Subtype-based microbial analysis in non-small cell lung cancer

Hye Jin Jang, MD1,2, Eunkyung Lee3, Young-Jae Cho, MD, MPH, PhD4; Sang Hoon Lee, MD, PhD1.

1Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea
2Department of internal medicine, Myongji Hospital, Hanyang University College of Medicine, Goyang-si, Gyeonggi-do, Republic of Korea
3Department of Biomedical Science and Engineering, Gwangju Institute of Science and Technology, Gwangju 61005, Republic of Korea
4Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam-si, Gyeonggi-do, Republic of Korea

*Correspondence:
Sang Hoon Lee, M.D, Ph.D.
Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Institute of Chest Diseases, Severance Hospital, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
Tel: +82-2-2228-1955; Fax: +82-2-393-6884
E-mail address: cloud9@yuhs.ac

Running title: Microbiome in lung cancer
Authors’ Contributions

Conceptualization: HJ Jang, YJ Cho and SH Lee

Methodology: HJ Jang, EK Lee, YJ Cho and SH Lee

Formal analysis: HJ Jang, EK Lee, YJ Cho and SH Lee

Data curation: HJ Jang, and SH Lee

Software: SH Lee

Validation: HJ Jang, YJ Cho, and SH Lee

Investigation: SH Lee

Writing – original draft preparation: HJ Jang, YJ Cho, and SH Lee

Writing – review and editing: HJ Jang, EK Lee, YJ Cho, and SH Lee

All authors read and approved the final manuscript.

Conflict of Interest

No potential conflicts of interest relevant to this article were reported.
ABSTRACT

**Background:** The human lung serves as a niche for a unique and dynamic bacterial community related to the development and aggravation of multiple respiratory diseases. Therefore, identifying the microbiome status is crucial to maintaining the microecological balance and maximizing the therapeutic effect on lung diseases. Therefore, we investigated the histological type-based differences in the lung microbiomes of patients with lung cancer.

**Materials and Methods:** We performed 16S rRNA sequencing to evaluate the respiratory tract microbiome present in bronchoalveolar lavage fluid. Patients with non-small cell lung cancer were stratified based on two main subtypes of lung cancer: adenocarcinoma and squamous cell carcinoma (SqCC).

**Results:** Among the 84 patients analyzed, 64 (76.2%) had adenocarcinoma, and 20 (23.8%) had SqCC. The α- and β-diversities showed significant differences between the two groups (p = 0.004 for Chao1, p = 0.001 for Simpson index, and p = 0.011 for PERMANOVA). *Actinomyces graevenitzii* was dominant in the SqCC group (LDA score, 2.46); the populations of *Haemophilus parainfluenza* (LDA score, 4.08), *Neisseria subflava* (LDA score, 4.07), *Porphyromonas endodontalis* (LDA score, 3.88), and *Fusobacterium nucleatum* (LDA score, 3.72) were significantly higher in the adenocarcinoma group.

**Conclusions:** Microbiome diversity is crucial for maintaining homeostasis in the lung environment, and dysbiosis may be related to the development and prognosis of lung cancer. The mortality rate was high, and the microbiome was not diverse in SqCC. Further large-scale studies are required to investigate the role of the microbiome in the development of different lung cancer types.

**Keywords:** Microbiome, Non-small cell lung cancer, Diversity
**Introduction**

The human body, including the gut, skin, and other mucosal environments, is colonized by several microorganisms (1), and the interaction between these microbiomes and the immune system has been continuously studied. In healthy individuals, this interaction can contribute to immune homeostasis and susceptibility to infectious and inflammatory diseases (2).

The human lung serves as a niche for a unique and dynamic bacterial community related to the development of multiple respiratory diseases, such as chronic obstructive pulmonary disease (3), tuberculosis (4), idiopathic pulmonary fibrosis, and cystic fibrosis (5). The lung microbiota is essential for barrier function, immune homeostasis, and anticancer immune surveillance via tumor antigenicity in healthy individuals (6). Bacteria may disrupt the cell cycle by toxin production, resulting in cell growth with alterations in protein expression controlling DNA repair, cell division, and apoptosis (7), thus influencing the host immune response against malignant cells and promoting disease and cancer development (8). These include *Papillomaviridae*, causing cervical cancer; *Helicobacter pylori*, contributing to non-cardia gastric cancer; and *Schistosoma hematobium*, responsible for bladder cancer (6).

