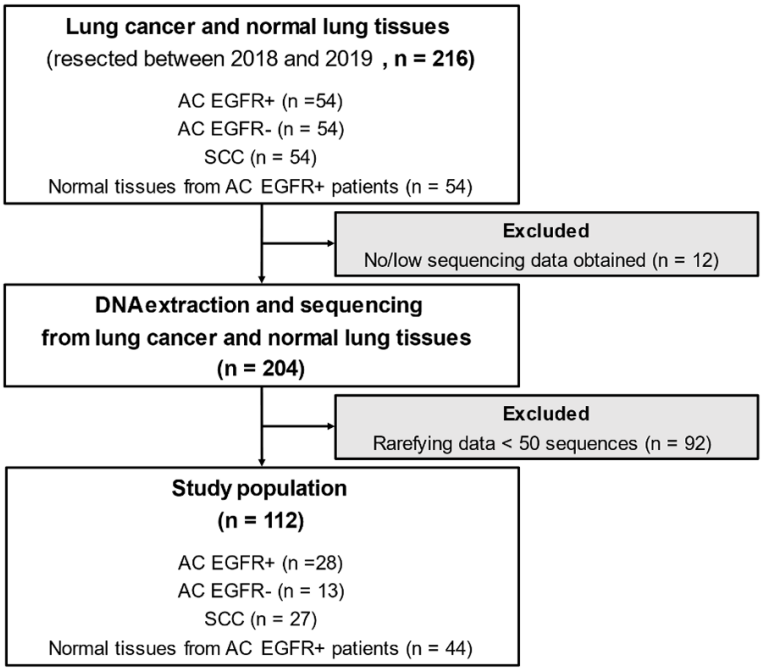

Figure S1. Study flow.

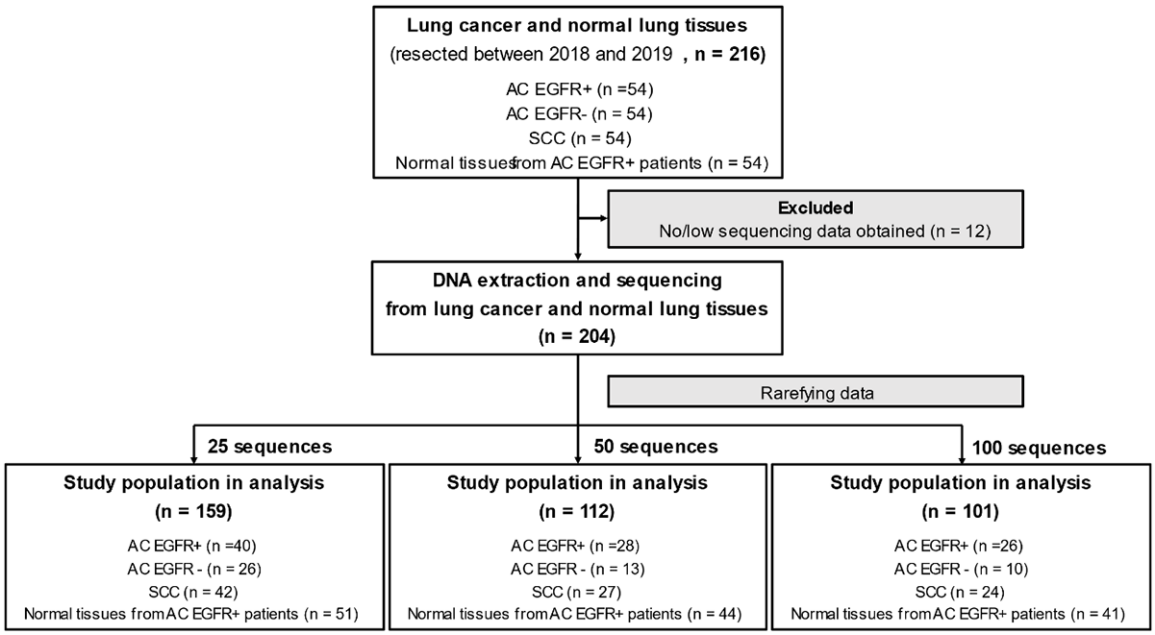

Figure S2. Study flow as rarefying samples at various sequences.

