


Variation in oral microbiome is associated with future risk of lung cancer among never-smokers

H Dean Hosgood ¹, Qiuyin Cai,² Xing Hua,³ Jirong Long,² Jianxin Shi,³ Yunhu Wan,³ Yaohua Yang,² Christian Abnet,³ Bryan A Bassig,³ Wei Hu,³ Bu-Tian Ji,³ Madelyn Klugman,¹ Yongbing Xiang,⁴ Yu-Tang Gao,⁴ Jason YY Wong,³ Wei Zheng,² Nathaniel Rothman,³ Xiao-Ou Shu,² Qing Lan³

► Additional material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/thoraxjnl-2020-215542>).

¹Albert Einstein College of Medicine, Bronx, New York, USA
²Vanderbilt University, Nashville, Tennessee, USA
³National Cancer Institute, Bethesda, Maryland, USA
⁴Shanghai Cancer Institute, Shanghai, China

Correspondence to

Dr H Dean Hosgood, Albert Einstein College of Medicine, Bronx, NY 10461, USA; dean.hosgood@einsteinmed.org

HDH, QC and XH contributed equally.
NR, X-OS and QL contributed equally.

Received 15 June 2020
Revised 9 September 2020
Accepted 6 October 2020
Published Online First
14 December 2020



► <http://dx.doi.org/10.1136/thoraxjnl-2020-216385>



© Author(s) (or their employer(s)) 2021. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Hosgood HD, Cai Q, Hua X, et al. *Thorax* 2021;**76**:256–263.

ABSTRACT

Objective To prospectively investigate whether diversity in oral microbiota is associated with risk of lung cancer among never-smokers.

Design and setting A nested case–control study within two prospective cohort studies, the Shanghai Women’s Health Study (n=74 941) and the Shanghai Men’s Health Study (n=61 480).

Participants Lifetime never-smokers who had no cancer at baseline. Cases were subjects who were diagnosed with incident lung cancer (n=114) and were matched 1:1 with controls on sex, age (≤ 2 years), date (≤ 30 days) and time (morning/afternoon) of sample collection, antibiotic use during the week before sample collection (yes/no) and menopausal status (for women).

Main outcomes and measures Metagenomic shotgun sequencing was used to measure the community structure and abundance of the oral microbiome in pre-diagnostic oral rinse samples of each case and control. Multivariable logistic regression models were used to estimate the association of lung cancer risk with alpha diversity metrics and relative abundance of taxa. The Microbiome Regression-Based Kernel Association Test (MiRKAT) evaluated the association between risk and the microbiome beta diversity.

Results Subjects with lower microbiota alpha diversity had an increased risk of lung cancer compared with those with higher microbial alpha diversity (Shannon: $p_{\text{trend}}=0.05$; Simpson: $p_{\text{trend}}=0.04$; Observed Species: $p_{\text{trend}}=0.64$). No case–control differences were apparent for beta diversity ($p_{\text{MiRKAT}}=0.30$). After accounting for multiple comparisons, a greater abundance of Spirochaetia (OR_{low} 1.00 (reference), OR_{medium} 0.61 (95% CI 0.32 to 1.18), OR_{high} 0.42 (95% CI 0.21 to 0.85)) and Bacteroidetes (OR_{low} 1.00 (reference), OR_{medium} 0.66 (95% CI 0.35 to 1.25), OR_{high} 0.31 (95% CI 0.15 to 0.64)) was associated with a decreased risk of lung cancer, while a greater abundance of the Bacilli class (OR_{low} 1.00 (reference), OR_{medium} 1.49 (95% CI 0.73 to 3.08), OR_{high} 2.40 (95% CI 1.18 to 4.87)) and Lactobacillales order (OR_{low} 1.00 (reference), OR_{medium} 2.15 (95% CI 1.03 to 4.47), OR_{high} 3.26 (95% CI 1.58 to 6.70)) was associated with an increased risk of lung cancer.

Conclusions Our prospective study of never-smokers suggests that lower alpha diversity was associated with a greater risk of lung cancer and the abundance of certain specific taxa was associated with altered risk, providing further insight into the aetiology of lung cancer in the absence of active tobacco smoking.

Key messages

What is the key question?

► Does the oral microbiota profile influence the risk of lung cancer among never-smokers?

What is the bottom line?

► In this prospective study of never-smoking lung cancer cases and controls, significant differences between cases and controls were observed in oral microbiome alpha diversity.

Why read on?

► When taken together, the limited but growing body of literature suggests that decreased microbial diversity and increased abundance of taxa within the Firmicutes phylum, and more specifically Lactobacillales, in the respiratory tract may be associated with an increased risk of lung cancer.

INTRODUCTION

Lung cancer is the leading cause of cancer-related death, with a significant portion of the disease attributed to active tobacco smoking. Lung cancer in non-smokers accounts for approximately 25% of lung cancer cases and is the seventh leading cause of cancer death globally.¹ Known lung cancer risk factors such as active tobacco smoke, secondhand tobacco smoke, radon, household air pollution, outdoor air pollution and family history of lung cancer do not fully account for the disease burden.

Bacteria are ubiquitous on and in the human body, and are particularly dense on the skin and in the oral cavity and gastrointestinal tract. Collectively, these bacteria (along with archaea, fungi and viruses) are referred to as the human microbiota, and they outnumber human cells, encoding 100-fold more genes than the human genome.² The diversity and abundance of bacterial communities within the body have led to the microbiota being referred to as the “11th organ”, with potential influences on human health and diseases.³ For example, associations have been observed between gastrointestinal tract cancers and the gut microbiome.⁴ A recent prospective study found that an increased abundance of oral commensal bacteria was associated with a decreased risk for head and neck squamous cell cancer, providing evidence that



microbiota outside the gut may also be associated with cancer risk.⁵ The oral microbiome has also been associated with future risk of pancreatic cancer.⁶

Bacterial communities have been detected in lung tissues and bronchoalveolar lavage samples,⁷ and may be associated with the risk of non-malignant and malignant respiratory diseases. For example, differences in lung bacterial flora have been found between patients with severe chronic obstructive pulmonary disease (COPD) and those with non-diseased lungs.⁸ Retrospective studies have reported decreases in alpha diversity, the number and distribution of distinct operational taxonomic units in a sample, to be related to disease severity in COPD, idiopathic pulmonary fibrosis and cystic fibrosis.⁹ Given that bacterial communities have been detected in the lung,⁷ respiratory tract microbiota may also play a role in the risk of lung cancer.¹⁰

In a small exploratory study, we conducted a bacterial community survey using 16S rRNA gene sequencing and found that there was a significant difference between lung cancer cases and controls in bacterial diversity in sputum samples from never-smokers.¹⁰

This study seeks to further elucidate the role the respiratory tract microbiome may play in lung cancer risk by leveraging one of the world's largest prospective cohorts of never-smoking women. The Shanghai Women's Health Study (SWHS), along with the Shanghai Men's Health Study (SMHS), collected oral rinse samples on a portion of the cohort members at baseline with sufficient follow-up time to allow for nested case-control studies of lung cancer. To assess the temporality between oral microbiome diversity and lung cancer risk, we evaluated bacterial diversity and abundance using metagenomic shotgun sequencing on non-invasive oral rinse samples that were collected among never-smokers at baseline as part of these two large prospective cohort studies.

