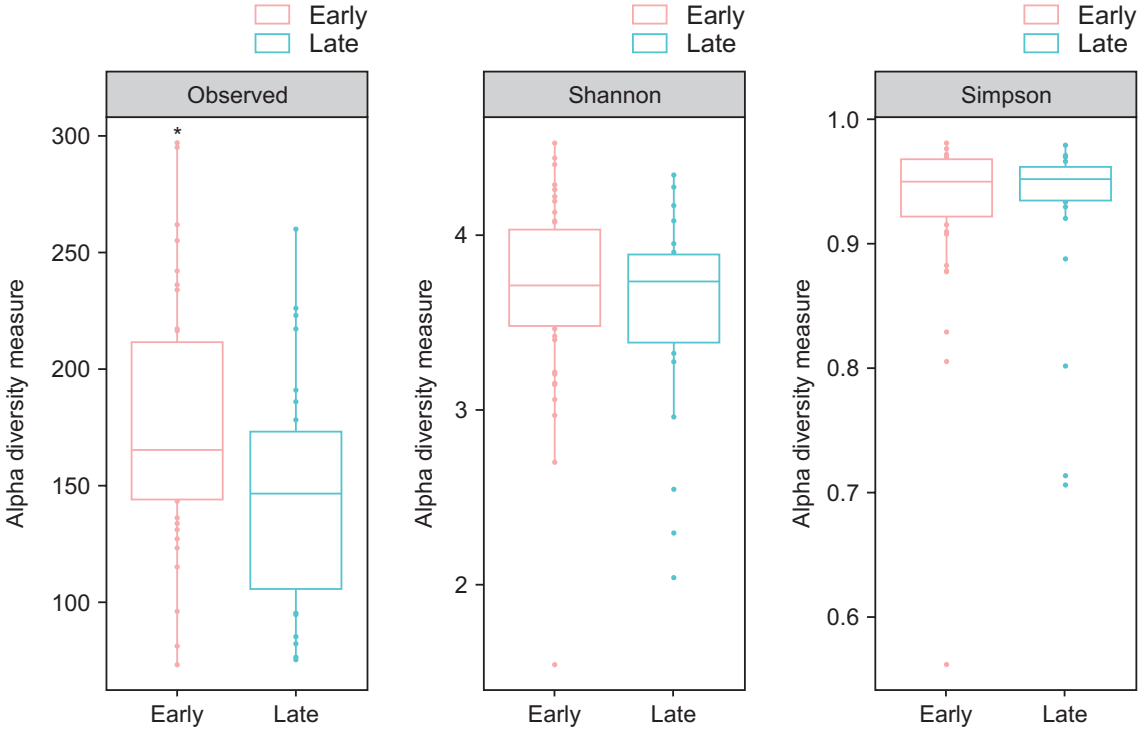
Supplementary Materials and Methods

1. 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 and 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 version 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. Shannon and Simpson alpha diversity indices were estimated[6]. To visualize the sample differences, β-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 microbiome taxonomic profiling (MTP), which is a CJ Bioscience's bioinformatic cloud platform.

Supplementary Figure S1. Comparison of the α-diversity in bronchoalveolar lavage (BAL) fluid microbiomes between early and late stage of non-small cell lung carcinoma. *p<0.05 by Wilcoxon test.



Supplementary References

1. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114-20.
2. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ 2016;4:e2584.
3. Myers EW, Miller W. Optimal alignments in linear space. Comput Appl Biosci 1988;4:11-7.
4. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol 2017;67:1613-7.
5. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 2011;27:2194-200.
6. Magurran AE. Measuring biological diversity. New York: John Wiley & Sons; 2013.
7. Beals EW. Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. In: MacFadyen A, Ford ED, editors. Advances in ecological research. London: Academic Press; 1984. p. 1-55.
8. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol 2011;12:R60.
9. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583-621.
10. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 2013;31:814-21.
11. Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Comput Biol 2009;5:e1000465.