Am J Cancer Res 2022;12(5):2350-2362
www.ajcr.us /ISSN:2156-6976/ajcr0141947

*Original Article*

# The microbiome of lung cancer tissue and its association with pathological and clinical parameters

Ock-Hwa Kim[1,2*], Bo-Yun Choi[3*], Dong Kwan Kim[4], Na Hyun Kim[1], Jin Kyung Rho[5], Woo Jun Sul[3#], Sei Won Lee[1#]

*[1]Department of Pulmonary and Critical Care Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea; [2]Division of Allergy and Pulmonology, Department of Internal Medicine, Chungnam National University Sejong Hospital, Sejong, Republic of Korea; [3]Department of Systems Biotechnology, Chung-Ang University, Anseong, Gyeonggi-do, Republic of Korea; [4]Department of Thoracic and Cardiovascular Surgery, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea; [5]Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea. [*]Co-first authors. [#]Equal contributors.*

Received January 22, 2022; Accepted May 2, 2022; Epub May 15, 2022; Published May 30, 2022

**Abstract:** Lung cancer is the primary cause of cancer-related deaths worldwide. Recently, although the microbiome has emerged as the key modulator of the carcinogenesis, it has not been evaluated in lung cancer. Here, we evaluated the microbial composition of lung cancer tissues according to the histologic type and genetic mutation, compared it with that of the adjacent normal lung tissues, and investigated the association between the lung microbiome and clinical parameters. We collected lung tissue samples from 162 patients with non-small cell lung cancer (NSCLC, 162 cancer and 54 adjacent normal tissues), surgically resected between January 2018 and December 2019, and analyzed their microbiome using 16S rRNA gene amplicon sequencing, the QIIME2 pipeline, and statistical analyses. NSCLC tissues had significantly lower alpha diversity than the normal tissues, and their microbial composition differed according to the histologic type and cancer genetic mutation. The genera *Romboutsia*, *Novosphingobium*, *Acinetobacter*, and *Prevotella* were significantly overrepresented in NSCLC tissues. Alpha diversity steadily declined from a normal to a more advanced stage, and microbial compositional differences were noted along with recurrence. *Stenotrophomonas* was the most predominant genus in the NSCLC tissues of patients with recurrence. The pathways related to the tricarboxylic acid cycle and L-glutamate and L-glutamine biosynthesis were predominant in adenocarcinoma, whereas those related to purine and pyrimidine nucleotide degradation and formaldehyde assimilation were predominant in squamous cell carcinoma. Our findings suggest that the altered lung cancer microbial composition might be associated with cancer initiation and/or progression.

**Keywords:** Lung cancer, normal lung tissue, microbiome, *Stenotrophomonas*

## Introduction

Lung cancer is the most frequently diagnosed cancer and the leading cause of cancer-related deaths worldwide, constituting approximately 11.6% of all cancer diagnoses and being responsible for 18.4% of all cancer-related deaths in 2020 [1]. Lung cancer is widely known as a complicated disease caused by the interactions between the host and the environment [2]. Among the diverse environmental risk factors, microbiota have emerged as key modulators of both the carcinogenic process and the immune response against cancer cells [3]. While healthy lungs were historically considered as a sterile environment, it has recently been suggested that certain microbiota exist in the lungs and microbiota changes are associated with the development of diseases, including lung cancer [4]. The microbiome can contribute to carcinogenesis via host inflammatory pathways, bacterial metabolites, and genotoxic pathways [4].

Previous studies on the lung cancer microbiome have identified the *Firmicutes*, *Proteobacteria*, and *Bacteroidetes* phyla, including the *Streptococcus*, *Neisseria*, and *Prevotella* genera, in lung cancer tissues [2, 5-10]. A study by Yu et al. found that the *Thermus* and *Legionella* genera are highly abundant in tissues from advanced stage lung cancer patients and in

patients who develop metastases, respectively, suggesting their role in tumor progression [8]. Another study showed that greater abundance of the families *Bacteroidaceae*, *Lachnospiraceae*, and *Ruminococcaceae* in the paired normal lung tissues is associated with reduced recurrence-free or disease-free survival and demonstrated the potential relationship between the normal lung microbiota and lung cancer prognosis [10].

Although there have been various studies to identify the lung cancer microbiome in respiratory samples such as saliva, sputum, bronchoalveolar lavage fluid, and bronchial brushing [2], little is known about the microbiota profile in lung cancer tissues, particularly the differences in the microbial composition according to the histologic type and/or genetic mutations. Moreover, few studies exist on the relationship between the microbiome and lung cancer prognosis.

Therefore, we aimed to identify the microbial composition of lung cancer tissues according to the histologic type and genetic mutations, compare it with that of the adjacent normal lung tissues, and investigate the association between the lung microbiome and the prognosis of lung cancer. Among lung cancers, non-small cell lung cancer (NSCLC) was selected for the analysis, which accounts for the majority (80-85%) of lung cancers and almost all surgically resected lung cancer cases, allowing for the collection of tissues from a large number of patients.

## Material and methods

### *Study population and sample collection*

We retrospectively and randomly selected 216 frozen lung tissue samples (162 cancer tissues and 54 adjacent normal tissues), which were surgically resected from 162 patients with NSCLC between January 2018 and December 2019 at the Asan Medical Center. Patients were included if they were 18 or more years of age and signed an informed consent form for the collection of tissues in the operating room. The exclusion criteria were as follows: patients with SCLC, NSCLC other than adenocarcinoma (AC) and squamous cell carcinoma (SCC), less than 18 years of age, and inability to provide informed consent.

We collected lung cancer samples by dividing them into AC and SCC, the most common subtypes of NSCLC. Then, the AC group was divided according to the presence of the epidermal growth factor receptor (EGFR) mutation, which is one of the most frequent mutations in NSCLC. Thus, the lung cancer samples were divided into the following three groups according to the histologic type and the presence of the EGFR mutation: AC with EGFR mutation (AC EGFR+), AC without EGFR mutation (AC EGFR-), and SCC. In total, 54 cancer tissues were included in each group, together with the paired normal lung tissues obtained from AC EGFR+ patients who were predicted to show relatively homogeneous features (Figure S1). All study patients consented to the collection of tissues in the operating room at the time of resection. Tissue samples were sterilely cut in the operating room, transferred to cryovials, and immediately snap-frozen in liquid nitrogen. Then, they were stored in the Bio-Resource Center at the Asan Medical Center until analysis. Clinical data, including baseline demographics, smoking status, postoperative stage, recurrence, and death, were retrospectively collected from medical records. The study protocol was approved by the Institutional Review Board of Asan Medical Center (IRB No. 2020-0194).

### *DNA extraction and bacterial 16S rRNA gene sequencing*

We extracted the bacterial gDNA from 216 lung tissue samples. DNA extraction was performed using the DNeasy PowerSoil Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. We cut each lung tissue using a flame-sterilized blade, and a piece of lung tissue (0.2 g) was transferred to a PowerBead tube. Horizontal vortexing was performed using Vortex Genie 2 (Scientific Industries, Bohemia, New York, USA) at the maximum speed. The gDNA concentration was measured using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc., Waltham, MA, USA). Polymerase chain reaction (PCR) amplification was performed using the primers targeting the V4-V5 region of the *16S rRNA* gene (forward: 5'-TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG CCA GCA GCY GCG GTA AN-3'; reverse: 5'-GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GCC GTC AAT TCN TTT RAG T-3'). The PCR conditions used were as follows: denaturation at 95°C for 3 min, followed by 30 cycles of ampli-

fication at 95°C for 30 s, 55°C for 30 s, and 72°C for 1 min. The PCR products were purified using AMPure XP beads (Beckman Coulter, High Wycombe, UK). The quality of the final purified product was assessed using a Nanodrop 2000 spectrophotometer. The final products were sequenced using an Illumina Miseq™ platform (Illumina, San Diego, CA, USA).

*Microbiome data analysis*

The lung tissue microbiome sequences were analyzed using the QIIME™ (Quantitative Insights Into Microbial Ecology) 2 pipeline (2019.7) [11]. Primer sequences of the raw sequences were removed using the Cutadapt [12] plugin. In this step, the reads that did not include bacterial primer sequences or that included low-quality primer sequences were removed. We denoised the sequences using the DADA2 plugin implemented in QIIME™ 2 and identified the bacterial amplicon sequence variants (ASVs). The ASVs were classified using the SILVA database [13]. Chloroplast, mitochondrial, and unassigned ASVs were removed using a taxa filter. We aligned the ASVs using the phylogeny align-to-tree-mafft-fasttree plugin and performed microbial diversity analyses using rarefying samples to an even depth of 50 sequences per sample.

