Lower airway bacterial microbiome may influence recurrence after resection of early-stage non–small cell lung cancer

Santosh K. Patnaik, MD, PhD,[a] Eduardo G. Cortes, MS,[b] Eric D. Kannisto, MS,[a]
Achamaporn Punnanitinont, BA,[a] Samjot S. Dhillon, MD,[c] Song Liu, PhD,[b] and Sai Yendamuri, MD[a]

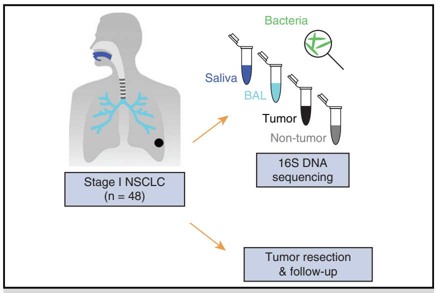
ABSTRACT

Objective: The lower airway bacterial microbiome influences carcinogenesis and response to immunotherapy in non–small cell lung cancer (NSCLC). We investigated the association of this microbiome with recurrence in early NSCLC.

Methods: Microbiomes of presurgery bronchoalveolar lavage (BAL) and saliva, and resected stage I NSCLC tumor and adjacent lung tissues of 48 patients were examined by *16S* gene sequencing. Tumor gene expression was measured by RNA sequencing.

Results: Spatial relationships of the different biospecimen types was reflected in their microbiomes, with microbiomes of BAL intermediate to those of saliva and lung tissue. BAL and saliva microbiomes were less dissimilar in patients with high α-amylase levels in BAL, indicating oral aspiration as a source of lower airway microbiota. BAL microbiomes of patients with recurrence within 32 months of surgery differed from those without recurrence during $\geq$32 months of follow-up ($n = 18$ each), despite no difference for age, sex, smoking history, and tumor histology and grade. The recurrence-associated BAL microbiome signature was present in 16 of the 18 recurrence cases but in only two of the others. Signature presence was associated with shorter recurrence-free survival (log-rank test $P < .001$; hazard ratio = 14.5), and greater expression in tumors of genes for cell proliferation and epithelial mesenchymal transition. Immune cellular composition of the tumor microenvironment was not different between patients with and without the signature.

Conclusions: Presurgery composition of lower airway microbiome may be associated with recurrence of early NSCLC. This association may reflect an influence of the microbiome on tumor biology. (J Thorac Cardiovasc Surg 2021;161:419-29)



Microbiome affects lung cancer outcome.

CENTRAL MESSAGE

Bacterial community in the lower airway may have an influence on the biology of lung cancer or the immune response to it. This influence can have an effect on disease recurrence after tumor resection.

PERSPECTIVE

It is now clear that human beings exist in an intimate symbiotic relationship with bacteria. It is also evident that this relationship is altered in and perhaps causative to human pathologic processes. Our study suggests that the bacterial community in the lower airway may have an influence on recurrence of early-stage lung cancer. Engineering this bacterial community, for example, with antibiotics, may be an approach that is worthy of investigation to develop adjunctive therapies for this disease.

See Commentaries on pages 430 and 432.

Lung cancer is a major cause of cancer related deaths in the United States and around the world. Non–small cell lung cancer (NSCLC), the subtype that constitutes the overwhelming majority of these cases, has a disappointing cure rate. Surgery, which is the only treatment that can be offered to stage I disease, has a 5-year survival of 73%.[1] Thus, a significant proportion of patients with early-stage NSCLC have recurrence, mostly

From the Departments of [a]Thoracic Surgery, [b]Biostatistics & Bioinformatics, and [c]Medicine, Roswell Park Comprehensive Cancer Center, Buffalo, NY.

This work was supported by intramural research funding to Dr Dhillon and by institutional support grant P30-CA016056 to RPCCC from National Cancer Institute.

Dr Patnaik and Mr Cortes contributed equally to this article.

Present address of Dr Dhillon: Sacramento Medical Center, Sacramento, Calif.

Read at the International Thoracic Surgical Oncology Summit 2019, New York, New York, September 27-28, 2019.

Received for publication Oct 17, 2019; revisions received Dec 25, 2019; accepted for publication Jan 18, 2020; available ahead of print March 23, 2020.

Address for reprints: Sai Yendamuri, MD, Roswell Park Comprehensive Cancer Center, Elm and Carlton Sts, Buffalo, NY 14263 (E-mail: sai.yendamuri@roswellpark.org).

0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2020.01.104

Abbreviations and Acronyms

BAL	= bronchoalveolar lavage
CI	= confidence interval
GSEA	= gene set enrichment analysis
LCBRN	= Lung Cancer Biospecimen Resource Network
NSCLC	= non–small cell lung cancer
OTU	= operational taxonomic unit
RABMS	= recurrence-associated BAL microbiome signature
rRNA	= ribosomal RNA

Scanning this QR code will take you to the article title page to access supplementary information.

within 3 years after surgery.[2] Identification of such biomarkers of recurrence beyond clinical and pathologic variables may enable the design and deployment of novel therapeutic strategies.

The healthy lung parenchyma has traditionally been considered sterile, but recent studies show that it harbors a bacterial community.[3-5] This is in keeping with the natural physiology of the lung as it is exposed to the external environment with every breath. The lung is also constantly subjected to microaspiration of oral secretions that are rich in bacteria.[6] In addition, tobacco in cigarettes has a high bacterial and fungal content that can be transferred in a viable form to lungs.[7,8] The significant advancement in detection of bacterial species by affordable high-throughput nucleic acid sequencing of the *16S* ribosomal RNA (rRNA) gene has enabled the comprehensive examination of this relationship.[9] Although alterations in the diversity and abundance of bacterial species of the lung microbiome have been noted for diseases such as chronic obstructive lung disease and cystic fibrosis,[10-12] their association with cancer is just being elucidated. Recent studies have pointed out the relationship between the respiratory microbiome and characteristics of the corresponding lung cancers.[13,14] In this study, we sought to examine the association of the airway microbiome composition with recurrence after resection of early-stage NSCLC. In addition, we sought to examine the association between recurrence-associated lower microbiome patterns with cellular pathways and the stromal immune composition of tumors assessed by gene expression analyses.

METHODS

Study Approval and Role of Funding Agencies

This retrospective study was conducted under protocol BDR 075016 (March 3, 2017) of the institutional review board of Roswell Park Comprehensive Cancer Center, Buffalo, NY. Funding agencies played no role in data interpretation.

Biospecimens and Clinical Data

Presurgery bronchoalveolar lavage (BAL) fluid and saliva samples, DNA extracted from frozen surgically resected lung tumor and adjacent non-tumor lung tissues, and clinical data of 48 patients were obtained from the Lung Cancer Biospecimen Resource Network (LCBRN).[15] Tissue and BAL and saliva fluid samples had been collected at 3 different academic medical centers in Charleston, SC (n = 14), Charlottesville, Va (n = 18), and St Louis, Mo (n = 16), and stored and processed at LCBRN's coordination center at University of Virginia, Charlottesville, Va. All patients were accrued during 2011-2014 and had pathologic stage I NSCLC, which was treated by resection without neoadjuvant therapy (Tables 1 and E1). Twenty-four patients were known to have recurrence after resection. The other group of 24 patients without known recurrence was chosen to match the recurrence group for various demographic and clinicopathologic variables. DNA of non-tumor tissues could not be obtained for 10 patients. Total RNA from the tumor tissues was also procured for 39 cases for which it was available. Additional details about the procedures for biospecimen collection and DNA/RNA isolation by LCBRN, and measurement of α-amylase activity levels of BAL fluid samples are in the Appendix E1. Standard operating procedures used by LCBRN are available online at the repository's web site (http://lungbio.sites.virginia.edu/standard-operating-procedures; last accessed on December 13, 2019).

Generation and Analyses of 16S Sequencing Data

A 0.46-kb amplicon of *16S* rRNA V3-V4 region was amplified for sequencing; detailed methods are in the Appendix E1. Raw sequencing data were deposited in the European Nucleotide Archive (study identification PRJEB29934). Data were processed with QIIME[16] software (version 1.9.1) with open-reference operational taxonomic unit (OTU)-picking workflow[17] and PyNAST[18] sequence alignment and uclust[19] aligned sequence clustering methods, as detailed in the Appendix E1. OTUs were assigned a taxonomy as per their representative *16S* gene sequence based on Greengenes[20] bacterial *16S* rRNA gene database (version 13.5; 97% sequence identity cut-off). Resulting count data for 1422 OTUs and 190 samples were used for further analyses in R (version 3.6; R Foundation for Statistical Computing, Vienna, Austria) using functions provided with phyloseq[21] and vegan packages (versions 1.28.0 and 2.5-4, respectively). Analyses of alpha and beta diversity are detailed in the Appendix E1. For 2-group comparison of abundance of microbial taxa, samples with total OTU count $\leq$500 were removed, and OTU count data were agglomerated at the genus level for differential abundance analysis with DESeq2[22] Bioconductor package (version 1.20.0) after excluding genera that had non-zero counts in <1/6th of the analyzed samples. Relative log expression normalization[22] and Wald significance testing based on local dispersion estimates were used in the analyses. Resulting *P* values were adjusted for multitesting with the Benjamini–Hochberg method. OTUs with adjusted $P < .05$ and absolute fold-change >1 were deemed differentially abundant. Count data for these OTUs, agglomerated at the genus level, were used for Dirichlet-multinomial mixtures modeling[23] to identify microbiome signatures of group phenotypes; DirichletMultinomial Bioconductor package

TABLE 1. Characteristics of patients of the study

	All (n = 48)	Recurrence (n = 18)*	No recurrence (n = 18)	*P* value†
Age, y, at diagnosis, mean; SD	65.4; 8.6	66.2; 7.7	64.7; 9.7	.61
Time, mo after surgery, mean; SD				
For recurrence	24.6; 16.8	14.6; 8.4		
For last follow-up	43.9; 17.2	34.5; 17.7	50.3; 9.8	<.01
Sex, n (%)				1.00
Female	25 (52)	10 (56)	10 (56)	
Male	23 (48)	8 (44)	8 (44)	
Race, n (%)				1.00
White	44 (92)	16 (88)	17 (94)	
African-American	4 (8)	2 (11)	1 (6)	
Treating institution, n (%)				
University of SC	14 (29)	3 (17)	6 (33)	.59
University of Virginia	18 (38)	8 (44)	6 (33)	
Washington University	16 (33)	7 (39)	6 (33)	
Smoking history, n (%)				.13
Current	18 (38)	5 (28)	11 (61)	
Past	26 (54)	11 (61)	6 (33)	
Never	4 (8)	2 (11)	1 (6)	
BAL α-amylase,‡ n (%)				.51
Low	25 (52)	10 (56)	7 (39)	
High	23 (48)	8 (44)	11 (61)	
Tumor histology, n (%)				.72
Adenocarcinoma	34 (71)	11 (61)	13 (72)	
Squamous cell ca.	13 (27)	6 (33)	5 (28)	
Adenosquamous ca.	1 (2)	1 (6)	0 (0)	
Tumor pathologic stage, n (%)				.72
Ia	27 (56)	11 (61)	13 (72)	
Ib	21 (44)	7 (39)	5 (28)	
Tumor size, cm in radioimaging, mean; SD	2.5; 1.1	2.4; 0.9	2.5; 1.2	.82

SD, Standard deviation; SC, South Carolina; *BAL*, bronchoalveolar lavage; *ca.*, carcinoma. *Recurrence within 32 months of lung cancer surgery; *No recurrence* cases had no recurrence during follow-up of ≥32 months. †In comparison of *Recurrence* and *No recurrence* groups using 2-tailed standard *t* and Fisher exact tests, respectively, for continuous and categorical variables. ‡Enzyme activity of BAL fluid samples: *low*, median or less among all 48 patients; *high*, greater than median.

(version 1.24.1) with goodness of fit assessed by Laplace criteria was used for this.

Generation and Analyses of Tumor Gene Expression

RNA of 39 tumors was sequenced as described in the Appendix E1. Raw sequencing data was deposited in the European Nucleotide Archive (study identification PRJEB29932). Tumor-infiltrating cells were estimating from gene expression data using CIBERSORT[24] and EPIC.[25] For 2-group comparison of tumor gene expression, DESeq2[22] Bioconductor package (version 1.20.0) was used. Genes that did not have expression value of >1 TPM in at least half of the analyzed samples were excluded from analyses. Relative log expression normalization and Wald significance testing based on local dispersion estimates were used in the analyses. Resulting *P* values were adjusted for multitesting with the Benjamini–Hochberg method. For gene set enrichment analysis, gene set enrichment analysis (GSEA)[26] software (version 3.0) and mSigDb[27] Hallmark and C2 CP:Reactome biological process gene set collections (version 6.2) were used. Classic pre-ranked GSEA[26] method with genes ranked by *P* values in DESeq2-based differential expression analyses was used. Absolute normalized enrichment score ≥2 and false discovery rate <25%, as suggested for the GSEA method, were required to consider a gene set as significantly enriched. Categories of Reactome sets were identified as their top-level pathway nodes in the Reactome pathway hierarchy.

RESULTS

Generation of Bacterial Microbiome Data

High-throughput sequencing of the 0.46-kb V3-V4 hypervariable region of prokaryotic *16S* rRNA gene was performed to profile bacterial microbiomes of presurgery oral (saliva) and lower respiratory (BAL fluid), and resected lung tumor and adjacent non-tumor tissue samples of 48 patients with stage I NSCLC (Tables 1 and E1). The experiment design for generation of *16S* data is discussed in the Appendix E2. Of the 129,000 to 305,000 sequencing reads that were obtained on average for each type of sample, 44% to 80% were usable for mapping against the Greengenes *16S* sequence database (Figure 1, *A*). For saliva, BAL fluid, and control samples, sequences of ≥85% of the usable reads could be matched with entries in the