METHODS

Study population

Our study population consisted of two nested case-control studies within the Shanghai Women's Health Study (SWHS) and Shanghai Men's Health Study (SMHS). Briefly, both the SWHS and SMHS cohorts are population-based prospective cohort studies consisting of >60 000 subjects (SWHS: n=74 941 women; SMHS: n=61 480 men). Enrolment for the SWHS was between 1996 and 2000, and enrolment for the SMHS was between 2002 and 2006. Both cohorts have high participation rates (SWHS=92.7%; SMHS=74.0%).^{11 12} In-person interviews were administered at baseline to obtain information on demographics, lifestyle and dietary habits, medical history and other characteristics. All study participants provided written informed consent before being interviewed, and the study protocols were approved by the institutional review boards of all participating institutions. Cohort members are followed for cancer diagnosis through in-person follow-up surveys administered every 2–3 years and annual record linkage with the Shanghai Cancer Registry and Vital Statistics Unit. All incident lung cancer cases who were lifetime never-smokers were eligible for the current study. For each case (SWHS: n=90; SMHS: n=24), a matched never-smoking control, who donated a mouth rinse (buccal cells) sample at baseline enrolment, was identified. Controls were individually matched on sex, age (≤ 2 years), date (≤ 30 days) and time (morning/afternoon) of sample collection, antibiotic use during the week before sample collection (yes/no) and menopausal status at the time of the sample collection (for women). A total of 114 case-control pairs were included in the study.

DNA extraction and shotgun metagenomic sequencing

DNA was isolated from baseline mouth rinse samples using DNeasy PowerSoil Kit (Qiagen) following the manufacturer's instructions. Sequencing libraries were prepared using the TruePrep DNA Library Prep Kit V2 or Nextera XT DNA Library Preparation Kit (Illumina), following the protocols provided by the manufacturer. Sequencing was performed at paired-end 150 bp using the Illumina HiSeq System at the BGI Americas. Sequencing failed in one case subject due to low DNA yield, leaving 113 case-control matched pairs with sequencing data for matched analyses, and 113 cases and 114 controls for unmatched analyses. Sequencing led to the median (mean; range) number of raw reads being 34 670 505 base pairs (bp) per sample (34 987 730 bp per sample; 23 232 030–51 362 960 bp).

Sequence data processing

The processing of the raw sequencing data was conducted using KneadData (<https://bit-bucket.org/biobakery/kneaddata>), by which bases and reads of low quality and reads mapped onto human genome (GRCh37/hg19) were removed. The clean non-human sequencing reads were then processed by MetaPhlan2¹³ to profile the composition of microbial communities. A total of 1032 taxa (L1=4, L2=14, L3=25, L4=45, L5=80, L6=148, L7=377, L8=339) were identified through these procedures. Three alpha diversity metrics (Shannon index, Simpson index and Observed Species) were calculated by rarefaction at 5000 reads/sample.

We then carried out the following analytical pipeline to assess beta diversity: (1) The 227 raw metagenomes were processed by Fastqmc by trimming position with quality <20 and discarding reads shorter than 90nt; (2) human DNA was removed by using Bowtie2 programme (version 2.3.4.3) against hg19 reference database for each sample; (3) the MetaPhlan2 program¹³ (version 2.7.8) was used on profiling metagenome to the species level; and (4) function profiler HUMAnN2 program¹³ (version 0.11.1) was used to generate pathway reports and gene family reports. Through this pipeline, we calculated the beta diversity (Bray Curtis distance matrix) between metagenomes by rarefaction at 5000 reads/sample.

Statistical analysis

Logistic regression models were used to evaluate the association between microbiome metrics (ie, alpha diversities, relative abundance of taxa, PCoAs of beta diversity) and risk of incident lung cancer. These logistic models were adjusted for matching factors (antibiotic use, sex, age, menopause status (for women), sample collection time) as well as other potential confounders (ie, education). The odds ratios (OR) were calculated based on comparison between the groups (high/medium/low) derived from the 1/3 and 2/3 quantiles of the distribution of the control samples (eg, tertile cut points: Shannon=3.51, 3.77; Simpson=0.95, 0.96; Observed Taxon=122.33, 142.00), along with a p_{trend} based on the Wald test ($H_0: \beta_x = 0, H_1: \beta_x \neq 0$) for the continuous variable in the logistic regression models. We also explored the associations between alpha diversity and time to diagnosis using Cox proportional hazards models adjusting for the same covariates as the logistic regression models. For these exploratory analyses, the proportional hazards assumptions were evaluated using the Schoenfeld individual test for Shannon index ($p=0.32$), Simpson index ($p=0.48$) and Observed Species ($p=0.04$). The association between beta diversity and the risk of incident lung cancer (as well as time to diagnosis) was evaluated using the Microbiome Regression-Based Kernel Association Test (MiRKAT)¹⁴ with the same adjusted covariates as the logistic models.

