareschi DV, Shi WY, Chen SB, Esfahani MS, et al. Detection and surveillance of bladder cancer using urine tumor DNA. Cancer Discov. 2018.

40. Maritschnegg E, Wang Y, Pecha N, Horvat R, Van Nieuwenhuysen E, Vergote I, et al. Lavage of the Uterine Cavity for Molecular Detection of Mullerian Duct Carcinomas: A Proof-of-Concept Study. J Clin Oncol. 2015;33(36):4293–300. [PubMed: 26552420]

41. Shah RH, Scott SN, Brannon AR, Levine DA, Lin O, Berger MF. Comprehensive mutation profiling by next-generation sequencing of effusion fluids from patients with high-grade serous ovarian carcinoma. Cancer cytopathology. 2015;123(5):289–97. [PubMed: 25655233]

42. Nair N, Camacho-Vanegas O, Rykunov D, Dashkoff M, Camacho SC, Schumacher CA, et al. Genomic Analysis of Uterine Lavage Fluid Detects Early Endometrial Cancers and Reveals a Prevalent Landscape of Driver Mutations in Women without Histopathologic Evidence of Cancer: A Prospective Cross-Sectional Study. PLoS Med. 2016;13(12):e1002206. [PubMed: 28027320]

43. Silvestri GA, Vachani A, Whitney D, Elashoff M, Porta Smith K, Ferguson JS, et al. A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. N Engl J Med. 2015;373(3):243–51. [PubMed: 25981554]

44. Yoshida K, Gowers KHC, Lee-Six H, Chandrasekharan DP, Coorens T, Maughan EF, et al. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature. 2020;578(7794):266–72. [PubMed: 31996850]

45. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature. 2019.

46. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science (New York, NY). 2018;362(6417):911–7.

47. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science (New York, NY). 2015;348(6237):880–6.

48. Moore L, Leongamornlert D, Coorens THH, Sanders MA, Ellis P, Dentro SC, et al. The mutational landscape of normal human endometrial epithelium. Nature. 2020;580(7805):640–6. [PubMed: 32350471]

49. Zviran A, Schulman RC, Shah M, Hill STK, Deochand S, Khamnei CC, et al. Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. Nature medicine. 2020;26(7):1114–24.

50. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature communications. 2017;8(1):1324.

51. Rolfo C, Mack P, Scagliotti GV, Aggarwal C, Arcila ME, Barlesi F, et al. Liquid Biopsy for Advanced NSCLC: A Consensus Statement From the International Association for the Study of Lung Cancer. J Thorac Oncol. 2021;16(10):1647–62. [PubMed: 34246791]
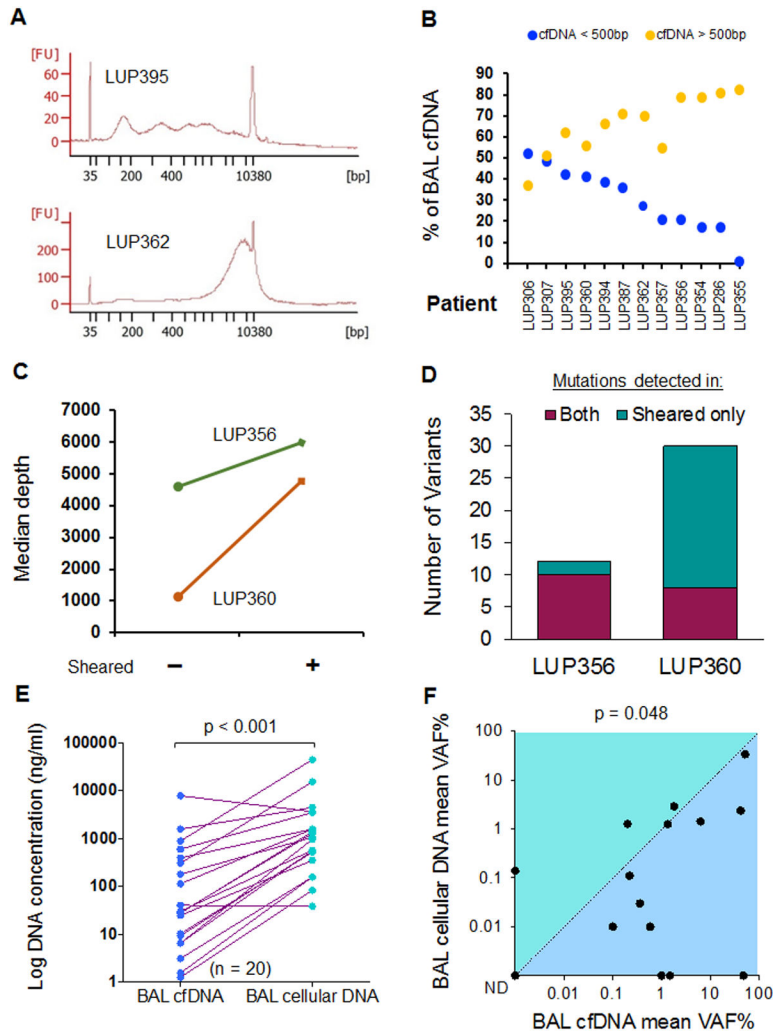
**Figure 1. Characteristics of bronchoalveolar lavage DNA from lung cancer**