Lung cancer is one of the most common cancers worldwide and is a leading cause of cancer-related morbidity (9). Lung cancer is a complicated disease caused by interactions between host and environmental factors (10). Among these risk factors, the microbiome plays a vital role in maintaining microecological balance and regulating host immune responses (11).

Recent studies using next-generation sequencing have revealed that the lung microbiome in patients with lung cancer differs from that in healthy individuals and that these microbiomes play a crucial role in immunity as well as cancer (12). Therefore, understanding the mechanisms by which microbes present in the airways can influence lung cancer development
and treatment could be beneficial to predicting cancer risk and improving treatment efficacy and safety (13).

Non-small cell lung carcinoma (NSCLC) is mainly classified into squamous cell carcinoma (SqCC), adenocarcinoma, and large cell carcinoma. The most common type of lung cancer is adenocarcinoma, which comprises approximately 40–60% of all lung cancer cases, whereas SqCC comprises 25–30% of all lung cancer cases (14). Large cell (undifferentiated) carcinomas account for 5–10% of lung cancers and are relatively rare. Several genetic alterations, including activating mutations in \textit{EGFR}, \textit{ALK}, \textit{ROS1}, \textit{RAS}, and \textit{BRAF}, have been identified as drivers of tumorigenesis in NSCLC (15). Although overall survival has improved owing to the use of targeted therapy for NSCLC patients with driver mutations and the introduction of immunotherapy, resistance eventually develops after using targeted anticancer agents for a certain period; additionally, targeted therapy for SqCC is limited (16). Furthermore, only a small portion of patients with lung cancer achieve optimal and sustained efficacy from immunotherapy. There is no effective anticancer drug for patients who are resistant to immunotherapy, and the prognosis is poor (17). Therefore, identifying the microbiome status is crucial to maintaining the microecological balance within the lungs, which could prevent the progression of lung cancer and maximize the effect of immunotherapy in NSCLC.

In this study, we aimed to investigate the microbial differences in patients with NSCLC according to the different subtypes of SqCC and adenocarcinoma, which are predominant in NSCLC.

\textbf{Methods}

\textit{Patient recruitment and sample collection}
A total of 84 patients who were pathologically diagnosed with NSCLC were recruited from June 1, 2018, to June 31, 2020. Patients admitted for a lung cancer diagnosis in two tertiary hospitals—the Severance Hospital and Bundang Seoul Hospital, in South Korea—. A bronchoscopy specialist collected Bronchoalveolar lavage (BAL) fluid samples using a sterile bronchoscope.

Sample collection
Before bronchoscopy, all patients washed their mouths twice with a sterile saline solution. The patients were given topical anesthesia (lidocaine) using a nebulizer. Subsequently, patients were administered with midazolam and fentanyl, as recommended (18). The bronchoscope was placed into the mouth of the patients and then placed into the lungs. BAL fluid was obtained according to a standardized protocol: when the bronchoscope arrived in the “involved” airway containing lung masses or lung nodules, the bronchi were flushed with 30–50 mL of sterile saline (0.9%). Approximately 15 mL of BAL fluid samples were obtained from each patient for sequencing analysis. BAL fluid samples were immediately placed at −70 °C in a freezer, and DNA extraction was conducted within 24 h.

DNA Extraction, Polymerase Chain Reaction (PCR) Amplification, and Sequencing
According to the manufacturer’s instructions, total DNA was extracted using the Maxwell RSC PureFood GMO and Authentication Kit (Promega). We performed PCR amplification using fusion primers which were targeting the V3–V4 regions of the 16S rRNA gene of the extracted DNA. The fusion primers 341F (‘AATGATACGCGACACCGAGATCTACAC-XXXXXXXTCGTCGGCAGCGAC-AGATGTGTATAAGAGACAG-CCTACGGGNGGCWGCAG-3’) and 805R (5′-CAAGCAGAAGACGACGATACAGAGATGCTCGTGGGCTCGAGAT-XXXXXXXGTCTCGTGGGCTCGAG-AGATGTGTATAAGAGACAG-3′) were used.
GACTACHVGGGTATCTAATCC-3′) were used for bacterial amplification; the underlined sequence indicates the target region of the primer. We performed amplification under the following conditions: initial denaturation at 95 °C for 3 min followed by 25 cycles of denaturation at 95 °C for 30 s, primer annealing at 55 °C for 30 s, extension at 72 °C for 30 s, and a final elongation step at 72 °C for 5 min.

