

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 0.32¹. Paired-end sequence data were merged together after the QC step using the fastq_mergepairs command of VSEARCH version 2.13.4² with default parameters.

We then trimmed the primers using the alignment algorithm of Myers and Miller³ 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 VSEARCH². The EzBioCloud 16S rRNA database⁴ was used for taxonomic assignment of the obtained 16S rRNA sequences using the 'usearch_global' command of VSEARCH², followed by a more precise pairwise alignment³. Chimeric reads were filtered from reads with <97% similarity via reference-based chimeric detection using the UCHIME algorithm⁵ 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' command² 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 estimated⁶. To visualize the sample differences, β -diversity distances were calculated using the method described by Bray-Curtis⁷. Taxonomic biomarkers and functional biomarkers were identified using statistical comparison algorithms (linear discriminant analysis [LDA] effect size [LEFse]⁸ and Kruskal-Wallis H test⁹). 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.

