Supplementary Material

Bioinformatics for bacterial community characterization

Raw sequence data (FASTQ files) from 16S rDNA sequencing were demultiplexed and quality assessed using the q2-demux plugin. Then, denoising, filtering, and chimera removal were performed with DADA2 pipeline (via q2-dada2 plugin),[1] thus identifying all amplicon sequence variants (ASVs)[2] and their relative abundance in each sample. To minimize the number of spurious ASVs, those unique sequences with a total abundance lower than 7 reads across all samples were filtered out.[3] ASVs were firstly aligned and then used to construct a phylogenetic tree by following the align-to-tree-mafft-fasttree pipeline[4-5] from the q2-phylogeny plugin. ASVs were taxonomically classified by using the classify-sklearn naïve Bayes taxonomy classifier (via q2-feature-classifier plugin)[6] against the Silva 132 database.[7] Sequences not assigned to any taxa or classified as Eukaryote or Archaea, were filtered out. Diversity analysis was done using q2-diversity plugin, after samples were normalized via rarefaction (subsampled without replacement). Diversity analysis comprised alpha diversity metrics (Chao1, Shannon index, and Faith-pd,[8] which measure microbiota degree of diversity) and beta diversity metrics (unweighted UniFrac[9] and weighted UniFrac,[10] which measure microbiota composition differences between samples, whileup-weighting differences in ASVs phylogenetic distance). Unweighted UniFrac reports differences in the presence or absences of ASVs, while weighted UniFrac also reports differences in the abundance of ASVs.

Bioinformatics for fungal community characterization

Raw sequence data from ITS1 sequencing were demultiplexed and quality assessed using the q2-demux plugin. Then, q2-itsxpress plugin[11] was used to quality filter and

trim the ITS region from sequences. After that, denoising, merging, and ASVs calling were done by DADA2 pipeline (via q2-dada2 plugin), thus identifying all ASVs and their relative abundance in each sample. Very low abundance ASVs (total n<7) were filtered out, as was done for bacteria. ASVs were taxonomically classified by using the classify-sklearn naïve Bayes taxonomy classifier, against the UNITE 7.2 database.[12] Diversity analysis was done using q2-diversity plugin, after samples were rarefied (subsampled without replacement). Diversity analysis comprised alpha diversity metrics (Chao1 and Shannon index), and beta diversity metrics Bray-Curtis distance,[13] which measure non-phylogenetic microbiome composition differences between samples).

Quality and throughput of the sequencing process

Negative controls sequencing results are detailed in the supplementary Table 1. 16s rDNA sequencing in the 125 samples yielded 7,336,127 reads whereas the 4 negative controls 53, 220, 447, and 881 reads, discarding high contamination levels. After quality filtering, chimera removal, and discard of very low frequency sequences, a total of 4,897,627 reads were finally obtained. After taxonomic assignment, non-bacterial sequences were removed and 4,464,212 reads of 4434 ASVs were finally obtained. Three samples from affected bronchi (patients 23, 24, and 25) and four samples from control bronchi (controls 2, 3, 7, and 16) were not included in downstream analysis due to sequencing depth requirements (<1000 reads per sample), and samples were rarefied to 1007 sequences to maximized the retention of samples and preserved representative diversity measures, according to alpha rarefaction plot.

The 50 samples using for mycobiome determination by ITS1 amplification and massive sequencing yielded 6,034,038 reads and the negative control performed 5,036 reads. After quality filtering, trimming, merging, and discard of very low frequency

sequences, 3,535,951 reads of 515 ASVs were finally obtained. One saliva sample from a control, which read count was very low, was not included in downstream analyses due to sampling depth requirements (>1000 reads). Rarefaction was set to 13,571 sequences per sample, since that sampling depth conserved all samples and reached a *plateau* according to alpha rarefaction plot inspection. The negative control was fully inspected to identify which taxa were present.

Statistical analysis

Statistical differences in mean alpha diversity metrics between patients and controls were calculated by Kruskal-Wallis test.[14] Differences in microbiota composition between samples were assessed and plotted by performing Principal Coordinates Analysis (PCoA) based on the beta diversity metrics. Permutational multivariate analysis of variance (PERMANOVA)[15] was performed to determine which factors explained statistically significant variance in microbiota composition. All statistical tests were conducted via q2-diversity plugin from QIIME2. To determine which specific taxa explained beta diversity differences, differential abundance analyses were performed only in variables that yielded statistically significant differences in beta diversity analysis. For that purpose, linear discriminate analysis effect size (LEfSe) was used for testing taxonomic comparisons.[16] LEfSe uses the common tests for statistical significance (Kruskal-Wallis test and pairwise Wilcoxon test) with linear discriminate analysis for taxa selection. Alpha value for the factorial Kruskal-Wallis test was 0.01 and threshold on the logarithmic LDA score for discriminative taxa was set to 4.0. ROC curves were plotted by IBM SPSS Statistics software (Version 22.0. Armonk, NY: IBM Corp).

To get an overview of the complexity of the microbiome in lung cancer disease, we have designed the taxon interaction network for the different data (healthy lung, healthy saliva, affected lung, diseased contralateral lung, diseased saliva and diseased stool) Networks have nodes and links. The nodes represent the different taxa and the links represent that both taxa have appeared in the same patient sample. In this case, we have used weighted networks to represent the microbiome. The networks have been created as follows: for each patient, we select the taxa with non-zero abundances. Then, we build a complete network between these taxa. The link weight has been obtained as the product of the probability of the presence of taxa. Finally, we have added the networks of each patient in a global network. This analysis are available at https://github.com/JJ-Lab/Cancer_Lung_Microbiota website.

References

1. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 2016;**13:**581-3. doi: 10.1038/nmeth.3869.
2. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 2017;11:2639-2643. doi: 10.1038/ismej.2017.119.
3. Wang J, Zhang Q, Wu G, Zhang C, Zhang M, Zhao L. Minimizing spurious features in 16S rRNA gene amplicon sequencing. PeerJ Preprints 2018;**6**:e26872v1.
4. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;**30**:3059-66. doi: 10.1093/nar/gkf436.

5. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 2010:**5**:e9490. doi: 10.1371/journal.pone.0009490.
6. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, *et al.* Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 2018;**6**:90. doi: 10.1186/s40168-018-0470-z.
7. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2013;**41**:D590-6. doi: 10.1093/nar/gks1219.
8. Faith DP. Conservation evaluation and phylogenetic diversity. Biological Conservation 1992;**61**:1-10.
9. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005;**71**:8228-35. doi: 10.1128/AEM.71.12.8228-8235.2005.
10. Lozupone CA, Hamady M, Kelley ST and Knight R. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol 2007;**73**:1576-85. doi: 10.1128/AEM.01996-06.
11. Rivers AR, Weber KC, Gardner TG, Liu S, Armstrong SD. ITSxpress: Software to rapidly trim internally transcribed spacer sequences with quality scores for marker gene analysis. F1000Res 2018;**7**:1418. doi: 10.12688/f1000research.15704.1.
12. UNITE Community (2017): UNITE QIIME release. Version 01.12.2017. UNITE Community. https://doi.org/10.15156/BIO/587481.

13. Bray JR, Curtis JT. An ordination of upland forest communities of southern Wisconsin. Ecological Monographs 1957;**27**:325-49.

14. Kruskal W, Wallis W. Use of ranks in one-criterion variance analysis. J Amer Statist Assoc 1952;**47**:583-621.

15. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol 2001;**26**:32-46.

16. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biol 2011;**12**:R60. doi: 10.1186/gb-2011-12-6-r60.