Table 1 Characteristics of lung cancer cases and controls from the prospective Shanghai Women's Health Study (SWHS) and Shanghai Men's Health Study (SMHS)

	SWHS			SMHS			Combined cohorts			
	Cases		Controls	Cases		Controls	Cases		Controls	
	n	%	n	%	n	%	n	%	n	%
Total	90	100	90	100	24	100	24	100	114	100
Sex										
Men	0	0	0	0	24	100	24	100	24	21
Women	90	100	90	100	0	0	0	0	90	79
Age*										
≤61 years	49	54	49	54	10	42	10	42	59	52
>61 years	41	46	41	46	14	58	14	58	55	48
Mean (±SD)	58.0 (±8.77)		58.2 (±8.77)		64.2 (±7.55)		64.3 (±7.77)		59.3 (±8.86)	
Use of antibiotics†										
Yes	11	12	11	12	3	13	0	0	14	11
No	79	88	79	88	21	88	24	100	100	88
Smoking status										
Never-smoker	90	100	90	100	24	100	24	100	114	100
Current	0	0	0	0	0	0	0	0	0	0
Secondhand smoke exposure†										
Yes	64	71	64	71	Data not collected				Not applicable	
No	26	29	26	29	Data not collected				Not applicable	
Family history of lung cancer										
Yes	5	6	4	4	1	4	0	0	6	4
No	85	94	86	96	23	96	24	100	108	95
Educational attainment										
≥College	13	14	16	18	6	25	11	46	19	17
High school	17	19	19	21	11	46	9	38	28	25
Middle school	23	26	24	27	5	21	3	13	28	25
≤Elementary	37	41	31	34	2	8	1	4	39	34

*Based on median age (years) in controls.

†Not collected in SMHS.

‡In 7 days prior to sample collection.

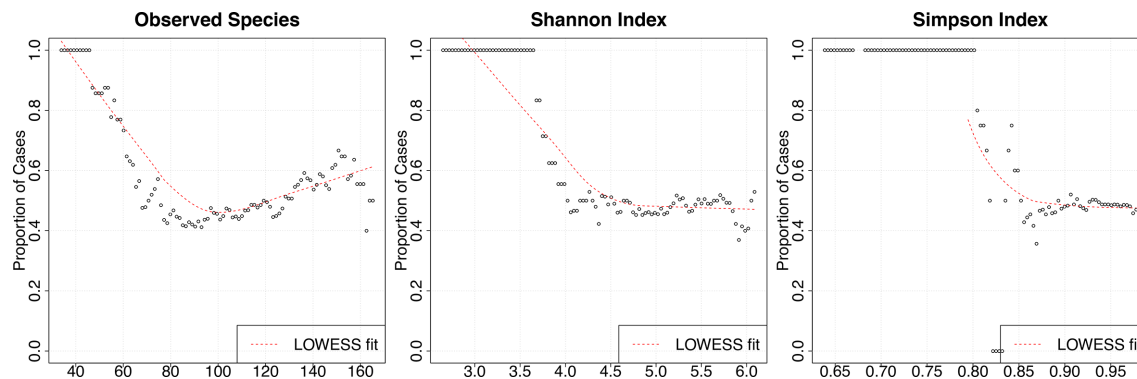


Figure 1 Alpha diversity associated with risk of lung cancer in two prospective cohorts of never-smokers. LOWESS fit of local proportion of cases. For each dot, the X axis is the value of alpha diversity and the Y axis is the proportion of cases among the patients within the local range (± 0.1 * total range) of the corresponding value in the X axis. The red dashed line is LOWESS fit of those dots. LOWESS, locally weighted least squares.

While case–control matching was successful for sex, age, timing of sample collection and menopausal status (for women), there was discordance in matching for antibiotic use. For three SMHS cases who used antibiotics, controls who used antibiotics were unavailable. To assess the impact of this discordance, we evaluated the alpha diversity metrics among all case–control matched pairs, only the pairs successfully matched on antibiotic use, and among all subjects (ie, breaking the match). Given that the results were similar in all three scenarios (online supplemental eTable 1), we opted to break the match to maximise the data available for analyses and proceeded with unconditional logistic regression models, adjusted for the matching factors (antibiotic use, sex, age, menopause status (for women), sample collection time) as well as other potential confounders (ie, education). P values < 0.05 were considered significant for the tests of alpha diversities. Significance testing for taxa was restricted to the “Bacteria” kingdom with prevalence > 0.1 (a total of 534

taxa were included in the tests: L1=1, L2=8, L3=17, L4=24, L5=36, L6=72, L7=197, L8=179). To account for multiple comparisons when evaluating individual taxa, false discovery rates (FDR) were calculated at each phylogenetic level from L2 (phylum) to L7 (species) and FDR < 0.10 were considered significant. Statistical analyses were conducted with R v.3.5.1.

RESULTS

Cases and controls were similar with respect to sex, age, family history of lung cancer and educational attainment (table 1). The median time to diagnosis among cases was 7.2 years (95% CI 0.7 to 13.1) and the median follow-up time among controls was 13.9 years (95% CI 8.7 to 14.6). Among our study population, the microbiota diversity of cases and controls differed when measured by Shannon and Simpson ($p_{\text{trend}}: p_{\text{shannon}} = 0.05$; $p_{\text{simpson}} = 0.04$), but not Observed species ($p_{\text{trend}}: p_{\text{observed species}} = 0.64$)