Alpha diversity was evaluated using Chao1, the Shannon index, and phylogenetic diversity [14, 15]. Beta diversity was assessed using unweighted UniFrac distances [16]. The microbial compositional differences between groups were assessed using principal coordinates analysis (PCoA) with unweighted UniFrac distances. The unweighted intragroup UniFrac distances were represented by box-and-whiskers plots. Differential taxonomy was identified using linear discriminant analysis (LDA) effect size (LEfSe) [17]. We used the neutral community model-based dominance test to determine the cause of the lung microbiome differences between the normal and cancer groups [18]. Additionally, we conducted a stratified sub-analysis, based on the postoperative stage and the recurrence of lung cancer, to address the association between the lung microbiome and lung cancer prognosis [10]. The Phylogenetic Investigation of Communities by Reconstruction of Unobserved States 2 (PICRUSt2) v.2.1.3-b software, which predicts the gene family abundance, was used to predict the functional profiles from the 16S rRNA data [19, 20]. The ASV table resulting from bacterial community analysis and the representative sequences were used, and the predicted functional profiles were obtained using the picrust2_pipeline.py. We used the MetaCyc pathway, which was inferred using PICRUSt2.

*Statistical analysis*

All clinical data are presented as mean ± standard deviation or median (interquartile range [IQR]) for continuous variables and as numbers (%) for categorical variables. Data categorized according to the cancer type were compared using one-way analysis of variance or Kruskal-Wallis test (for continuous variables) and the $\chi^2$ or Fisher's exact test (for categorical variables). If there was a significant difference among the three values, a *post hoc* analysis for multiple comparisons was performed using the Bonferroni correction. All analyses of the clinical data were performed using SPSS software (Version 24.0; SPSS, Chicago, IL, USA).

We performed an analysis of similarity to identify the differences in the microbial communities. To identify any significant differences in alpha diversity and the UniFrac dissimilarities, we performed the Wilcoxon rank-sum test or *t*-test in R. The Kruskal-Wallis rank-sum test was used to distinguish the significant differential abundance among different groups in LEfSe analysis. The Kaplan-Meier method was used to estimate the cumulative rate of recurrence. All significance tests were two-sided, and *p*-values < 0.05 indicated statistical significance.

## Results

*Clinical characteristics of the study population*

The mean patient age was 65.3 years (range 21-83), with a preponderance of male (64.8%) patients. Upon comparing the three groups according to the cancer type, the proportions of males and smokers were significantly greater in the SCC group than in the AC EGFR+ and AC EGFR- groups (P < 0.001). Furthermore, the SCC group had a significantly higher median number of pack-years (39 pack-years, P < 0.001). Otherwise, there were no significant differences in age, postoperative stage, nodal stage, recurrence, and death among the three groups. **Table 1** presents the clinical characteristics of the 162 study patients.

**Table 1.** Clinical characteristics of the study population

| | AC EGFR+ (n = 54) | AC EGFR- (n = 54) | SCC (n = 54) | *p*-value | | | |
|---|---|---|---|---|---|---|---|
| | | | | | AC EGFR+ *vs.* AC EGFR- | AC EGFR- *vs.* SCC | SCC *vs.* AC EGFR+ |
| Age, yr | 64.9 ± 9.3 | 64.2 ± 9.2 | 66.7 ± 9.4 | 0.358 | | | |
| Male gender, n (%) | 28 (51.9%) | 25 (46.3%) | 52 (96.3%) | < 0.001 | 0.564 | < 0.001 | < 0.001 |
| Body mass index (kg/m$^2$) | 25.0 ± 2.5 | 25.0 ± 3.4 | 24.7 ± 3.0 | 0.882 | | | |
| Smoking status | | | | | | | |
| Non-smoker, n (%) | 30 (55.6%) | 31 (57.4%) | 3 (5.6%) | < 0.001 | 0.846 | < 0.001 | < 0.001 |
| Ever smoker, n (%) | 24 (44.4%) | 23 (42.6%) | 51 (94.4%) | | | | |
| Pack-years, median (IQR) | 0 (0-30) | 0 (0-27) | 39 (20-46) | < 0.001 | 0.671 | < 0.001 | < 0.001 |
| Stage, n (%) | | | | | | | |
| I–II | 41 (75.9%) | 44 (81.5%) | 45 (83.3%) | 0.603 | | | |
| III-IV | 13 (24.1%) | 10 (18.5%) | 9 (16.7%) | | | | |
| Nodal stage | | | | | | | |
| N0 | 39 (72.2%) | 45 (83.3%) | 39 (72.2%) | 0.296 | | | |
| N1-2 | 15 (27.8%) | 9 (16.7%) | 15 (27.8%) | | | | |
| Follow-up duration months, median (IQR) | 27.5 (24.0-29.0) | 22.0 (18.8-25.3) | 25.0 (11.8-27.0) | < 0.001 | < 0.001 | 0.193 | 0.004 |
| Recurrence, n (%) | 16 (29.6%) | 13 (24.1%) | 9 (16.7%) | 0.280 | | | |
| Death, n (%) | 2 (3.7%) | 1 (1.9%) | 6 (11.1%) | 0.085 | | | |

*Microbial diversity analysis*

The lung tissue microbiome of the 216 tissues was analyzed using 16S rRNA gene amplicon sequencing. Twelve samples that had no or low sequencing data were excluded. In 204 samples, a total of 8,552,251 paired-end reads were merged into 326,385 reads from the tissue samples, resulting in an average of 1,599 merged reads per sample. As the rarefied depth was 50, we calculated the alpha diversity and beta diversity using 112 samples (Figure S1).

The alpha diversity indexes, including Chao1, Shannon, and phylogenetic diversity, were significantly higher in the normal group than in the cancer groups ($P < 0.0001$, $P < 0.05$, and $P < 0.01$, respectively, **Figure 1A**). When comparing the three histologic cancer types, Chao1 and phylogenetic diversity were higher in the normal group than in the AC and SCC groups (**Figure 1B**). The comparison among the four groups in accordance with the histologic type and EGFR mutation showed that the normal group had greater diversity than all the other groups, as assessed via the Chao1 index (**Figure 1C**). When we performed sub-analysis by rarefying samples to the depth of 25 and 100 sequences, according to the study flow of Figure S2, not only the Chao1 index but also the phylogenetic diversity was greater in the normal group than in all the other groups (Figure S3).

Unweighted UniFrac distances were calculated to evaluate the similarity of microbiota profiles. The PCoA based on the unweighted UniFrac distance revealed significant compositional differences among different groups, indicating a more heterogeneous microbial distribution in the cancer group (Figure S4B and S4C). The results of the sub-analysis, wherein samples were rarefied to a depth of 25 and 100 sequences, were similar (Figure S3). Among the cancer groups, there was also a compositional difference between AC and SCC (Figure S4D). The box-and-whisker plot of unweighted intra-group UniFrac distances showed that the cancer group had a significantly greater dissimilarity among samples than the normal group, in the following order: SCC, AC, and normal group (Figure S4E-S4H).

*Differential abundance analysis*

To detect the taxa with differential abundance among the different groups, we used LEfSe with an LDA value of 3.0. By comparing the cancer and normal groups, 25 bacterial taxa at the genus level with significant abundance differences were identified. The genera *Romboutsia*, *Christensenellaceae R-7 group*, *Novosphingobium*, *Acinetobacter*, *Rhizobium*, and *Prevotella* were significantly overrepresented in the cancer group, whereas other taxa, including *Staphylococcus*, *Burkholderiaceae*, and *Cutibacterium,* were enriched in the normal group. The taxa enriched in the different groups