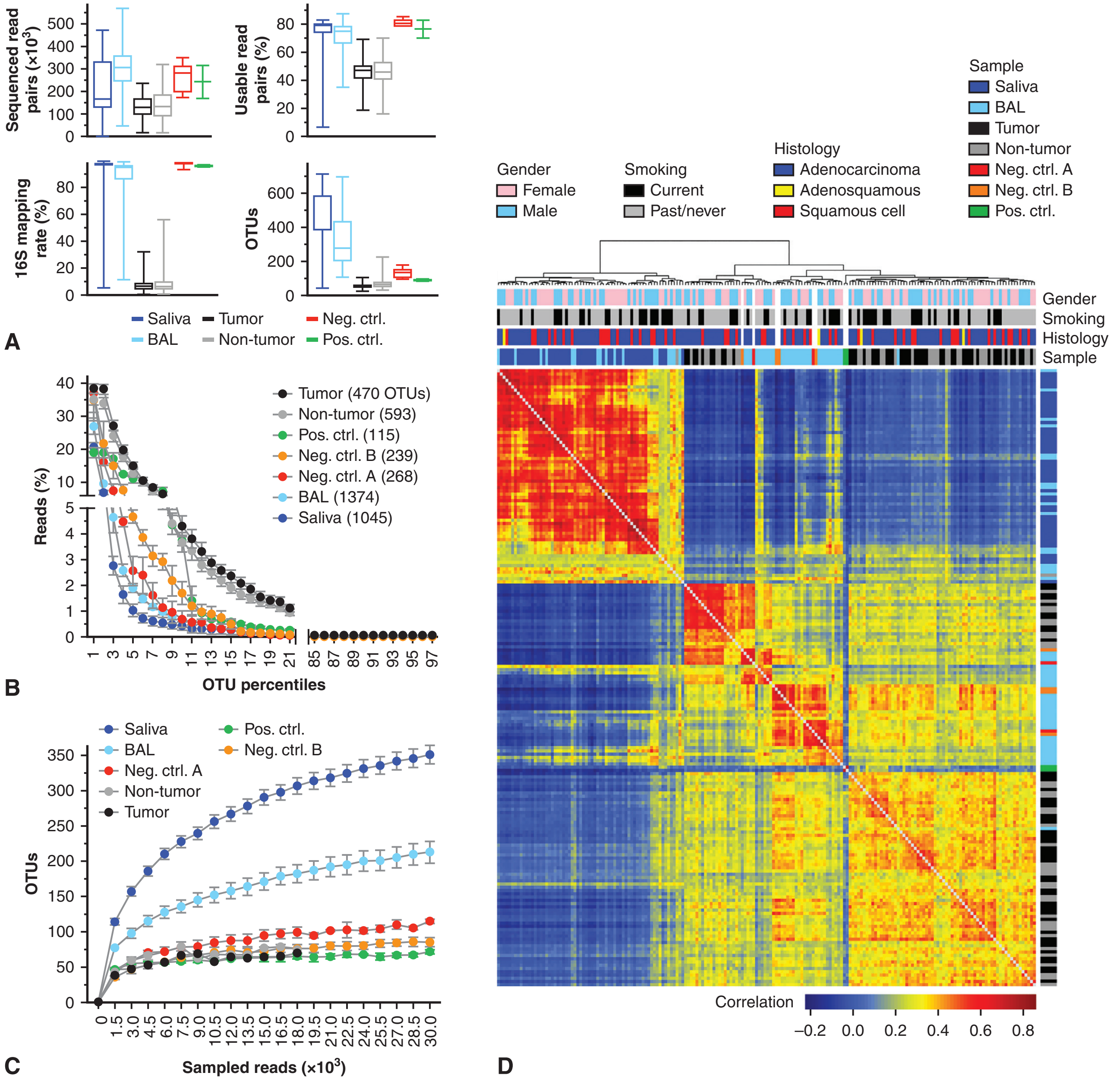


FIGURE 1. 16S DNA sequencing. Biospecimens of different types from 48 patients with stage I non–small cell lung cancer as well as various control samples were subjected to *16S* DNA sequencing for bacterial microbiome characterization. A, *Boxplots* show the number of raw sequencing reads, their fractions that were usable for mapping against the *16S* sequence database, the percentage of usable reads that were identified as bacterial *16S* sequences, and the number of unique bacterial OTUs that were identified among the *16S* sequences. The plots depict the median, 25th and 75th percentile, and minimum and maximum values for presurgery saliva (n = 48) and BAL (n = 48) fluids, and surgically resected lung tumor (n = 48) and adjacent non-tumor (n = 38) tissues of the 48 patients. Values are also shown for 2 positive (Pos.) and 6 negative controls (Neg. ctrl.). Negative control of types A (n = 2) and B (n = 4) refer to controls used for DNA extraction and *16S* PCR steps, respectively. B, Representation of OTUs among *16S* read sequence data. OTUs are binned by percentiles. Total unique OTUs for each type of sample are noted in the legend. C, Rarefaction curves for each of the different types of samples were generated by identifying unique OTUs in *16S* read data that was subsampled at varying depths. Means and standard deviations are plotted in panels B and C. D, Intersample Spearman correlations of OTU counts are shown with a heatmap annotated with patient characteristics. Complete linkages of Spearman distances (1 – Spearman coefficient) were used to generate the dendrograms for sample clustering. *OTUs*, Organizational taxonomic units; *BAL*, bronchoalveolar lavage.

database. As expected from their high host and low bacterial DNA load, the mapping rate was significantly lower (8%) for tissue samples (Figure 1, *A*), a majority of whose reads matched instead with host (human) mitochondrial *16S* gene sequence. With clustering of mapped reads at ≥97% sequence similarity into OTUs, an average of 473 (standard deviation = 158) and 322 (167) bacterial OTUs were respectively identified in the saliva and BAL fluids (Figure 1, *A*). In contrast, there were on average 56 (standard deviation = 17) and 68 (33) OTUs, respectively, in the tumor and non-tumor tissue samples. For all samples, most of the *16S* read sequences arose from a small fraction of the OTUs present in them (Figure 1, *B*). A total of 1414 OTUs were identifiable in the clinical samples, and 98%, 96%, 81%, and 29% of the OTUs could be assigned a taxonomy down to order, family, genus, and species levels, respectively. Bacterial *16S* sequences were observed in negative controls for both DNA extraction and *16S* PCR (Figures 1, *A*, and E1), indicating presence of exogenous, contaminating bacterial DNA in the experimental preparations and reagents. This contamination may be significant for the tissue samples, which likely had low bacterial biomass. However, the negative controls had 3- to 5-fold fewer detectable OTUs compared with saliva and BAL fluids (Figure 1, *A*), and rarefaction analysis by subsampling of mapped reads also showed that any contribution of contaminants to the bacterial profiles generated for these samples was minor (Figure 1, *C*). There is currently no suitable guideline or computational approach for handling of contaminating *16S* sequences,[28] and the negative controls' data were therefore ignored. Analysis of data from the duplicate positive controls, whose OTU count values had a Spearman correlation coefficient of 0.73, suggested that the DNA isolation and/or *16S* PCR methods that were used for this study had efficiency biases for and against various bacterial taxa (Figure E2). Existence of such bias is well-known in microbial sequencing studies.[29] A heatmap of Spearman correlations among all samples of this study for their OTU counts, with unsupervised clustering of the samples by their Spearman distances, is shown in Figure 1, *D*. Clustering together of samples to a moderate degree by their type (saliva, BAL fluid, or tissue), but not sex, smoking history, or cancer histology, was noticeable. Heatmaps and sample clustering for each of the 4 types of biospecimens are individually shown in Figure E3.

Spatial Relationship of the Different Sample Types Is Reflected in Their Microbiomes

Relative abundance of bacterial taxa at the class level among the microbiomes of the different sample types is shown in Figure 2, *A*. Saliva microbiomes were characterized by high abundance of bacteria of classes *Bacteroidia* and *Clostridia* and low abundance of *Alphaproteobacteria*, *Betaproteobacteria*, and *Gammaproteobacteria*. The latter 3 classes were more prevalent in tumor and non-tumor tissues, and their prevalence in BAL fluids was intermediate to saliva and tissues. Bacterial genera that were found to be differentially abundant in 2-group comparisons of the various sample types are listed in Table E2. Alpha diversity (species richness) of the microbiomes was measured using evenly subsampled (rarefied) data as OTU count, and Shannon and inverse Simpson indices. All 3 diversity measurements were significantly greater (1.5- to 2-fold) for saliva and BAL fluids compared with tumor or non-tumor tissues (Wilcoxon rank-sum or signed rank test $P > .05$; Figure 2, *B*). Diversities of saliva and BAL fluids, and of tumor and non-tumor tissues were similar. Beta diversity assessments using the Bray–Curtis and Jensen–Shannon indices showed that whereas tumor and non-tumor tissue microbiomes were similar, they both were significantly different from saliva microbiomes (Adonis test $P < .05$; Figure 2, *C*). BAL fluid microbiomes had a high dispersion, with some being similar to saliva microbiomes and some being more similar to tissue microbiomes. These observations show that the spatial relationship of the different sample types is reflected in their microbiomes, with microbiomes of BAL intermediate to those of saliva and lung tissue. Any association of treating institution with diversity measurements for any of the biospecimen types was not observed.

Oral Sourcing of Lower Respiratory Bacteria

In an assay for α-amylase, which is secreted in saliva but not respiratory fluids, we detected significant levels of enzyme activity in BAL fluid samples of some patients (Table 1), indicating the presence of oral secretion in the BAL fluids. This could happen physiologically because of microaspiration of oral contents, or iatrogenically during BAL procedure. The former scenario suggests that the oral cavity may be an important microbe seeding source for lower airways. Alpha and beta diversity comparison of saliva and BAL fluid microbiomes showed that indeed the 2 types of microbiomes were more similar in patients with high BAL α-amylase levels (Figure 2, *D* and *E*).

Microbiome Differences by Recurrence Status

Most recurrence following resection of stage I NSCLC tumors occurs within 2.5 to 3 years of surgery.[2] Among this study's patients, 18 each had recurrence within 32 months or did not have recurrence during follow-up of ≥32 months. Age, sex, race, smoking history, tumor histology, size, and stage, and treating institution of these 2 groups of patients were similar (Table 1). To compare the microbiomes of the 2 groups, differential abundance analysis was performed using OTU count data at the genus level (Table 2). Abundances of 2 bacterial genera were significantly different by ≥2-fold in the presurgery saliva

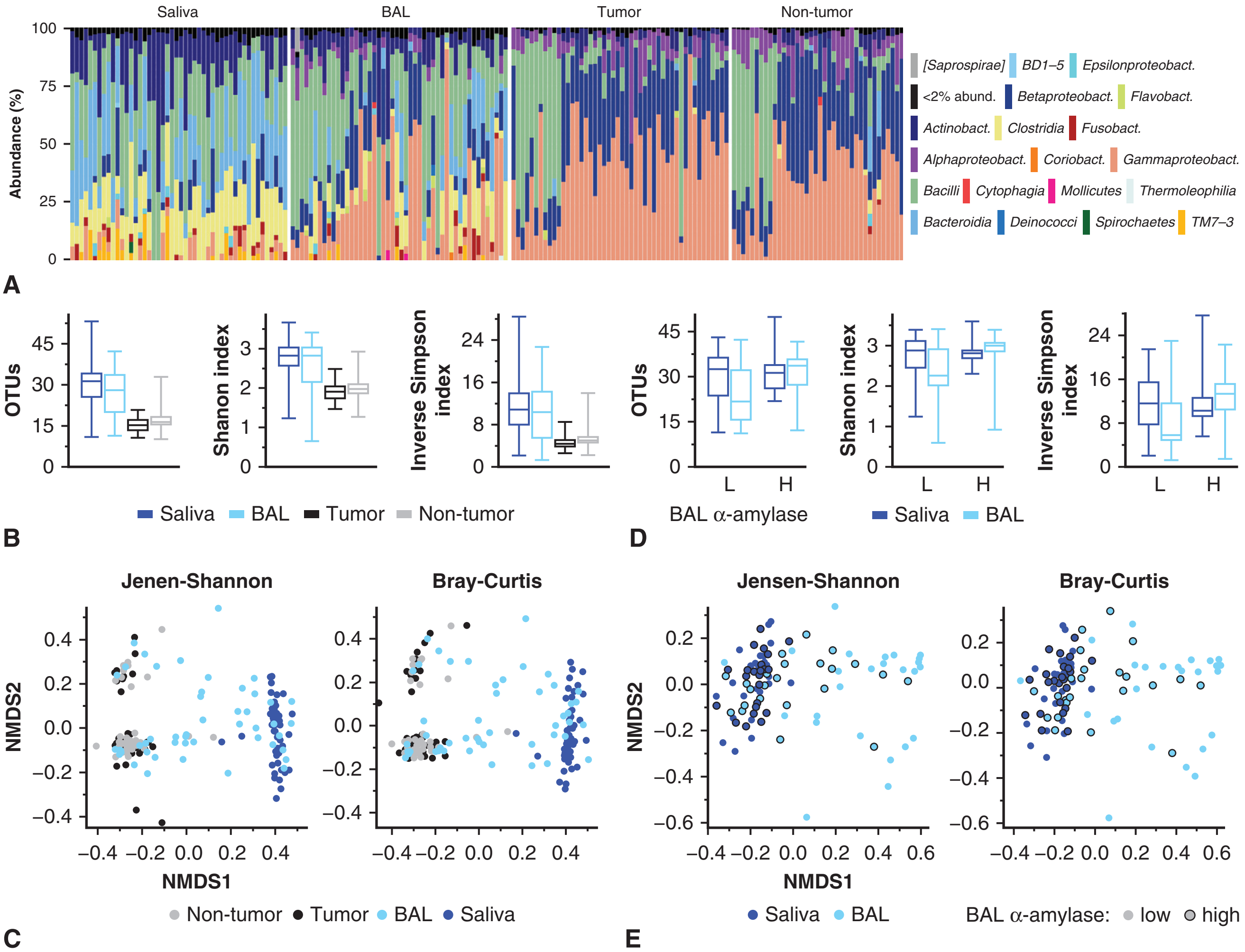


FIGURE 2. Bacterial microbiomes of the 4 types of biospecimens. A, The stacked *barplots* show the relative abundance of bacterial classes among the microbiomes of presurgery saliva (n = 48) and BAL (n = 48) fluids, and surgically resected lung tumor (n = 48) and adjacent non-tumor (n = 38) tissues of the 48 patients with stage I non–small cell lung cancer of this study. Samples are arranged by their patient identifiers. Bacterial classes with <2% abundance within a sample are grouped as one category. B, Alpha diversity characteristics of microbiomes of the 4 types of biospecimens are shown with *boxplots* of OTUs and Shannon and inverse Simpson indices. The plots depict the median, 25th and 75th percentile, and minimum and maximum values for the 36 to 48 samples of each type of biospecimen. C, Beta diversity characteristics of microbiomes of the 4 types of biospecimens are shown with NMDS plots of Jensen–Shannon and Bray–Curtis distance measurements among the 36 to 48 samples of each type of biospecimen. D and E, Alpha and beta diversity characteristics of saliva and BAL fluid samples of patients with low (n = 25) or high (n = 23) α-amylase activity in BAL. *Boxplots* depict the median, 25th and 75th percentile, and minimum and maximum values for alpha diversity data. OTU count data were subsampled to an even depth for the alpha and beta diversity measurements. *OTUs*, Organizational taxonomic units; *BAL*, bronchoalveolar lavage; *NMDS*, non-metric multidimensional scaling.

microbiomes with Wald test $P < .05$ (adjusted for multiple testing), with *Delftia* and *Bifidobacterium* genera respectively more and less prevalent in recurrence compared to no-recurrence patients. For tumor tissues, *Staphylococcus* had a significantly greater abundance in the recurrence group, whereas it was reduced for *Bacillus* and *Anaerobacillus*. No genus was differentially abundant in non-tumor tissue microbiomes. Nineteen genera were differentially abundant in case of presurgery BAL fluid microbiomes. They included the 5 differentially abundant genera of saliva and tumor tissue microbiomes. Abundances of *Sphingomonas*, *Psychromonas*, and *Serratia* genera were increased the most in the recurrence group, whereas abundances of *Cloacibacterium*, *Geobacillus*, and *Brevibacterium* were reduced the most (Table 2). A heatmap of relative abundances of the 19 genera in the BAL fluid microbiomes is shown in Figure 3, *A*.