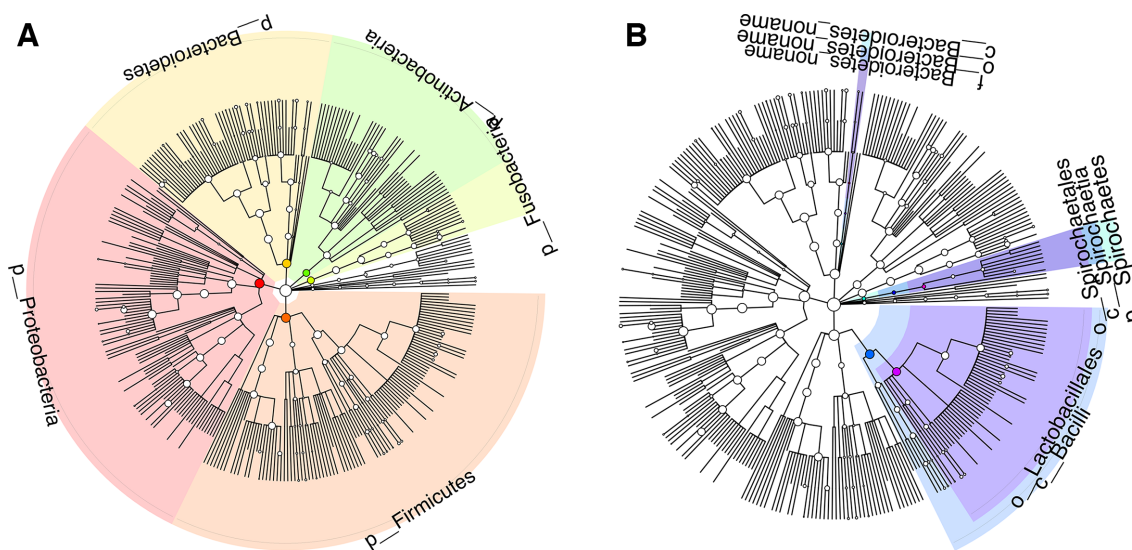


Figure 2 Oral microbiome composition in never-smoking lung cancer cases and controls in the Shanghai Men’s Health Study (SMHS) and Shanghai Women’s Health Study (SWHS). (A) Distribution and description of the microbiome community in all never-smokers (lung cancer cases: $n=113$; controls: $n=114$). Size of circles is representative of relative abundance of taxa in the phylogenetic tree. Colours from red to green represent taxa within the five phyla with the greatest relative abundance (Proteobacteria: 30.3% controls, 30.3% cases; Firmicutes: 26.6% controls, 28.7% cases; Bacteroidetes: 21.7% controls, 19.3% cases; Fusobacteria: 9.3% controls, 9.3% cases; Actinobacteria: 8.7% controls, 9.1% cases). Uncoloured taxa are within rare phyla. The size of the wedge reflects the number of unique phylogenetic attributes in each phylum, not the relative abundance. (B) Summary of the results of the case–control analyses. The blue and purple colours represent the six taxa significantly associated with risk of lung cancer (FDR < 0.10) when adjusted for sex, age, sample collection time, menopause status (for women), education and antibiotic use. FDR, false discovery rate.

Table 2 Abundance of taxa associated (FDR <0.10) with risk of lung cancer in two prospective cohorts of never-smokers*

Taxa†	Medium vs low abundance				High vs low abundance				P _{trend}	FDR
	OR	95% CI		P value	OR	95% CI		P value		
Spirochaetes (p), Spirochaetia (c)	0.61	0.32	1.18	0.14	0.42	0.21	0.85	0.02	0.01	0.08
Bacteroidetes (p), Bacteroidetes (c)	0.66	0.35	1.25	0.21	0.31	0.15	0.64	0.002	0.002	0.03
Bacteroidetes (p), Bacteroidetes (c), Bacteroidetes (o)	0.66	0.35	1.25	0.21	0.31	0.15	0.64	0.002	0.002	0.02
Bacteroidetes (p), Bacteroidetes (c), Bacteroidetes (o), Bacteroidetes (f)	0.66	0.35	1.25	0.21	0.31	0.15	0.64	0.002	0.002	0.06
Firmicutes (p), Bacilli (c)	1.49	0.73	3.08	0.28	2.40	1.18	4.87	0.02	0.01	0.08
Firmicutes (p), Bacilli (c), Lactobacillales (o)	2.15	1.03	4.47	0.04	3.26	1.58	6.70	0.001	0.002	0.02

*Adjusted for sex, age, sample collection time, menopause status (for women), education and antibiotic use.

†Indicated by phylogenetic tree attributes (p=phylum; c=class; o=order; f=family).

FDR, false discovery rate.

(see online supplemental eTable 2). These case-control differences were similar when excluding subjects who used antibiotics 7 days prior to sample collection (p_{trend} : $p_{\text{shannon}}=0.07$; $p_{\text{simpson}}=0.04$; $p_{\text{observed species}}=0.96$) and those with antibiotic use and who were diagnosed ≤ 2 years of sample collection (p_{trend} : $p_{\text{shannon}}=0.09$; $p_{\text{simpson}}=0.06$; $p_{\text{observed species}}=0.84$) (see online supplemental eTable 2). Similar associations were observed in each cohort/sex (see online supplemental eTable 3). No case-control differences among the individual vectors or overall matrix were observed for beta diversity (MiRKAT $p=0.30$; see online supplemental eTable 4).

Overall, subjects with increased Shannon and Simpson diversity tended to have a decreased risk of lung cancer compared with subjects who had lower microbial diversity (figure 1). Similar results were also observed for the association between alpha diversity and time to diagnosis (Observed species: $p=0.17$, HR (per 10-species increase) 0.91 (95% CI 0.79 to 1.04); Shannon index: $p=0.043$, HR (per 1-unit increase) 0.72 (95% CI 0.52 to 0.99); Simpson index: $p=0.011$, HR (per 0.1-unit increase) 0.61 (95% CI 0.42 to 0.89)). Beta diversity was not associated with time to diagnosis (MiRKAT $p=0.55$). Associations at the phylogenetic tree levels (figure 2) showed that the relative abundance of six taxa was associated with lung cancer risk (FDR <0.10) (table 2). Within the Spirochaetes phylum, an increased abundance of class Spirochaetia ($p_{\text{trend}}=0.01$; FDR=0.08) was associated with a decreased risk (OR_{low} 1.00 (reference), OR_{medium} 0.61 (95% CI 0.32 to 1.18), OR_{high} 0.42 (95% CI 0.21 to 0.85)). Within the Firmicutes phylum, an increased abundance of the Bacilli class (OR_{low} 1.00 (reference), OR_{medium} 1.49 (95% CI 0.73 to 3.08), OR_{high} 2.40 (95% CI 1.18 to 4.87); $p_{\text{trend}}=0.01$; FDR=0.08) and, more specifically, the Lactobacillales order (OR_{low} 1.00 (reference), OR_{medium} 2.15 (95% CI 1.03 to 4.47), OR_{high} 3.26 (95% CI 1.58 to 6.70); $p_{\text{trend}}=0.002$; FDR=0.02) was associated with an increased risk of lung cancer. Bacteroidetes was associated with a decreased lung cancer risk (OR_{low} 1.00 (reference), OR_{medium} 0.66 (95% CI 0.35 to 1.25), OR_{high} 0.31 (95% CI 0.15 to 0.64)) at the class ($p_{\text{trend}}=0.002$; FDR=0.03), order ($p_{\text{trend}}=0.002$; FDR=0.02) and family ($p_{\text{trend}}=0.002$; FDR=0.06) levels. These six taxa remained associated with risk of lung cancer when restricting only to subjects with no antibiotic use in the 7 days prior to sample collection (see online supplemental eTable 5). Similar patterns were observed for all six taxa and risk of lung cancer among the younger (≤ 61 years) and older (> 61 years) age groups, as well as women (table 3). Among the limited sample size of men ($n=24$ controls, 23 cases), only an increased abundance of Spirochaetia was associated with risk of lung cancer (table 3).