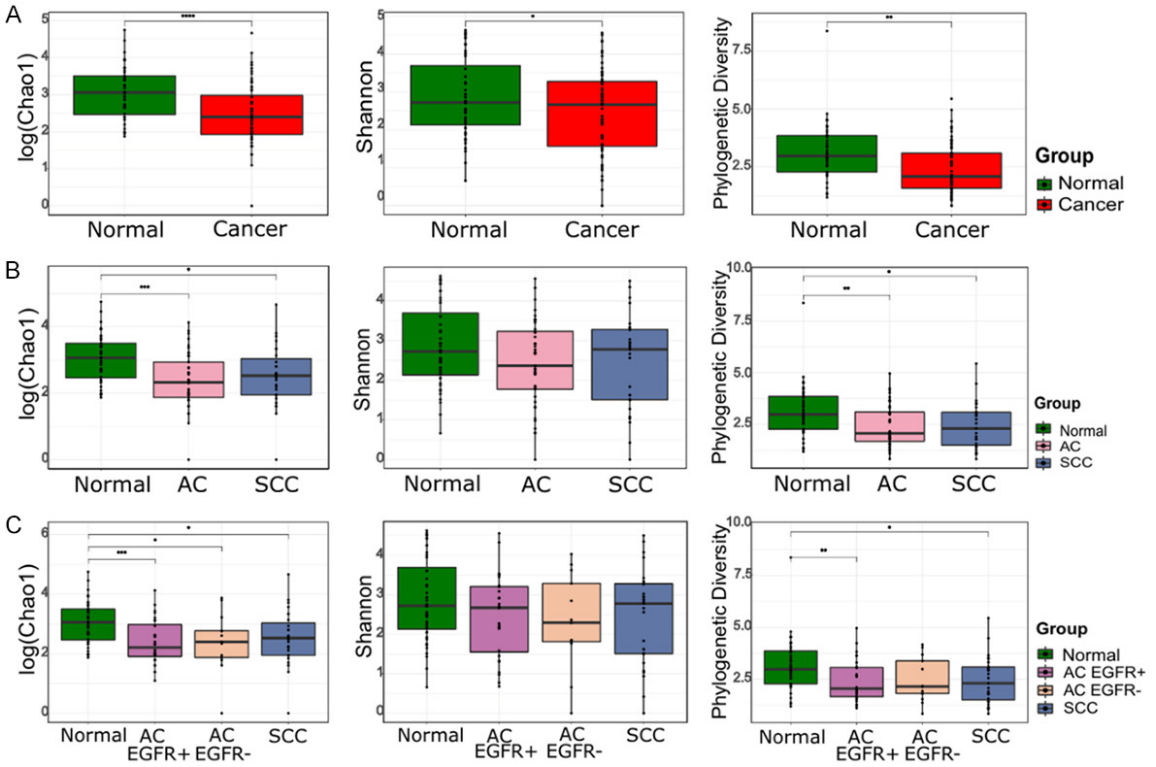


**Figure 1.** Comparison of the alpha diversity indexes (Chao1, Shannon, and phylogenetic diversity) between lung cancer and normal lung tissues. (A) Normal vs. Cancer, (B) Normal vs. AC vs. SCC, and (C) Normal vs. AC EGFR+ vs. AC EGFR- vs. SCC. Box plot represents the minimum, first quartile, median, third quartile, and maximum of the alpha indexes.

are displayed in **Figure 2**. Notably, compared with the normal group, *Haemophilus influenzae* (*H. influenzae*) was enriched in the SCC group (**Figure 2**). In total, four ASVs were classified as species *H. influenzae* and there were nine samples containing these ASVs. Seven of them (77.8%, 7/9) were in the SCC group, one in the normal group, and the other in the AC group. Table S1 presents the ASV count classified as species *H. influenzae* in each sample. We selected two SCC tissues with the most abundant *H. influenzae* ASVs. The tissues were disrupted with stainless steel beads, plated, and cultured on Chocolate agar in 5% $CO_2$ at 37°C. After three days, we observed colonies in one of the two SCC tissues. We picked the colonies, spread them onto a fresh Chocolate agar plate using the dilution streaking method [21], and incubated the plate at 37°C and 5% $CO_2$ for three days. We could identify *H. influenzae* from the pure culture using 16S rRNA sequencing.

*Neutral community model-based dominance test*

We used a neutral community model-based dominance test to determine the cause of the lung microbiome difference between the normal and cancer groups (Figure S5). The goodness-of-fit ($R^2$) values were 0.74 and 0.64 for the normal and cancer groups, respectively. Considering that an $R^2$ value close to 1 implies that the composition of the lung microbiome is consistent with the neutral processes of dispersal and ecological drift [18], the microbial community of the normal group was a better fit to the neutral model than that of the cancer group. Next, we determined whether any ASVs deviated from the predictions of the neutral model. While some ASVs, including the genera *Micrococcus*, *Romboutsia*, *Novosphingobium*, *Bacteroides*, *Parasutterella*, and *Sphingobium*, were overrepresented, *H. influenzae* species and *Burkholderiaceae* family were underrepresented in the cancer group.

*Microbiome analysis according to the postoperative lung cancer stage*

To address the association between the lung microbiome and the prognosis of lung cancer, microbial diversity and taxonomy were analyzed by dividing the samples according to the

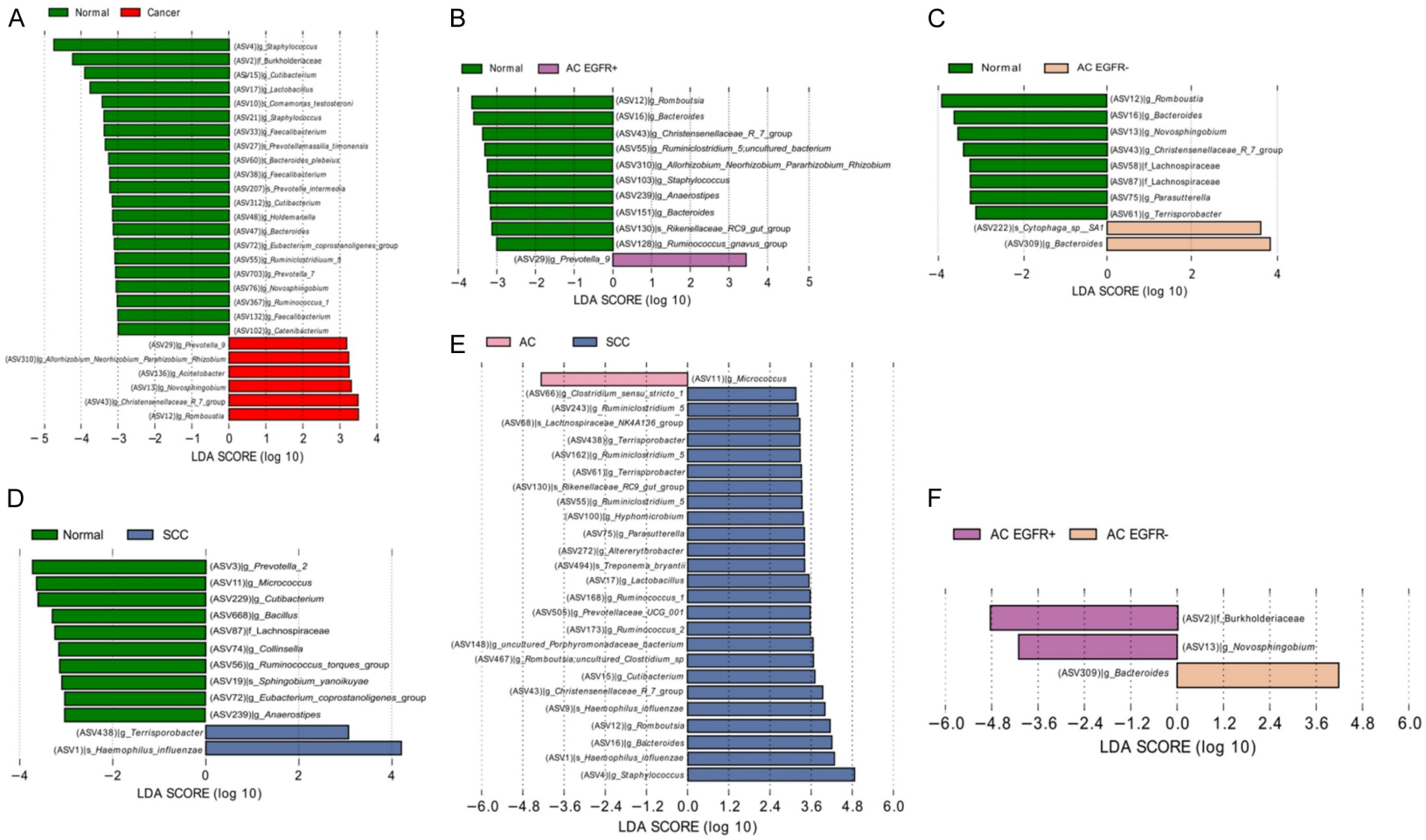


**Figure 2.** The linear discriminant analysis effect size plots of bacterial communities in cancer and normal lung tissues. Linear discriminant analysis scores were calculated, with higher scores indicating a greater effect size (significance determined by an LDA score > 3.0 and a $p < 0.05$ in the Kruskal-Wallis rank-sum test). Taxonomic categories include f = family, g = genus, and s = species. (A) Normal vs. Cancer, (B) Normal vs. AC EGFR+, (C) Normal vs. AC EGFR-, (D) Normal vs. SCC, (E) SCC vs. AC, and (F) AC EGFR+ vs. AC EGFR-.

