

RESEARCH

Open Access



House dust metagenome and pulmonary function in a US farming population

Mikyeong Lee^{1*}, Abhishek Kaul², James M. Ward³, Qiyun Zhu^{4,5}, Marie Richards⁶, Ziyue Wang⁷, Antonio González⁸, Christine G. Parks¹, Laura E. Beane Freeman⁹, David M. Umbach⁷, Alison A. Motsinger-Reif⁷, Rob Knight^{8,10,11,12} and Stephanie J. London¹

Abstract

Background Chronic exposure to microorganisms inside homes can impact respiratory health. Few studies have used advanced sequencing methods to examine adult respiratory outcomes, especially continuous measures. We aimed to identify metagenomic profiles in house dust related to the quantitative traits of pulmonary function and airway inflammation in adults. Microbial communities, 1264 species (389 genera), in vacuumed bedroom dust from 779 homes in a US cohort were characterized by whole metagenome shotgun sequencing. We examined two overall microbial diversity measures: richness (the number of individual microbial species) and Shannon index (reflecting both richness and relative abundance). To identify specific differentially abundant genera, we applied the Lasso estimator with high-dimensional inference methods, a novel framework for analyzing microbiome data in relation to continuous traits after accounting for all taxa examined together.

Results Pulmonary function measures (forced expiratory volume in one second (FEV₁), forced vital capacity (FVC), and FEV₁/FVC ratio) were not associated with overall dust microbial diversity. However, many individual microbial genera were differentially abundant (p -value < 0.05 controlling for all other microbial taxa examined) in relation to FEV₁, FVC, or FEV₁/FVC. Similarly, fractional exhaled nitric oxide (FeNO), a marker of airway inflammation, was unrelated to overall microbial diversity but associated with differential abundance for many individual genera. Several genera, including *Limosilactobacillus*, were associated with a pulmonary function measure and FeNO, while others, including *Moraxella* to FEV₁/FVC and *Stenotrophomonas* to FeNO, were associated with a single trait.

Conclusions Using state-of-the-art metagenomic sequencing, we identified specific microorganisms in indoor dust related to pulmonary function and airway inflammation. Some were previously associated with respiratory conditions; others were novel, suggesting specific environmental microbial components contribute to various respiratory outcomes. The methods used are applicable to studying microbiome in relation to other continuous outcomes.

Keywords Microbiome, Microbiota, Metagenome, Whole genome sequencing, Spirometry, Respiratory function tests

*Correspondence:

Mikyeong Lee
mikyeong.lee@nih.gov

Full list of author information is available at the end of the article



This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Chronic respiratory illnesses pose a major public health burden [1]. Although exposure to microorganisms inside homes has been linked to respiratory health [2], data in adults are limited. A few studies have examined associations of microbial composition with asthma and allergies [3–5]. Continuous outcomes have been largely ignored. No large studies of adult respiratory outcomes that capture environmental microbial exposure using state-of-the-art whole metagenome shotgun sequencing have been reported.

Pulmonary function is a continuous measure of the physiologic state of the lungs in health and disease. Lower pulmonary function associates with poor health-related quality of life [6] and predicts mortality, independently of other risk factors [7]. Compared to genetic [8, 9] and epigenetic [10] factors, less is known regarding potential impacts of exposures to microbial components inside homes on pulmonary function. Lower pulmonary function in asthmatic individuals was associated with exposure to house dust endotoxin, a generic measure of gram-negative bacteria [11]. Some studies reported no significant associations of adult pulmonary function with bacterial or fungal components [12], while others found associations with specific microorganisms [13] or moldiness [14]. These studies measured microbial agents using quantitative PCR (qPCR) which limits the number of microorganisms under investigation [12–14]. Metagenomics provides the opportunity to explore a broad spectrum of microorganisms, including ones that cannot be cultured in a laboratory setting.

Fractional exhaled nitric oxide (FeNO) is a quantitative, noninvasive measure of airway inflammation. In children, exposure to endotoxin in house dust was related to lower FeNO [15]. FeNO levels associated with diverse indoor fungal communities among 55 adults [16]. We find no large sequencing-based study of indoor microbial profiles and FeNO in adults.

We used whole metagenome shotgun sequencing to comprehensively profile microorganisms in house dust from 779 households in the Agricultural Lung Health Study (ALHS), a case-control study of asthma nested within a US farming cohort. We investigated associations of pulmonary function and FeNO with diversity of microbial communities inside homes and with abundance of individual microbial taxa. We implemented a novel framework for analyzing microbiome data in relation to continuous health outcomes.

Methods

Study population

Participants were enrolled in the Agricultural Lung Health Study (ALHS), a case-control study of current

asthma nested within the Agricultural Health Study (AHS), a cohort of farmers and spouses of farmers in North Carolina (NC) and Iowa (IA) [17]. Details have been described previously (data version P3REL201209.00) [11, 18]. In brief, the ALHS enrolled 3301 participants (1223 asthma cases and 2078 noncases) in 2009–2013. Of these, 2871 received a home visit at which bedroom dust was collected. Of a simple random sample ($N = 1000$) chosen for our previous 16S rRNA amplicon sequencing study, 879 samples passed quality control and were included in our previous dust microbiome analyses [19]. These 879 samples were subjected to whole metagenome shotgun sequencing for more accurate comprehensive characterization of microbial communities.

Respiratory outcomes

Measurement of pulmonary function and FeNO in ALHS has been described previously [20]. In brief, trained staff measured prebronchodilator spirometric parameters, including the forced expiratory volume in the first second (FEV₁) and forced vital capacity (FVC), during in-home visits using an EasyOne[®] spirometer (NDD Medical Technologies, Chelmsford, MA, USA) according to American Thoracic Society guidelines. Participants were asked to avoid use of bronchodilators for at least 6 h before the visit; only 3% of participants failed to comply. We calculated FEV₁/FVC as a proportion ranging from 0 to 1. FeNO was measured using NIOX MINO (Aerocrine AB, Solna, Sweden) following manufacturer guidelines in duplicate and then averaged. Values below the limit of detection (LOD; < 5 ppb; 5.1%) were assigned to LOD/ $\sqrt{2} = 3.5$ ppb.

House dust collection and whole genome shotgun sequencing

Trained field technicians collected dust samples using a DUSTREAM[™] Collector (Indoor Biotechnologies, Inc., Charlottesville, VA, USA) from participants' bedrooms during home visits [11]. Technicians vacuumed a one square yard area on the sleeping surface and on the floor next to the bed. Details on dust samples and DNA extraction were previously described [19]. Extracted DNA samples were sent to Center for Microbiome Innovation, University of California San Diego, for whole genome shotgun metagenomic sequencing using Illumina NovaSeq (Illumina, Inc., San Diego, CA, USA). Processing included (1) trimming of low-quality sequence reads, duplicates, and adapters based on FastQC results (v0.11.5) [21] and (2) identification and removal of potential contaminant sequence reads, not from microbial genomes but from host genomic sources (human, cow, pig, chicken, turkey, horse, goat, sheep, dog, cat, and

dust mite) (Table S1) plus PhiX, a spike-in control in an Illumina experiment, using Bowtie2 [22] and KneadData (v0.7.10) [23]. We obtained taxonomic classification of sequences using Kraken2 (v2.1.1) [24] and generated abundance (counts) for each taxon using Bracken (v2.5.0) [25] with RefSeq genomes for bacteria, archaea, eukaryotes, fungi, viruses, and plasmids and NCBI taxonomy information. Additionally, we identified and removed sequences related to potential contamination from sample collection and laboratory reagents (168 taxa) (Table S2) using negative “blank” controls of sterile water and the decontam R package (v1.10.0) [26]. We used both frequency-based (the default threshold of 0.1) and prevalence-based (a stricter threshold of 0.5) methods. For further analyses, we excluded 98 samples having sequence reads < 1000 and taxa having < 0.0005% of the total number of sequence reads across all samples [27, 28] or assigned to Eukaryota and viruses with limited RefSeq genome databases available, leaving microbial abundance data for 1264 species (389 genera) in 781 samples.