The PCR product was identified by performing 1% agarose gel electrophoresis and visualized using the Gel Doc system (BioRad, Hercules, CA, USA). The amplified products were purified using a CleanPCR kit (CleanNA; Waddinxveen, Netherlands). The purified products at the same concentrations were pooled, and short fragments (non-target products) were cleared using CleanPCR. We assessed quality and product size with a Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA, USA) using a DNA 7500 chip. Following the manufacturer’s instructions, mixed amplicons were pooled, and sequencing was performed at CJ Bioscience, Inc. (Seoul, Korea) using an Illumina MiSeq Sequencing system (Illumina, San Diego, CA, USA). Detailed methods are described in the supplementary files.

Statistical Analysis
Categorical variables are reported as numbers (percentages). Continuous variables with normal distribution are reported as mean ± standard deviation, while variables with abnormal distribution are reported as median with interquartile ranges (IQR, 25th–75th percentiles). Depending on the normality of distribution, categorical variables were compared using the chi-square test, and continuous variables were compared using either an independent t-test or Mann–Whitney U test. P values < 0.05 were considered statistically significant. All statistical analyses were performed with IBM SPSS Statistics version 25.0 (IBM, Armonk, NY, USA).

Ethics Approval and Patient Consent
The study protocol was approved by the Institutional Review Board of Severance Hospital, South Korea (IRB No. 4-0018-0313) and Seoul National University Bundang Hospital, South Korea (IRB No. B-1610/365-302). The study design was approved by the appropriate ethical review committee, and we obtained informed consents was from all participants.

Results

Patient Characteristics

The patients were divided into two groups according to their pathologic subtype: adenocarcinoma and SqCC. The baseline characteristics of the two groups are listed in Table 1. The adenocarcinoma group had 64 patients (76.2%), and 20 (23.8%) belonged to the SqCC group. The mean age was 66.7 ± 11.2 years, and the patients in the SqCC group were older (p = 0.033). Male gender (51.6% vs. 95.0%) and smoking history (46.9% vs. 90.0%) were dominant in the SqCC group, and the mortality rate was significantly higher in the SqCC group (6.3% vs. 25%, p = 0.035).

Table 1. Demographics and clinical characteristics of patients

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Adenocarcinoma</th>
<th>SqCC</th>
<th>Total</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>64 (76.2)</td>
<td>20 (23.8)</td>
<td>84 (100.0)</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>65.2 ± 10.5</td>
<td>71.5 ± 11.1</td>
<td>66.7 ± 11.2</td>
<td>0.033</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male, n (%)</td>
<td>33 (51.6)</td>
<td>19 (95.0)</td>
<td>52 (61.9)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>31 (48.4)</td>
<td>1 (5.0)</td>
<td>32 (38.1)</td>
<td></td>
</tr>
<tr>
<td>Smoking history</td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>Current or former, n (%)</td>
<td>30 (46.9)</td>
<td>18 (90.0)</td>
<td>48 (57.1)</td>
<td></td>
</tr>
<tr>
<td>Never, n (%)</td>
<td>34 (53.1)</td>
<td>2 (10.0)</td>
<td>36 (42.9)</td>
<td></td>
</tr>
<tr>
<td>Smoking amount (pack-years)</td>
<td>28.2 ± 15.0</td>
<td>37.0 ± 14.9</td>
<td>31.5 ± 15.4</td>
<td>0.055</td>
</tr>
<tr>
<td>Neutrophil-Lymphocyte ratio</td>
<td>4.76 (1.83, 6.27)</td>
<td>2.87 (1.95, 5.31)</td>
<td>3.05 (1.91, 5.85)</td>
<td>0.423</td>
</tr>
<tr>
<td>Stage, n (%)</td>
<td>25/8/14/17</td>
<td>2/2/9/7</td>
<td>27/10/23/24</td>
<td></td>
</tr>
</tbody>
</table>
Taxonomy Composition in Patients with Lung Cancer Based on Pathological Subtypes

Figures 1a and b depict the differences between lung microbiomes according to the pathological type. The dominant phyla in the adenocarcinoma group were Bacteroidetes (40.8%), Proteobacteria (24.9%), Firmicutes (24.1%), Fusobacteria (6.0%), and Actinobacteria (2.8%). In the SqCC group, Bacteroidetes (35.0%), Firmicutes (29.3%), Proteobacteria (27.8%), Fusobacteria (3.8%), and Actinobacteria (3.3%) were dominant.