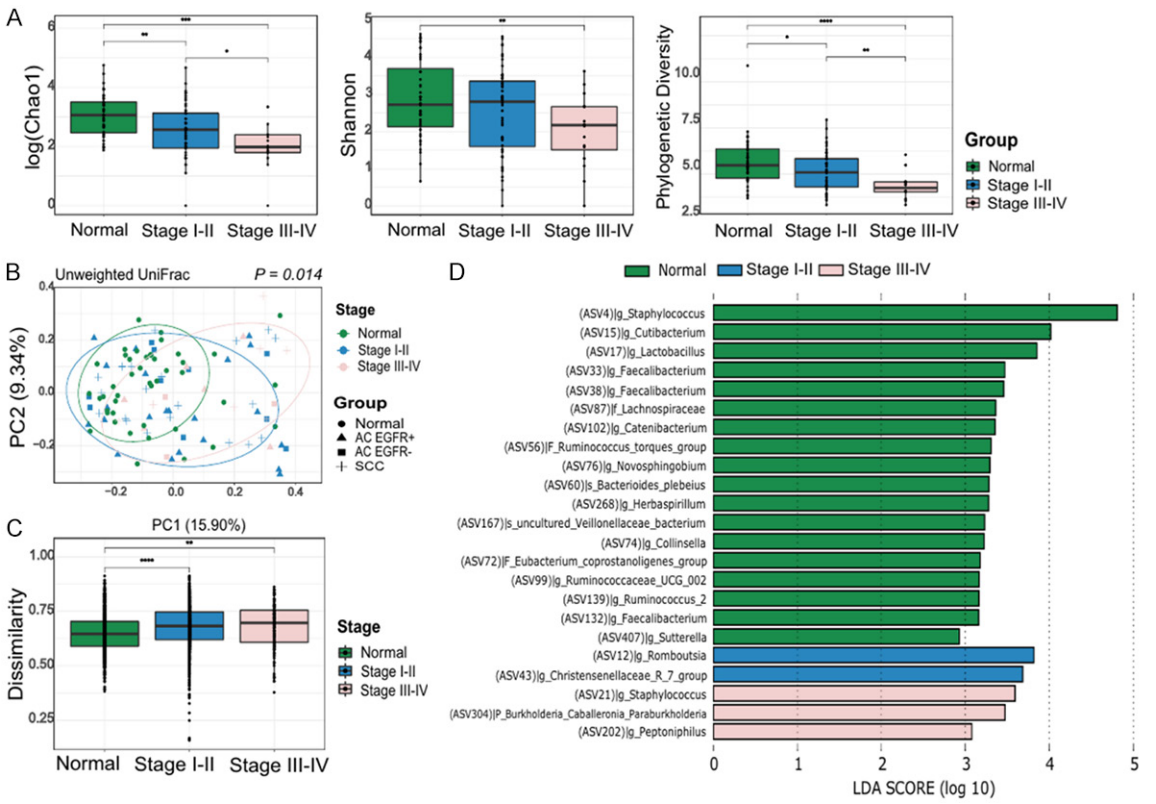


**Figure 3.** Microbiome analysis according to the postoperative lung cancer stage. (A) Comparison of the alpha diversity indexes (Chao1, Shannon, and phylogenetic diversity), (B) Principal coordinates analyses of the unweighted UniFrac distance, (C) Box-and-whisker plot of unweighted intragroup UniFrac distances, and (D) LEfSe analysis of the bacterial communities.

postoperative stage: stages I-II and stages III-IV of TNM classification. The alpha diversity, which was evaluated via Chao1, and the phylogenetic diversity steadily declined from normal to stages I-II and to stages III-IV (**Figure 3A**). Compositional differences were noted based on the beta diversity analysis. The PCoA plot revealed that the microbial distribution differed according to the postoperative lung cancer stage (P = 0.014, **Figure 3B**). Although there was no statistical significance, lung cancer at an advanced stage tended to have greater dissimilarities (**Figure 3C**). In the LEfSe analysis of three subgroups according to the postoperative lung cancer stage, the *Staphylococcus*, *Burkholderia-Caballeronia-Paraburkholderia*, and *Peptoniphilus* genera were enriched in the advanced stage lung cancer group (**Figure 3D**).

*Microbiome analysis according to lung cancer recurrence*

Next, we analyzed the microbial diversity and taxonomy by dividing the samples according to recurrence during the follow-up period. The alpha diversity indexes did not differ according to recurrence (**Figure 4A**). Compositional differences were noted based on beta diversity analysis. The PCoA plot revealed that the microbial distribution differed with recurrence, presenting a more heterogeneous microbial distribution in the cancer with recurrence group (P = 0.002, **Figure 4B**). Considering the unweighted intragroup UniFrac distance in **Figure 4C**, the dissimilarity among the samples steadily increased in the following order: normal, lung cancer without recurrence, and lung cancer with recurrence. In the LEfSe analysis of the three subgroups according to recurrence, several differential microbial taxa, including the *Stenotrophomonas*, *Bacteroides*, and *Peptoniphilus* genera, were noted in the cancer with recurrence group (**Figure 4D**). We used the Kaplan-Meier method to estimate the cumulative rate of recurrence according to the presence of the genus *Stenotrophomonas*, which was identified as the most predominant in the

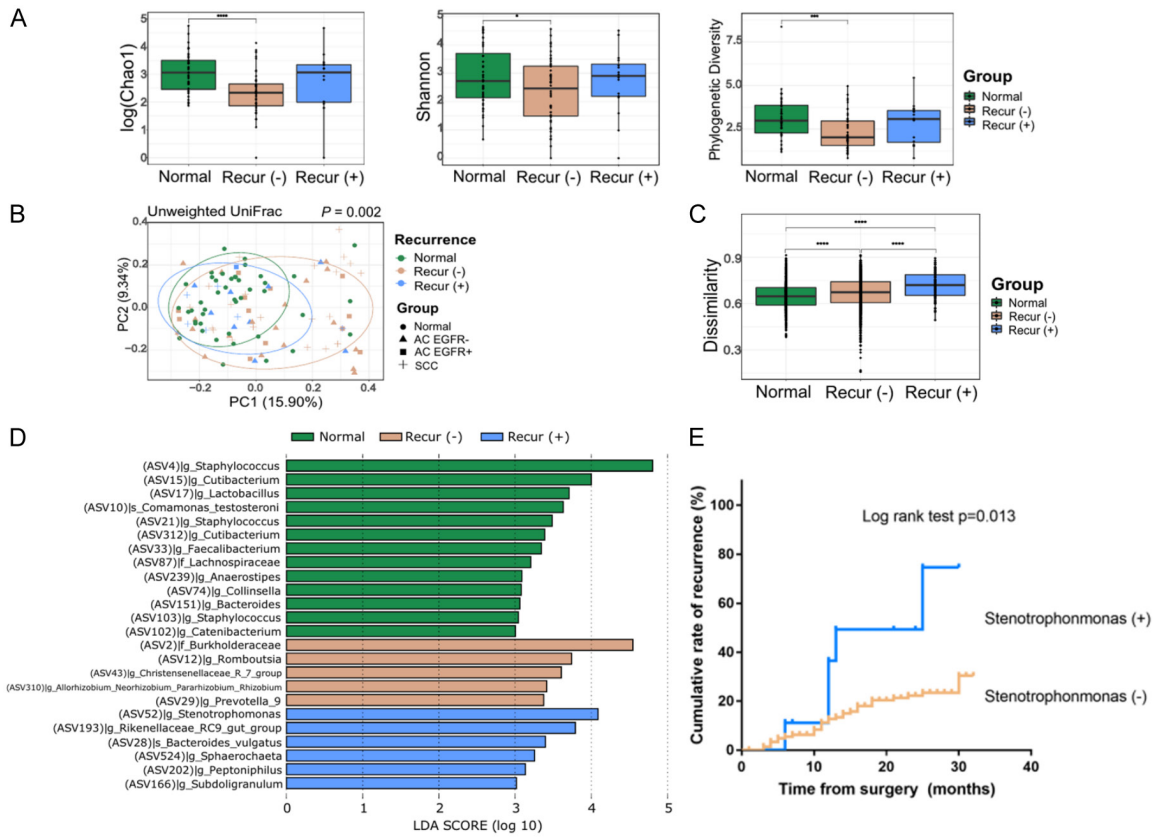


**Figure 4.** Microbiome analysis according to the recurrence of lung cancer. (A) Comparison of the alpha diversity indexes (Chao1, Shannon, and phylogenetic diversity), (B) Principal coordinates analyses of the unweighted UniFrac distance, (C) Box-and-whisker plot of unweighted intragroup UniFrac distance, (D) LEfSe analysis of the bacterial communities, and (E) The cumulative rate of recurrence according to the presence of the *Stenotrophomonas* genus.

lung cancer with recurrence group. The cumulative rate of recurrence was significantly higher in the groups with *Stenotrophomonas* than in those without *Stenotrophomonas* (P = 0.013, **Figure 4E**).