**(A)** cfDNA fragment size distribution for two representative cell free BAL (BAL cfDNA) samples. FU = Fluorescence units. BP = base pairs.

**(B)** DNA fragment size for BAL cfDNA stratified by a 500 bp threshold (orange > 500 bp, blue < 500 bp) for n = 12 patients (x-axis).

**(C)** Median sequencing depth obtained from BAL cfDNA samples with and without shearing.

**(D)** Number of mutations detected in both sheared and un-sheared samples or in sheared samples alone for the two patients from C. A total of 42 mutations were detected in sheared samples versus 18 in unsheared (p < 0.001).

**(E)** BAL DNA concentrations in 20 lung cancer patients for BAL cfDNA and BAL cellular DNA.

**(F)** Comparison of mean VAF% for BAL cfDNA and BAL cellular DNA for 15 patients. ND = not detected. P-value is displayed above the plot.
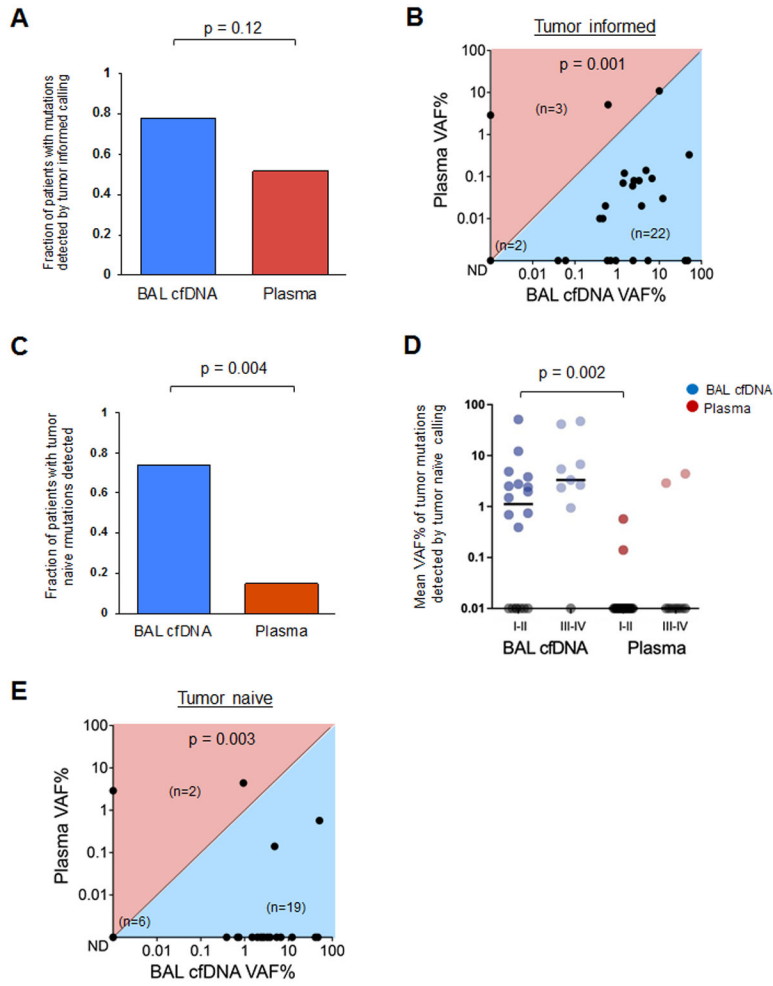
**Figure 2. Bronchoalveolar lavage identifies more tumor mutations than plasma**

**(A)** Fraction of patients with tumor-derived mutations identified in BAL cfDNA (blue) compared to plasma cfDNA (red) using the tumor informed approach for 27 patients (Supplementary Table 5).