Overall microbial diversity in relation to respiratory outcomes

We calculated two measures of overall microbial diversity within each sample (alpha diversity): richness (the number of individual microbial species) and the Shannon index [29], which reflects both richness and relative abundances of each species. Using linear regression, we evaluated associations of the diversity measures with pulmonary function parameters (FEV₁, FVC, and FEV₁/FVC) or FeNO. Due to its negatively skewed distribution, Shannon index was exponentially transformed before association analyses. Covariates for pulmonary function were age, age squared, sex, height, height squared, weight (for FVC only), cigarette smoking (former or current, both relative to never), cigarette pack-years, asthma status (yes/no), state of residence (NC/IA), and ancestry (European/not based on genome-wide genetic information, except for one sample whose information was filled with self-reported race as White). Models for FeNO included abovementioned covariates except age squared and height squared. As sensitivity analyses, we additionally adjusted for season of dust collection (winter/not) which showed associations with overall microbial diversity in ALHS [19].

To avoid bias due to different sequencing depths among samples, abundance data were rarefied to the minimum number of sequences (975) across samples before assessing microbial diversity. After excluding two participants without smoking pack-years, 779 were included in association analyses. FeNO was available for 767 participants. We used R version 4.1.0 to summarize characteristics of the study population and perform association analyses of

microbial diversity. We used functions *specnumber* and *diversity* in the vegan R package (v2.6.2) [30] to calculate the richness and the Shannon index, respectively. We set p -value < 0.05 as the threshold for statistical significance in diversity analyses.

Individual microbial taxa differentially abundant in relation to respiratory outcomes

To examine differential abundance of individual taxa in relation to pulmonary function or FeNO, we applied statistical inference techniques that provide accurate tests of hypotheses in large-scale data sets with high-dimensional predictors. We analyzed microbial abundance at the genus level to overcome the sparsity when examining individual species. To lessen the impact of extreme sequence reads on regression models, we used winsorization [31]; for each genus, sequence reads for the samples with five largest numbers of sequence reads were set to the sixth largest number of sequence reads. We converted abundance read counts to relative abundances for each genus and centered and scaled predictor and response variables to remove the need for an intercept term in the regression models. The same covariates were included as in our diversity analyses. We estimated coefficients with the Lasso estimator and tested hypotheses regarding predictor-response associations with a post-selection inference methodology [32, 33]. This method determined whether each microbial taxon (the predictor) was differentially abundant in relation to a respiratory outcome (the response variable) while controlling for all other microbial taxa using Lasso estimation to shrink coefficients of unimportant predictors toward zero. This procedure produces a test of significance of each taxon; but critically, each taxon is tested after removing the effects of all other taxa. Typical one-taxon-at-a-time testing does not involve either shrinkage of unimportant coefficients or adjustment for other taxa. Accordingly, we used a cutoff of p -value < 0.05 for statistical significance.

We used R version 3.4.0 for computation with optimization of the Lasso estimator performed by the software *mosek* wrapped through the R package *Rmosek*, which implemented fivefold cross-validation to choose the Lasso regularizer.

Results

Of the 779 participants, 60% were male. Participants were 62 years old on average and from NC (32%) or IA (68%) (Table 1). As expected, individuals with asthma exhibited statistically significantly lower lung function and higher FeNO than noncases (Table S3).

After quality-control filtering, our house dust metagenome data included 173,766,690 sequence reads, with 223,064 sequence reads per sample on average. The taxa

Table 1 Characteristics of study participants ($N = 779$)

Characteristic	N (%) or mean \pm SD
Age, years	62 \pm 11
Sex	
Male	468 (60)
Female	311 (40)
Height, cm	171 \pm 10
Weight, kg	90 \pm 21
Smoking status	
Never	506 (65)
Former	227 (29)
Current	46 (6)
Pack-years in ever smokers	21 \pm 23
Asthma status	
Yes	295 (38)
No	484 (62)
State of residence	
North Carolina	246 (32)
Iowa	533 (68)
Ancestry	
European	746 (96)
Not	33 (4)
Season of dust collection	
Winter	178 (23)
Not	601 (77)
Pulmonary function parameters	
FEV ₁ , ml	2649 \pm 853
FVC, ml	3620 \pm 1036
FEV ₁ /FVC	0.73 \pm 0.10
FeNO ^a , ppb	2.8 \pm 0.7

^a Fractional exhaled nitric oxide, available in 767 participants. Values were natural log (ln) transformed for normality

identified were 1264 species assigned to 389 genera. Most species (1260; 99.7%) were from 13 distinct phyla within Bacteria; only a few were from Archaea (Table S4). Of the 13 bacterial phyla, four predominated: Proteobacteria (39%), Actinobacteria (33%), Firmicutes (16%), and Bacteroidetes (11%). At the genus level, the genera *Staphylococcus* (phylum Firmicutes), *Pseudomonas* (phylum Proteobacteria), *Brevibacterium* (phylum Actinobacteria) were more abundant than other genera, and the three account for about 30% of the dust microbial communities. The three most abundant species were *Brevibacterium aurantiacum* (phylum Actinobacteria), *Cutibacterium acnes* (phylum Actinobacteria), and *Staphylococcus aureus* (phylum Firmicutes); each made up about 5–7% of the microbial communities. Microbial community compositions at the phylum level varied by sample (Fig. 1).

For overall diversity measures, the average richness was 241 (SD 40), and Shannon H index was 4.4 (SD 0.43)

(Figure S1). Overall microbial diversity in house dust was not significantly related to respiratory outcomes. For both richness and Shannon index, higher diversity was related to higher pulmonary function parameters (FEV₁ and FEV₁/FVC), but these associations were not statistically significant (Table S5). Similar patterns were seen for FeNO.

Many individual microbial genera were differentially abundant in relation to pulmonary function. Of the 389 genera examined, 76 were related to one or more pulmonary function parameters (p -value $<$ 0.05) in analyses that adjusted for all other microbial taxa (Fig. 2, Table 2). Most were from the bacterial phyla Actinobacteria ($N = 25$), Proteobacteria ($N = 23$), or Firmicutes ($N = 17$). Slightly more genera showed positive than negative associations: 55% for FEV₁, 62% for FVC, and 71% for FEV₁/FVC. Of the 76 genera, 22 were associated with two parameters, including *Ilumatobacter* (phylum Actinobacteria), *Chroococcidiopsis* (phylum Cyanobacteria), and *Anaerobutyrium* (phylum Firmicutes). The remaining 54 genera were uniquely associated with a specific parameter, including *Streptococcus* (phylum Firmicutes) and *Moraxella* (phylum Proteobacteria) (Table S6). Significant genera were largely from more abundant phyla including Proteobacteria and Actinobacteria (Fig. 3). Notably, of two genera from phylum Acidobacteria examined in this work, one *Luteitalea* showed significant associations with FEV₁ and FEV₁/FVC.

For FeNO, we found 30 differentially abundant genera (p -value $<$ 0.05) after controlling for all other microbial taxa, including *Stenotrophomonas* (phylum Proteobacteria) (Table 3). Notably, seven of these, including the bacterial genus *Limosilactobacillus* (phylum Firmicutes), were also related to a pulmonary function parameter (Table S6).

When we interrogated our data at the species level, we found 189 (of 1264) species (106 genera) significantly related to lung function and 80 species (51 genera) significantly related to FeNO (p -value $<$ 0.05 after accounting for all species examined together, Table S7). Of genera related to lung function and/or FeNO in the species level results, 63 containing 145 species were not significant in the genus level results. Of the 76 genera related to lung function in our genus level association results, 47 (62%) genera contained one or more significant species in the species level results (Table S8). Of the 30 genera associated with FeNO in the genus level results, 19 (63%) genera contained species exhibiting significant associations in the species level results (Table S9).