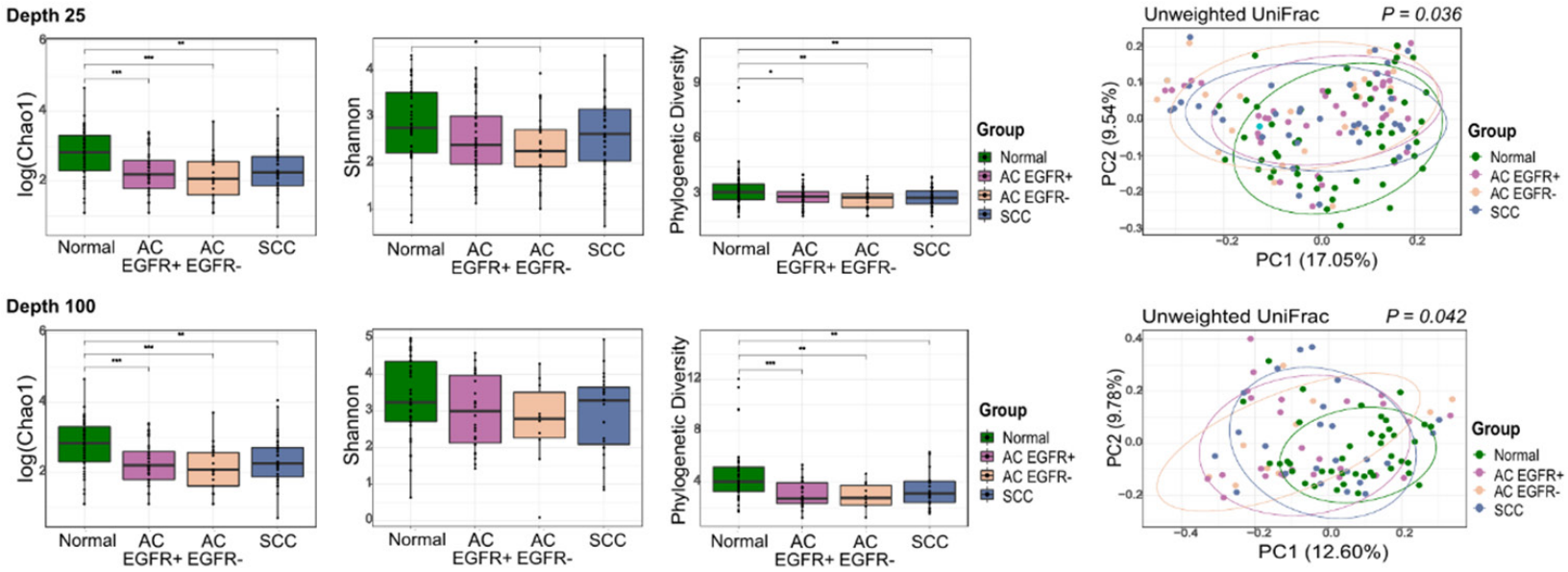

Figure S3. Comparison of the alpha and beta diversities between lung cancer and normal lung tissues by rarefying samples to the depth of 25 and 100 sequences.