Gene/pathway-based analyses identified a total of 879 029 non-redundant genes mapped to 422 pathways, of which 338 333 genes and 313 pathways had a prevalence of > 0.1 . The associations between the relative abundance of these 313 pathways and the risk of lung cancer yielded a higher than anticipated number of significant associations (red points in online supplemental eFigure 1A). When exploring this finding further, we found that the risk of lung cancer was associated with the microbiota reads rate (defined as the proportion of microbiota reads in total reads) ($p=0.016$, OR per 10% increase of the rate 1.23 (95% CI 1.07 to 1.42); see online supplemental eFigure 2). After adjusting for the potential confounding by microbiota reads rate, the observed associations between pathways and lung cancer risk were as expected (blue points in online supplemental eFigure 1A). Similar results were also observed in testing the presence/absence of those 338 333 genes (online supplemental eFigure 1B). Of note, adjustment for microbiota reads rate did not meaningfully modify our alpha diversity results (see online supplemental eTable 1).

DISCUSSION

We conducted a nested case-control study of lung cancer among never-smoking women and never-smoking men in Shanghai, China and found that decreased oral microbiota alpha diversity, but not beta diversity, was associated with an increased risk of lung cancer. This is the first report of a prospective study of the oral microbiome and lung cancer risk among never-smokers. The robustness of our findings is exemplified by the associations remaining after excluding individuals who used antibiotics during the 7 days prior to sample collection, as well as after excluding individuals who were diagnosed with lung cancer shortly (eg, within 2 years) after the baseline sample was collected. In addition, we observed that among never-smokers, increased relative abundance within the Bacteroidetes and Spirochaetes phyla was associated with a reduced risk of lung cancer, whereas increased abundance within the Firmicutes phylum was associated with an increased risk of lung cancer. Our observed associations highlight the importance of the microbial richness, and the potential relevance of rarer taxa, in relation to the risk of lung cancer.

In a recent retrospective case-control study of never-smoking women in Xuanwei, China ($n=45$ cases, $n=45$ controls), we also observed that an increased risk of lung cancer was associated with lower alpha diversity compared with higher alpha diversity in sputum samples.¹⁵ The literature, although limited, also supports our observations. For example, when comparing paired malignant versus non-malignant lung cancer tumour tissues, the malignant tissues had lower alpha diversity.¹⁶ Furthermore,

Table 3 Abundance of taxa associated with risk of lung cancer* in two prospective cohorts of never-smokers stratified by sex/cohort and age

Taxat	Medium vs low abundance				High vs low abundance				Medium vs low abundance				High vs low abundance				
	OR	95% CI		P value	OR	95% CI		P trend	OR	95% CI		P value	OR	95% CI		P trend	
		LCI	UCI			LCI	UCI			LCI	UCI			LCI	UCI		
Men/Shanghai Men's Health Study																	
24 controls vs 23 cases																	
Spirochaetes (p), Spirochaetia (c)	0.63	0.13	3.07	0.57	0.05	<0.01	0.73	0.03	0.04	0.57	0.27	1.21	0.14	0.53	0.25	1.13	0.10
Bacteroidetes (p), Bacteroidetes (c)	1.02	0.21	4.91	0.98	0.20	0.02	1.56	0.12	0.18	0.59	0.29	1.20	0.14	0.30	0.13	0.68	<0.01
Bacteroidetes (p), Bacteroidetes (c), Bacteroidetes (o)	1.02	0.21	4.91	0.98	0.20	0.02	1.56	0.12	0.18	0.59	0.29	1.20	0.14	0.30	0.13	0.68	<0.01
Bacteroidetes (p), Bacteroidetes (c), Bacteroidetes (o), Bacteroidetes (f)	1.02	0.21	4.91	0.98	0.20	0.02	1.56	0.12	0.18	0.59	0.29	1.20	0.14	0.30	0.13	0.68	<0.01
Firmicutes (p), Bacilli (c)	0.46	0.08	2.74	0.39	0.25	0.04	1.73	0.16	0.16	1.88	0.82	4.32	0.14	3.70	1.64	8.35	<0.01
Firmicutes (p), Bacilli (c), Lactobacillales (o)	2.32	0.44	12.10	0.32	0.64	0.10	4.04	0.64	0.56	2.00	0.87	4.60	0.10	4.58	2.02	10.38	<0.01
Age ≤61																	
59 controls vs 59 cases																	
Spirochaetes (p), Spirochaetia (c)	0.51	0.20	1.33	0.17	0.47	0.18	1.27	0.14	0.14	0.65	0.25	1.68	0.37	0.27	0.09	0.83	0.02
Bacteroidetes (p), Bacteroidetes (c)	0.76	0.31	1.87	0.55	0.25	0.09	0.72	0.01	0.01	0.53	0.20	1.35	0.18	0.33	0.11	0.99	0.04
Bacteroidetes (p), Bacteroidetes (c), Bacteroidetes (o)	0.76	0.31	1.87	0.55	0.25	0.09	0.72	0.01	0.01	0.53	0.20	1.35	0.18	0.33	0.11	0.99	0.04
Bacteroidetes (p), Bacteroidetes (c), Bacteroidetes (o), Bacteroidetes (f)	0.76	0.31	1.87	0.55	0.25	0.09	0.72	0.01	0.01	0.53	0.20	1.35	0.18	0.33	0.11	0.99	0.04
Firmicutes (p), Bacilli (c)	0.89	0.34	2.34	0.81	1.60	0.63	4.12	0.33	0.31	3.39	1.04	11.08	0.04	5.00	1.47	17.02	0.01
Firmicutes (p), Bacilli (c), Lactobacillales (o)	1.41	0.54	3.69	0.48	1.97	0.74	5.24	0.17	0.17	4.24	1.27	14.09	0.02	6.84	2.08	22.47	<0.01

*Adjusted for age, sex, sample collection time, menopause status (for women), education and antibiotic use (excluding adjustment for sex and menopause status in the stratified analysis when not applicable).

† Indicated by phylogenetic tree attributes (p=phylum; c=class; o=order; f=family).