When we examined our 16S data [19], of the 76 genera related to lung function in our metagenome data, 31 were present. Of these, six, including *Moraxella* (phylum Proteobacteria), were significant (Table S10). Of the 30

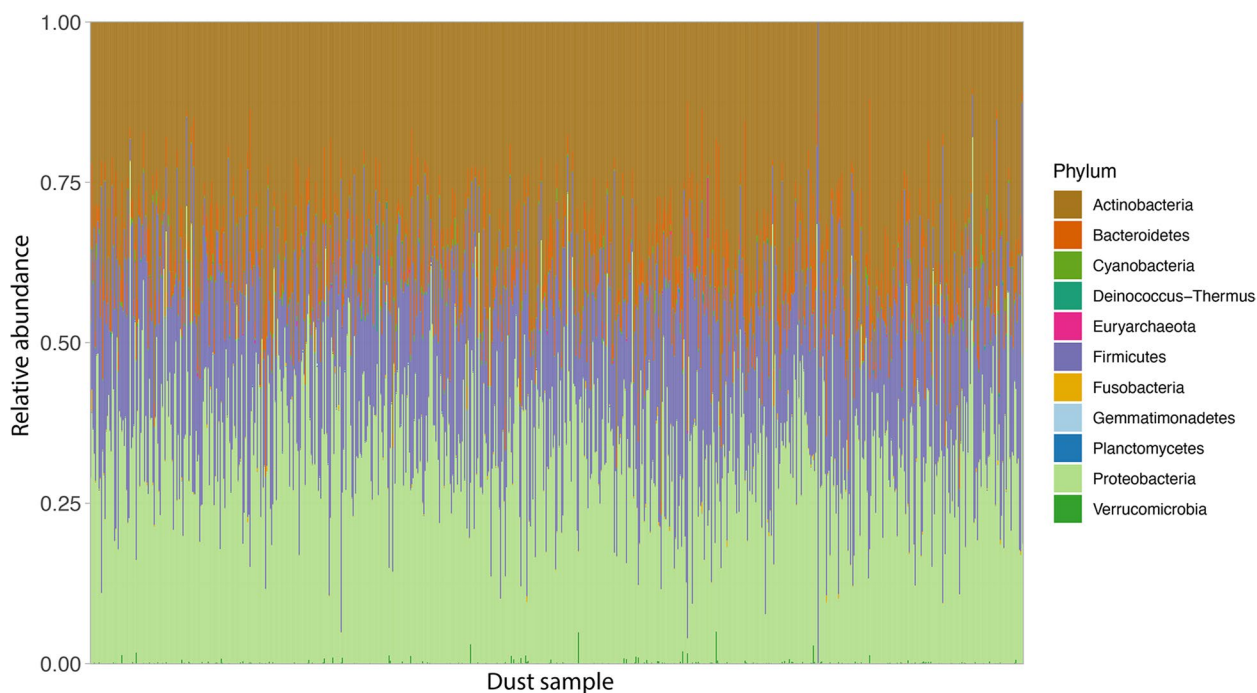


Fig. 1 Phylum level summary of microbial taxa across all dust samples ($n = 779$). This figure shows the phylum level summary of relative abundance in each sample. The x-axis indicates house dust samples examined, and the y-axis represents relative abundance at the phylum level

genera associated with FeNO in our metagenome data, 11 were present, and 1 genera *Stenotrophomonas* (phylum Proteobacteria) showed significance (Table S11).

Discussion

To our knowledge, this is the first large study to assess metagenome profiles in house dust using whole metagenome shotgun sequencing to examine associations of house dust microbiota with respiratory outcomes in adults. By comprehensively profiling microorganisms in house dust, we identified many individual microbial genera differentially abundant in relation to pulmonary function and/or airway inflammation. Among the genera that we identified as related to pulmonary function, several have been linked to pathogenesis of lung diseases previously, but others have not. Our results suggest that chronic exposure to specific microorganisms indoors may play a role in occupants' respiratory outcomes.

House dust contains diverse microbial profiles. In our study, bacteria from four phyla (Proteobacteria, Actinobacteria, Firmicutes, and Bacteroidetes) predominated; each phylum included > 10% of the total species. This finding is similar to that generated using 16S rRNA amplicon sequencing [19]. Our whole metagenome sequencing provides more accurate identification of microbial species and adds nonbacterial microorganisms not targeted in 16S technology [34].

There are few data on associations between microbial communities measured in house dust and pulmonary function and none using whole metagenome sequencing. While we were not able to identify an independent indoor dust microbiome dataset to replicate our findings, we found some of our results overlapping with findings from a recent oral microbiome study of pulmonary function in Norwegian adults [35]. That study used 16S rRNA amplicon sequencing data and examined categorized levels of pulmonary function (low vs normal) and airway inflammation (eosinophilic inflammation vs normal). From a look-up analysis of our differentially abundant genera in their results, we were able to validate some of our findings [35]. Of genera identified for FEV₁ and FVC in our dust data, the genus *Achromobacter* (phylum Proteobacteria) showed same directional (inverse) associations with the two lung function parameters ($FDR < 0.05$) in their oral microbiome data. Among genera we identified for FeNO, the genus *Janibacter* (phylum Actinobacteria) was negatively associated with FeNO in the oral microbiome data. The overlap between our findings and theirs provides partial validation for our findings.

Notably, several genera we identified as differentially abundant in relation to pulmonary function or airway inflammation have been linked to pathogenesis of lung diseases. The genus *Streptococcus* (phylum Firmicutes; family Streptococcaceae) was inversely associated