*Pathway analysis predicted using PICRUSt2*

To better understand how the bacterial functional profiles differed according to the histologic type of cancer, we used PICRUSt2, a bioinformatic analysis tool, to predict the functional metagenomes from 16S rRNA gene profiling data. We found 18 differentially abundant MetaCyc pathways among the three groups (with an LDA score > 2.9) (**Figure 5**). The pathways related to the tricarboxylic acid (TCA) cycle, sulfur compound metabolism, fatty acid salvage, and L-glutamate and L-glutamine biosynthesis were predominantly observed in the AC group, whereas those related to adenine and adenosine salvage, acetylene degradation, purine and pyrimidine nucleotide degradation, L-arginine degradation, formaldehyde assimilation, biotin biosynthesis, and 2-carboxy-1,4-naphthoquinol biosynthesis were predominant in the SCC group.

## Discussion

Although an increasing number of studies have profiled the microbiome of respiratory samples from lung cancer patients, the microbiota profiles of lung cancer tissues and their association with clinical and pathological parameters remain unexplored. One of the key findings in our study was that the alpha diversity was significantly higher in normal lung tissues than in the cancer tissues, and the microbial composition differed according to the histologic type of cancer and the EGFR mutation. When we conducted a stratified sub-analysis to address the association between the lung microbiome and lung cancer prognosis, microbial compositional

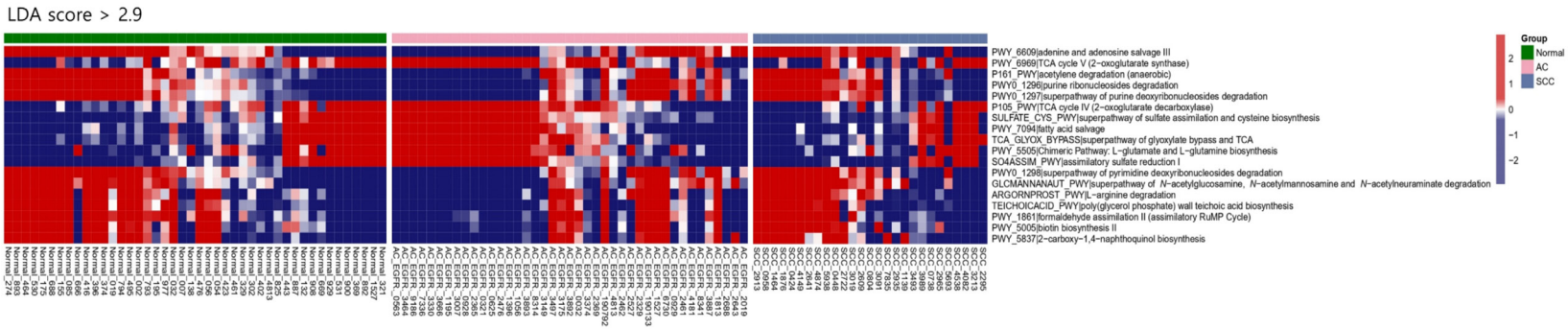


**Figure 5.** Heatmap of the differentially presented pathways according to the histologic type of cancer predicted using PICRUSt2. Red represents a high abundance, whereas blue represents a low abundance. Significance was determined based on an LDA score > 2.9 and $P < 0.05$.

differences were noted according to the postoperative lung cancer stage and the presence of recurrence. Finally, we found several inferred metagenomic functional pathways enriched in different histologic types of cancer from the 16S rRNA data.

Greathouse et al. showed a similar alpha diversity between non-tumor adjacent tissues and tumor tissues [6]. Another study showed that the alpha diversity was significantly lower in tumor tissues than in non-malignant lung tissues, but no significant differences were observed in the overall composition (beta diversity) [8]. We revealed the compositional differences between the normal and cancer groups, as well as a decreased bacterial diversity in lung cancer tissues compared to that in normal lung tissues. This suggested that the dysbiosis and altered microbial composition may be associated with cancer pathogenesis.

Only a few studies have reported the microbiome profile in lung cancer tissues. Liu et al. observed that lung cancer patients had a higher prevalence of *Firmicutes* (*Streptococcus*) and *Bacteroidetes* (*Prevotella*) than emphysema-only patients [22]. In our study, *Provotella_9* was also predominant in cancer tissues, particularly in the AC EGFR+ group. Another study by Apopa et al. found that the predominant phyla in the formalin-fixed lung samples embedded in paraffin were *Actinobacteria*, *Proteobacteria*, and *Bacteroidetes*. The *Cyanobacteria* phylum was consistently observed in AC samples [22]. We found that the *Romboutsia*, *Christensenellaceae R-7 group*, *Novosphingobium*, *Acinetobacter*, *Rhizobium*, and *Prevotella* genera were significantly overrepresented in the cancer group, with some ASVs occurring more frequently than predicted using the neutral community model. A previous study reported that the *Novosphingobium* genus is present in more severe chronic obstructive pulmonary disease and increases inflammation in a smoke exposure mouse model [23]. It might also promote carcinogenesis by creating an inflammatory milieu. Previously, 16S rRNA gene sequencing and transcriptome analyses of bronchial brushing samples revealed that airway *Veillonella*, *Streptococcus*, and *Prevotella* are associated with the upregulation of extracellular signal-regulated kinase and phosphoinositide 3-kinase signaling pathways, which regulate cell proliferation, survival, and differentiation in the airway [24].

Here, *H. influenzae* was enriched in the SCC group compared to the normal group, and we demonstrated the presence of *H. influenzae* from a pure culture of SCC lung tissues. Ochoa et al. found that the exposure of the airway to smoke particulates and non-typeable *H. influenzae* promoted lung cancer cell proliferation via the release of interleukin-6 and other inflammatory cytokines, which further activated the signal transducer and activator of transcription 3 and nuclear factor kB pathways in the airway epithelium [25]. Additionally, several studies reported that airway inflammation induced by non-typeable *H. influenzae* provides a tumor microenvironment that favors lung tumor promotion and progression [26-28]. Thus, lung microbiota could lead to carcinogenesis via several mechanisms, such as the creation of an inflammatory milieu, metabolic effects of dysbiosis, and genotoxicity [4, 29], and *H. influenzae* may be associated with these mechanisms.

We performed a stratified sub-analysis according to the postoperative stage and the recurrence of lung cancer to address the association between the lung microbiome and lung cancer prognosis. Although Kovaleva et al. did not find any significant differences in alpha diversity between tumors at different stages, they reported a tendency of alpha diversity increase in tumors at later stages [30]. In contrast, here, several alpha diversity indexes steadily declined from normal to more advanced stages, reflecting a trend toward better prognosis with higher alpha diversity. Moreover, a previous pilot study demonstrated that several microbiota in normal lung tissues were associated with recurrence-free survival; however, the association between tumor tissue microbiome and survival was not revealed [10]. We found that there was a microbial compositional difference along with the presence of recurrence during the follow-up period, with the cancer with recurrence group presenting a more heterogeneous microbial distribution. Representatively, the *Stenotrophomonas* genus, which is known as a significant pathogen, particularly in patients with obstructive lung cancer [31, 32], was identified as the most pre-

dominant genus in the lung cancer with recurrence group.

We further investigated the inferred metabolic/metagenomic function of the lung microbiome from 16S rRNA data of lung tissues in different histologic types of lung cancer. Cheng et al. performed 16S rRNA sequencing to show that the pathways of ribosome, pyrimidine, and purine metabolism were overrepresented in the bronchoalveolar lavage samples of lung cancer patients [33]. Furthermore, Apopa et al. showed that the peroxisome proliferator-activated receptors signaling pathway and the D-glutamine and D-glutamate metabolism were more enriched in AC than in SCC [22]. In our study, pathways such as the TCA cycle, and L-glutamate and L-glutamine biosynthesis were predominant in the AC group. Emerging evidence demonstrates that certain cancer cells, particularly those with deregulated oncogene and tumor suppressor expression, rely heavily on the TCA cycle for energy production and macromolecule synthesis [34]. Lung cancer cells also require glutamine as a nitrogen source for various anabolic processes in cancer and as an anaplerotic carbon source that replenishes the TCA intermediates [35]. Moreover, the pathway related to purine and pyrimidine nucleotide degradation was predominant in the SCC group. Increased nucleotide metabolism, which is a critical pathway for DNA replication, RNA synthesis, and cellular bioenergetics, suggests uncontrolled growth of tumors and is a hallmark of cancer [36]. The microbiome of lung cancer might play a role in these associated pathways. Further studies of functional microbiomic approaches, focused on the metabolic activity and function of the lung microbiota, are needed.