Association of Recurrence With a Presurgery BAL Fluid Microbiome Signature

By using Dirichlet-multinomial mixture modeling[23] on the OTU count data at the genus-level for their 19

TABLE 2. Bacterial genera with significantly different abundances between recurrence and no-recurrence cases*

Genus	Phylum	Class	Log$_2$(FC)†	Adjusted *P*‡
Saliva (n = 17, 18)				
Delftia	*Proteobacteria*	*Betaproteobacteria*	3.1	2.61E-02
Bifidobacterium	*Actinobacteria*	*Actinobacteria*	–3.7	2.61E-02
Bronchoalveolar lavage fluid (n = 18, 18)				
Sphingomonas	*Proteobacteria*	*Alphaproteobacteria*	9.4	3.10E-05
Psychromonas	*Proteobacteria*	*Gammaproteobacteria*	9.3	5.84E-04
Serratia	*Proteobacteria*	*Gammaproteobacteria*	6.4	2.02E-02
Stenotrophomonas	*Proteobacteria*	*Gammaproteobacteria*	4.9	3.10E-05
Mycoplasma	*Tenericutes*	*Mollicutes*	4.3	5.56E-03
Delftia	*Proteobacteria*	*Betaproteobacteria*	4.1	5.84E-04
Lautropia	*Proteobacteria*	*Betaproteobacteria*	3.7	1.08E-02
Staphylococcus	*Firmicutes*	*Bacilli*	–3.0	1.34E-02
Microbacterium	*Actinobacteria*	*Actinobacteria*	–4.3	1.17E-02
Halomonas	*Proteobacteria*	*Gammaproteobacteria*	–5.0	2.97E-04
Agrobacterium	*Proteobacteria*	*Alphaproteobacteria*	–5.7	2.11E-02
Bifidobacterium	*Actinobacteria*	*Actinobacteria*	–6.7	3.25E-02
Anaerobacillus	*Firmicutes*	*Bacilli*	–7.0	3.78E-03
Anoxybacillus	*Firmicutes*	*Bacilli*	–7.5	1.70E-04
Thermicanus	*Firmicutes*	*Bacilli*	–7.8	1.70E-04
Bacillus	*Firmicutes*	*Bacilli*	–7.9	9.29E-06
Brevibacterium	*Actinobacteria*	*Actinobacteria*	–8.2	4.35E-03
Geobacillus	*Firmicutes*	*Bacilli*	–9.0	3.10E-05
Cloacibacterium	*Bacteroidetes*	*Flavobacteriia*	–28.3	2.53E-27
Tumor (n = 17, 17)				
Staphylococcus	*Firmicutes*	*Bacilli*	4.8	1.79E-03
Bacillus	*Firmicutes*	*Bacilli*	–6.1	4.99E-04
Anaerobacillus	*Firmicutes*	*Bacilli*	–6.4	1.79E-03

Log2(FC), Log$_2$-transformed fold-change. *In 2-group comparison of saliva, bronchoalveolar lavage fluid, or tumor samples with DESeq2 package in R. Only samples with total operational taxonomic unit count >500 were analyzed. Genera are arranged by decreasing fold-change values, and annotated with taxonomies at phylum and class levels. †Estimating the size of difference in abundance of the genus between the 2 groups (recurrence vs no recurrence). ‡*P* values in Wald test adjusted for multiple testing with the Benjamini–Hochberg method.

differentially abundant genera, the presurgery BAL fluid microbiomes could be fitted the best in a model with 3 Dirichlet components (bacterial community states). Weightages of each of the 19 genera in the components are listed in Table E3. One of the 3 components, referred henceforth as recurrence-associated BAL microbiome signature (RABMS), was present in BAL fluid microbiomes of 16 (89%) patients of the recurrence group and 2 (11%) of the no-recurrence group. Any association of RABMS with age, sex, race, smoking history, tumor histology, size, or stage, or BAL α-amylase level was not evident (Figure 3, *A*; $P > .05$ in standard *t* or Fisher exact tests). The 3 treating institutions were similarly represented among the RABMS+ and RABMS– groups of patients (Fisher exact test $P = .19$). In leave-one-out cross-validation of Dirichlet-multinomial mixture as a Bayesian classifier to identify recurrence or no-recurrence grouping of the BAL fluid microbiomes, accuracy was 89% and area under receiver operating characteristic curve was 0.77 (95% confidence interval [CI], 0.62-0.93; Figure 3, *B*). As expected, patients with RABMS + BAL fluid microbiomes had worse recurrence-free survival in Kaplan–Meier analysis compared with those with RABMS– microbiomes, with a hazard ratio of 14.5 (95% CI, 5.5-38.0; log-rank test $P < .001$; Figure 3, *C*).

Presurgery BAL Fluid Microbiome Signature Is Associated With Tumor Expression of Genes for Cell Proliferation, Immunity, and Signaling

To obtain mechanistic insights on the association of RABMS with cancer recurrence, we examined the gene expression in resected tumor specimens of RABMS+ and RABMS– patients (n = 14 each). Expression of 25 and 8 genes, respectively, was significantly up- and down-regulated by ≥2-fold in RABMS+ compared with RABMS– tumors with Wald test $P < .05$ (adjusted for multiple testing). These genes included those encoding enzymes such as arginine deiminase and steroid alpha-reductase, and immune-related proteins like and CXC-motif chemokine ligands (Table E4).

For a better understanding of the gene expression differences, gene set enrichment analyses using mSigDb Hallmark and Reactome gene set collections were performed. Significantly enriched expression with normalized

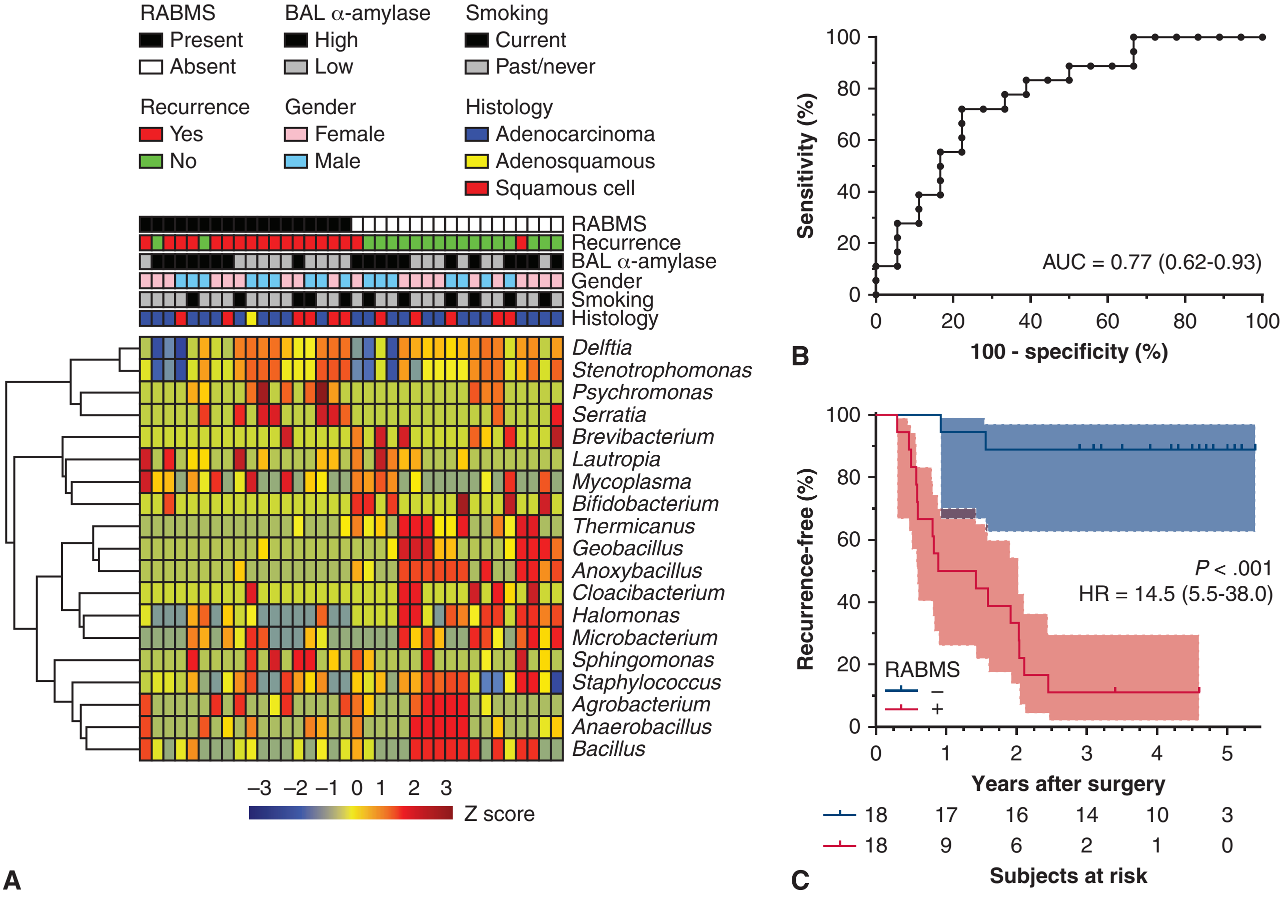


FIGURE 3. Recurrence-associated lower airway microbiome signature. A, Heatmap of relative abundance of the 19 bacterial genera with significant difference for abundance between presurgery BAL samples of patients with stage I non–small cell lung cancer with and without recurrence after surgical resection of tumors (n = 18 each). Abundance values are Z-scaled across patients and colored as per the scale shown beneath the heatmap. The heatmap is annotated with patient characteristics, including BAL α-amylase activity level and presence of an RABMS. The dendrogram on *left* indicates hierarchical clustering of the genera by Euclidean distance and complete linkage metrics. Log$_2$-transformed abundance values (fractions) relative to all bacterial genera identified in samples were used for the analysis. B, Receiver operating characteristic curve in leave-one-out cross-validation of the predictive value of RABMS for recurrence after tumor resection. All 36 subjects were used for the internal cross-validation. AUC and its 95% confidence interval are noted. C, Kaplan–Meier recurrence-free survival curves of patients with and without RABMS. *Shaded regions* indicate 95% confidence intervals. Subjects at risk at different time points, and log-rank test *P* value and HR and its 95% confidence interval are noted. *BAL*, Bronchoalveolar lavage; *RABMS*, recurrence-associated presurgery BAL microbiome signature; *AUC*, area under curve; *HR*, hazard ratio.

enrichment score ≥2 and false discovery rate <0.25 for 14 of the 50 Hallmark gene sets, including those for cell proliferation and epithelial mesenchymal transition, was noted in RABMS+ tumors, whereas no set was enriched in RABMS– tumors (Figure 4, *A*; Table E5). Four of the 14 sets enriched in RABMS+ tumors are related to cell cycle, and 2 are related to cellular ATP generation. Among the 1499 Reactome gene sets, enrichment of expression was seen for 84 and 14 sets, respectively, in RABMS+ and RABMS– tumors (Figure 4, *A*; Table E5). Thirty-seven (44%), 12 (14%), and 8 (10%) of the 84 sets enriched in RABMS+ tumors respectively belong to the categories of cell cycle, metabolism, and immune system (Figure 4, *B*). A majority (10; 71%) of the 14 Reactome sets enriched in RABMS– tumors belong to the signal transduction category, whereas 3 others are related to immune system. Of note, the Reactome PD1 signaling gene set is enriched in RABMS– tumors. Immune gene sets enriched in the RABMS+ group include those related to HIV infection and B-cell receptor signaling. We also used the tumor gene expression data to characterize the cellular composition of the tumor microenvironment through quantitative estimation of infiltrating fibroblasts and various types of immune cells. Abundances of these cell types in RABMS+ and RABMS– tumors were similar in analysis using a standard *t* test with correction for multiple testing ($P \geq .05$).

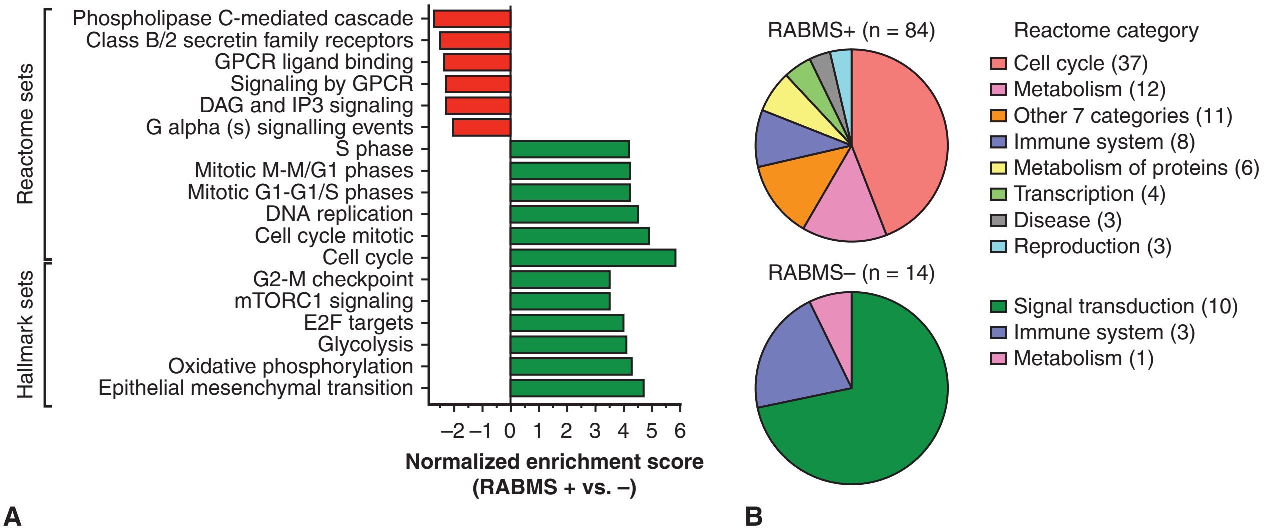


FIGURE 4. Tumor gene expression enrichments among patients with and without recurrence-associated lower airway microbiome signature. A, Tumor gene expression of patients with stage I non–small cell lung cancer with (n = 18) and without (n = 18) RABMS was compared to identify mSigDb Hallmark and C2:CP Reactome gene sets with significant enrichment for expression in either group of tumors compared with the other. Six each of the gene sets with greatest enrichments observed in the comparisons are shown. No Hallmark set was enriched in the RABMS– group. Absolute normalized enrichment score ≥2 and false discovery rate <0.25 in enrichment analyses with classic preranked Gene Set Enrichment Analysis method were used to deem significance. B, The *pie charts* depict the categories of the 84 and 14 Reactome gene sets whose expression is significantly enriched among RABMS+ and RABMS– patients, respectively. *RABMS*, Recurrence-associated presurgery BAL microbiome signature.

DISCUSSION

This study demonstrates, for the first time, an association between the composition of the lower airway microbiome and recurrence after resection of stage I NSCLC (Figure 5). The lower airway microbiome is the most proximate microbiome to the lung parenchyma. Several studies have described the enrichment of specific microbial populations in patients with lung cancer compared with controls.[4,30] A study by Tsay and colleagues[14] demonstrated that both *Veillonella* and *Streptococcus* species were increased in the lower airway microbiome of patients with lung cancer (n = 39) when compared with controls (n = 36). Interestingly, pathway analysis performed on gene expression data obtained by RNA sequencing of the tumors demonstrated a relationship between the airway microbiome and up-regulation of the phosphoinositide

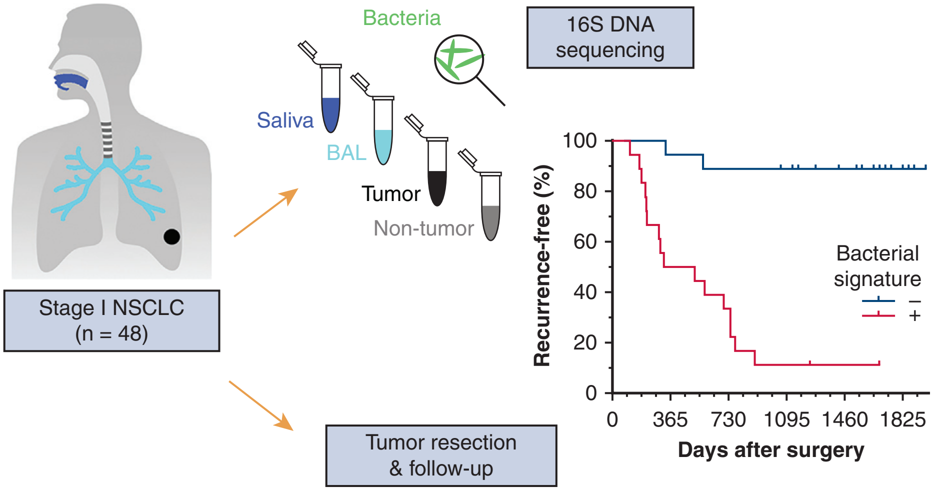


FIGURE 5. Examining the association of bacterial communities of the body on lung cancer recurrence. Presurgery oral (saliva), airway (bronchioalveolar lavage fluid), and surgically resected lung cancer tumor and adjacent normal tissue microbiomes of 48 patients with stage I non–small cell lung cancer were evaluated using *16S* ribosomal RNA gene sequencing. Presence of a bacterial signature in presurgery airway microbiome was associated with cancer recurrence. *NSCLC*, Non–small cell lung cancer; *BAL*, bronchoalveolar lavage.