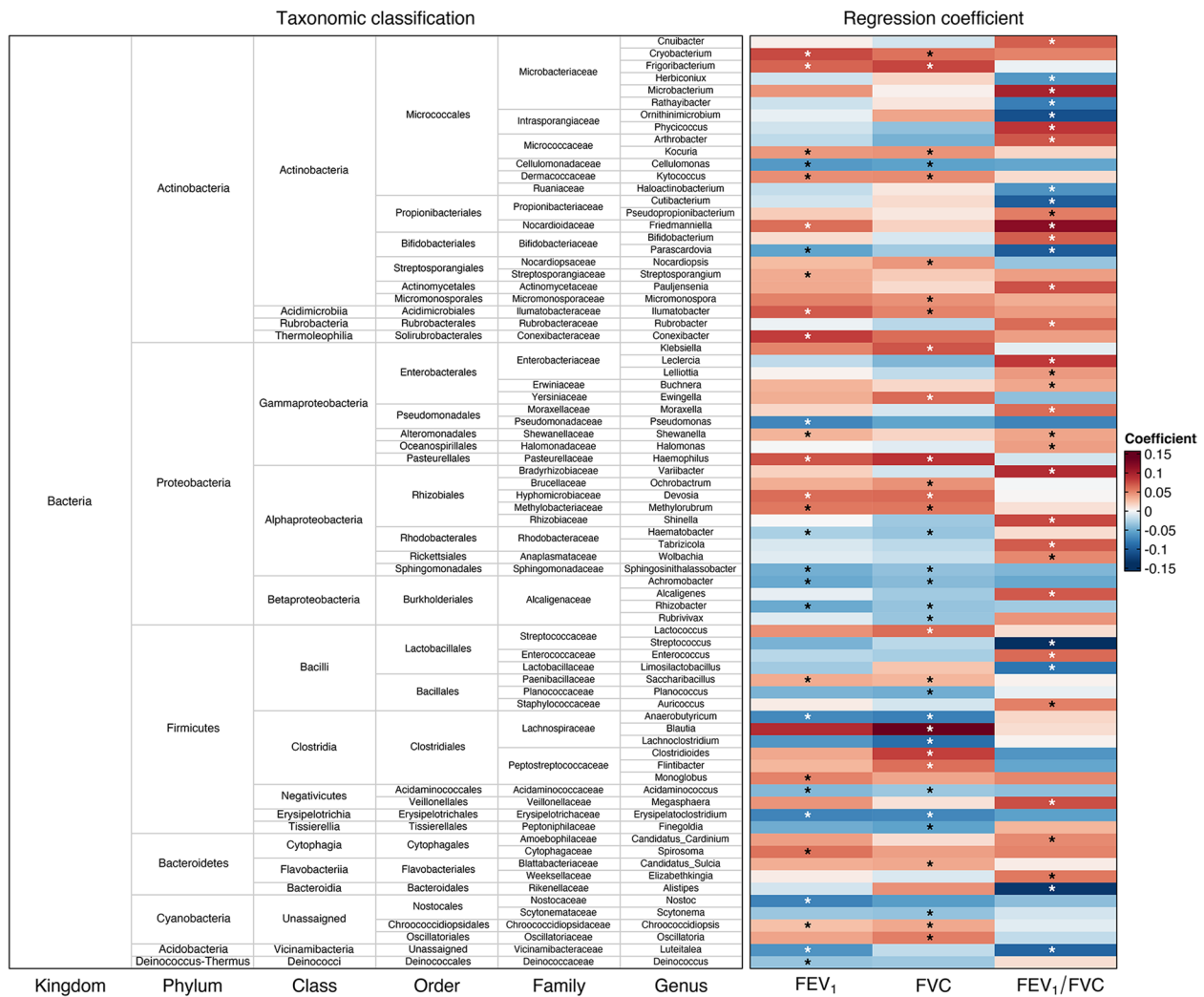


Fig. 2 Heatmap of genera differentially abundant in relation to one or more pulmonary function parameters. The three rightmost columns visualize regression coefficients with statistical significance separately by pulmonary function parameter (FEV₁, FVC, or FEV₁/FVC). The six leftmost columns provide taxonomic classification (from kingdom to genus) for each taxon. Coding for *p*-value threshold is * for *p*-value < 0.05 after controlling for all other microbial taxa examined

with FEV₁/FVC in our data; it contains several species, including *Streptococcus pneumoniae* and *Streptococcus pyogenes*, well known to cause pneumonia [36]. We also found the genus *Moraxella* (phylum Proteobacteria; family Moraxellaceae) related to FEV₁/FVC. *Moraxella* is a genus of gram-negative bacteria and includes *Moraxella catarrhalis*, frequently observed in sputum of COPD patients and related to asthma exacerbations [37, 38]. *M. catarrhalis*, a known pathogen, is not generally considered an environmental microbe because it requires a human host to survive; however, it can live in dried sputum in the environment for up to 3 weeks [39]. Our identification of an association of this organism in dust with a pulmonary outcome hints at potential interactions

between the human and indoor dust microbiomes. An additional genus positively associated with FEV₁/FVC was *Bifidobacterium*. A recent review recognized potential contributions of probiotics, including the genus *Bifidobacterium*, in management of respiratory diseases [40]. The genus *Limosilactobacillus* associated with both FEV₁/FVC and FeNO in our data includes *Limosilactobacillus reuteri*, which also has probiotic properties [41]. We observed the genus *Stenotrophomonas* (phylum Proteobacteria; family Xanthomonadaceae) significantly related to FeNO. The genus includes *Stenotrophomonas maltophilia*, a common multidrug-resistant organism related to severe lung infections in individuals with cystic fibrosis [42, 43].

Table 2 Genera differentially abundant (p -value < 0.05 controlling for all other microbial taxa examined) in relation to pulmonary function parameters

Phylum	Family	Genus	Coefficient	p -value
Trait: FEV₁				
Acidobacteria	Vicinamibacteraceae	<i>Luteitalea</i>	-0.066	0.012
Actinobacteria	Bifidobacteriaceae	<i>Parascardovia</i>	-0.055	0.012
Actinobacteria	Cellulomonadaceae	<i>Cellulomonas</i>	-0.062	0.004
Actinobacteria	Conexibacteraceae	<i>Conexibacter</i>	0.081	0.020
Actinobacteria	Dermacoccaceae	<i>Kytococcus</i>	0.046	0.038
Actinobacteria	Ilumatobacteraceae	<i>Ilumatobacter</i>	0.068	0.001
Actinobacteria	Microbacteriaceae	<i>Cryobacterium</i>	0.076	0.013
Actinobacteria	Microbacteriaceae	<i>Frigoribacterium</i>	0.064	0.005
Actinobacteria	Micrococcaceae	<i>Kocuria</i>	0.043	0.040
Actinobacteria	Nocardioideaceae	<i>Friedmanniella</i>	0.059	0.023
Actinobacteria	Streptosporangiaceae	<i>Streptosporangium</i>	0.035	0.031
Bacteroidetes	Cytophagaceae	<i>Spirosoma</i>	0.058	0.028
Cyanobacteria	Chroococciidiopsidaceae	<i>Chroococciopsis</i>	0.026	0.046
Cyanobacteria	Nostocaceae	<i>Nostoc</i>	-0.076	0.036
Deinococcus-Thermus	Deinococcaceae	<i>Deinococcus</i>	-0.037	0.049
Firmicutes	Acidaminococcaceae	<i>Acidaminococcus</i>	-0.043	0.022
Firmicutes	Erysipelotrichaceae	<i>Erysipelatoclostridium</i>	-0.075	0.029
Firmicutes	Lachnospiraceae	<i>Anaerobutyricum</i>	-0.072	0.045
Firmicutes	Paenibacillaceae	<i>Saccharibacillus</i>	0.034	0.049
Firmicutes	Peptostreptococcaceae	<i>Monoglobus</i>	0.051	0.022
Proteobacteria	Alcaligenaceae	<i>Achromobacter</i>	-0.052	0.005
Proteobacteria	Alcaligenaceae	<i>Rhizobacter</i>	-0.052	0.004
Proteobacteria	Hyphomicrobiaceae	<i>Devosia</i>	0.060	0.031
Proteobacteria	Methylobacteriaceae	<i>Methylorubrum</i>	0.055	0.004
Proteobacteria	Pasteurellaceae	<i>Haemophilus</i>	0.071	0.020
Proteobacteria	Pseudomonadaceae	<i>Pseudomonas</i>	-0.074	0.047
Proteobacteria	Rhodobacteraceae	<i>Haematobacter</i>	-0.029	0.030
Proteobacteria	Shewanellaceae	<i>Shewanella</i>	0.031	0.008
Proteobacteria	Sphingomonadaceae	<i>Sphingosinithalassobacter</i>	-0.049	0.010
Trait: FVC				
Actinobacteria	Cellulomonadaceae	<i>Cellulomonas</i>	-0.057	0.003
Actinobacteria	Dermacoccaceae	<i>Kytococcus</i>	0.047	0.025
Actinobacteria	Ilumatobacteraceae	<i>Ilumatobacter</i>	0.051	0.015
Actinobacteria	Microbacteriaceae	<i>Cryobacterium</i>	0.058	0.040
Actinobacteria	Microbacteriaceae	<i>Frigoribacterium</i>	0.078	0.001
Actinobacteria	Micrococcaceae	<i>Kocuria</i>	0.044	0.026
Actinobacteria	Micromonosporaceae	<i>Micromonospora</i>	0.045	0.044
Actinobacteria	Nocardiopsaceae	<i>Nocardiopsis</i>	0.043	0.026
Bacteroidetes	Blattabacteriaceae	<i>Candidatus_Sulcia^a</i>	0.036	0.049
Cyanobacteria	Chroococciidiopsidaceae	<i>Chroococciopsis</i>	0.036	0.007
Cyanobacteria	Oscillatoriaceae	<i>Oscillatoria</i>	0.053	0.033
Cyanobacteria	Scytonemataceae	<i>Scytonema^a</i>	-0.037	0.044
Firmicutes	Acidaminococcaceae	<i>Acidaminococcus</i>	-0.034	0.024
Firmicutes	Erysipelotrichaceae	<i>Erysipelatoclostridium</i>	-0.073	0.014
Firmicutes	Lachnospiraceae	<i>Anaerobutyricum</i>	-0.078	0.006
Firmicutes	Lachnospiraceae	<i>Blautia</i>	0.139	0.038
Firmicutes	Lachnospiraceae	<i>Lachnoclostridium</i>	-0.089	0.032