This study has several limitations. First, this study has a cross-sectional design, from which it is difficult to infer causality. Second, only a single race (Asian) was included, and the microbiota composition can differ according to racial characteristics [37, 38]. Third, although there were various subtypes of EGFR mutations, including exon 19 deletions and the L858R point mutation, we did not investigate which mutation occurred in each AC EGFR+ tissue sample. Finally, further studies with validation cohorts are needed for more confirmative results.

In conclusion, we showed that the lung cancer tissues had significantly lower alpha diversity than the normal tissues and that the microbial composition differed according to the cancer histologic type and a genetic mutation. Then, we revealed that alpha diversity steadily declined from normal to more advanced stages, and there were microbial compositional differences according to the presence of recurrence. Finally, we found several inferred metagenomic functional pathways enriched in different histologic types of lung cancer. These findings suggest that the altered microbial composition in lung cancer might be associated with cancer initiation and/or progression.

## Acknowledgements

The biospecimen and data used in this study were provided by the Asan Bio-Resource Center, Korea Biobank Network (2020-05(202)). This study was supported by a grant (2020-IF0013) from the Asan Institute for Life, Asan Medical Center, Seoul, Korea, the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020-R1A2C1008431, SWL) and a grant of the Korea Health Promotion R&D Project, funded by the Ministry of Health & Welfare, Republic of Korea (HS21C0096).

## Disclosure of conflict of interest

None.

**Address correspondence to:** Dr. Woo Jun Sul, Department of Systems Biotechnology, Chung-Ang University, Anseong, Gyeonggi-Do 17546, Republic of Korea. Tel: 82-31-670-4707; Fax: 82-31-675-3108; E-mail: sulwj@cau.ac.kr; Dr. Sei Won Lee, Department of Pulmonary and Critical Care Medicine, Asan Medical Center, University of Ulsan College of Medicine, 88, Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea. Tel: 82-2-3010-3990; Fax: 82-2-3010-6968; E-mail: seiwon@amc.seoul.kr

## References

[1] WHO report on cancer: setting priorities, investing wisely and providing care for all. Geneva: World Health Organization 2020.

[2] Liu NN, Ma Q, Ge Y, Yi CX, Wei LQ, Tan JC, Chu Q, Li JQ, Zhang P and Wang H. Microbiome dysbiosis in lung cancer: from composition to therapy. NPJ Precis Oncol 2020; 4: 33.

[3] Ramirez-Labrada AG, Isla D, Artal A, Arias M, Rezusta A, Pardo J and Galvez EM. The influence of lung microbiota on lung carcinogenesis, immunity, and immunotherapy. Trends Cancer 2020; 6: 86-97.
[4] Maddi A, Sabharwal A, Violante T, Manuballa S, Genco R, Patnaik S and Yendamuri S. The microbiome and lung cancer. J Thorac Dis 2019; 11: 280-291.
[5] Jin C, Lagoudas GK, Zhao C, Bullman S, Bhutkar A, Hu B, Ameh S, Sandel D, Liang XS, Mazzilli S, Whary MT, Meyerson M, Germain R, Blainey PC, Fox JG and Jacks T. Commensal microbiota promote lung cancer development via γδ T cells. Cell 2019; 176: 998-1013.
[6] Greathouse KL, White JR, Vargas AJ, Bliskovsky VV, Beck JA, von Muhlinen N, Polley EC, Bowman ED, Khan MA, Robles AI, Cooks T, Ryan BM, Padgett N, Dzutsev AH, Trinchieri G, Pineda MA, Bilke S, Meltzer PS, Hokenstad AN, Stickrod TM, Walther-Antonio MR, Earl JP, Mell JC, Krol JE, Balashov SV, Bhat AS, Ehrlich GD, Valm A, Deming C, Conlan S, Oh J, Segre JA and Harris CC. Interaction between the microbiome and TP53 in human lung cancer. Genome Biol 2018; 19: 123.
[7] Liu Y, O'Brien JL, Ajami NJ, Scheurer ME, Amirian ES, Armstrong G, Tsavachidis S, Thrift AP, Jiao L, Wong MC, Smith DP, Spitz MR, Bondy ML, Petrosino JF and Kheradmand F. Lung tissue microbial profile in lung cancer is distinct from emphysema. Am J Cancer Res 2018; 8: 1775-1787.
[8] Yu G, Gail MH, Consonni D, Carugno M, Humphrys M, Pesatori AC, Caporaso NE, Goedert JJ, Ravel J and Landi MT. Characterizing human lung tissue microbiota and its relationship to epidemiological and clinical features. Genome Biol 2016; 17: 163.
[9] Liu HX, Tao LL, Zhang J, Zhu YG, Zheng Y, Liu D, Zhou M, Ke H, Shi MM and Qu JM. Difference of lower airway microbiome in bilateral protected specimen brush between lung cancer patients with unilateral lobar masses and control subjects. Int J Cancer 2018; 142: 769-778.
[10] Peters BA, Hayes RB, Goparaju C, Reid C, Pass HI and Ahn J. The microbiome in lung cancer tissue and recurrence-free survival. Cancer Epidemiol Biomarkers Prev 2019; 28: 731-740.
[11] Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodriguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vazquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R and Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 2019; 37: 852-857.
[12] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 2011; 17: 10-12.
[13] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J and Glockner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2013; 41: D590-596.
[14] Gotelli NJ and Colwell RK. Estimating species richness. In: Magurran AE, McGill BJ, editors. Biological diversity: frontiers in measurement and assessment. Oxford: Oxford University Press; 2011. pp. 39-54.
[15] Magurran AE. Measuring biological diversity. Hoboken, NJ: Wiley; 2004.
[16] Lozupone C, Lladser ME, Knights D, Stombaugh J and Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J 2011; 5: 169-172.
[17] Huang D, Su X, Yuan M, Zhang S, He J, Deng Q, Qiu W, Dong H and Cai S. The characterization of lung microbiome in lung cancer patients with different clinicopathology. Am J Cancer Res 2019; 9: 2047-2063.
[18] Venkataraman A, Bassis CM, Beck JM, Young VB, Curtis JL, Huffnagle GB and Schmidt TM. Application of a neutral community model to assess structuring of the human lung microbiome. MBio 2015; 6: e02284-14.
[19] Kim HJ, Kim JJ, Myeong NR, Kim T, Kim D, An S, Kim H, Park T, Jang SI, Yeon JH, Kwack I and Sul WJ. Segregation of age-related skin microbiome characteristics by functionality. Sci Rep 2019; 9: 16748.
[20] Xia F, Zhou X, Liu Y, Li Y, Bai X and Zhou X. Composition and predictive functional analysis