3-kinases pathway. This is similar to changes in cell-cycle pathways associated with recurrence-associated BAL microbiome found in our study (Figure 4, *A* and *B*). Similarly, Peters and colleagues[31] demonstrated the impact of microbial diversity of non-tumor lung tissue on oncologic outcomes of lung cancer. Specifically, they found that taxa belonging to genus *Sphingomonas* were associated with disease-free survival, as was found in our study. Although these and our own study identify an association of microbiome with lung cancer, its causality remains to be identified. A causative role of the airway microbiome is supported by experimental data from Le Noci and colleagues[32] that demonstrate differences in tumor growth by manipulation of the microbiome. If such role can be confirmed, then engineering of the airway microbiome with antibiotics or specific bacteria will be a modality worth exploring for adjuvant therapy.

An interesting finding in our study is the closer correlation between the salivary and bronchoalveolar microbial compositions of patients with a greater salivary amylase activity level in BAL fluid (Figure 2, *D* and *E*). This is consistent with the hypothesis that most of the airway microbiome originates from seeding from the oral cavity. This may explain the observation by several investigators that the salivary microbiome is associated with the lung cancer state.[33] It is not a reach to assume that microaspiration may have an impact on the airway microbiome as well as the immune environment of the lung. In an analysis of BAL fluids from 49 patients, Segal and colleagues[34] demonstrate that approximately one half of the subjects have evidence of microaspiration leading to increased lung inflammation. It is conceivable that this may have long-term oncologic consequences, similar to the impact of microaspiration in interstitial lung disease and chronic obstructive pulmonary disease.[35,36] Microbiome-induced inflammation may be a link that ties these observations together; this is supported by the association of immune pathway changes with recurrence-associated BAL fluid microbiome patterns seen in the current study. Although the authors are not aware of specific studies that have examined the relationship of tumor recurrence with cytokine levels in the BAL fluid, several studies have described associations of BAL cytokine levels with the lung cancer disease state.[37-39] Whether the microbiome influences the inflammatory milieu or vice versa remains to be studied.

Although the results of our study are exciting, several limitations exist. The first is that the number of patients and samples included are modest. We did not have additional patients, or data from another study to validate our findings, especially the prognostic value of RABMS. Another limitation is that the samples were not collected specifically for microbiome studies. Maintenance of sterility during BAL procedure or surgery, sample handling, and antibiotic use in the days before sample collection are some of the factors that we have no information on. Although a greater concern with descriptive analyses, this limitation should have a lower impact on the comparison of the microbiome between patients with and without recurrence, as it is difficult to imagine a scenario in which patients with recurrence have systematically different contaminants than patients without. In recognition of this limitation, we did not examine the association of lung tissue microbiome (normal and tumor), as the abundance of the microbiome in these 2 types of specimens was low and is likely to be influenced by contamination issue to a greater degree. Also, the patients were treated at hospitals spread over 3 states. Although we did not discern any association of treating institution with patient recurrence or microbial profiles, it is possible that our findings are affected by the different microbial environments to which the patients were exposed. Although we did not observe differences between our study's groups for many characteristics that are associated with survival in lung cancer, such as histology, sex, and smoking history (Table 1), the association of BAL microbiome with recurrence could have arisen because of group differences for other factors such as quality of BAL procedure or surgery. Notwithstanding these limitations, we believe that our study offers novel insights into the composition of the lower airway microbiome associated with recurrence and potential mechanisms that may explain this association. These observations may inform future mechanistic studies to explore this association.

CONCLUSIONS

Presurgery composition of lower airway microbiome may be associated with recurrence of early NSCLC. This association may reflect an influence of the microbiome on tumor biology.

Webcast

You can watch a Webcast of this AATS meeting presentation by going to: https://aats.blob.core.windows.net/media/ITSOS19/ME%20-%20Friday/ITSOS2019_092719_Lower%20Airway%20Microbiome%20May%20Influence%20Recurrence%20After%20Resection%20Of%20Early%20Stage%20Non%20Small%20Cell%20Lung%20Cancer_Saikrishna%20Yendamuri.mp4.

ROSWELL PARK CANCER INSTITUTE

Lower Airway Bacterial Microbiome may Influence Recurrence after Resection of Early-stage Non-small Cell Lung Cancer

Sai Yendamuri MD, FACS
Professor and Chair,
Department of Thoracic Surgery

Saikrishna Yendamuri

Lower Airway Microbiome May Influence Recurrence After Resection of Early Stage Non-Small Cell Lung CancerOverall Survival Update from the PACIFIC Trial

Conflict of Interest Statement

Authors have nothing to disclose with regard to commercial support.

References

1. Goldstraw P, Crowley J, Chansky K, et al. The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM classification of malignant tumours. *J Thorac Oncol*. 2007;2:706-14.
2. van den Berg LL, Klinkenberg TJ, Groen HJ, Widder J. Patterns of recurrence and survival after surgery or stereotactic radiotherapy for early stage NSCLC. *J Thorac Oncol*. 2015;10:826-31.
3. D'Journo XB, Bittar F, Trousse D, et al. Molecular detection of microorganisms in distal airways of patients undergoing lung cancer surgery. *Ann Thorac Surg*. 2012;93:413-22.
4. Lee SH, Sung JY, Yong D, et al. Characterization of microbiome in bronchoalveolar lavage fluid of patients with lung cancer comparing with benign mass like lesions. *Lung Cancer*. 2016;102:89-95.
5. Yu G, Gail MH, Consonni D, et al. Characterizing human lung tissue microbiota and its relationship to epidemiological and clinical features. *Genome Biol*. 2016;17:163.
6. Gleeson K, Eggli DF, Maxwell SL. Quantitative aspiration during sleep in normal subjects. *Chest*. 1997;111:1266-72.
7. Pauly JL, Paszkiewicz G. Cigarette smoke, bacteria, mold, microbial toxins, and chronic lung inflammation. *J Oncol*. 2011;2011:819129.
8. Pauly JL, Waight JD, Paszkiewicz GM. Tobacco flakes on cigarette filters grow bacteria: a potential health risk to the smoker? *Tob Control*. 2008;17(suppl 1):i49-52.
9. Kim D, Hofstaedter CE, Zhao C, et al. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome*. 2017;5:52.
10. Huang YJ, Sethi S, Murphy T, Nariya S, Boushey HA, Lynch SV. Airway microbiome dynamics in exacerbations of chronic obstructive pulmonary disease. *J Clin Microbiol*. 2014;52:2813-23.
11. Mammen MJ, Sethi S. COPD and the microbiome. *Respirology*. 2016;21:590-9.
12. Huang YJ, LiPuma JJ. The microbiome in cystic fibrosis. *Clin Chest Med*. 2016;37:59-67.
13. Greathouse KL, White JR, Vargas AJ, et al. Interaction between the microbiome and TP53 in human lung cancer. *Genome Biol*. 2018;19:123.
14. Tsay JJ, Wu BG, Badri MH, et al. Airway microbiota is associated with upregulation of the PI3K pathway in lung cancer. *Am J Respir Crit Care Med*. 2018;198:1188-98.
15. Moskaluk CA. The LCBRN: a biospecimen resource for lung cancer biomarker and discovery science. Presented at: 104th Annual Meeting of the AACR; 2013; Washington, DC.
16. Hall M, Beiko RG. 16S rRNA gene analysis with QIIME2. *Methods Mol Biol*. 2018;1849:113-29.
17. Rideout JR, He Y, Navas-Molina JA, et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ*. 2014;2:e545.
18. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*. 2010;26:266-7.
19. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26:2460-1.
20. DeSantis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*. 2006;72:5069-72.
21. McMurdie PJ, Holmes S. Phyloseq: a bioconductor package for handling and analysis of high-throughput phylogenetic sequence data. *Pac Symp Biocomput*. 2012;235-46.
22. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.
23. Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One*. 2012;7:e30126.
24. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol Biol*. 2018;1711:243-59.
25. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife*. 2017;6:e26476.
26. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102:15545-50.
27. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;1:417-25.
28. Karstens L, Asquith M, Davin S, et al. Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments. *mSystems*. 2019;4:e00290-19.
29. Brooks JP, Edwards DJ, Harwich MD Jr, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol*. 2015;15:66.
30. Liu HX, Tao LL, Zhang J, et al. 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;142:769-78.
31. Peters BA, Hayes RB, Goparaju C, Reid C, Pass HI, Ahn J. The microbiome in lung cancer tissue and recurrence-free survival. *Cancer Epidemiol Biomarkers Prev*. 2019;28:731-40.
32. Le Noci V, Guglielmetti S, Arioli S, et al. Modulation of pulmonary microbiota by antibiotic or probiotic aerosol therapy: a strategy to promote immunosurveillance against lung metastases. *Cell Rep*. 2018;24:3528-38.
33. Hosgood HD III, Sapkota AR, Rothman N, et al. The potential role of lung microbiota in lung cancer attributed to household coal burning exposures. *Environ Mol Mutagen*. 2014;55:643-51.
34. Segal LN, Clemente JC, Tsay JC, et al. Enrichment of the lung microbiome with oral taxa is associated with lung inflammation of a Th17 phenotype. *Nat Microbiol*. 2016;1:16031.
35. Kim Y, Lee YJ, Cho YJ, et al. Association between pepsin in bronchoalveolar lavage fluid and prognosis of chronic fibrosing interstitial lung disease. *Tohoku J Exp Med*. 2018;246:147-53.
36. Biswas A, Mehta HJ, Folch EE. Chronic obstructive pulmonary disease and lung cancer: inter-relationships. *Curr Opin Pulm Med*. 2018;24:152-60.
37. Chyczewska E, Mroz RM, Kowal E. TNF-alpha, IL-1 and IL-6 concentration in bronchoalveolar lavage fluid (BALF) of non-small cell lung cancer (NSCLC). *Rocz Akad Med Bialymst*. 1997;42(suppl 1):123-35.
38. Erdogan B, Uzaslan E, Budak F, et al. The evaluation of soluble Fas and soluble Fas ligand levels of bronchoalveolar lavage fluid in lung cancer patients. *Tuberk Toraks*. 2005;53:127-31.
39. Schmid S, Le UT, Haager B, et al. Local Concentrations of CC-chemokine-ligand 18 correlate with tumor size in non-small cell lung cancer and are elevated in lymph node-positive disease. *Anticancer Res*. 2016;36:4667-71.

Key Words: 16S sequencing, biomarker, gene expression, lung cancer, lung microbiome, recurrence

APPENDIX E1. ADDITIONAL METHODS

Biospecimen Characteristics

Average (range; standard deviation [SD]) percentage cellularity, necrosis, and stromal content values of the tumor tissues were determined by the Lung Cancer Biospecimen Resource Network as 57 (25-80; 14), 5 (0-25; 7), and 44 (10-90; 21), respectively. For the samples procured by the repository, normal saline was used for collection of 20 to 40 mL of bronchoalveolar lavage (BAL) fluids, which were centrifuged at 1300*g* for 25 minutes, following which supernatants were collected and stored at –80°C or in liquid nitrogen. Salivette tubes (Sarstedt, Newton, NC) were used for collecting saliva, which were similarly stored. DNA and RNA were separately extracted from the same tissue samples using TissueLyser II, 5-mm stainless steel beads, and AllPrep DNA/RNA/Protein Mini kit (QIAGEN, Valencia, Calif), and nucleic acid concentrations of the preparations were quantified by spectrophotometry. The mean RNA integrity number value of the tumor RNA preparations, measured by Bioanalyzer assay (Agilent, Santa Clara, Calif), was 7.4 (range = 5.7-8.7; SD = 0.8).

Isolation of DNA From BAL and Saliva Samples

QIAamp UCP Pathogen Mini kit and Pathogen Lysis L tubes (QIAGEN) were used to extract DNA from BAL fluid (0.4 mL) and saliva (0.2 mL) samples. The protocol "Pretreatment of Pathogen DNA from 400 μL of Whole Blood (Mechanical Pre-lysis Protocol)" provided with the kit was used. The kit's ATL buffer with reagent Dx was added to samples to make their volume to 0.4 mL before mechanical lysis. The 2 types of samples were processed in separate batches, and each batch included a negative control for DNA extraction, an empty, sterile, nuclease-free 1.5-mL microcentrifuge tube. DNA was also isolated in each of the 2 batches from a positive control, ZymoBIOMICS Microbial Community Standard (product D6300; Zymo, Irvine, Calif). DNA concentrations were measured with TapeStation D1000 ScreenTape (Agilent). The average DNA yield from BAL fluid and saliva samples (n = 48 each) was 198 ng (SD = 636) and 88 ng (SD = 259), respectively, with DNA yield too low for measurement for 42% and 19% of the DNA preparations.

Measurement of α-Amylase Activity

A colorimetric assay kit (product K711-100; BioVision, Milpitas, Calif), with ethylidene-*4*-nitrophenyl-*α*-*D*-maltoheptaoside as substrate and a sensitivity of 0.2 mU (0.2 nmole nitrophenol generated from substrate per minute at 25°C and pH 7.2), was used to measure α-amylase activity levels of BAL fluid samples. Reactions were performed for 1 hour at 25°C in a volume of 120 μL with 20 μL of sample or amylase standards provided with the kit. Absorbance of the reaction mixes at 405 nm was measured on a Synergy H1m microplate reader (BioTek Instruments, Winooski, Vt). Activity levels of the 48 samples ranged from 0 to 471 mU/mL (mean = 52; SD = 113; median = 8; Table E1).