Table 2 (continued)

Phylum	Family	Genus	Coefficient	p-value
Firmicutes	Paenibacillaceae	<i>Saccharibacillus</i>	0.030	0.047
Firmicutes	Peptoniphilaceae	<i>Finegoldia</i>	-0.056	0.044
Firmicutes	Peptostreptococcaceae	<i>Clostridioides</i>	0.080	0.011
Firmicutes	Peptostreptococcaceae	<i>Flintibacter</i>	0.059	0.020
Firmicutes	Planococcaceae	<i>Planococcus</i> ^a	-0.050	0.014
Firmicutes	Streptococcaceae	<i>Lactococcus</i>	0.060	0.004
Proteobacteria	Alcaligenaceae	<i>Achromobacter</i>	-0.042	0.028
Proteobacteria	Alcaligenaceae	<i>Rhizobacter</i>	-0.036	0.025
Proteobacteria	Alcaligenaceae	<i>Rubrivivax</i>	-0.035	0.050
Proteobacteria	Brucellaceae	<i>Ochrobactrum</i> ^a	0.045	0.038
Proteobacteria	Enterobacteriaceae	<i>Klebsiella</i>	0.071	0.026
Proteobacteria	Hyphomicrobiaceae	<i>Devosia</i>	0.061	0.023
Proteobacteria	Methylobacteriaceae	<i>Methylorubrum</i>	0.056	0.006
Proteobacteria	Pasteurellaceae	<i>Haemophilus</i>	0.085	0.004
Proteobacteria	Rhodobacteraceae	<i>Haematobacter</i>	-0.036	0.007
Proteobacteria	Sphingomonadaceae	<i>Sphingosinithalassobacter</i>	-0.038	0.026
Proteobacteria	Yersiniaceae	<i>Ewingella</i>	0.061	0.046
Trait: FEV₁/FVC				
Acidobacteria	Vicinamibacteraceae	<i>Luteitalea</i>	-0.096	0.016
Actinobacteria	Actinomycetaceae	<i>Pauljensenia</i>	0.072	0.026
Actinobacteria	Bifidobacteriaceae	<i>Bifidobacterium</i>	0.067	0.005
Actinobacteria	Bifidobacteriaceae	<i>Parascardovia</i>	-0.103	0.003
Actinobacteria	Intrasporangiaceae	<i>Ornithinimicrobium</i>	-0.112	0.004
Actinobacteria	Intrasporangiaceae	<i>Phycococcus</i>	0.084	0.047
Actinobacteria	Microbacteriaceae	<i>Cnuibacter</i>	0.066	0.042
Actinobacteria	Microbacteriaceae	<i>Herbiconiux</i>	-0.064	0.016
Actinobacteria	Microbacteriaceae	<i>Microbacterium</i> ^a	0.099	0.020
Actinobacteria	Microbacteriaceae	<i>Rathayibacter</i>	-0.079	2.1 × 10 ⁻⁴
Actinobacteria	Micrococcaceae	<i>Arthrobacter</i>	0.070	0.043
Actinobacteria	Nocardiodaceae	<i>Friedmanniella</i>	0.117	0.001
Actinobacteria	Propionibacteriaceae	<i>Cutibacterium</i>	-0.101	0.011
Actinobacteria	Propionibacteriaceae	<i>Pseudopropionibacterium</i>	0.053	0.049
Actinobacteria	Ruaniaceae	<i>Haloactinobacterium</i>	-0.066	0.022
Actinobacteria	Rubrobacteraceae	<i>Rubrobacter</i>	0.061	0.012
Bacteroidetes	Amoebophilaceae	<i>Candidatus_Cardinium</i>	0.048	0.049
Bacteroidetes	Rikenellaceae	<i>Alistipes</i>	-0.132	0.025
Bacteroidetes	Weeksellaceae	<i>Elizabethkingia</i>	0.055	0.045
Firmicutes	Enterococcaceae	<i>Enterococcus</i>	0.061	0.008
Firmicutes	Lactobacillaceae	<i>Limosilactobacillus</i> ^a	-0.085	0.037
Firmicutes	Staphylococcaceae	<i>Auricoccus</i>	0.052	0.022
Firmicutes	Streptococcaceae	<i>Streptococcus</i>	-0.144	0.011
Firmicutes	Veillonellaceae	<i>Megasphaera</i>	0.073	0.049
Proteobacteria	Alcaligenaceae	<i>Alcaligenes</i>	0.069	0.012
Proteobacteria	Anaplasmataceae	<i>Wolbachia</i>	0.049	0.027
Proteobacteria	Bradyrhizobiaceae	<i>Variibacter</i>	0.091	0.002
Proteobacteria	Enterobacteriaceae	<i>Leclercia</i>	0.083	0.029
Proteobacteria	Enterobacteriaceae	<i>Lelliottia</i> ^a	0.042	2.7 × 10 ⁻⁴
Proteobacteria	Erwiniaceae	<i>Buchnera</i>	0.036	0.026
Proteobacteria	Halomonadaceae	<i>Halomonas</i>	0.040	0.012

Table 2 (continued)

Phylum	Family	Genus	Coefficient	p-value
Proteobacteria	Moraxellaceae	<i>Moraxella</i>	0.060	0.009
Proteobacteria	Rhizobiaceae	<i>Shinella</i>	0.076	0.023
Proteobacteria	Rhodobacteraceae	<i>Tabrizicola</i>	0.068	0.014
Proteobacteria	Shewanellaceae	<i>Shewanella</i>	0.037	0.006

Association results were obtained from linear regression with each microbial taxon as the predictor and a pulmonary function parameter as the outcome while controlling for all other microbial taxa. Therefore, a cutoff of p -value < 0.05 was used for statistical significance. Covariates were age, age squared, sex, height, height squared (weight for FVC only), cigarette smoking (former or current, both relative to never), pack-years of cigarette smoking, asthma status (yes/no), state of residence (NC/IA), and ancestry (European/not). For each taxon, a positive regression coefficient means that a change in the relative abundance results in a change in the measure trait in the same direction. Negative coefficients imply an inverse association between the relative abundance and the measured trait