The abundance of Fusobacteria differed between the adenocarcinoma and SqCC groups (Wilcoxon test, p = 0.024, Figure 2a), whereas that of Firmicutes, Bacteroidetes, and Proteobacteria were not statistically significant (Figure 2b–d). The ACE, Shannon, and Simpson indices were evaluated to estimate α-diversity in the lung microbiome, which summarizes the structure of an ecological community with respect to its richness (number of taxonomic groups), evenness (distribution of abundances of the groups), or both (19). The operational taxonomic units (OTUs) of both groups exhibited statistically significant differences (p = 0.006; Figure 3a). Species richness differed between the two groups based on Chao1 (p = 0.004, Figure 3b), Simpson (p = 0.001, Figure 3c), and Shannon indices (p = 0.0002, Figure 3d).

Principal coordinate analysis (PCoA) was performed to examine the similarity between the bacterial communities in each group. The Bray–Curtis distance was calculated to estimate the β-diversity in the lung taxonomy community structure in patients with NSCLC, which provides a measure of the degree to which samples differ from one another and can help elucidate the aspects of microbial ecology that are not apparent from the composition of individual samples.
A significant difference according to the pathological type was observed between the groups (p = 0.011, Figure 4).

We performed LEfSe analysis to further evaluate the differences in these dominant genera between patients with NSCLC in the adenocarcinoma and SqCC groups, suggesting that the genus *Actinomyces*, which belongs to the phylum Actinobacteria, was significantly more abundant in the SqCC group (Wilcoxon test, p = 0.049, Figure 5). The genus with the greatest influence on the distinction between the two groups, with linear discriminant analysis (LDA) score of 4.08, was *Haemophilus*, belonging to the Proteobacteria phylum (Figure 5).

*Neisseria subflava*, belonging to Proteobacteria, was dominant in the adenocarcinoma group with an LDA score of 4.07 (Wilcoxon test, p = 0.029). These were followed by *Porphyromonas* and *Fusobacterium* with LDA scores of 3.88 (p < 0.001) and 3.72 (p < 0.001), respectively, in the adenocarcinoma group (Figure 5).

**Discussion**

In this study, we characterized lung cancer microbiota by analyzing BAL in patients with NSCLC from two hospitals using 16S rRNA sequencing. The diversity of microbiome was significantly reduced in SqCC comparing with adenocarcinoma.

Recent research has suggested strong associations between lung cancer and specific microorganisms through different mechanisms, including induction of host inflammatory pathways, production of bacterial toxins that alter host genomic stability, and release of cancer-promoting microbial metabolites (21). For example, *Veillonella* is present in patients with lung cancer and plays a role in NSCLC pathogenesis (22). It also plays a role in the increased infiltration of inflammatory cells (Th17 cells) and upregulation of the extracellular signal-regulated kinase/phosphoinositide 3-kinase (ERK/PI3K) pathway in bronchial epithelial
Notably, PI3K pathway upregulation was previously shown to be an early pathogenic event in NSCLC, regulating cell proliferation, survival, differentiation, and invasion (24). Concordantly, our study further revealed significant differences in the microbiome in NSCLC according to different pathologic types (adenocarcinoma and SqCC) exhibiting different prognoses.

Fusobacteria was a significantly abundant phylum in the adenocarcinoma group, and Proteobacteria was abundant in both groups; however, the abundance of certain species was different. *Haemophilus parainfluenza*, belonging to Proteobacteria, was the most dominant species in the adenocarcinoma group and is known to cause various invasive, chronic, and recurrent diseases (25). A previous study revealed that the abundance of *H. parainfluenza* significantly differed between healthy individuals and patients with lung cancer (26). We further found that it was more abundant in the adenocarcinoma group than in the SqCC group. *Porphyromonas endodontalis*, belonging to Bacteroidetes, is known to be elevated in the sputum of patients with early lung cancer (27). In our study, the proportion of early lung cancer (stage I/II) was higher in patients with adenocarcinoma (51.6% vs. 20.0%). *Porphyromonas gingivalis*, which belongs to the same phylum as *P. endodontalis*, promotes the survival and proliferation of epithelial cells by increasing PI3K/Akt signaling shortly after an infection, resulting in the inhibition of intrinsic apoptosis and increased expression of cancer stem cell markers CD44 and CD133(28). Although *P. endodontalis* has not yet been well studied, our study revealed its enhanced abundance in patients with adenocarcinoma, similar to *P. gingivalis* in oral cancer.