of bacterial communities inhabiting Chinese Cordyceps insight into conserved core microbiome. BMC Microbiol 2019; 19: 105.
[21] Poje G and Redfield RJ. General methods for culturing Haemophilus influenzae. In: Herbert MA, Hood DW, Moxon ER, editors. Haemophilus influenzae protocols. Totowa, NJ: Humana Press; 2003. pp. 51-56.
[22] Apopa PL, Alley L, Penney RB, Arnaoutakis K, Steliga MA, Jeffus S, Bircan E, Gopalan B, Jin J, Patumcharoenpol P, Jenjaroenpun P, Wongsurawat T, Shah N, Boysen G, Ussery D, Nookaew I, Fagan P, Bebek G and Orloff MS. PARP1 is up-regulated in non-small cell lung cancer tissues in the presence of the cyanobacterial toxin microcystin. Front Microbiol 2018; 9: 1757.
[23] Rutebemberwa A, Stevens MJ, Perez MJ, Smith LP, Sanders L, Cosgrove G, Robertson CE, Tuder RM and Harris JK. Novosphingobium and its potential role in chronic obstructive pulmonary diseases: insights from microbiome studies. PLoS One 2014; 9: e111150.
[24] Goto T. Airway microbiota as a modulator of lung cancer. Int J Mol Sci 2020; 21: 3044.
[25] Ochoa CE, Mirabolfathinejad SG, Ruiz VA, Evans SE, Gagea M, Evans CM, Dickey BF and Moghaddam SJ. Interleukin 6, but not T helper 2 cytokines, promotes lung carcinogenesis. Cancer Prev Res (Phila) 2011; 4: 51-64.
[26] King PT and Sharma R. The lung immune response to nontypeable Haemophilus influenzae (Lung Immunity to NTHi). J Immunol Res 2015; 2015: 706376.
[27] Moghaddam SJ, Ochoa CE, Sethi S and Dickey BF. Nontypeable Haemophilus influenzae in chronic obstructive pulmonary disease and lung cancer. Int J Chron Obstruct Pulmon Dis 2011; 6: 113-123.
[28] Sriram KB, Cox AJ, Sivakumaran P, Singh M, Watts AM, West NP and Cripps AW. Nontypeable Haemophilus influenzae detection in the lower airways of patients with lung cancer and chronic obstructive pulmonary disease. Multidiscip Respir Med 2018; 13: 11.
[29] Xu N, Wang L, Li C, Ding C, Li C, Fan W, Cheng C and Gu B. Microbiota dysbiosis in lung cancer: evidence of association and potential mechanisms. Transl Lung Cancer Res 2020; 9: 1554-1568.
[30] Kovaleva O, Podlesnaya P, Rashidova M, Samoilova D, Petrenko A, Zborovskaya I, Mochalnikova V, Kataev V, Khlopko Y, Plotnikov A and Gratchev A. Lung microbiome differentially impacts survival of patients with non-small cell lung cancer depending on tumor stroma phenotype. Biomedicines 2020; 8: 349.
[31] Safdar A and Rolston KV. Stenotrophomonas maltophilia: changing spectrum of a serious bacterial pathogen in patients with cancer. Clin Infect Dis 2007; 45: 1602-1609.
[32] Brooke JS. Stenotrophomonas maltophilia: an emerging global opportunistic pathogen. Clin Microbiol Rev 2012; 25: 2-41.
[33] Cheng C, Wang Z, Wang J, Ding C, Sun C, Liu P, Xu X, Liu Y, Chen B and Gu B. Characterization of the lung microbiome and exploration of potential bacterial biomarkers for lung cancer. Transl Lung Cancer Res 2020; 9: 693-704.
[34] Anderson NM, Mucka P, Kern JG and Feng H. The emerging role and targetability of the TCA cycle in cancer metabolism. Protein Cell 2018; 9: 216-237.
[35] Vanhove K, Derveaux E, Graulus GJ, Mesotten L, Thomeer M, Noben JP, Guedens W and Adriaensens P. Glutamine addiction and therapeutic strategies in lung cancer. Int J Mol Sci 2019; 20: 252.
[36] Siddiqui A and Ceppi P. A non-proliferative role of pyrimidine metabolism in cancer. Mol Metab 2020; 35: 100962.
[37] Carbone C, Piro G, Di Noia V, D'Argento E, Vita E, Ferrara MG, Pilotto S, Milella M, Cammarota G, Gasbarrini A, Tortora G and Bria E. Lung and gut microbiota as potential hidden driver of immunotherapy efficacy in lung cancer. Mediators Inflamm 2019; 2019: 7652014.
[38] Gupta VK, Paul S and Dutta C. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. Front Microbiol 2017; 8: 1162.