^a Also associated with FeNO. See Table 3. All genera from kingdom Bacteria

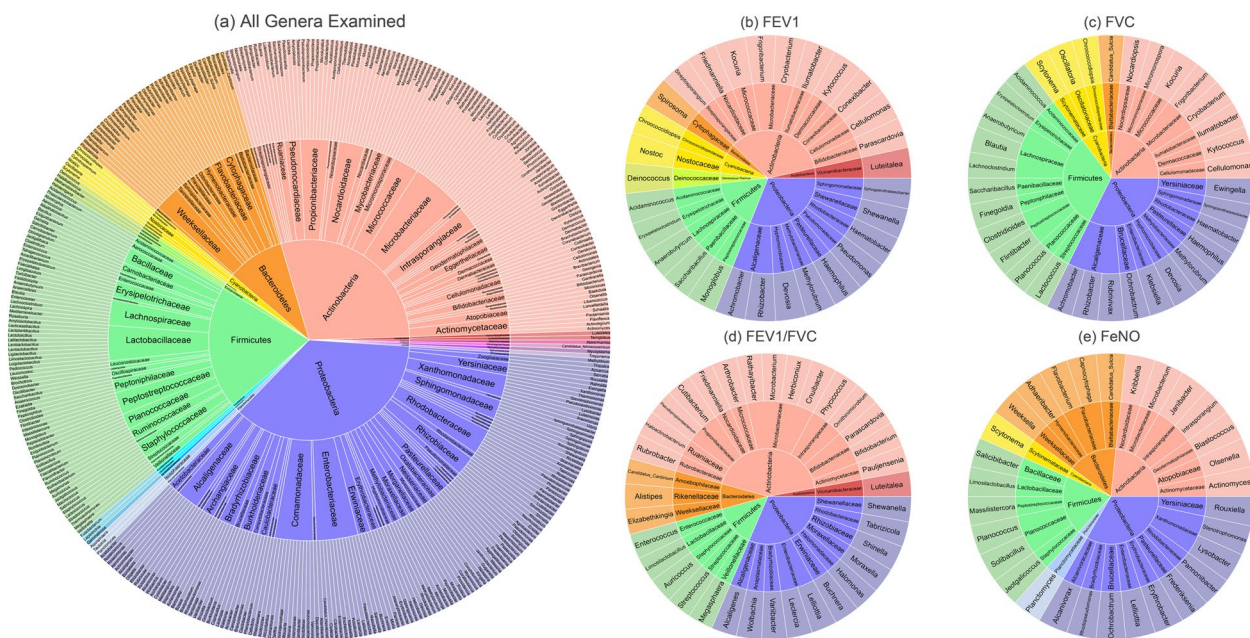


Fig. 3 Sunburst plots visualizing membership of microbial taxa significantly associated with a health outcome at different taxonomic levels. Each sunburst plot displays a hierarchical summary of phylum-, family-, and genus-level taxonomic name (from center to edge respectively) for **a** all genera examined and genera differentially abundant (p -value < 0.05 controlling for all other microbial taxa examined) in relation to each outcome: **b** FEV₁, **c** FVC, **d** FEV₁/FVC, and **e** FeNO

Findings from earlier studies of pulmonary function and microbial components using older technologies are limited in the number of microbial agents investigated [11–14]. Most studies analyzed bacterial or fungal agents with qPCR, but results are not comparable across studies because the microbial components examined differed [12–14]. We generated a comprehensive set of metagenomic profiles in house dust using whole metagenome sequencing and identified specific microbial signatures related to pulmonary function and airway inflammation.

Sparsity due to excessive zero counts in microbiome data poses challenges in statistical analysis. Applying an additional filtering step to remove rare taxa and analyzing

differential abundance data at the genus level enabled us to reduce spurious associations potentially induced by sparsity. Analysis at the species level results in greater sparsity. For some genera, we were able to identify specific species contributing to the significant associations at the genus level. We also found genera significant in the genus level results, but the species level analysis gave no significant species within that genus. These genus level associations were not driven by a single species but by the combined effects of different species.

Whole metagenome sequencing can capture microbial community composition with higher accuracy than 16S technology. In the same population, we identified

Table 3 Genera differentially abundant (p -value < 0.05 controlling for all other microbial taxa examined) in relation to FeNO

Phylum	Family	Genus	Coefficient	p -value
Actinobacteria	Actinomycetaceae	<i>Actinomyces</i>	-0.125	0.010
Actinobacteria	Atopobiaceae	<i>Olsenella</i>	0.106	0.019
Actinobacteria	Geodermatophilaceae	<i>Blastococcus</i>	-0.066	0.039
Actinobacteria	Intrasporangiaceae	<i>Intrasporangium</i>	-0.145	3.4×10^{-4}
Actinobacteria	Intrasporangiaceae	<i>Janibacter</i>	-0.069	0.037
Actinobacteria	Microbacteriaceae	<i>Microbacterium</i> ^a	-0.106	0.034
Actinobacteria	Nocardioideaceae	<i>Kribbella</i>	-0.081	0.012
Bacteroidetes	Blattabacteriaceae	<i>Candidatus_Sulcia</i> ^a	-0.073	0.033
Bacteroidetes	Flavobacteriaceae	<i>Capnocytophaga</i>	0.132	0.011
Bacteroidetes	Flavobacteriaceae	<i>Flavobacterium</i>	0.089	0.038
Bacteroidetes	Hymenobacteraceae	<i>Adhaeribacter</i>	-0.077	0.036
Bacteroidetes	Weeksellaceae	<i>Weeksella</i>	0.039	0.024
Cyanobacteria	Scytonemataceae	<i>Scytonema</i> ^a	0.074	0.046
Firmicutes	Bacillaceae	<i>Salicibibacter</i>	0.098	0.030
Firmicutes	Lactobacillaceae	<i>Limosilactobacillus</i>	-0.087	0.027
Firmicutes	Peptostreptococcaceae	<i>Massilistercora</i>	-0.076	0.017
Firmicutes	Planococcaceae	<i>Planococcus</i> ^a	0.070	0.035
Firmicutes	Planococcaceae	<i>Solibacillus</i>	-0.081	4.9×10^{-4}
Firmicutes	Staphylococcaceae	<i>Jeotgalicoccus</i>	0.124	0.027
Planctomycetes	Planctomycetaceae	<i>Planctomyces</i>	0.079	0.011
Proteobacteria	Alcanivoraceae	<i>Alcanivorax</i>	0.096	3.2×10^{-5}
Proteobacteria	Bradyrhizobiaceae	<i>Rhodopseudomonas</i>	0.071	0.035
Proteobacteria	Brucellaceae	<i>Ochrobactrum</i> ^a	0.086	0.007
Proteobacteria	Enterobacteriaceae	<i>Lelliottia</i> ^a	-0.046	8.0×10^{-4}
Proteobacteria	Erythrobacteraceae	<i>Erythrobacter</i>	-0.041	0.004
Proteobacteria	Pasteurellaceae	<i>Frederiksenia</i>	-0.151	6.4×10^{-5}
Proteobacteria	Rhodobacteraceae	<i>Pannonibacter</i>	0.075	0.023
Proteobacteria	Xanthomonadaceae	<i>Lysobacter</i>	0.096	0.011
Proteobacteria	Xanthomonadaceae	<i>Stenotrophomonas</i>	0.115	2.9×10^{-4}
Proteobacteria	Yersiniaceae	<i>Rouxiiella</i>	0.190	0.023

Association results were obtained from linear regression with each microbial taxon as the predictor and FeNO as the outcome while controlling for all other microbial taxa. Therefore, a cutoff of p -value < 0.05 was used for statistical significance. Covariates were age, sex, height, cigarette smoking (former or current, both relative to never), pack-years of cigarette smoking, asthma status (yes/no), state of residence (NC/IA), and ancestry (European/not). For each taxon, a positive regression coefficient means that a change in the relative abundance results in a change in the measure trait in the same direction. Negative coefficients imply an inverse association between the relative abundance and the measured trait

^a Also associated with at least one pulmonary function parameter (FEV₁, FVC, or FEV₁/FVC) (see Table 2). All genera from kingdom Bacteria

many more microbial taxa and significant associations using whole genome sequencing microbiome data [34] compared to 16S [19]. Similarly, in this investigation of continuous health outcomes, we found some genera significant in our metagenome data were also related to lung function and/or FeNO in our 16S data (p -value < 0.05 after accounting for all genus level taxa examined together). Given the different microbial taxa and compositions captured using the two sequencing methods and the unique nature of our statistical method of considering all identified taxa in the same model, validation of these microorganisms identified using metagenome

sequencing in our 16S data is reassuring and useful for comparison with prior 16S literature.

In a usual differential abundance analysis, researchers examine each microbial taxon one at a time and thus execute statistical testing with correction for the number of individual taxa included in the analysis. In contrast, the statistical method we implemented examines each microbial taxon after accounting for all microbial taxa identified in the data. The dust microbiome is inherently a mixture, and individuals are exposed to all of the identified organisms simultaneously. Therefore, our method of accounting for all the other taxa in dust may be more

appropriate and can provide additional insights into understanding the impact of this complex environmental exposure on human health.