**(B)** Mean Variant Allele Frequency (mean VAF%) in BAL and plasma cfDNA using the tumor informed approach to identify tumor-derived variants. Each dot represents one patient. Numbers of patients in each half of the plot are shown in parentheses within the plot along with the number of samples with no tumor derived mutations detected (ND).

**(C)** Driver Mutations identified in BAL cfDNA (blue) compared to plasma cfDNA (red) using the tumor naive calling approach for 27 patients (Supplemental Table 7).

**(D)** Mean VAF% of mutant DNA detected in BAL cfDNA (blue) or plasma cfDNA (red) by stage using the tumor naïve calling strategy (p = 0.002). Each dot represents one patient and samples with no mutations detected (ND) are shown on the x-axis in gray.

**(E)** Mean VAF% in BAL and plasma cfDNA using the tumor naïve approach. Each dot represents one patient. Numbers of patients in each half of the plot are shown in parentheses within the plot along with the number of samples with no tumor derived mutations detected (ND).
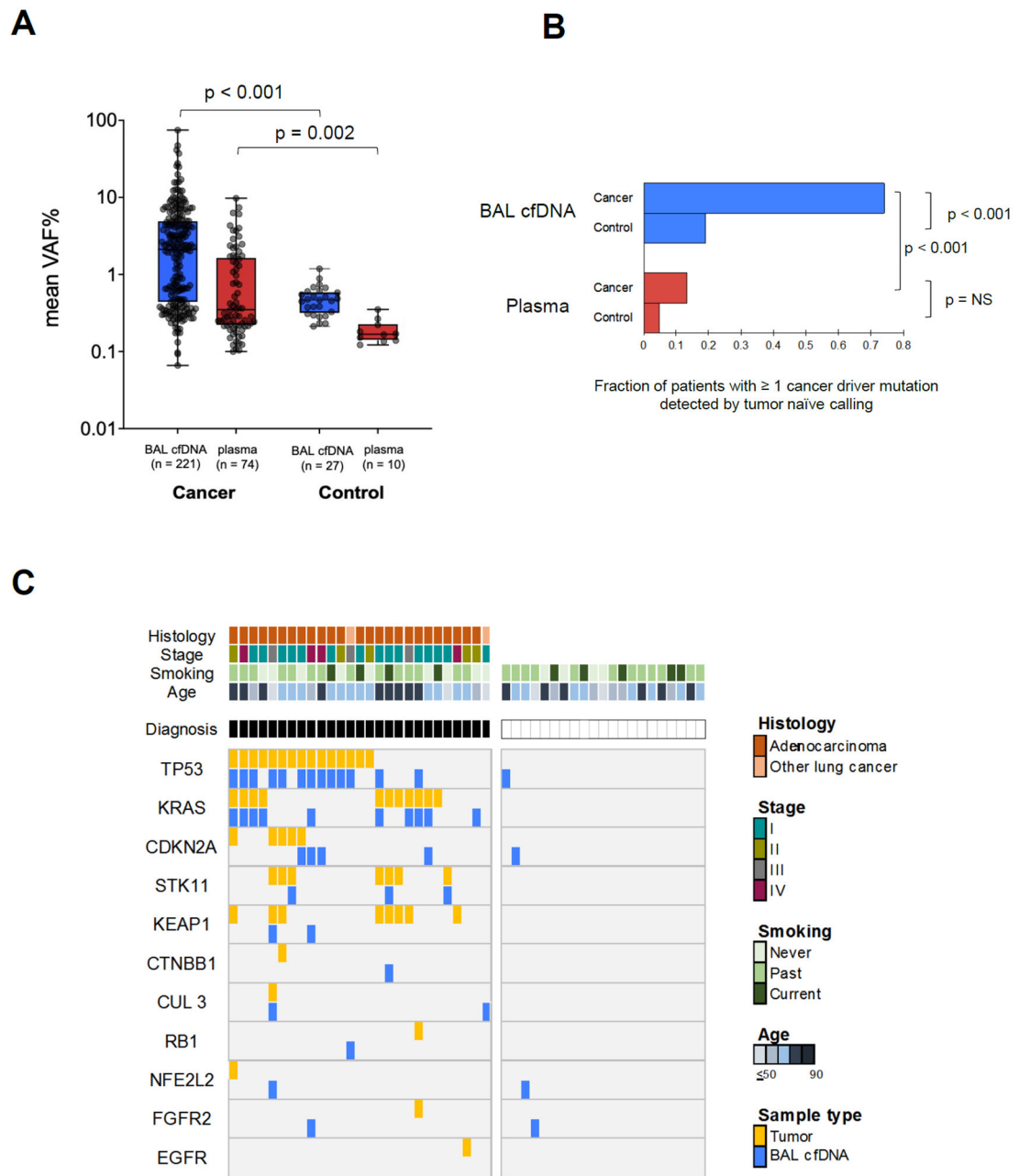
**Figure 3. Tumor mutation profiling of bronchoalveolar lavage fluid in at risk controls identifies field cancerization**