LCI, lower bound of 95% CI; UCI, upper bound of 95% CI.

among patients with COPD, which is a risk factor for lung cancer, significantly decreased alpha diversity in sputum has been shown to be associated with increased severity of certain types of COPD exacerbations.¹⁷

We also observed that the abundance of specific taxa was associated with an increased risk of lung cancer. Notably, the current study replicates our previous observation that a greater abundance of Lactobacillales in sputum was associated with lung cancer in Xuanwei, China.¹⁰ Our findings regarding the abundance within the Bacteroidetes and Firmicutes phyla are in agreement with prior literature assessing the relationship between the microbiome and lung disease. These two phyla are common in clinically stable COPD,⁹ with the abundance of Firmicutes shown to be increased in COPD cases compared with controls in several studies, including one report using lung tissue samples that demonstrated an increase in genera within Lactobacillales.^{8, 18} An increased relative abundance of Firmicutes in bronchoalveolar lavage fluid has also been associated with lung cancer.¹⁹ A lower abundance of class Bacteroidetes among lung cancer cases compared with controls has been reported in a small case-control study conducted in China.²⁰ Interestingly, the risk of head and neck squamous cell carcinoma was recently associated with an increased abundance of Bacteroidetes and Firmicutes in the oral microbiome.⁵ When taken together, this limited body of literature suggests that decreased microbial diversity and increased abundance of taxa within the Firmicutes phylum, and more specifically Lactobacillales, in the respiratory tract may be associated with an increased risk of lung cancer. Although little is currently known about the potential aetiological mechanisms, it is hypothesised that these microbial changes may lead to carcinogenesis by their production of DNA-damaging metabolites.¹⁸ Exposure to known lung cancer risk factors may also decrease the overall diversity or increase abundance of specific taxa.¹⁶ Although our study did not directly assess the functionality of specific microbiota, our observed association between lung cancer risk and Lactobacillales is biologically plausible given that Lactobacillales has been shown in laboratory-based studies to: (1) have antiviral activities; (2) have antimicrobial activities; and (3) be involved in the detoxification of carcinogens, particularly those that play a role in lung cancer (ie, polycyclic aromatic hydrocarbons).^{21–23} Population-based studies are needed to assess the biological function of Lactobacillales in the respiratory tract.

Our study has a number of strengths and limitations. First, our nested case-control design used samples that were collected years prior to diagnosis of lung cancer in two cohorts that used the identical sample collection and analysis protocol and had similar results. Further, we analysed only lifetime never-smokers which eliminates the possibility of our results being driven by the influences active tobacco smoking may have on the oral microbiota.¹⁶ Finally, we were able to integrate critical covariate data, also collected as baseline, to adjust for potential confounders including antibiotic use in the 7 days prior to sample collection. Given that microbiomes have been shown to vary by geographical location, a limitation of our study is the homogeneity of our study population (eg, all from Shanghai). Further, we only had a single sample from each subject for analysis, so we were not able to assess temporal variation and microbiota stability in our subjects. While our study provides evidence that variation in the oral microbiome plays a role in lung cancer risk, the interpretation of our study must be done while considering the caveat that our findings are from a single time point in a single geographical location. Future studies should focus on samples collected at various time points to assess the temporal variation

of the oral microbiome, and should include other ethnicities/races from a variety of locations to evaluate the generalisability of our findings.

CONCLUSION

In this study we found that lower bacterial alpha diversity was associated with a greater risk of lung cancer among never-smokers. Further, taxa in the Bacteroidetes, Spirochaetes and Firmicutes phyla were also associated with altered risk. To the best of our knowledge, this is the first prospective study to evaluate the oral microbiome and risk of lung cancer in never-smokers. Given the novelty of our findings and our limited sample size, replication studies are essential. Overall, our results highlight the need for further research on the role microbiota of the oropharyngeal and respiratory tract play in respiratory diseases.

Contributors QL, QC, WZ, X-OS and NR conceived and designed the study. YX, Y-TG, YY and B-TJ performed sample collection and data harmonisation. XH, JL, JS and YW conducted the data analysis. HDH, QC, MK and QL drafted the initial manuscript. All authors contributed to the interpretation of the results and manuscript preparation.

Funding Part of this work was funded by UM1CA182910, UM1CA173640 and R01CA207466. Sample preparation was performed at the Survey and Biospecimen Shared Resource, which is supported in part by the Vanderbilt-Ingram Cancer Center (P30 CA068485). The microbiome data processing analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University.

Disclaimer The funding sources had no involvement in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication. The researchers confirm independence from the funders and that authors had access to the data and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. A limited dataset may be available upon reasonable request.

ORCID iD

H Dean Hosgood <http://orcid.org/0000-0003-4151-1133>

REFERENCES

- Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers--a different disease. *Nat Rev Cancer* 2007;7:778–90.
- Ley RE, Peterson DA, Gordon JL. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 2006;124:837–48.
- Young VB. The role of the microbiome in human health and disease: an introduction for clinicians. *BMJ* 2017;356:j831.
- Mima K, Ogino S, Nakagawa S, et al. The role of intestinal bacteria in the development and progression of gastrointestinal tract neoplasms. *Surg Oncol* 2017;26:368–76.
- Hayes RB, Ahn J, Fan X, et al. Association of oral microbiome with risk for incident head and neck squamous cell cancer. *JAMA Oncol* 2018;4:358–65.
- Fan X, Alekseyenko AV, Wu J, et al. Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. *Gut* 2018;67:120–7.
- Cui L, Morris A, Huang L, et al. The microbiome and the lung. *Ann Am Thorac Soc* 2014;11(Suppl 4):S227–32.
- Sze MA, Dimitriu PA, Hayashi S, et al. The lung tissue microbiome in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2012;185:1073–80.
- Faner R, Sibila O, Agustí A, et al. The microbiome in respiratory medicine: current challenges and future perspectives. *Eur Respir J* 2017;49. doi:10.1183/13993003.02086-2016. [Epub ahead of print: 12 Apr 2017].
- Hosgood HD, 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.
- Zheng W, Chow W-H, Yang G, et al. The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *Am J Epidemiol* 2005;162:1123–31.
- Shu X-O, Li H, Yang G, et al. Cohort profile: the Shanghai Men's Health Study. *Int J Epidemiol* 2015;44:810–8.
- Segata N, Waldron L, Ballarini A, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 2012;9:811–4.