The high-dimensional metagenome data analysis method we implemented addresses the limited availability of statistical analysis methods suitable for continuous outcomes and metagenome data also allowing for adjustment for covariates. While it is relatively new in the context of metagenome data analysis, it is built upon a strong statistical foundation [32, 33]. Using this high-dimensional method in the analysis of a continuous outcome is a strength of our study and should be useful to researchers studying impacts of the microbiome on other continuous outcomes and will advance metagenome analysis.

In our dust metagenome sequencing data, we observed 3% of sequence reads mapped to *Dermatophagoides farinae*, a known dust mite associated with asthma symptoms [44]. Examining relative abundance of this organism in relation to lung function and airway inflammation is of interest.

Recent studies suggest impacts of air pollutants on indoor microbiome in urban populations [45, 46]. Due to lack of air pollution data at the time of home visit, we are not able to assess how air pollution plays a role in associations between indoor microbiome and respiratory health. We also note that the expected variability in air pollution in this predominantly rural population would limit the ability to detect impacts on the indoor microbiome. Future research is warranted to disentangle the complex relationships between air pollution, indoor microbiome, and respiratory health.

This study has limitations. We analyzed a single dust sample from each house to reflect the usual home condition and measured only bedroom dust. Exposure assessment would be improved by collecting samples at multiple time points. Improved exposure assessment could enhance our ability to robustly detect associations between dust microbial composition and the outcomes under study. Because most individuals spend a large portion of their day in the bedroom, it is a highly relevant single location to sample. That our study population is from a US farming cohort could potentially limit the generalizability of our findings; however, because farm exposures contribute to higher microbial diversity [19, 47] in house dust, our farming population might improve our ability to identify associations with health outcomes. Because of the unique characteristics of our study — house dust metagenome characterized using whole genome sequencing together with adult respiratory outcomes — we could not identify a replication population. Therefore, we looked up taxa we identified in literature and found associations of some with lung diseases. We were unable to

examine absolute bacterial load using sequencing data. Although we increased the number of microorganisms examined by using whole metagenome sequencing, sequencing methods do not allow absolute quantification of microbial abundance. Nevertheless, our differential abundance method allowed identification of directions of associations. Though it would be useful, we lack matching human microbiome data. Besides, potentially impacting the human microbiome, however, the house dust metagenome is an environmental exposure that can directly impact health outcomes. Thus, the associations we observed are of interest. Finally, the cross-sectional observational study design limited our ability to draw causal inferences. Associations could reflect influences of microbial exposures on respiratory outcomes or alternatively reflect the influence of occupants' respiratory health on the microbial composition of house dust.

Key strengths of our study include the large sample size as the first indoor metagenomic study of pulmonary function in adults. We generated comprehensive metagenome profiles using whole genome shotgun sequencing. Compared to the older 16S rRNA amplicon sequencing which obtains information on operational taxonomic units (OTUs) based on sequence similarity, shotgun metagenomic sequencing can lead to more accurate detection of microorganisms by direct sequencing fragments of the genome. We applied several quality-control steps and filtering criteria to improve the quality of metagenome abundance data before association analyses; in particular, we removed poor quality sequence reads and rare microbial taxa. Our dataset with 252,595 (25%) zero microbial counts across all samples was much less sparse than typically seen for metagenome data [48]. Excluding rare taxa, reducing zero counts, and removing extreme outliers by winsorization should reduce false-positive findings. By using whole genome shotgun sequencing, we also captured nonbacterial profiles: 23 species from four phyla within Eukaryota, 4 from three phyla within Archaea, and 3 from two viral phyla. We applied recently developed inference methods that allowed us to examine associations between a high-dimensional predictor (house dust metagenome) and respiratory outcomes while accounting for the many taxa examined together. Because humans are exposed to the entire community of microorganisms, it is crucial to use an appropriate high-dimensional modeling approach like ours that probes the complex relationships among microorganisms. In addition to pulmonary function, we analyzed FeNO, a marker of airway inflammation. With no prior whole metagenome sequencing-based studies of microbial exposure, our findings of individual genera associated with pulmonary function and FeNO could

inform mechanisms between exposure to microbes indoors and respiratory inflammation.

Our study fills the knowledge gap in the current literature by validating previously known disease-related microorganisms in the study of continuous measures reflecting respiratory health as well as identifying novel associations. Microbes related to several lung diseases appear to correlate with lung function and/or airway inflammation regardless of disease status. Validation in additional population studies reinforced by mechanistic studies could provide the basis for actionable guidelines for farmers and public health professionals.

We found microbial signatures in house dust associated with continuous measures of pulmonary function and airway inflammation in adults. Although overall microbial diversity was not significantly related to pulmonary function, many specific genera were differentially abundant in relation to pulmonary function and/or airway inflammation. Further investigation of the genera identified could inform contributions of exposure to indoor microorganisms to respiratory health. This comprehensive investigation of microbial signatures in house dust and adult respiratory outcomes could help elucidate complex mechanisms of chronic exposure to house dust and respiratory health across the life course.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40168-024-01823-y>.

Additional file 1: Online methods: 1. Identification of differentially abundant taxa. 2. Whole genome shotgun metagenomic sequencing and quality control steps.

Additional file 2: Supplemental figure: **Figure S1.** Distributions of alpha diversity measures: richness and exponentially transformed Shannon H index. Supplemental tables: **Table S1.** Reference genomes used to evaluate host contaminant related sequence reads. **Table S2.** Taxonomic classification of 168 taxa removed after evaluation of potential contaminant DNA sequence using the decontam R package. **Table S3.** Pulmonary function parameters and airway inflammation by asthma status. **Table S4.** Taxonomic classification of the 1264 species, from Bacteria and Archaea, included in our metagenome analysis of house dust. **Table S5.** Overall microbial diversity in relation to pulmonary function parameters and FeNO. **Table S6.** Taxa differentially abundant (P-value<0.05 after accounting for all other taxa examined together) in relation to one or more pulmonary function parameters and/or FeNO. **Table S7.** Taxa differentially abundant (P-value<0.05 after accounting for all other taxa examined together) in relation to one or more pulmonary function parameters and/or FeNO: Species level analysis results. **Table S8.** Species level association results for the 76 genera related to lung function parameters (p-value<0.05) in the genus level differential abundance analysis results. **Table S9.** Species level association results for the 30 genera related to FeNO (P<0.05) in the genus level differential abundance analysis results. **Table S10.** Associations of lung function parameters with 31 genera examined in our 16S data. **Table S11.** Associations of FeNO with 11 genera examined in our 16S data.

Additional file 3: Online Supplement: Complete association results.

Acknowledgements

We appreciate study participants for their contribution to this research. We thank Drs. Shyamal D. Peddada of NIEHS for useful discussions and insight toward development of the statistical analysis; Frank Day of NIEHS for expert computational assistance; Jane Hoppin, ScD (North Carolina State University, Raleigh, NC, USA), for her important contribution to the Agricultural Lung Health Study during her tenure at NIEHS; Gail Ackermann (University of California San Diego) for assistance with metadata; and Greg Humphrey (University of California San Diego) for laboratory work.

Authors' contributions

ML and SJL conceived and designed the study. ML performed statistical analyses and wrote the first draft of the manuscript. AK supervised statistical analyses. QZ, AGP, and RK contributed to generation of whole genome metagenome sequencing data and initial quality control of the sequencing data. ZW and AMR contributed to additional quality control steps of the sequencing data. All authors including JMW, MR, CGP, LEBF, and DMU contributed to interpretation of the results and/or revision of the manuscript. All authors approved the final version.

Funding

Open access funding provided by the National Institutes of Health This work was supported by the Intramural Research Program of the National Institutes of Health, the National Institute of Environmental Health Sciences (Z01-ES049030 and Z01-ES102385) and the National Cancer Institute (Z01-CP010119B), and by American Recovery and Reinvestment Act funds.

Availability of data and materials

Metagenome data used for this analysis are available at the Sequence Read Archive (SRA) under project number PRJNA975673 (<https://www.ncbi.nlm.nih.gov/sra/>). Complete association results are in the [Online supplement](#).