**(A)** Mean VAF% (y-axis, $\log_{10}$ scale) for mutations detected in lung cancer patients compared to risk-matched controls for BAL cfDNA (p<0.001) and plasma cfDNA (p=0.002) by the tumor naïve calling approach (Supplemental Table 8). Box plots depict median values, 25-75% interquartile range, and minimal and maximal values. The number of variants represented in the scatter plots are denoted in the x-axis labels. Blue = BAL cfDNA. Red = plasma. P-values are displayed above the plot.

**(B)** Fraction of cancer patients and controls with driver mutations detected in BAL (blue) or plasma cfDNA (red). NS = not significant.

**(C)** Oncoprint of lung cancer driver genes in 27 lung cancer cases and 21 risk-matched controls for BAL cfDNA profiles (Supplemental Table 8). Tumor DNA results are shown for cancer patients. Each column denotes one patient and each row a driver mutation. For mutations found in tumors, BAL cfDNA mutations are only shown if they are identical. Clinical characteristics (tumor histology, tumor stage according to AJCC VIII guidelines, smoking status, and age in years) are displayed. NSCLC = Non-small cell lung cancer.
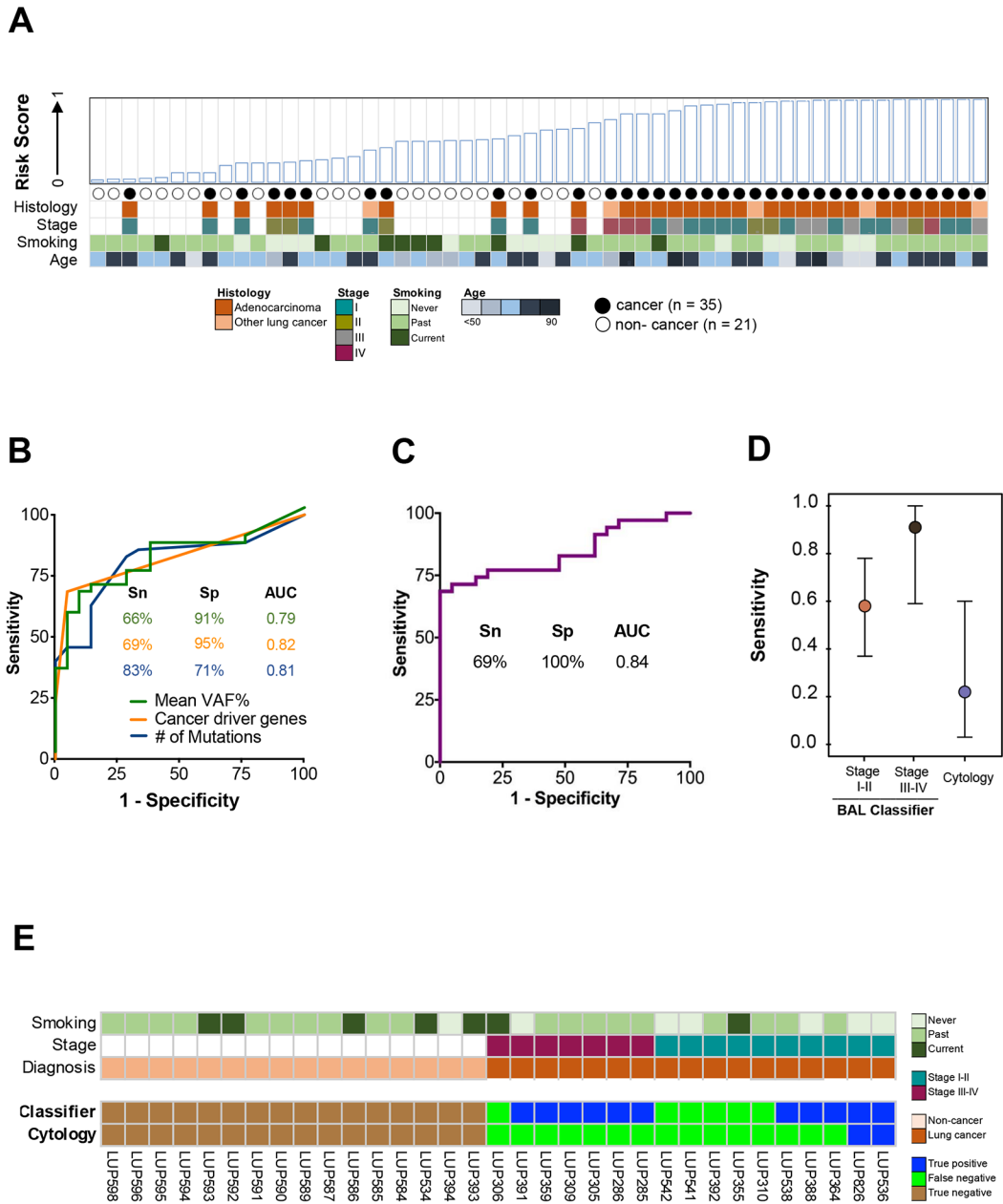
**Figure 4. Derivation and performance of a BAL cfDNA genomic classifier for lung cancer diagnosis**