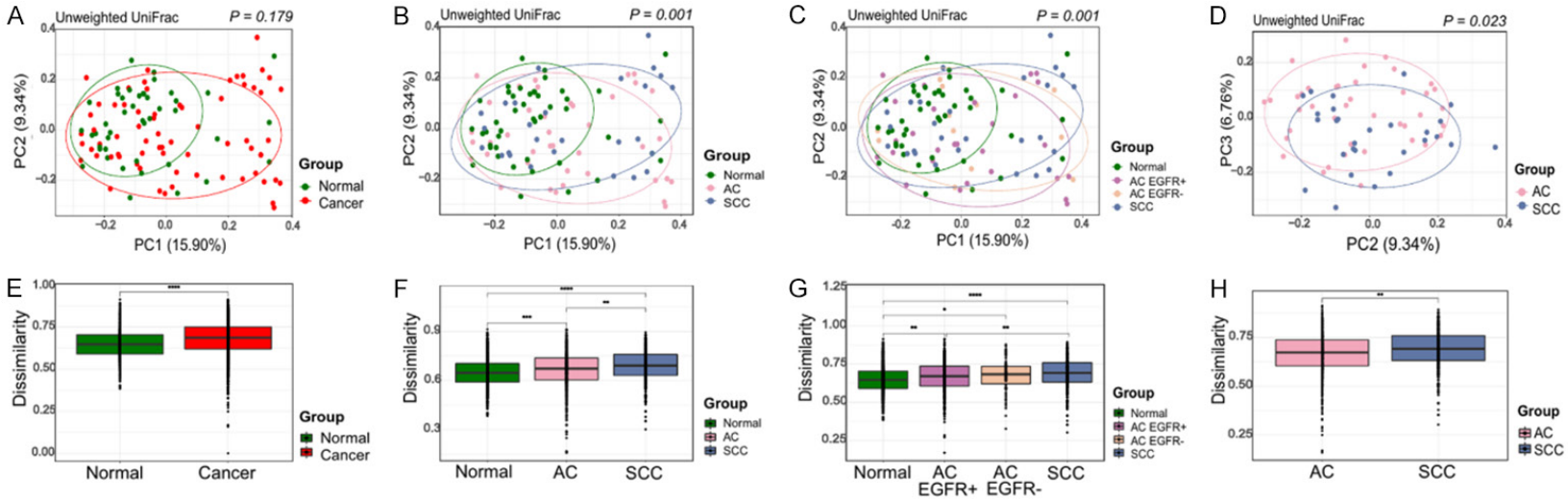

Figure S4. Comparison of the beta diversity between lung cancer and normal lung tissues. (A-D) Principal coordinates analyses of the unweighted UniFrac distance between different groups. (A) Normal vs. Cancer, (B) Normal vs. AC vs. SCC, (C) Normal vs. AC EGFR+ vs. AC EGFR- vs. SCC, and (D) AC vs. SCC. (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.

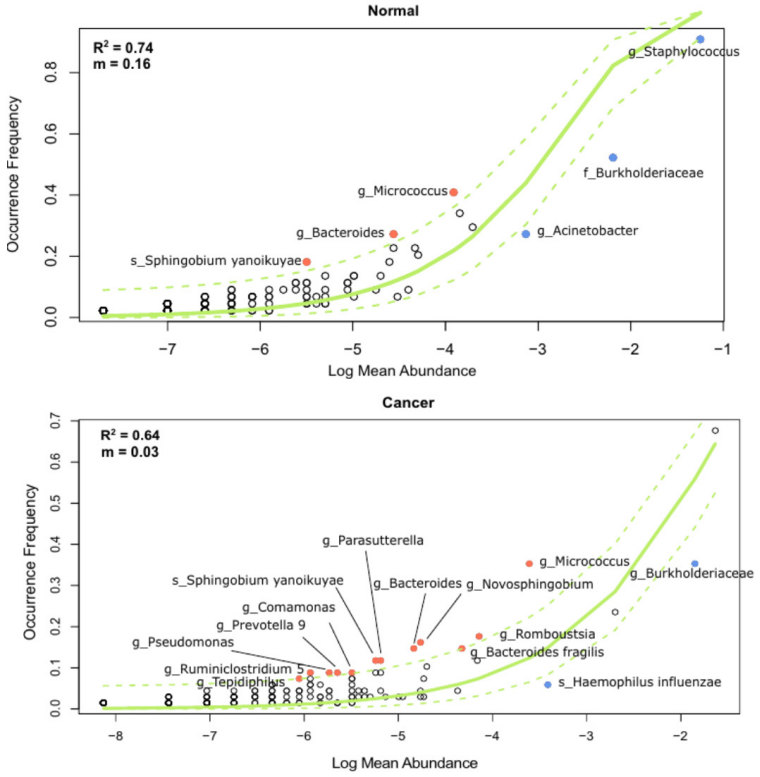

Figure S5. The neutral community model-based dominance analysis. The theoretical and observed relationships between the log mean relative abundance of a species and the occurrence frequency were compared. Each dot represents a different amplicon sequence variant (ASV), and the solid green line represents 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.