*Neisseria* species are associated with infections such as septic arthritis (29) and meningitis (30) among them, *N. subflava* is increased in patients with oral cavity cancer(31). *Fusobacterium nucleatum* was also increased in oral cancers (32). It promotes oral cavity cancer via direct interaction with oral epithelial cells through Toll-like receptors, increasing the expression of
TLR2 in OSCC (Oral Squamous Cell Carcinoma) cells and IL-6 in cells and a mouse model (33). *F. nucleatum* in esophageal cancer tissues has been associated with shorter survival, suggesting its potential role as a prognostic biomarker, which might also contribute to aggressive tumor behavior by activating chemokines, such as chemokine (C-C motif) ligand 20 (34). The oral microbiome is correlated with the lung microbiome because the lungs are directly connected to the oral cavity (35); thus, we can infer that *N. subflava* and *F. nucleatum* are associated with the development of lung cancer, especially adenocarcinoma. The microbiome was diverse in the adenocarcinoma group and similar to the microbiota of oral cavity cancer, suggesting that the oral microbial environment plays a crucial role in patients with genetic susceptibility to lung cancer development.

SqCC is associated with male smokers (36) and certain pathogens, such as that causing tuberculosis (37). In patients with SqCC, *Actinomyces graevenitzii* was much more abundant than in the adenocarcinoma group. *A. graevenitzii* is a component of the oropharyngeal flora and is associated with pulmonary infections in some cases (38); however, it has not yet been comprehensively studied. Our study revealed that it is related to SqCC.

Previous studies have reported differences in the clinical characteristics and treatment methods between the two subtypes (39). Wang *et al.* (40) analyzed clinical characteristics of 48,296 patients with lung cancer by dividing them into adenocarcinoma and SqCC groups. The two groups differed significantly in many clinical characteristics, including age, sex, clinical/pathological stage, and treatment. Notably, they reported that the stage-specific five-year overall survival rate differed significantly between the two groups even after propensity score matching, suggesting that the two cancers should be analyzed separately to provide a precise outcome. Similarly, our study identified significantly different microbiomes and PCoA plots depending on the cancer subtype. Furthermore, the microbiome was significantly more diverse in the adenocarcinoma group than that in the SqCC group. Also, there was significant
difference of $\alpha$-diversity between early (stage 1, 2 and 3a) and late stage (stage 3b and 4) of NSCLC (Supplemental Figure 1). As the microbiome plays a vital role in barrier function and immune surveillance, this low microbial diversity may also be related to the poor prognosis of SqCC.

Although diet and antibiotics usage, which might influence the microbial composition, were not investigated, a specialized bronchoscopist conducted sterile bronchoscopy, and we used 16S rRNA sequencing during the first diagnosis, which was prospective, informative, and sensitive compared with other conventional methods.

**Conclusions**

We identified that the microbial $\alpha$- and $\beta$-diversities were different between the two types of lung cancer. *H. parainfluenzae*, *N. subflava*, *P. endodontalis*, and *F. nucleatum* were significantly abundant in the adenocarcinoma group, whereas *A. graevenitzii* was abundant in the SqCC group. Prolonged colonization and subsequent inflammation may disrupt the normal host immune barrier, eventually leading to cancer development. Further large-scale studies are required to determine the role of the microbiome in the development of different lung cancer types.