16S DNA Sequencing and Data Processing

The *16S* sequencing method suggested in the 16S Metagenomic Library preparation guide of Illumina (San Diego, Calif) was followed. An ~464 bp amplicon of bacterial 16S rRNA V3-V4 region was amplified with forward and reverse primers of sequences TCGTCGGCAGCGTC-ad-CCTAC GGGNGGCWGCAG and GTCTCGTGGGCTCGG-ad-GACTACHVGGGTATCTAATCC, respectively (ad = AG ATGTGTATAAGAGACAG). DNA (25 ng) was subjected to 35 cycles of PCR with annealing temperature of 55°C and KAPA HiFi HotStart DNA polymerase. For samples for which 25 ng of DNA was unavailable, the maximum volume of a sample's DNA preparation was used. Amplified DNA was purified with AMPure XP beads and was indexed with Nextera XT index kit in an 8-cycle PCR to generate sequencing libraries. Libraries were purified with AMPure XP beads and 12 pmoles of each library was sequenced on a MiSeq instrument (Illumina) with v3 sequencing reagents to obtained paired 300 bp reads. Between 46 and 48 libraries were combined with a control PhiX sequencing library (Illumina; at 10% molar ratio) for each sequencing run; sequencing data were later demultiplexed using Casava (version 1.8.2; Illumina). Raw sequencing data were deposited in the European Nucleotide Archive (study identification PRJEB29934). De-multiplexed paired-end sequencing data of all 190 libraries were co-processed for quality filtering followed by joining of paired reads using scripts provided with QIIME[15] software (version 1.9.1). To assign to each joined read sequence an operational taxonomic unit (OTU), QIIME's subsampled open-reference OTU-picking workflow was used with PyNAST sequence alignment and uclust aligned sequence clustering methods. A true value was used for the *enable_rev_strand_match* option for the alignment step, and sequence clustering was at 97% similarity. Sequences that were chimeric were removed with ChimeraSlayer. OTUs represented by <2 sequences or <0.001% of sequences of all samples were discarded. OTUs were assigned a taxonomy as per their representative *16S* gene sequence based on Greengenes bacterial *16S* rRNA gene database (version 13.5; 97% sequence identity cut-off). OTUs (n = 151) whose kingdom-level taxonomy was not bacteria, or which were of mitochondrion or chloroplast origin were removed. The resulting sequencing count data for 1422 OTUs and 190 samples was used for further analyses.

Alpha and Beta Diversity Analyses

Subsamplings (rarefactions) of OTU count data were performed with replacement, and to a depth equal to the minimum total OTU count among the data's samples for alpha and beta diversity measurements. Ten iterations of rarefactions were performed, and measurements obtained with each iteration of rarefied data were combined for further analyses. In case of distance measurements, *fuse* function in analogue package (version 0.17-3) was used for such combination; for other measurements, combination was by averaging. Unpaired and paired sample 2-group comparisons of alpha diversity measurements were with Wilcoxon rank-sum and signed rank tests, respectively. For beta diversity (distance) measurements, adonis permutational multivariate analysis of variance was used to compare groups; dispersion (variance) within groups of samples was quantified with *betadisper* function and compared by permutational analysis of variance.

Total RNA Sequencing and Data Processing

Sequencing libraries were prepared from 300 ng RNA of 39 Lung Cancer Biospecimen Resource Network tumor tissue samples using Illumina TruSeq Stranded Total RNA Library Prep Gold kit (with rRNA depletion) with 15 PCR cycles. All libraries were sequenced in one run on an Illumina NextSeq 500 instrument using reagents of NextSeq 500/550 High Output v2 kit (150 cycles) to obtain paired 76 bp reads. An average of 19.1 million sequencing read pairs (SD = 2.1) were obtained for each library. Raw sequencing data was filtered with Trimmomatic 0.35 to remove adapter and poor-quality sequences and mapped against the GRCh38 reference genome and transcriptome using HISAT2 (020516 release). The mean overall read mapping rate was 84.1% (SD = 6.3). Uniquely mapped reads and Ensembl gene information (release 81) were used to generate gene-level mapped read counts with Subread featureCounts 1.5.0-p1, with an average feature assignment rate of 68.6% (SD = 14.3). The final gene expression dataset for the 39 tumors was created by converting the count values to transcripts per million using total gene exon length values generated by featureCounts. Ensembl gene identifiers without a Human Genome Organization Gene Nomenclature Committee gene symbol were removed from the dataset, and transcripts per million values for identifiers with same symbol were summarized to a single value by addition. Raw sequencing data was deposited in the European Nucleotide Archive (study identification PRJEB29932).

APPENDIX E2. CONSIDERATIONS OF EXPERIMENT DESIGN FOR GENERATION OF BACTERIAL MICROBIOME DATA

To minimize environmental and cross-sample contamination, DNA of saliva and BAL fluids was extracted in separate batches by one person at a single facility using DNA-free reagents. Contamination during collection of specimens and during tissue DNA extraction were beyond our control. To amplify the *16S* gene by PCR for sequencing, a high number of cycles (35) was used because bacterial load was anticipated to be very low in the tissue samples. The same number of cycles was used for the other 2 types of samples to avoid bias in polymerase chain reaction (PCR) artifacts (sequence errors and chimeras). Primer sequences, chosen from a set of widely used *16S* primers, and annealing temperature for PCR were selected after testing (data not shown) to maximize the yield of desired product and minimize nonspecific amplification of host DNA, which constituted almost all of the DNA prepared from tissues. DNA from all specimens of a patient was subjected to PCR, preparation of sequencing library using PCR products, and sequencing of libraries in the same batch, with patients randomized across 4 batches and tasks for each batch handled by the same one person at a single facility. Each batch included a negative control for PCR (no template). Two each of negative control for DNA extraction (no specimen) and positive control (mock community of 8 bacteria and 2 fungi at 2%-12% composition) were also included to assess contamination and artifact generation during the *16S* sequencing work.

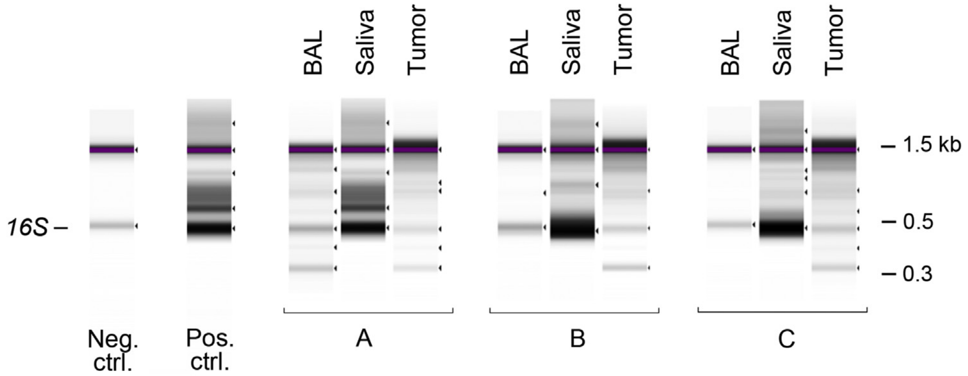


FIGURE E1. Example electrophoretograms of *16S* gene PCRs. Images from D1000 ScreenTape assay of PCRs are shown for a negative control (Neg. ctrl.; PCR without addition of DNA), a positive control (Pos. ctrl.; PCR of DNA from the mock microbial community standard), and for BAL fluid, saliva, and lung tumors of 3 patients (A-C). Molecular weights and the *16S* V3-V4 amplicon of ~0.46 kb are indicated. The band of ~1.5 kb is a marker DNA of the ScreenTape assay. *BAL*, Bronchoalveolar lavage.

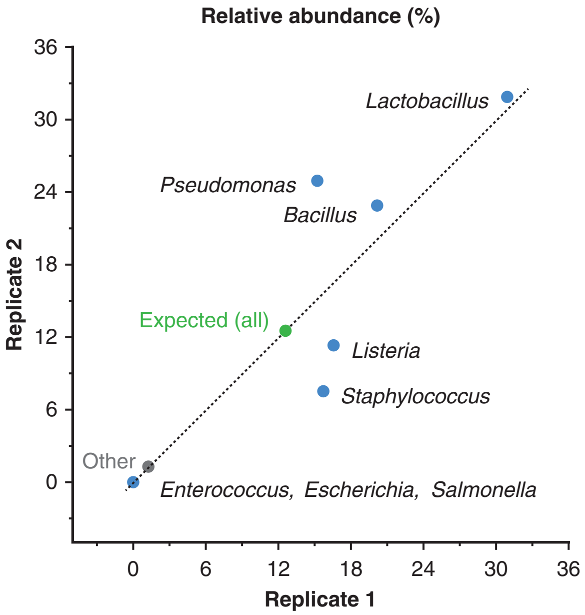


FIGURE E2. Genus-level compositions of the duplicate positive control samples. Relative abundance of operational taxonomic unit (OTU) values at genus level are shown for the duplicate positive control samples, which were aliquots of the ZymoBIOMICS microbial community standard (a mixture of 8 bacteria, each at 12.5% at genomic DNA level, and 2 fungi, each at 2%). The line of identity is plotted. A total of 115 OTUs were detected in the 2 samples.

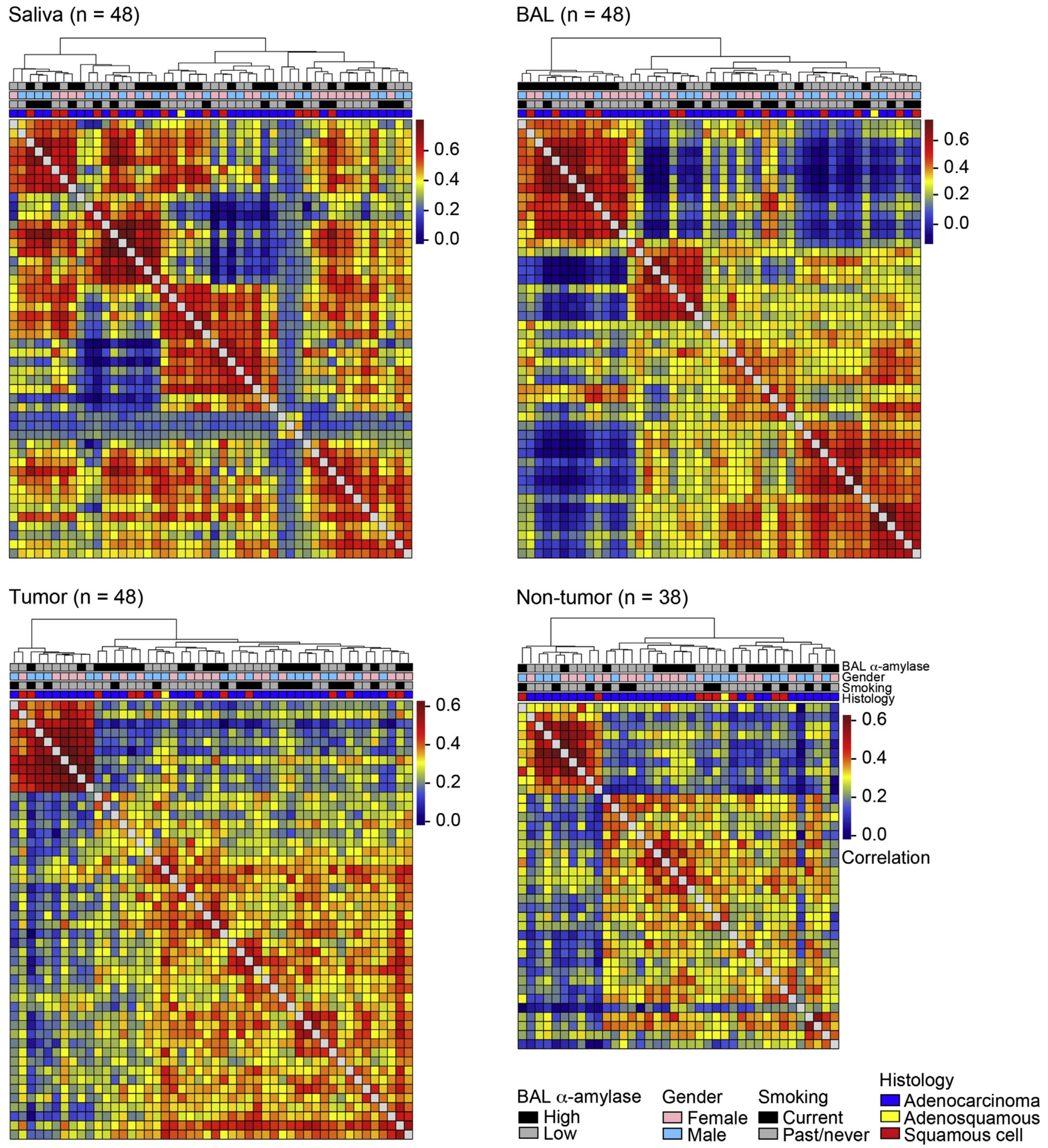


FIGURE E3. Intersample correlations of operational taxonomic unit (OTU) counts by type of biospecimens. Heatmaps are shown for Spearman correlation coefficient values for each of the 4 types of biospecimens. Complete linkages of Spearman distances (1 – Spearman coefficient) are used to generate the dendrograms for sample clustering. The heatmaps are annotated with patient characteristics, including α-amylase activity levels of BAL fluid samples. *BAL*, Bronchoalveolar lavage.

TABLE E1. Characteristics of individual patients of study

Patient	Age, y	Vitality	Sex	Race	Smoking history	Cancer history (organ)	Cancer histology	Tumor size,* cm	TNM T stage	Time to recurrence, d	Time to last follow-up, y	BAL α-amylase, mU/mL
S0012	51	A	M	W	C	Larynx	AC	4.4	T2a	2015	5.50	8.9
S0026	54	A	F	W	P	None	AC	5.0	T2a	NR	5.40	0.5
S0057	52	A	M	AA	C	None	AC	2.1	T1a	NR	1.75	0.0
S0094	67	A	M	W	C	None	AC	1.6	T2a	1168	5.50	2.0
S0100	69	A	F	W	N	Lymphoma	SCC	NA	T2a	NR	5.20	0.0
S0115	70	A	F	AA	P	Breast, lung	AC	1.5	T1a	571	3.00	57.8
S0144	64	A	M	W	C	None	SCC	3.3	T2a	NR	4.50	8.4
S0175	76	A	M	W	P	Prostate	AC	4.4	T2a	NR	3.50	2.5
S0179	59	D	F	W	P	Skin, vulva	AC	3.5	T2a	743	3.10	7.9
S0216	71	A	M	W	P	Prostate	AC	3.0	T2a	NR	2.25	1.5
S0217	83	A	F	W	P	Breast	AC	1.5	T1a	NR	3.90	6.4
S0224	58	D	F	AA	C	None	AC	1.8	T1a	294	3.40	4.0
S0233	66	A	F	W	C	Skin	AC	1.5	T1a	NR	3.20	20.2
S0237	67	D	F	W	C	None	AC	2.6	T2a	NR	0.60	0.5
V0001	68	A	M	W	C	None	SCC	4.3	T2a	1170	5.50	3.0
V0026	57	A	M	W	P	Prostate	AC	2.1	T1a	770	5.50	1.5
V0061	75	A	F	W	P	None	SCC	1.7	T1b	NR	5.10	5.4
V0076	49	A	M	W	P	None	AC	3.0	T1a	NR	5.00	0.0
V0090	76	D	M	W	P	Prostate	AC	1.4	T2a	1246	4.00	4.9
V0098	67	D	M	W	C	None	SCC	2.2	T2a	214	0.80	4.0
V0101	71	A	F	W	P	None	SCC	1.2	T1a	742	5.10	0.0
V0103	68	A	M	W	P	None	AS	2.3	T1b	581	4.60	4.0
V0111	78	D	F	W	P	None	AC	2.6	T2a	211	2.00	0.0
V0131	65	A	M	W	C	Bladder	AC	3.2	T2a	NR	4.70	206.9
V0132	71	A	F	W	P	Lung, thyroid	AC	1.1	T1a	NR	4.60	437.5
V0148	70	A	M	W	C	None	AC	1.6	T1a	1282	4.70	59.3
V0213	55	A	M	W	C	None	SCC	3.0	T2a	301	4.60	0.5
V0216	58	A	F	W	N	None	AC	3.0	T2a	1083	4.30	4.4
V0224	57	D	M	W	P	None	AC	4.6	T1b	217	2.20	0.0
V0236	68	A	M	W	C	None	SCC	1.1	T1a	NR	4.30	41.0
V0276	71	A	F	W	N	None	AC	3.2	T2a	337	4.00	11.9
V0287	56	A	F	W	P	None	AC	2.0	T1a	NR	3.10	15.8
W0022	75	D	F	W	C	None	SCC	2.2	T1b	170	1.10	97.3
W0023	71	D	F	W	C	None	SCC	3.0	T2a	1373	5.70	49.9
W0044	55	A	F	W	C	None	AC	2.8	T1b	NR	4.80	7.9
W0063	56	A	F	W	P	None	AC	2.2	T1a	1473	5.00	5.9
W0089	45	A	M	W	P	None	SCC	2.2	T1a	NR	4.60	50.4
W0094	56	D	M	W	C	Lymphoma	AC	2.0	T1a	184	2.10	20.7
W0107	67	D	F	W	N	Lymphoma	AC	2.0	T2a	896	3.70	417.3
W0111	65	D	M	W	P	None	AC	2.0	T1a	111	1.00	1.0
W0139	71	D	M	W	P	None	SCC	3.8	T2a	324	1.20	470.6
W0217	71	A	M	AA	P	Prostate	AC	NA	T1a	NR	4.20	253.3