- 14 Zhao N, Chen J, Carroll IM, *et al.* Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am J Hum Genet* 2015;96:797–807.
- 15 Hosgood HD, Mongodin EF, Wan Y, *et al.* The respiratory tract microbiome and its relationship to lung cancer and environmental exposures found in rural China. *Environ Mol Mutagen* 2019;60:617–23.
- 16 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.
- 17 Wang Z, Singh R, Miller BE, *et al.* Sputum microbiome temporal variability and dysbiosis in chronic obstructive pulmonary disease exacerbations: an analysis of the COPDMAP study. *Thorax* 2018;73:331–8.
- 18 Mao Q, Jiang F, Yin R, *et al.* Interplay between the lung microbiome and lung cancer. *Cancer Lett* 2018;415:40–8.
- 19 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.
- 20 Yan X, Yang M, Liu J, *et al.* Discovery and validation of potential bacterial biomarkers for lung cancer. *Am J Cancer Res* 2015;5:3111–22.
- 21 Lili Z, Junyan W, Hongfei Z, *et al.* Detoxification of cancerogenic compounds by lactic acid bacteria strains. *Crit Rev Food Sci Nutr* 2018;58:2727–42.
- 22 Biliavska L, Pankivska Y, Povnitsa O, *et al.* Antiviral activity of exopolysaccharides produced by lactic acid bacteria of the genera *Pediococcus*, *Leuconostoc* and *Lactobacillus* against human adenovirus type 5. *Medicina* 2019;55. doi:10.3390/medicina55090519. [Epub ahead of print: 22 Aug 2019].
- 23 Ren D, Zhu J, Gong S, *et al.* Antimicrobial characteristics of lactic acid bacteria isolated from homemade fermented foods. *Biomed Res Int* 2018;2018:5416725.

eTable 1: Comparison of Alpha Diversity p-trend results from conditional and unconditional logistic regression models, among lung cancer cases and controls, in two prospective cohorts of never-smokers.

Alpha diversity index metric	All 113 case-control pairs		Excluding 3 pairs unsuccessfully matched on Antibiotic Use	
	Conditional Logistic Regression	Unconditional Logistic Regression [¥]	Conditional Logistic Regression	Unconditional Logistic Regression [¥]
Shannon	0.07	0.04	0.08	0.04
Shannon*	0.02	0.02	0.03	0.02
Simpson	0.04	0.03	0.05	0.03
Simpson*	0.03	0.03	0.04	0.03
Observed species	0.96	0.64	0.97	0.55
Observed species*	0.07	0.05	0.08	0.05

[¥]adjusted for matching factors (sex, age, sample collection time, menopause status (for women), and antibiotic use); *adjusted for sequencing reads per sample.

eTable 2: Alpha Diversity between lung cancer cases and controls, in two prospective cohorts of never-smokers[†].

α -diversity metric	Medium vs Low Diversity				High vs Low Diversity				P _{trend}
	Odds Ratio	95%CI		P	Odds Ratio	95%CI		P	
		Lower	Upper			Lower	Upper		
All subjects (114 controls vs 113 cases)									
Shannon	0.62	0.32	1.21	0.16	0.73	0.37	1.46	0.38	0.05
Simpson	0.55	0.28	1.07	0.08	0.74	0.38	1.44	0.37	0.04
Observed Species	0.81	0.41	1.60	0.55	1.06	0.55	2.08	0.85	0.64
Only subjects with no antibiotic use in 7 days prior to sample collection (103 controls vs 99 cases)[†]									
Shannon	0.59	0.29	1.20	0.15	0.73	0.36	1.49	0.39	0.07
Simpson	0.47	0.23	0.96	0.04	0.71	0.35	1.42	0.33	0.04
Observed Species	0.86	0.42	1.77	0.69	1.22	0.60	2.46	0.58	0.96
Only subjects with no antibiotic use in 7 days prior to sample collection, and sample collected ≥ 2 years before cancer (103 controls vs 86 cases)[†]									
Shannon	0.61	0.29	1.28	0.19	0.73	0.34	1.57	0.43	0.09
Simpson	0.47	0.22	1.02	0.05	0.72	0.34	1.52	0.39	0.06
Observed Species	0.81	0.38	1.73	0.59	1.11	0.53	2.33	0.79	0.84

[†]Adjusted for sex, age, sample collection time, menopause status (among women), education, and antibiotic use (when appropriate).

eTable 3: Alpha Diversity, stratified by sex/cohort, between lung cancer cases and controls, in two prospective cohorts of never-smokers†.

α-diversity metric†	Men / Shanghai Men's Health Study									Women / Shanghai Women's Health Study								
	Medium vs Low Diversity				High vs Low Diversity					Medium vs Low Diversity				High vs Low Diversity				
	OR ¹	95%CI		P	OR	95%CI		P	P _{trend}	OR	95% CI		P	OR	95% CI		P	P _{trend}
LCI ²		UCI ³	LCI			UCI	LCI				UCI	LCI			UCI			
All subjects																		
24 controls vs 23 cases										90 controls vs 90 cases								
Shannon	0.41	0.08	1.99	0.27	0.31	0.04	2.67	0.29	0.23	0.60	0.27	1.30	0.19	0.80	0.38	1.71	0.57	0.10
Simpson	0.48	0.09	2.44	0.38	1.23	0.23	6.63	0.81	0.23	0.48	0.22	1.04	0.06	0.65	0.31	1.38	0.26	0.06
Observed species	1.20	0.28	5.10	0.81	1.08	0.15	7.88	0.94	0.55	0.79	0.36	1.75	0.57	1.14	0.55	2.39	0.72	0.98
Only subjects with no antibiotic use in 7 days prior to sample collection																		
24 controls vs 20 cases										79 controls vs 79 cases								
Shannon	0.41	0.08	1.99	0.27	0.31	0.04	2.67	0.29	0.23	0.58	0.25	1.32	0.19	0.82	0.37	1.81	0.62	0.15
Simpson	0.48	0.09	2.44	0.38	1.23	0.23	6.63	0.81	0.23	0.42	0.18	0.97	0.04	0.63	0.29	1.38	0.25	0.07
Observed species	1.20	0.28	5.10	0.81	1.08	0.15	7.88	0.94	0.55	0.83	0.35	1.95	0.67	1.33	0.61	2.91	0.48	0.65

†Adjusted for age, sample collection time, menopause status (for women), education, and antibiotic use (excluding adjustment for antibiotic use in the stratified analysis); ¹Odds Ratio, ²Lower bound of 95% confidence interval (CI); ³Upper bound of 95% confidence interval (CI).

eTable 4: Beta Diversity¹ between lung cancer cases and controls, in two prospective cohorts of never-smokers †.