**Funding:** This Study was Supported by a 2021-Grant from The Korean Academy of Tuberculosis and Respiratory Diseases

**Acknowledgments:** Not applicable
References


**FIGURE LEGENDS**

**Figure 1.** Taxonomic composition of the microbiome community between subgroups. **a.** Dominant phyla based on the types of non-small cell lung carcinoma (NSCLC). **b.** Dominant genera based on the types of NSCLC.
Figure 2. Differential abundances of phyla. a. Fusobacteria, b. Firmicutes, c. Bacteroidetes, and d. Proteobacteria between the adenocarcinoma and squamous cell carcinoma (SqCC) groups.
Figure 3. Comparison of the $\alpha$-diversity in bronchoalveolar lavage (BAL) fluid microbiomes between the adenocarcinoma and SqCC groups. **a.** Number of OTUs, **b.** Chao1 index, **c.** Simpson index, and **d.** Shannon index. OTUs, operational taxonomic units; upper box = 2nd quartile; midline = median; lower box = 3rd quartile; whiskers = highest and lowest quartiles.
Figure 4. PCoA plot based on the Bray–Curtis distance of the BAL fluid microbiome between the adenocarcinoma and SqCC groups.
Figure 5. LEfSe analysis of the dominant genera between adenocarcinoma and SqCC groups. LDA, linear discriminant analysis; LEfSe, LDA effect size. The blue and red bars indicate the taxa identified in greater relative abundance in patients with adenocarcinoma and SqCC, respectively.
Subtype-based microbial analysis in non-small cell lung cancer

Hye Jin Jang, MD\textsuperscript{1,2}, Eunkyung Lee\textsuperscript{3}, Young-Jae Cho, MD, MPH, PhD\textsuperscript{4}; Sang Hoon Lee, MD, PhD\textsuperscript{1}.

\textsuperscript{1}Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea

\textsuperscript{2}Department of internal medicine, Myongji Hospital, Hanyang University College of Medicine, Goyang-si, Gyeonggi-do, Korea

\textsuperscript{3}Department of Biomedical Science and Engineering, Gwangju Institute of Science and Technology, Gwangju 61005, Republic of Korea

\textsuperscript{4}Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam-si, Gyeonggi-do, Republic of Korea

*Correspondence:*

Sang Hoon Lee, M.D, Ph.D.
Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Institute of Chest Diseases, Severance Hospital, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea

Tel: +82-2-2228-1955; Fax: +82-2-393-6884

E-mail address: cloud9@yuhs.ac
Supplemental methods

Data Analysis Pipeline

A quality check (QC) was performed for processing raw reads and low-quality reads (<Q25) were excluded using Trimmomatic version 0.32.[1] Paired-end sequence data were merged together after the QC step using the fastq_mergepairs command of VSEARCH version 2.13.4[2] with default parameters.

We then trimmed the primers using the alignment algorithm of Myers & Miller[3] at a similarity cut-off of 0.8. Non-specific amplicons that did not encode 16S rRNA were detected using the ‘nhmmer’ function of the HMMER software package ver. 3.2.1 with hidden Markov model profiles. Unique reads were extracted, and redundant reads were clustered with the unique reads using the derep_fulllength command of VSEARCH.[2] The EzBioCloud 16S rRNA database[4] was used for taxonomic assignment of the obtained 16S rRNA sequences using the ‘usearch_global ‘command of VSEARCH,[2] followed by a more precise pairwise alignment[3]. Chimeric reads were filtered from reads with <97% similarity via reference-based chimeric detection using the UCHIME algorithm[5] and the non-chimeric 16S rRNA database from EzBioCloud. After chimeric-read filtering, the reads that were not identified at the species level (with <97% similarity) using the EzBioCloud database, were compiled, and the ‘cluster_fast’ command[2] was used to perform de-novo clustering to generate additional operational taxonomic units (OTUs). Finally, OTUs with single reads (singletons) were omitted from further analysis. The secondary analysis, which included diversity calculation and biomarker discovery, was conducted using in-house programs of CJ Bioscience, Inc (Seoul, South Korea). Shannon[6] and Simpson[6] alpha diversity indices were estimated. To visualize the sample differences, beta diversity distances were calculated using the method described by Bray–Curtis.[7] Taxonomic biomarkers and functional biomarkers were identified using statistical comparison algorithms (linear discriminant analysis [LDA] Effect Size [LEFse])[8]
and Kruskal–Wallis H Test[9]). To analyze the microbial community’s functional capabilities, functional profiling was conducted using PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states)[10] and MinPath (Minimal set of Pathways)[11]. All aforementioned analyses were performed using EzBioCloud 16S-based MTP, which is a CJ Bioscience’s bioinformatic cloud platform.
Supplemental Figure Legend

Supplemental Figure 1. Comparison of the α-diversity in bronchoalveolar lavage (BAL) fluid microbiomes between early and late stage of NSCLC.

*<0.05 by "Wilcoxon Test"
References