(Continued)

TABLE E1. Continued

Patient	Age, y	Vitality	Sex	Race	Smoking history	Cancer history (organ)	Cancer histology	Tumor size,* cm	TNM T stage	Time to recurrence, d	Time to last follow-up, y	BAL α-amylase, mU/mL
W0283	65	A	M	W	P	None	AC	1.9	T1a	NR	3.40	38.5
W0299	66	A	F	W	C	Breast	AC	2.9	T1b	NR	2.90	18.8
W0309	67	A	F	W	P	None	AC	2.0	T1b	701	2.50	12.3
W0314	80	D	F	W	P	None	SCC	1.0	T1a	519	1.80	63.2
W0317	67	A	F	W	P	None	AC	1.7	T1a	NR	3.10	44.0
W0320	77	A	F	W	P	None	AC	3.9	T2a	1265	3.50	24.2

TNM, Tumor, node, and metastasis staging system; *BAL*, bronchoalveolar lavage fluid; *A*, alive; *M*, male; *W*, white; *C*, current tobacco smoker; *AC*, adenocarcinoma; *F*, female; *P*, past tobacco smoker; *NR*, no known recurrence at last follow-up; *AA*, African-American; *N*, never smoked tobacco; *SCC*, squamous cell carcinoma; *NA*, data unavailable; *D*, deceased; *AS*, adenosquamous carcinoma. *In computed tomography or positron emission tomography imaging.

TABLE E2. Bacterial genera with significantly different abundances between types of biospecimens[*]

Greengenes ID	Phylum	Class	Order	Family	Genus	Log$_2$(FC)[†]	*P* value	Adjusted *P*[‡]
Saliva vs BAL (n = 45 pairs)								
543942	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Moraxellaceae*	*Acinetobacter*	−15.5	1.22E-128	7.95E-127
1088265	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Propionibacteriaceae*	*Propionibacterium*	−12.4	2.54E-75	8.24E-74
525199	*Proteobacteria*	*Betaproteobacteria*	*Burkholderiales*	*Comamonadaceae*	*Delftia*	−10.7	2.52E-69	5.46E-68
967275	*Proteobacteria*	*Gammaproteobacteria*	*Xanthomonadales*	*Xanthomonadaceae*	*Stenotrophomonas*	−10.1	1.33E-61	2.17E-60
963779	*Proteobacteria*	*Alphaproteobacteria*	*Rhizobiales*	*Brucellaceae*	*Ochrobactrum*	−11.5	9.14E-59	1.19E-57
646549	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonas*	−9.2	1.64E-48	1.77E-47
1934300	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Bacillus*	−12.2	2.18E-39	2.02E-38
1013670	*Proteobacteria*	*Gammaproteobacteria*	*Oceanospirillales*	*Halomonadaceae*	*Halomonas*	−13.8	7.46E-33	6.06E-32
866280	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Micrococcaceae*	*Rothia*	4.2	1.22E-25	8.81E-25
226338	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Enterococcaceae*	*Enterococcus*	−17.6	4.52E-25	2.94E-24
4451251	*Actinobacteria*	*Coriobacteriia*	*Coriobacteriales*	*Coriobacteriaceae*	*Atopobium*	5.9	6.10E-22	3.61E-21
757622	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Veillonella*	3.7	7.09E-20	3.84E-19
642525	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Selenomonas*	3.8	1.19E-16	5.96E-16
530164	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Porphyromonadaceae*	*Porphyromonas*	4.3	2.36E-15	1.09E-14
749837	*Firmicutes*	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Oribacterium*	4.8	4.20E-15	1.82E-14
938948	*Fusobacteria*	*Fusobacteriia*	*Fusobacteriales*	*Fusobacteriaceae*	*Fusobacterium*	3.6	8.22E-15	3.34E-14
949789	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Carnobacteriaceae*	*Granulicatella*	3.6	3.71E-14	1.42E-13
696234	*Proteobacteria*	*Alphaproteobacteria*	*Rhizobiales*	*Rhizobiaceae*	*Agrobacterium*	−22.3	4.21E-14	1.52E-13
437105	*Proteobacteria*	*Betaproteobacteria*	*Burkholderiales*	*Oxalobacteraceae*	*Ralstonia*	−22.1	7.69E-14	2.63E-13
530206	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Prevotellaceae*	*Prevotella*	3.3	4.33E-13	1.41E-12
27737	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Geobacillus*	−19.3	6.15E-11	1.90E-10
439457	*Firmicutes*	*Bacilli*	*Bacillales*	*[Thermicanaceae]*	*Thermicanus*	−19.3	7.04E-11	2.08E-10
1616059	*Proteobacteria*	*Epsilonproteobacteria*	*Campylobacterales*	*Campylobacteraceae*	*Campylobacter*	3.1	6.84E-10	1.93E-09
965129	*Proteobacteria*	*Alphaproteobacteria*	*Sphingomonadales*	*Sphingomonadaceae*	*Sphingomonas*	−17.6	2.79E-09	7.56E-09
1051517	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Anoxybacillus*	−17.5	3.47E-09	9.01E-09
693231	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Anaerobacillus*	−17.4	3.84E-09	9.59E-09
	Firmicutes	*Bacilli*	*Lactobacillales*	*Enterococcaceae*	*Vagococcus*	3.0	4.57E-09	1.10E-08
403701	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Dialister*	3.4	8.03E-09	1.87E-08
1089121	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Actinomycetaceae*	*Actinomyces*	2.7	2.44E-08	5.39E-08
511378	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Megasphaera*	3.3	2.49E-08	5.39E-08
586968	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	4.2	3.30E-08	6.92E-08
344593	*Proteobacteria*	*Betaproteobacteria*	*Neisseriales*	*Neisseriaceae*	*Neisseria*	2.6	7.29E-07	1.48E-06
4154872	*Bacteroidetes*	*Flavobacteriia*	*Flavobacteriales*	*[Weeksellaceae]*	*Cloacibacterium*	−14.3	1.28E-06	2.53E-06
1076316	*Firmicutes*	*Bacilli*	*Bacillales*	*Staphylococcaceae*	*Staphylococcus*	−2.2	2.56E-06	4.90E-06
4417749	*Bacteroidetes*	*Flavobacteriia*	*Flavobacteriales*	*Flavobacteriaceae*	*Capnocytophaga*	2.5	9.11E-06	1.69E-05
851938	*Firmicutes*	*Erysipelotrichi*	*Erysipelotrichales*	*Erysipelotrichaceae*	*Bulleidia*	2.7	1.07E-05	1.93E-05
1040713	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Corynebacteriaceae*	*Corynebacterium*	−2.7	1.35E-05	2.37E-05
968675	*Proteobacteria*	*Gammaproteobacteria*	*Pasteurellales*	*Pasteurellaceae*	*Haemophilus*	2.0	1.89E-05	3.23E-05
714766	*Firmicutes*	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Moryella*	2.5	2.27E-05	3.78E-05
1042850	*Fusobacteria*	*Fusobacteriia*	*Fusobacteriales*	*Leptotrichiaceae*	*Leptotrichia*	2.0	2.33E-04	3.79E-04

(Continued)

TABLE E2. Continued

Greengenes ID	Phylum	Class	Order	Family	Genus	Log_2(FC)†	*P* value	Adjusted *P*‡
71146	*Spirochaetes*	*Spirochaetes*	*Spirochaetales*	*Spirochaetaceae*	*Treponema*	2.6	3.00E-04	4.76E-04
122517	*Firmicutes*	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Catonella*	3.0	4.47E-04	6.92E-04
254476	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Schwartzia*	4.0	5.81E-04	8.78E-04
4443201	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Porphyromonadaceae*	*Tannerella*	2.6	6.16E-04	9.10E-04
851704	*Firmicutes*	*Clostridia*	*Clostridiales*	*[Tissierellaceae]*	*Parvimonas*	1.9	1.95E-03	2.82E-03
4352772	*Proteobacteria*	*Gammaproteobacteria*	*Pasteurellales*	*Pasteurellaceae*	*Aggregatibacter*	1.8	6.95E-03	9.81E-03
971907	*Proteobacteria*	*Gammaproteobacteria*	*Pasteurellales*	*Pasteurellaceae*	*Actinobacillus*	1.8	1.09E-02	1.51E-02
1084417	*Proteobacteria*	*Betaproteobacteria*	*Burkholderiales*	*Burkholderiaceae*	*Lautropia*	2.0	3.00E-02	3.98E-02
Saliva vs non-tumor lung (n = 33 pairs)								
525199	*Proteobacteria*	*Betaproteobacteria*	*Burkholderiales*	*Comamonadaceae*	*Delftia*	−12.1	2.24E-115	1.21E-113
646549	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonas*	−11.3	1.02E-87	2.75E-86
967275	*Proteobacteria*	*Gammaproteobacteria*	*Xanthomonadales*	*Xanthomonadaceae*	*Stenotrophomonas*	−10.7	1.93E-76	3.48E-75
757622	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Veillonella*	9.0	1.38E-67	1.86E-66
530206	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Prevotellaceae*	*Prevotella*	8.6	2.25E-54	2.43E-53
963779	*Proteobacteria*	*Alphaproteobacteria*	*Rhizobiales*	*Brucellaceae*	*Ochrobactrum*	−10.4	2.51E-52	2.26E-51
866280	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Micrococcaceae*	*Rothia*	8.6	6.46E-52	4.99E-51
1082294	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Streptococcaceae*	*Streptococcus*	7.5	1.73E-48	1.16E-47
1089121	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Actinomycetaceae*	*Actinomyces*	7.5	1.10E-31	6.57E-31
949789	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Carnobacteriaceae*	*Granulicatella*	7.5	2.65E-28	1.43E-27
938948	*Fusobacteria*	*Fusobacteriia*	*Fusobacteriales*	*Fusobacteriaceae*	*Fusobacterium*	7.3	9.87E-28	4.85E-27
642525	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Selenomonas*	6.8	9.20E-23	4.14E-22
530164	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Porphyromonadaceae*	*Porphyromonas*	7.0	8.03E-22	3.34E-21
968675	*Proteobacteria*	*Gammaproteobacteria*	*Pasteurellales*	*Pasteurellaceae*	*Haemophilus*	6.1	1.20E-21	4.63E-21
1616059	*Proteobacteria*	*Epsilonproteobacteria*	*Campylobacterales*	*Campylobacteraceae*	*Campylobacter*	6.4	2.67E-20	9.63E-20
4451251	*Actinobacteria*	*Coriobacteriia*	*Coriobacteriales*	*Coriobacteriaceae*	*Atopobium*	6.4	3.36E-18	1.13E-17
511378	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Megasphaera*	6.1	3.76E-17	1.19E-16
749837	*Firmicutes*	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Oribacterium*	6.1	7.90E-14	2.37E-13
344593	*Proteobacteria*	*Betaproteobacteria*	*Neisseriales*	*Neisseriaceae*	*Neisseria*	5.5	2.28E-13	6.47E-13
4310396	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*[Paraprevotellaceae]*	*[Prevotella]*	6.0	7.23E-13	1.86E-12
1088265	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Propionibacteriaceae*	*Propionibacterium*	−8.4	7.16E-13	1.86E-12
1042850	*Fusobacteria*	*Fusobacteriia*	*Fusobacteriales*	*Leptotrichiaceae*	*Leptotrichia*	5.0	2.21E-11	5.42E-11
851938	*Firmicutes*	*Erysipelotrichi*	*Erysipelotrichales*	*Erysipelotrichaceae*	*Bulleidia*	5.1	1.17E-08	2.74E-08
586968	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	5.4	3.34E-08	7.52E-08
4417749	*Bacteroidetes*	*Flavobacteriia*	*Flavobacteriales*	*Flavobacteriaceae*	*Capnocytophaga*	5.3	6.05E-08	1.31E-07
696234	*Proteobacteria*	*Alphaproteobacteria*	*Rhizobiales*	*Rhizobiaceae*	*Agrobacterium*	−16.8	1.84E-07	3.83E-07
714766	*Firmicutes*	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Moryella*	4.4	1.09E-06	2.18E-06
403701	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Dialister*	3.7	1.13E-03	2.18E-03
1076316	*Firmicutes*	*Bacilli*	*Bacillales*	*Staphylococcaceae*	*Staphylococcus*	2.4	1.48E-03	2.75E-03
543942	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Moraxellaceae*	*Acinetobacter*	−8.6	7.64E-03	1.37E-02
1040713	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Corynebacteriaceae*	*Corynebacterium*	1.8	1.84E-02	3.21E-02
693231	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Anaerobacillus*	−7.2	2.55E-02	4.30E-02

(Continued)