**(A)** BAL risk score based on eleven genomic features (Supplemental Table 9; **see methods**). Case-control status and relevant clinico-pathologic variables are indicated. NSCLC = Non-small cell lung cancer.

**(B)** Individual performance of the top three features contributing to the BAL cfDNA classifier at the optimal cut point (Youden's value). Mean VAF% (green line), cancer driver genes (orange line), and number of mutations (blue line). Sn = Sensitivity. Sp = Specificity. AUC = Area under the curve.

**(C)** Performance of the BAL cfDNA classifier using all eleven features at the optimal cut point (Supplemental Table 9). Sp = Specificity. AUC = Area under the curve.

**(D)** Sensitivity of the BAL genomic classifier and BAL cytology at a specificity of 100% for diagnosing lung cancer stratified by stage. Point estimates are represented by a circle and 95% confidence intervals by whiskers. Analysis based on 17 cancer patients and 16 controls who had both cytology and BAL genomics scores available.

**(E)** Patient level comparison of BAL cytology and BAL genomic classifier risk scores using a 100% specificity threshold. Seventeen cancer patients and 16 non-cancer patients that had both cytology and a risk score assigned to them are arranged on the x-axis by stage and cancer status.

**Table 1.**

Tumor mutation characteristics by fluid type

| Variant Statistic | Tumor n = 34 | BAL cfDNA n = 31 | Plasma n = 34 |
|---|---|---|---|
| Tumor variants detected [*] | 34 (100%) | 25 (81%) | 16 (47%) |
| Mean number of variants | 5.1 | 1.9 | 0.91 |
| Median number of variants | 3.5 | 1.0 | 0.0 |
| Mean VAF% | 12.9 | 6.6 | 2.5 |
| Median VAF% | 4.2 | 2.4 | 0.09 |
| Mean VAF%, drivers only | 17.4 | 9.0 | 3.3 |
| Median VAF%, drivers only | 8.9 | 2.4 | 0.08 |

[*] Including indels, see Supp Table 5; p < 0.05 level using a Monte-Carlo approach described in the methods.