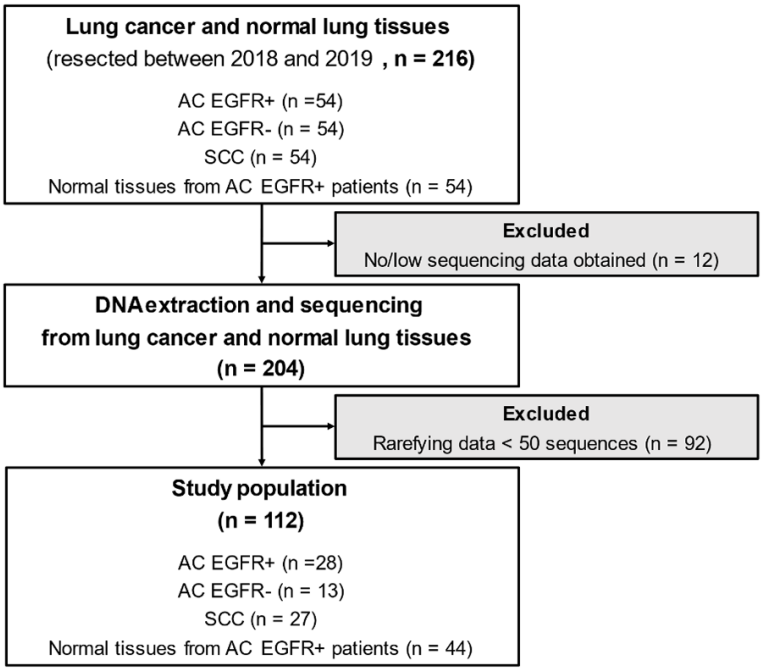


**Figure S1.** Study flow.

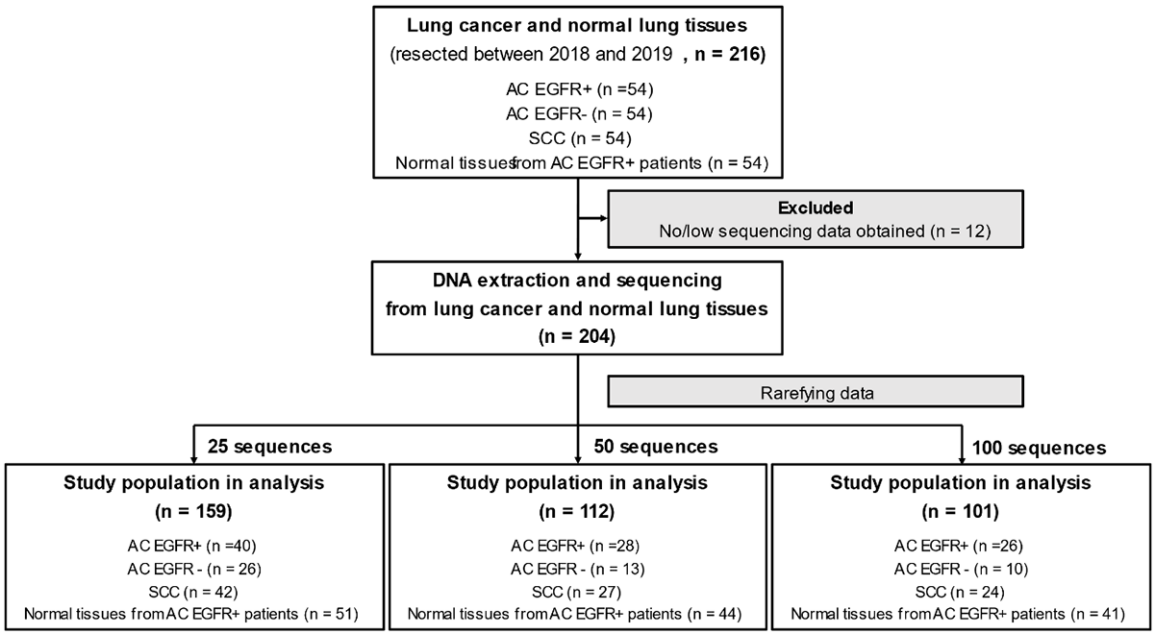


**Figure S2.** Study flow as rarefying samples at various sequences.

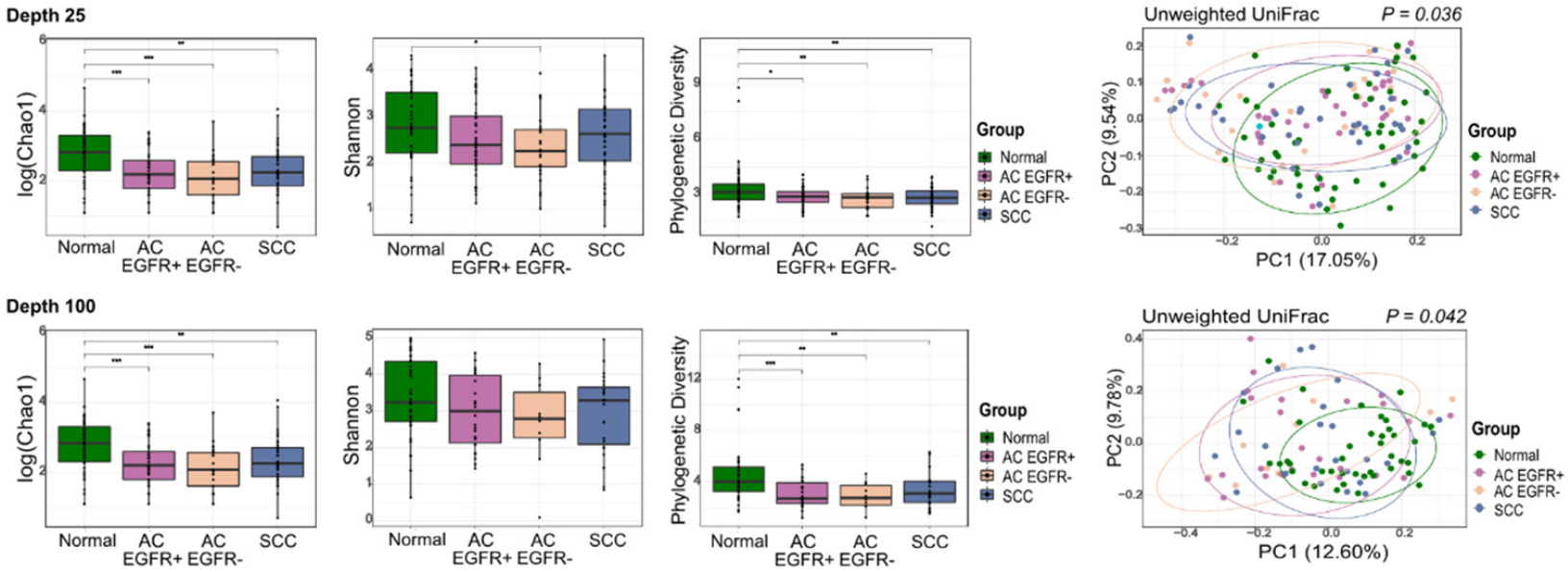


**Figure S3.** Comparison of the alpha and beta diversities between lung cancer and normal lung tissues by rarefying samples to the depth of 25 and 100 sequences.

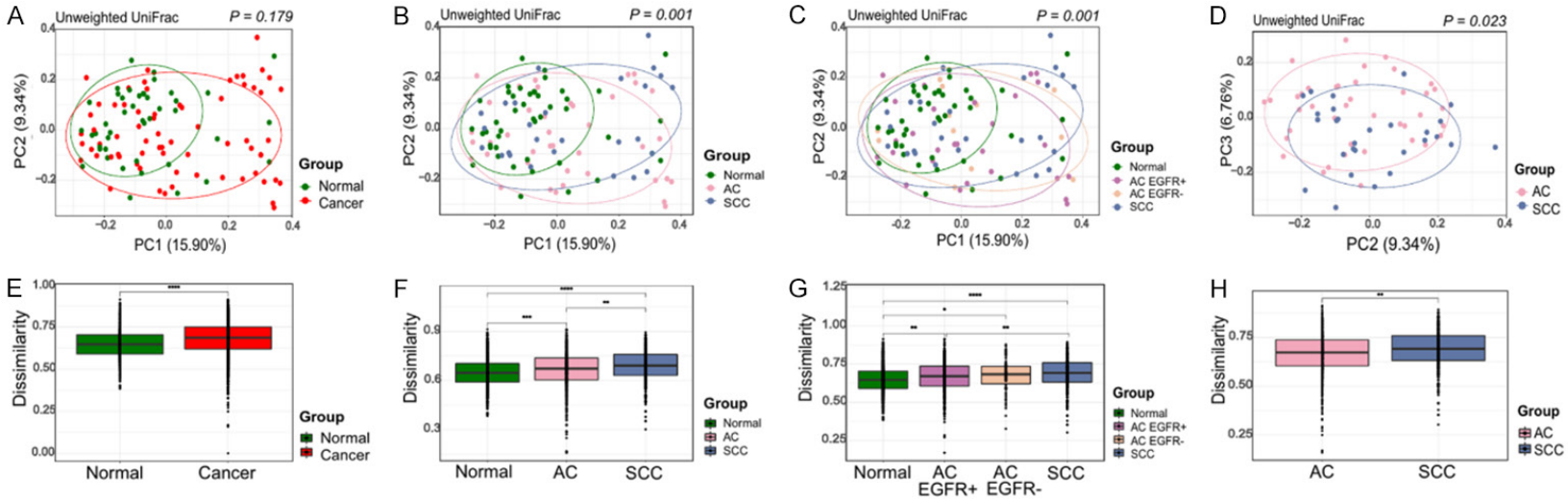

**Figure S4.** Comparison of the beta diversity between lung cancer and normal lung tissues. (A-D) Principal coordinates analyses of the unweighted UniFrac distance between different groups. (A) Normal vs. Cancer, (B) Normal vs. AC vs. SCC, (C) Normal vs. AC EGFR+ vs. AC EGFR- vs. SCC, and (D) AC vs. SCC. (E-H) Box-and-whisker plot of unweighted intragroup UniFrac distance between each sample and all other samples from the same group. (E) Normal vs. Cancer, (F) Normal vs. AC vs. SCC, (G) Normal vs. AC EGFR+ vs. AC EGFR- vs. SCC, and (H) AC vs. SCC.

**Table S1.** The ASV count classified as species *H. influenzae* in each sample

| Group | Patient age/sex | ASV1 | ASV9 | ASV24 | ASV1011 |
|---|---|---|---|---|---|
| Normal | 80/F | 0 | 0 | 0 | 5 |
| AC | 58/F | 225 | 0 | 0 | 0 |
| SCC | 72/M | 796 | 4,008 | 820 | 0 |
| SCC | 71/M | 6 | 0 | 0 | 0 |
| SCC | 67/M | 0 | 15 | 0 | 0 |
| SCC | 66/M | 3 | 0 | 0 | 0 |
| SCC | 83/M | 0 | 54 | 0 | 0 |
| SCC | 71/F | 29,116 | 0 | 0 | 0 |
| SCC | 59/M | 6 | 0 | 0 | 0 |

ASV, amplicon sequence variant; AC, adenocarcinoma; SCC, squamous cell carcinoma.

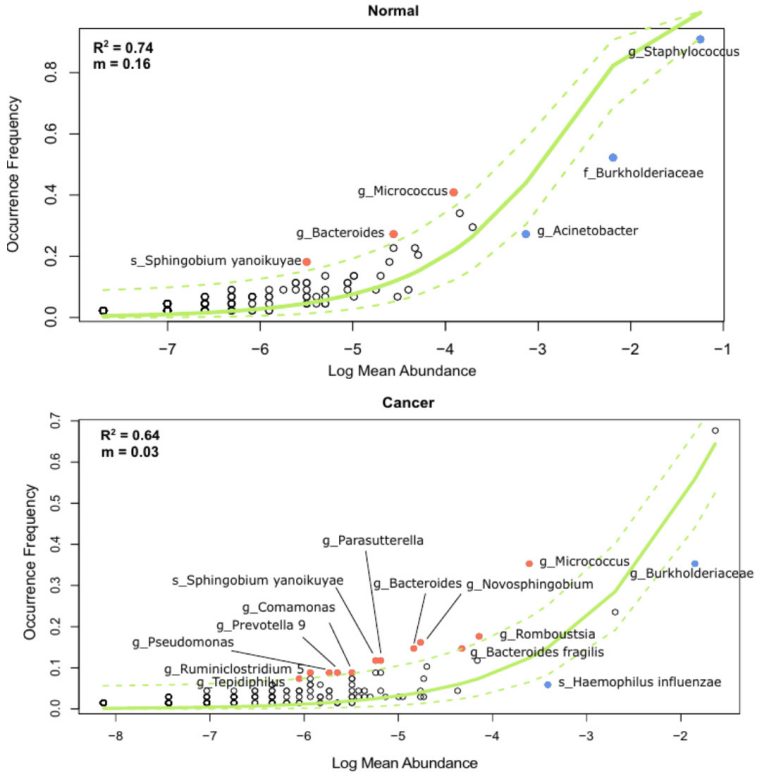


**Figure S5.** The neutral community model-based dominance analysis. The theoretical and observed relationships between the log mean relative abundance of a species and the occurrence frequency were compared. Each dot represents a different amplicon sequence variant (ASV), and the solid green line represents the best fit to the neutral mode. The dashed lines indicate the 95% confidence intervals for the neutral model prediction. The ASVs that occurred more frequently than predicted using the model are shown in orange, whereas those that occurred less frequently than predicted are shown in blue.