TABLE E2. Continued

Greengenes ID	Phylum	Class	Order	Family	Genus	Log_2(FC)†	*P* value	Adjusted *P*‡
851704	*Firmicutes*	*Clostridia*	*Clostridiales*	*[Tissierellaceae]*	*Parvimonas*	3.3	2.71E-02	4.44E-02
Saliva vs tumor (n = 43 pairs)								
525199	*Proteobacteria*	*Betaproteobacteria*	*Burkholderiales*	*Comamonadaceae*	*Delftia*	−11.6	2.63E-139	1.52E-137
967275	*Proteobacteria*	*Gammaproteobacteria*	*Xanthomonadales*	*Xanthomonadaceae*	*Stenotrophomonas*	−11.0	6.82E-127	1.98E-125
646549	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonas*	−11.1	4.58E-102	8.85E-101
757622	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Veillonella*	9.5	2.35E-94	3.40E-93
530206	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Prevotellaceae*	*Prevotella*	9.5	3.44E-87	4.00E-86
1082294	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Streptococcaceae*	*Streptococcus*	8.7	1.93E-84	1.86E-83
866280	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Micrococcaceae*	*Rothia*	9.8	5.86E-68	4.86E-67
1089121	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Actinomycetaceae*	*Actinomyces*	8.6	1.55E-55	1.13E-54
938948	*Fusobacteria*	*Fusobacteriia*	*Fusobacteriales*	*Fusobacteriaceae*	*Fusobacterium*	8.3	3.08E-41	1.98E-40
963779	*Proteobacteria*	*Alphaproteobacteria*	*Rhizobiales*	*Brucellaceae*	*Ochrobactrum*	−8.9	9.30E-41	5.40E-40
1616059	*Proteobacteria*	*Epsilonproteobacteria*	*Campylobacterales*	*Campylobacteraceae*	*Campylobacter*	7.9	1.33E-36	7.01E-36
949789	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Carnobacteriaceae*	*Granulicatella*	8.2	6.88E-36	3.33E-35
642525	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Selenomonas*	7.3	1.98E-35	8.82E-35
530164	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Porphyromonadaceae*	*Porphyromonas*	7.6	9.77E-31	4.05E-30
968675	*Proteobacteria*	*Gammaproteobacteria*	*Pasteurellales*	*Pasteurellaceae*	*Haemophilus*	7.7	6.10E-30	2.36E-29
4451251	*Actinobacteria*	*Coriobacteriia*	*Coriobacteriales*	*Coriobacteriaceae*	*Atopobium*	6.9	5.76E-28	2.09E-27
1088265	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Propionibacteriaceae*	*Propionibacterium*	−9.0	2.09E-23	7.14E-23
344593	*Proteobacteria*	*Betaproteobacteria*	*Neisseriales*	*Neisseriaceae*	*Neisseria*	6.0	2.26E-21	7.29E-21
511378	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Megasphaera*	7.3	7.40E-20	2.26E-19
4310396	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*[Paraprevotellaceae]*	*[Prevotella]*	6.5	1.32E-19	3.82E-19
1042850	*Fusobacteria*	*Fusobacteriia*	*Fusobacteriales*	*Leptotrichiaceae*	*Leptotrichia*	6.0	2.20E-17	6.07E-17
749837	*Firmicutes*	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Oribacterium*	6.4	1.35E-16	3.56E-16
586968	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	6.8	9.96E-14	2.51E-13
851938	*Firmicutes*	*Erysipelotrichi*	*Erysipelotrichales*	*Erysipelotrichaceae*	*Bulleidia*	5.9	1.00E-11	2.43E-11
965129	*Proteobacteria*	*Alphaproteobacteria*	*Sphingomonadales*	*Sphingomonadaceae*	*Sphingomonas*	−18.5	6.72E-10	1.56E-09
4417749	*Bacteroidetes*	*Flavobacteriia*	*Flavobacteriales*	*Flavobacteriaceae*	*Capnocytophaga*	5.4	9.81E-10	2.19E-09
714766	*Firmicutes*	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Moryella*	4.4	3.70E-09	7.96E-09
1076316	*Firmicutes*	*Bacilli*	*Bacillales*	*Staphylococcaceae*	*Staphylococcus*	3.8	4.46E-09	9.23E-09
1108350	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Microbacteriaceae*	*Microbacterium*	−14.1	5.43E-09	1.09E-08
1934300	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Bacillus*	−6.1	9.74E-06	1.88E-05
403701	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Dialister*	3.6	2.18E-04	4.08E-04
1013670	*Proteobacteria*	*Gammaproteobacteria*	*Oceanospirillales*	*Halomonadaceae*	*Halomonas*	−5.9	5.06E-03	9.17E-03
851704	*Firmicutes*	*Clostridia*	*Clostridiales*	*[Tissierellaceae]*	*Parvimonas*	3.7	6.75E-03	1.19E-02
	Firmicutes	*Bacilli*	*Lactobacillales*	*Enterococcaceae*	*Vagococcus*	2.3	1.05E-02	1.80E-02
4443201	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Porphyromonadaceae*	*Tannerella*	3.3	1.21E-02	2.01E-02
437105	*Proteobacteria*	*Betaproteobacteria*	*Burkholderiales*	*Oxalobacteraceae*	*Ralstonia*	−7.1	1.79E-02	2.88E-02
543942	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Moraxellaceae*	*Acinetobacter*	−6.7	2.52E-02	3.94E-02

(Continued)

TABLE E2. Continued

Greengenes ID	Phylum	Class	Order	Family	Genus	Log$_2$(FC)†	*P* value	Adjusted *P*‡
BAL vs non-tumor lung (n = 36 pairs)								
1082294	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Streptococcaceae*	*Streptococcus*	4.9	7.09E-27	3.97E-25
580117	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	*Mycoplasmataceae*	*Mycoplasma*	25.7	3.46E-16	9.68E-15
757622	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Veillonella*	3.8	2.06E-15	3.84E-14
1076316	*Firmicutes*	*Bacilli*	*Bacillales*	*Staphylococcaceae*	*Staphylococcus*	5.8	4.58E-13	6.41E-12
1040713	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Corynebacteriaceae*	*Corynebacterium*	5.0	2.52E-10	2.83E-09
530206	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Prevotellaceae*	*Prevotella*	3.1	4.20E-09	3.92E-08
866280	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Micrococcaceae*	*Rothia*	2.6	9.49E-08	7.59E-07
1051517	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Anoxybacillus*	15.9	4.25E-07	2.98E-06
439457	*Firmicutes*	*Bacilli*	*Bacillales*	*[Thermicanaceae]*	*Thermicanus*	15.8	4.98E-07	3.10E-06
27737	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Geobacillus*	15.8	5.54E-07	3.10E-06
4310396	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*[Paraprevotellaceae]*	*[Prevotella]*	3.7	3.91E-06	1.99E-05
1088265	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Propionibacteriaceae*	*Propionibacterium*	3.4	4.83E-06	2.25E-05
971907	*Proteobacteria*	*Gammaproteobacteria*	*Pasteurellales*	*Pasteurellaceae*	*Actinobacillus*	13.5	1.83E-05	7.89E-05
968675	*Proteobacteria*	*Gammaproteobacteria*	*Pasteurellales*	*Pasteurellaceae*	*Haemophilus*	2.5	2.75E-05	1.03E-04
4352772	*Proteobacteria*	*Gammaproteobacteria*	*Pasteurellales*	*Pasteurellaceae*	*Aggregatibacter*	13.2	2.73E-05	1.03E-04
	Proteobacteria	*Gammaproteobacteria*	*Alteromonadales*	*Psychromonadaceae*	*Psychromonas*	12.8	4.66E-05	1.63E-04
646549	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonas*	−2.3	6.20E-05	1.83E-04
4154872	*Bacteroidetes*	*Flavobacteriia*	*Flavobacteriales*	*[Weeksellaceae]*	*Cloacibacterium*	12.6	6.18E-05	1.83E-04
1089121	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Actinomycetaceae*	*Actinomyces*	2.6	5.95E-05	1.83E-04
965129	*Proteobacteria*	*Alphaproteobacteria*	*Sphingomonadales*	*Sphingomonadaceae*	*Sphingomonas*	−7.2	1.48E-04	4.16E-04
949789	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Carnobacteriaceae*	*Granulicatella*	2.5	3.79E-04	1.00E-03
511378	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Megasphaera*	4.4	3.93E-04	1.00E-03
4451251	*Actinobacteria*	*Coriobacteriia*	*Coriobacteriales*	*Coriobacteriaceae*	*Atopobium*	9.1	2.62E-03	6.38E-03
749837	*Firmicutes*	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Oribacterium*	4.9	2.95E-03	6.88E-03
525199	*Proteobacteria*	*Betaproteobacteria*	*Burkholderiales*	*Comamonadaceae*	*Delftia*	−1.5	9.46E-03	2.12E-02
1013670	*Proteobacteria*	*Gammaproteobacteria*	*Oceanospirillales*	*Halomonadaceae*	*Halomonas*	−2.2	1.72E-02	3.70E-02
BAL vs tumor (n = 45 pairs)								
1082294	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Streptococcaceae*	*Streptococcus*	5.6	7.49E-39	3.97E-37
757622	*Firmicutes*	*Clostridia*	*Clostridiales*	*Veillonellaceae*	*Veillonella*	4.4	1.97E-23	2.80E-22
580117	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	*Mycoplasmataceae*	*Mycoplasma*	29.3	2.11E-23	2.80E-22
27737	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Geobacillus*	29.4	1.89E-23	2.80E-22
530206	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Prevotellaceae*	*Prevotella*	4.2	1.19E-18	1.27E-17
971907	*Proteobacteria*	*Gammaproteobacteria*	*Pasteurellales*	*Pasteurellaceae*	*Actinobacillus*	24.0	3.60E-16	3.18E-15
866280	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Micrococcaceae*	*Rothia*	4.2	1.47E-15	1.11E-14
968675	*Proteobacteria*	*Gammaproteobacteria*	*Pasteurellales*	*Pasteurellaceae*	*Haemophilus*	5.1	2.18E-15	1.44E-14
1076316	*Firmicutes*	*Bacilli*	*Bacillales*	*Staphylococcaceae*	*Staphylococcus*	5.8	1.42E-14	7.53E-14
1040713	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Corynebacteriaceae*	*Corynebacterium*	6.2	1.33E-14	7.53E-14
226338	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Enterococcaceae*	*Enterococcus*	16.1	1.69E-14	8.13E-14
1051517	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Anoxybacillus*	19.7	2.26E-11	1.00E-10
344593	*Proteobacteria*	*Betaproteobacteria*	*Neisseriales*	*Neisseriaceae*	*Neisseria*	4.9	1.69E-10	6.88E-10

(Continued)

TABLE E2. Continued

Greengenes ID	Phylum	Class	Order	Family	Genus	Log_2(FC)†	*P* value	Adjusted *P*‡
439457	*Firmicutes*	*Bacilli*	*Bacillales*	*[Thermicanaceae]*	*Thermicanus*	17.0	6.99E-09	2.64E-08
4310396	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*[Paraprevotellaceae]*	*[Prevotella]*	4.0	3.55E-08	1.26E-07
1089121	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Actinomycetaceae*	*Actinomyces*	3.7	8.69E-08	2.88E-07
543942	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Moraxellaceae*	*Acinetobacter*	3.4	1.11E-06	3.47E-06
4352772	*Proteobacteria*	*Gammaproteobacteria*	*Pasteurellales*	*Pasteurellaceae*	*Aggregatibacter*	14.1	1.55E-06	4.55E-06
1088265	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Propionibacteriaceae*	*Propionibacterium*	2.4	3.41E-05	9.52E-05
4451251	*Actinobacteria*	*Coriobacteriia*	*Coriobacteriales*	*Coriobacteriaceae*	*Atopobium*	10.5	3.63E-04	9.62E-04
949789	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Carnobacteriaceae*	*Granulicatella*	2.8	4.46E-04	1.13E-03
686789	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Carnobacteriaceae*	*Carnobacterium*	9.3	1.62E-03	3.90E-03
646549	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonas*	−1.6	1.83E-03	4.22E-03
4154872	*Bacteroidetes*	*Flavobacteriia*	*Flavobacteriales*	*[Weeksellaceae]*	*Cloacibacterium*	8.7	3.21E-03	7.09E-03
437105	*Proteobacteria*	*Betaproteobacteria*	*Burkholderiales*	*Oxalobacteraceae*	*Ralstonia*	−2.9	4.30E-03	9.11E-03
696234	*Proteobacteria*	*Alphaproteobacteria*	*Rhizobiales*	*Rhizobiaceae*	*Agrobacterium*	3.8	4.48E-03	9.13E-03
530164	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Porphyromonadaceae*	*Porphyromonas*	1.8	7.33E-03	1.44E-02
967275	*Proteobacteria*	*Gammaproteobacteria*	*Xanthomonadales*	*Xanthomonadaceae*	*Stenotrophomonas*	−1.3	1.20E-02	2.28E-02
938948	*Fusobacteria*	*Fusobacteriia*	*Fusobacteriales*	*Fusobacteriaceae*	*Fusobacterium*	2.3	1.93E-02	3.41E-02
1616059	*Proteobacteria*	*Epsilonproteobacteria*	*Campylobacterales*	*Campylobacteraceae*	*Campylobacter*	2.2	1.88E-02	3.41E-02

FC, Fold-change; *BAL*, bronchoalveolar lavage fluid. *In paired 2-group comparison of OTU count data agglomerated at genus level for calculation of Wald test *P* and effect size values with DESeq2 package in R. Only samples with total OTU count >500 were analyzed. Genera are arranged by increasing adjusted *P* values and identified by Greengenes database identifiers and taxonomies at different levels. Database identifiers are unavailable for some genera. No genus was different in tumor vs non-tumor lung comparison. †Log2-transformed fold-change, estimating the size of difference in abundance of the genus between the 2 groups (group 1 compared with 2). Absolute $\log_2$(FC) >1 and adjusted $P < .05$ were required to deem significance. ‡Raw *P* values adjusted for multiple testing with the Benjamini–Hochberg method.

TABLE E3. Taxonomic contributions to the 3 Dirichlet components of the model to fit cancer recurrence with bronchoalveolar lavage fluid microbiome*

Genus	Phylum	Class	m1	m2	m3	Mean
Staphylococcus	*Firmicutes*	*Bacilli*	4.30E-01	2.61E-01	1.65E-01	3.45E-01
Halomonas	*Proteobacteria*	*Gammaproteobacteria*	4.91E-01	2.69E-01	5.98E-02	3.09E-01
Mycoplasma	*Tenericutes*	*Mollicutes*	1.40E-02	2.16E-01	6.65E-02	7.95E-02
Lautropia	*Proteobacteria*	*Betaproteobacteria*	8.09E-03	3.37E-02	1.28E-02	3.26E-02
Anaerobacillus	*Firmicutes*	*Bacilli*	4.66E-03	2.40E-02	1.67E-01	2.93E-02
Bifidobacterium	*Actinobacteria*	*Actinobacteria*	9.28E-03	1.37E-02	5.79E-03	2.30E-02
Bacillus	*Firmicutes*	*Bacilli*	4.80E-03	4.15E-02	3.57E-03	2.30E-02
Delftia	*Proteobacteria*	*Betaproteobacteria*	5.17E-03	4.44E-03	3.80E-01	2.00E-02
Geobacillus	*Firmicutes*	*Bacilli*	2.17E-03	1.85E-02	3.83E-02	1.87E-02
Brevibacterium	*Actinobacteria*	*Actinobacteria*	3.62E-03	7.08E-03	6.47E-02	1.78E-02
Psychromonas	*Proteobacteria*	*Gammaproteobacteria*	1.71E-03	2.06E-02	1.62E-02	1.72E-02
Sphingomonas	*Proteobacteria*	*Alphaproteobacteria*	3.25E-03	1.51E-02	5.24E-03	1.46E-02
Stenotrophomonas	*Proteobacteria*	*Gammaproteobacteria*	4.10E-03	1.48E-02	1.91E-03	1.36E-02
Serratia	*Proteobacteria*	*Gammaproteobacteria*	1.27E-03	1.73E-02	9.46E-03	1.32E-02
Cloacibacterium	*Bacteroidetes*	*Flavobacteriia*	6.50E-03	4.20E-03	1.42E-04	1.10E-02
Thermicanus	*Firmicutes*	*Bacilli*	6.52E-03	1.94E-04	1.42E-04	8.96E-03
Agrobacterium	*Proteobacteria*	*Alphaproteobacteria*	1.95E-03	1.45E-02	1.61E-04	8.37E-03
Microbacterium	*Actinobacteria*	*Actinobacteria*	2.21E-03	7.25E-03	1.85E-03	7.60E-03
Anoxybacillus	*Firmicutes*	*Bacilli*	5.88E-05	1.69E-02	1.89E-03	7.33E-03

*Weightage or contribution (in fraction) of each bacterial genus to each of the 3 Dirichlet components (m1-3) and their means are listed. Genera are annotated with taxonomies at the phylum and class levels, and arranged by decreasing mean values.