Model	P Value										
	PCoA1	PCoA2	PCoA3	PCoA4	PCoA5	PCoA6	PCoA7	PCoA8	PCoA9	PCoA10	MiRKAT
All (114 controls vs 113 cases)	0.53	0.06	0.45	0.62	0.93	0.19	0.41	0.44	0.07	0.98	0.30
Male (24 controls vs 23 cases)	0.81	0.13	0.27	0.26	0.63	0.68	0.89	0.68	0.66	0.30	0.61
Female (90 controls vs 90 cases)	0.52	0.18	0.13	0.38	0.97	0.09	0.20	0.36	0.08	0.52	0.22
Age ≤ 61 (59 controls vs 59 cases)	0.89	0.30	0.83	0.78	0.49	0.98	0.78	0.62	0.02	0.99	0.88
Age > 61 (55 controls vs 54 cases)	0.23	0.16	0.82	0.89	0.88	0.07	0.27	0.64	0.77	1.00	0.35
No Antibiotic (103 controls vs 99 cases)	0.64	0.03	0.46	0.68	0.71	0.17	0.55	0.58	0.07	0.85	0.19

¹ Bray Curtis distance matrix

† MiRKAT tests the association between each covariate and each beta diversity, adjusted for sex, age, sample collection time, menopause status (for women), education, and antibiotic use (menopause, sex, and antibiotic use excluded in stratified analysis if not applicable). The percentage of variance explained by each PCoA: PCoA1 (13.7 %), PCoA2 (11.6 %), PCoA3 (6.7 %), PCoA4 (5.8 %), PCoA5 (5.0 %), PCoA6 (3.4 %), PCoA7 (3.1 %), PCoA8 (2.9 %), PCoA9 (2.5 %), and PCoA10 (2.3 %).

Table 5: Abundance of taxa associated with risk of lung cancer[†] in two prospective cohorts of never-smokers, stratified by antibiotic use.

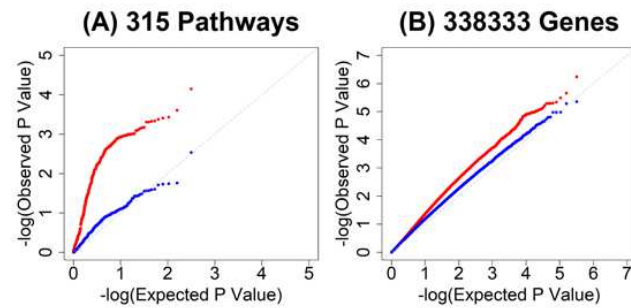
Taxa [‡]	Medium vs Low Abundance				High vs Low Abundance				P _{trend}	Medium vs Low Abundance				High vs Low Abundance				
	OR ¹	95%CI		P	OR	95%CI		P		OR	95% CI		OR	95% CI		P	P _{trend}	
		LCI ²	UCI ³			LCI	UCI				LCI	UCI		LCI	UCI			
All subjects										Only subjects with no antibiotic use in 7 days prior to sample collection								
114 controls vs 113 cases										103 controls vs 99 cases								
<i>Spirochaetes</i> (p), <i>Spirochaetia</i> (c)	0.61	0.32	1.18	0.14	0.42	0.21	0.85	0.02	0.01	0.68	0.34	1.37	0.28	0.40	0.19	0.85	0.02	0.02
<i>Bacteroidetes</i> (p), <i>Bacteroidetes</i> (c)	0.66	0.35	1.25	0.21	0.31	0.15	0.64	0.002	0.002	0.74	0.38	1.45	0.38	0.25	0.11	0.57	<0.01	<0.01
<i>Bacteroidetes</i> (p), <i>Bacteroidetes</i> (c), <i>Bacteroidetes</i> (o)	0.66	0.35	1.25	0.21	0.31	0.15	0.64	0.002	0.002	0.74	0.38	1.45	0.38	0.25	0.11	0.57	<0.01	<0.01
<i>Bacteroidetes</i> (p), <i>Bacteroidetes</i> (c), <i>Bacteroidetes</i> (o), <i>Bacteroidetes</i> (f)	0.66	0.35	1.25	0.21	0.31	0.15	0.64	0.002	0.002	0.74	0.38	1.45	0.38	0.25	0.11	0.57	<0.01	<0.01
<i>Firmicutes</i> (p), <i>Bacilli</i> (c)	1.49	0.73	3.08	0.28	2.40	1.18	4.87	0.02	0.01	2.01	0.94	4.29	0.07	2.76	1.30	5.85	<0.01	<0.01
<i>Firmicutes</i> (p), <i>Bacilli</i> (c), <i>Lactobacillales</i> (o)	2.15	1.03	4.47	0.04	3.25	1.58	6.70	0.001	0.002	2.60	1.20	5.62	0.02	3.61	1.67	7.78	<0.01	<0.01

[†]Adjusted for age, sex, sample collection time, menopause status (for women), education, and antibiotic use (excluding adjustment for antibiotic use in the stratified analysis when not applicable);

¹Odds Ratio, ²Lower bound of 95% confidence interval (CI); ³Upper bound of 95% confidence interval (CI).

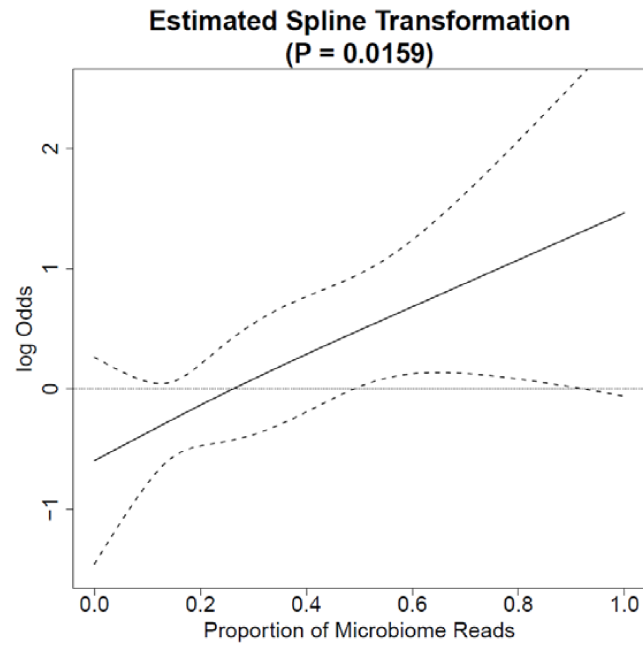
[‡]Indicated by phylogenetic tree attributes (p=phylum; c=class; o=order; f=family)

eFigure 1: QQ plots of P-values from logistic models of (A) relative abundance of 315 pathways and (B) Presence/absence of 338,333 genes with a prevalence of >0.1 associated with risk of lung cancer in two prospective cohorts of never-smokers[†].



[†]Red points: Adjusted for sex, age, sample collection time, menopause status (for women), education, and antibiotic use; Blue points: additionally adjusted the microbiota reads rate.

eFigure 2: Proportion of microbiota reads associated with risk of lung cancer in two prospective cohorts of never-smokers[†].



[†]Adjusted for sex, age, sample collection time, menopause status (for women), education, and antibiotic use