TABLE E4. Genes with significant difference for expression in tumors of patients with and without recurrence-associated bronchoalveaolar lavage fluid microbiome signature (RABMS)[*]

Gene	Description	Log$_2$(FC)[†]	*P*	Adjusted *P* value[‡]
PADI3	peptidyl arginine deiminase 3	8.3	4.94E-06	9.95E-03
TCN1	transcobalamin 1	4.9	7.91E-06	9.95E-03
KLK6	kallikrein related peptidase 6	4.1	2.33E-05	2.13E-02
SLCO4A1-AS1	SLCO4A1 antisense RNA 1	4.0	3.10E-06	9.95E-03
CYP4F11	cytochrome P450 family 4 subfamily F member 11	3.9	4.05E-06	9.95E-03
CYP4F3	cytochrome P450 family 4 subfamily F member 3	3.7	4.58E-06	9.95E-03
TRPM8	transient receptor potential cation channel subfamily M member 8	3.6	1.04E-05	1.17E-02
SLC7A11	solute carrier family 7 member 11	3.4	1.96E-08	3.94E-04
LRRC66	leucine rich repeat containing 66	3.3	7.42E-06	9.95E-03
GRIN2A	glutamate ionotropic receptor NMDA type subunit 2A	3.2	6.75E-06	9.95E-03
IGHE	immunoglobulin heavy constant epsilon	3.1	2.91E-05	2.44E-02
LINC01589	long intergenic non-protein coding RNA 1589	3.0	3.06E-07	2.05E-03
FLNC	filamin C	2.7	3.60E-06	9.95E-03
LINC01116	long intergenic non-protein coding RNA 1116	2.5	6.18E-05	4.24E-02
KCND2	potassium voltage-gated channel subfamily D member 2	2.4	2.63E-05	2.30E-02
CXCL1	C-X-C motif chemokine ligand 1	2.4	7.61E-06	9.95E-03
APCDD1L	APC down-regulated 1 like	2.2	6.17E-05	4.24E-02
SP6	Sp6 transcription factor	2.1	4.87E-05	3.63E-02
COL22A1	collagen type XXII alpha 1 chain	2.1	3.69E-06	9.95E-03
CMBL	carboxymethylenebutenolidase homolog	2.0	6.91E-05	4.49E-02
CXCL8	C-X-C motif chemokine ligand 8	2.0	3.80E-05	3.06E-02
GPR27	G protein-coupled receptor 27	1.9	7.37E-05	4.50E-02
ATP8B3	ATPase phospholipid transporting 8B3	1.9	5.93E-06	9.95E-03
LYPD1	LY6/PLAUR domain containing 1	1.9	1.33E-05	1.41E-02
PLAU	plasminogen activator, urokinase	1.8	1.14E-07	1.15E-03
SYDE2	synapse defective Rho GTPase homolog 2	–1.2	4.35E-05	3.37E-02
NPR1	natriuretic peptide receptor 1	–1.4	8.44E-05	5.00E-02
SHE	Src homology 2 domain containing E	–2.0	8.53E-06	1.01E-02
COLGALT2	collagen beta(1-O)galactosyltransferase 2	–2.2	6.31E-05	4.24E-02
RASL10B	RAS like family 10 member B	–2.5	1.92E-05	1.84E-02
CYP11A1	cytochrome P450 family 11 subfamily A member 1	–2.8	7.38E-05	4.50E-02
PCSK2	proprotein convertase subtilisin/kexin type 2	–3.6	1.41E-05	1.42E-02
SRD5A2	steroid 5 alpha-reductase 2	–3.7	7.75E-06	9.95E-03

FC, Fold-change. [*] In 2-group comparison of 14 each of lung tumors of patients with (+) and without (–) RABMS for calculation of Wald test *P* and effect size values with DESeq2 package in R. Genes are arranged by decreasing FC values, and identified by Human Genome Organization Gene Nomenclature Committee symbols and descriptions. [†] Log2-transformed fold-change, estimating the size of difference in gene expression between the 2 groups (RABMS+ compared with RABMS–). Absolute log$_2$(FC) > 1 and adjusted $P < .05$ were required to deem significance. [‡] Raw *P* values adjusted for multiple testing with the Benjamini–Hochberg method.

TABLE E5. mSigDb collection gene sets with significant enrichment of expression in tumors of patients with and without recurrence-associated bronchoalveaolar lavage fluid microbiome signature (RABMS)*

Gene set	ES†	NES	*P* value	FDR‡
Hallmark gene sets				
Enriched in RABMS+ compared to RABMS- tumors				
1. Epithelial mesenchymal transition	0.3	4.7	.00E+00	0.00E+00
2. Oxidative phosphorylation	0.3	4.3	.00E+00	0.00E+00
3. Glycolysis	0.3	4.1	.00E+00	0.00E+00
4. E2F targets	0.2	4.0	.00E+00	0.00E+00
5. MTORC1 signaling	0.2	3.5	.00E+00	0.00E+00
6. G2M checkpoint	0.2	3.5	.00E+00	0.00E+00
7. MYC targets V1	0.2	3.5	.00E+00	0.00E+00
8. TNFA signaling via NFKB	0.2	2.8	.00E+00	1.61E-04
9. Hypoxia	0.2	2.4	2.04E-03	1.15E-03
10. KRAS signaling up	0.1	2.4	.00E+00	1.45E-03
11. Reactive oxygen species pathway	0.3	2.3	2.00E-03	2.61E-03
12. Unfolded protein response	0.2	2.1	3.80E-03	6.69E-03
13. Adipogenesis	0.1	2.0	5.99E-03	1.09E-02
14. Apical junction	0.1	2.0	7.69E-03	1.15E-02
C2:CP Reactome gene sets				
Enriched in RABMS+ compared to RABMS– tumors				
1. Cell cycle	0.3	5.8	.00E+00	0.00E+00
2. Cell cycle mitotic	0.2	4.9	.00E+00	0.00E+00
3. DNA replication	0.3	4.5	.00E+00	0.00E+00
4. Mitotic G1 G1 S phases	0.3	4.2	.00E+00	0.00E+00
5. Mitotic M M G1 phases	0.3	4.2	.00E+00	0.00E+00
6. S phase	0.3	4.2	.00E+00	0.00E+00
7. Cell cycle checkpoints	0.3	4.2	.00E+00	0.00E+00
8. G1 S transition	0.3	4.1	.00E+00	0.00E+00
9. Chromosome maintenance	0.3	4.1	.00E+00	0.00E+00
10. Respiratory electron transport ATP synthesis by chemiosmotic coupling and heat production by uncoupling proteins	0.4	4.1	.00E+00	0.00E+00
11. RNA POL I promoter opening	0.4	4.1	.00E+00	0.00E+00
12. Amyloids	0.4	4.1	.00E+00	0.00E+00
13. Synthesis of DNA	0.4	4.0	.00E+00	0.00E+00
14. Deposition of new CENPA containing nucleosomes at the centromere	0.4	3.9	.00E+00	0.00E+00
15. Meiotic recombination	0.4	3.9	.00E+00	0.00E+00
16. Respiratory electron transport	0.4	3.8	.00E+00	0.00E+00
17. Regulation of mitotic cell cycle	0.4	3.8	.00E+00	0.00E+00
18. APC C CDC20 mediated degradation of mitotic proteins	0.4	3.8	.00E+00	0.00E+00
19. Telomere Maintenance	0.4	3.7	.00E+00	0.00E+00
20. Meiosis	0.3	3.7	.00E+00	0.00E+00
21. M G1 transition	0.3	3.7	.00E+00	0.00E+00
22. TCA cycle and respiratory electron transport	0.3	3.7	.00E+00	0.00E+00
23. Collagen formation	0.4	3.7	.00E+00	0.00E+00
24. RNA POL I transcription	0.3	3.6	.00E+00	0.00E+00
25. Metabolism of proteins	0.2	3.6	.00E+00	0.00E+00
26. Autodegradation of CDH1 by CDH1 APC C	0.4	3.6	.00E+00	0.00E+00
27. Class I MHC mediated antigen processing presentation	0.2	3.6	.00E+00	0.00E+00
28. APC C CDH1 mediated degradation of CDC20 and other APC C CDH1 targeted proteins in late mitosis early G1	0.4	3.6	.00E+00	0.00E+00
29. Packaging of telomere ends	0.4	3.5	.00E+00	0.00E+00
30. SCFSKP2 mediated degradation OF P27 P21	0.4	3.5	.00E+00	0.00E+00
31. VIF mediated degradation of APOBEC3G	0.4	3.5	.00E+00	0.00E+00
32. CDT1 association with the CDC6 ORC origin complex	0.4	3.5	.00E+00	0.00E+00
33. Extracellular matrix organization	0.3	3.4	.00E+00	0.00E+00

(Continued)

TABLE E5. Continued

Gene set	ES†	NES	*P* value	FDR‡
34. SCF beta TRCP mediated degradation of EMI1	0.4	3.4	.00E+00	0.00E+00
35. Cyclin E associated events during G1 S transition	0.4	3.4	.00E+00	0.00E+00
36. ORC1 removal from chromatin	0.4	3.4	.00E+00	0.00E+00
37. RNA POL I RNA POL III and mitochondrial transcription	0.3	3.3	.00E+00	0.00E+00
38. P53 independent G1 S DNA damage checkpoint	0.4	3.3	.00E+00	0.00E+00
39. CDK mediated phosphorylation and removal of CDC6	0.4	3.3	.00E+00	0.00E+00
40. Regulation of ornithine decarboxylase ODC	0.4	3.2	.00E+00	0.00E+00
41. Meiotic synapsis	0.3	3.2	.00E+00	0.00E+00
42. Signaling by WNT	0.4	3.2	.00E+00	0.00E+00
43. Assembly of the pre replicative complex	0.3	3.2	.00E+00	0.00E+00
44. ER phagosome pathway	0.4	3.2	.00E+00	0.00E+00
45. Regulation of apoptosis	0.4	3.2	.00E+00	0.00E+00
46. Transcription	0.2	3.2	.00E+00	0.00E+00
47. Glucuronidation	0.7	3.1	.00E+00	0.00E+00
48. Antigen processing ubiquitination proteasome degradation	0.2	3.1	.00E+00	0.00E+00
49. Cross presentation of soluble exogenous antigens endosomes	0.4	3.1	.00E+00	0.00E+00
50. Activation of NF KAPPAB in B cells	0.3	3.0	.00E+00	0.00E+00
51. Destabilization of MRNA by AUF1 HNRNP D0	0.4	3.0	.00E+00	0.00E+00
52. Antigen processing cross presentation	0.3	3.0	.00E+00	0.00E+00
53. P53 dependent G1 DNA damage response	0.3	3.0	.00E+00	0.00E+00
54. Autodegradation of the E3 UBIQUITIN ligase COP1	0.4	3.0	.00E+00	0.00E+00
55. HIV infection	0.2	2.9	.00E+00	2.36E-05
56. Host interactions of HIV factors	0.2	2.8	.00E+00	1.13E-04
57. Apoptosis	0.2	2.8	.00E+00	1.79E-04
58. Downstream signaling events of B cell receptor BCR	0.2	2.7	.00E+00	1.75E-04
59. Membrane trafficking	0.2	2.6	.00E+00	3.63E-04
60. Post translational protein modification	0.2	2.6	.00E+00	6.44E-04
61. G2 M checkpoints	0.3	2.5	.00E+00	7.16E-04
62. E2F mediated regulation of DNA replication	0.4	2.5	.00E+00	7.89E-04
63. Mitotic prometaphase	0.2	2.5	.00E+00	8.57E-04
64. Glucose metabolism	0.3	2.5	.00E+00	9.62E-04
65. DNA strand elongation	0.4	2.4	.00E+00	1.24E-03
66. Mitochondrial protein import	0.3	2.4	.00E+00	1.55E-03
67. Metabolism of carbohydrates	0.1	2.3	.00E+00	2.20E-03
68. Synthesis and interconversion of nucleotide DI and triphosphates	0.5	2.3	.00E+00	2.42E-03
69. Asparagine N linked glycosylation	0.2	2.3	.00E+00	2.91E-03
70. Phase II conjugation	0.2	2.3	.00E+00	3.45E-03
71. Cyclin A B1 associated events during G2 M transition	0.5	2.2	2.04E-03	4.34E-03
72. G0 and early G1	0.4	2.2	4.23E-03	4.55E-03
73. Activation of ATR in response to replication stress	0.3	2.2	1.96E-03	4.57E-03
74. Glycolysis	0.4	2.2	.00E+00	6.78E-03
75. APC CDC20 mediated degradation of NEK2A	0.4	2.2	.00E+00	6.85E-03
76. O linked glycosylation of mucins	0.3	2.2	2.03E-03	7.43E-03
77. Activation of the pre replicative complex	0.3	2.1	.00E+00	8.12E-03
78. Metabolism of amino acids and derivatives	0.1	2.1	.00E+00	9.24E-03
79. Antigen presentation folding assembly and peptide loading of class I MHC	0.4	2.1	7.97E-03	9.50E-03
80. Processing of capped intron containing pre mRNA	0.2	2.1	2.03E-03	1.01E-02
81. metabolism of nucleotides	0.2	2.1	1.96E-03	1.08E-02
82. GAP junction trafficking	0.4	2.1	1.95E-03	1.22E-02
83. Prefoldin mediated transfer of substrate to CCT TRIC	0.3	2.1	2.00E-03	1.25E-02
84. DNA repair	0.2	2.1	2.06E-03	1.30E-02
Enriched in RABMS– compared with RABMS+ tumors				
1. Phospholipase C mediated cascade	0.3	2.7	.00E+00	5.09E-03
2. Class B 2 secretin family receptors	0.3	2.5	.00E+00	1.44E-02
3. GPCR LIGAND BINDING	0.1	2.3	2.17E-03	3.42E-02

(Continued)

TABLE E5. Continued

Gene set	ES†	NES	*P* value	FDR‡
4. Signaling by GPCR	0.1	2.3	2.09E-03	3.35E-02
5. DAG and IP3 signaling	0.4	2.3	1.96E-03	2.89E-02
6. Phosphorylation of CD3 and TCR ZETA chains	0.5	2.1	3.94E-03	7.77E-02
7. Downstream signaling of activated FGFR	0.2	2.1	1.98E-03	6.72E-02
8. GPCR downstream signaling	0.1	2.1	5.91E-03	6.58E-02
9. PD1 signaling	0.4	2.1	1.93E-03	7.25E-02
10. FGFR ligand binding and activation	0.4	2.1	1.90E-03	6.72E-02
11. Regulation of insulin secretion by glucagon like PEPTIDE1	0.3	2.0	5.86E-03	6.50E-02
12. Signaling by FGFR mutants	0.3	2.0	.00E+00	6.38E-02
13. G alpha S signalling events	0.2	2.0	3.91E-03	5.95E-02
14. Generation of second messenger molecules	0.3	2.0	2.00E-03	6.29E-02

ES, Enrichment score; *NES*, normalized enrichment score; *FDR*, false discovery rate. *In classic pre-ranked gene set enrichment analysis of 14 each of lung tumors of patients with (+) and without (–) RABMS. Sets are arranged by decreasing NES. No Hallmark set was enriched in RABMS– compared to RABMS+ tumors. †Enrichment score. NES is generated from ES values by the mean division method; absolute NES >2 was required to deem significance. ‡False discovery rate (q-value); FDR < 0.25 was required to deem significance.