Declarations

Ethics approval and consent to participate

The study was approved by the Institutional Review Board at the National Institute of Environmental Health Sciences. Written informed consent was obtained from all participants.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Immunity Inflammation and Disease Laboratory, National Institute of Environmental Health Sciences (NIEHS), Durham, NC 27709, USA. ²Department of Mathematics and Statistics, Washington State University, Pullman, WA, USA. ³Integrative Bioinformatics Support Group, NIEHS, Durham, NC, USA. ⁴School of Life Sciences, Arizona State University, Tempe, AZ, USA. ⁵Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University, Tempe, AZ, USA. ⁶Westat, Durham, NC, USA. ⁷Biostatistics and Computational Biology Branch, NIEHS, Durham, NC, USA. ⁸Department of Pediatrics, University of California San Diego, La Jolla, CA, USA. ⁹Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA. ¹⁰Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA. ¹¹Department of Bioengineering, University of California San Diego, La Jolla, CA, USA. ¹²Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA.

Received: 26 September 2023 Accepted: 25 April 2024

Published online: 18 July 2024

References

- Collaborators GBDCRD. Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Respir Med.* 2020;8:585–96.
- Shan Y, Wu W, Fan W, Haahtela T, Zhang G. House dust microbiome and human health risks. *Int Microbiol.* 2019;22:297–304.

3. Fu X, Ou Z, Sun Y. Indoor microbiome and allergic diseases: from theoretical advances to prevention strategies. *Eco Environ Health*. 2022;1:133–46.
4. Fu X, Ou Z, Zhang M, Meng Y, Li Y, Wen J, Hu Q, Zhang X, Norback D, Deng Y, et al. Indoor bacterial, fungal and viral species and functional genes in urban and rural schools in Shanxi province, China-association with asthma, rhinitis and rhinoconjunctivitis in high school students. *Microbiome*. 2021;9:138.
5. Lee MK, Wyss AB, Carnes MU, Richards M, Parks CG, Beane Freeman LE, Thorne PS, Umbach DM, Azcarate-Peril MA, Peddada SD, London SJ. House dust microbiota in relation to adult asthma and atopy in a US farming population. *J Allergy Clin Immunol*. 2021;147:910–20.
6. Wen Y, Wang D, Zhou M, Zhou Y, Guo Y, Chen W. Potential effects of lung function reduction on health-related quality of life. *Int J Environ Res Public Health*. 2019;16(2):260.
7. Neas LM, Schwartz J. Pulmonary function levels as predictors of mortality in a national sample of US adults. *Am J Epidemiol*. 1998;147:1011–8.
8. Wyss AB, Sofer T, Lee MK, Terzikhan N, Nguyen JN, Lahousse L, Latourelle JC, Smith AV, Bartz TM, Feitosa MF, et al. Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for pulmonary function. *Nat Commun*. 2018;9:2976.
9. Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, Batinic C, Fawcett KA, Song K, Sakornsakolpat P, et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet*. 2019;51:481–93.
10. Lee M, Huan T, McCartney DL, Chittoor G, de Vries M, Lahousse L, Nguyen JN, Brody JA, Castillo-Fernandez J, Terzikhan N, et al. Pulmonary function and blood DNA methylation: a multi-ancestry epigenome-wide association meta-analysis. *Am J Respir Crit Care Med*. 2022;206:321–36.
11. Carnes MU, Hoppin JA, Metwali N, Wyss AB, Hankinson JL, O'Connell EL, Richards M, Long S, Freeman LE, Sandler DP, et al. House dust endotoxin levels are associated with adult asthma in a U.S. farming population. *Ann Am Thorac Soc*. 2017;14:324–31.
12. Tischer C, Zock JP, Valkonen M, Doekes G, Guerra S, Heederik D, Jarvis D, Norback D, Olivieri M, Sunyer J, et al. Predictors of microbial agents in dust and respiratory health in the ECRHS. *BMC Pulm Med*. 2015;15:48.
13. Juel Holst G, Porneki A, Lindgreen J, Thuesen B, Bonlokke J, Hyvarinen A, Elholm G, Ostergaard K, Loft S, Brooks C, et al. Household dampness and microbial exposure related to allergy and respiratory health in Danish adults. *Eur Clin Respir J*. 2020;7:1706235.
14. McSharry C, Vesper S, Wymer L, Howieson S, Chaudhuri R, Wright GR, Thomson NC. Decreased FEV1 % in asthmatic adults in Scottish homes with high Environmental Relative Moldiness Index values. *Clin Exp Allergy*. 2015;45:902–7.
15. Casas L, Tischer C, Wouters IM, Torrent M, Gehring U, Garcia-Esteban R, Thiering E, Postma DS, de Jongste J, Smit HA, et al. Early life microbial exposure and fractional exhaled nitric oxide in school-age children: a prospective birth cohort study. *Environ Health*. 2013;12:103.
16. Vandenberg LE, Enaud R, Urien C, Coron N, Girodet PO, Ferreira S, Berger P, Delhaes L. Type 2-high asthma is associated with a specific indoor mycobiome and microbiome. *J Allergy Clin Immunol*. 2021;147(1296–1305):e1296.
17. Alavanja MC, Sandler DP, McMaster SB, Zahm SH, McDonnell CJ, Lynch CF, Pennybacker M, Rothman N, Dosemeci M, Bond AE, Blair A. The Agricultural Health Study. *Environ Health Perspect*. 1996;104:362–9.
18. House JS, Wyss AB, Hoppin JA, Richards M, Long S, Umbach DM, Henneberger PK, Beane Freeman LE, Sandler DP, Long O'Connell E, et al. Early-life farm exposures and adult asthma and atopy in the Agricultural Lung Health Study. *J Allergy Clin Immunol*. 2017;140(249–256):e214.
19. Lee MK, Carnes MU, Butz N, Azcarate-Peril MA, Richards M, Umbach DM, Thorne PS, Beane Freeman LE, Peddada SD, London SJ. Exposures related to house dust microbiota in a U.S. farming population. *Environ Health Perspect*. 2018;126:067001.
20. White JD, Wyss AB, Hoang TT, Lee M, Richards M, Parks CG, Beane-Freeman LE, Hankinson JL, Umbach DM, London SJ. Residential wood burning and pulmonary function in the Agricultural Lung Health Study. *Environ Health Perspect*. 2022;130:87008.
21. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. 2010. Available at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
22. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–9.
23. Beghini F, McIver LJ, Blanco-Miguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife*. 2021;10:65088.
24. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019;20:257.
25. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*. 2017;3:e104.
26. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*. 2018;6:226.
27. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods*. 2013;10:57–9.
28. Navas-Molina JA, Peralta-Sanchez JM, Gonzalez A, McMurdie PJ, Vazquez-Baeza Y, Xu Z, Ursell LK, Lauber C, Zhou H, Song SJ, et al. Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol*. 2013;531:371–444.
29. Shannon CE. A mathematical theory of communication. *Bell System Technical Journal*. 1948;27:379–423.
30. Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, et al. *Vegan: Community Ecology Package*. R package version 2.6-4. 2022.
31. Davidov O, Jelsema CM, Peddada S. Testing for inequality constraints in singular models by trimming or winsorizing the variance matrix. *J Am Stat Assoc*. 2018;113:906–18.
32. Chernozhukov V, Hansen C, Spindler M. Valid post-selection and post-regularization inference: an elementary, general approach. *Annual Review of Economics*. 2015;7:649–88.
33. Ning Y, Liu H. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of Statistics*. 2017;45:158–95.
34. Wang Z, Dalton KR, Lee M, Parks CG, Beane Freeman LE, Zhu Q, Gonzalez A, Knight R, Zhao S, Motsinger-Reif AA, London SJ. Metagenomics reveals novel microbial signatures of farm exposures in house dust. *Front Microbiol*. 2023;14:1202194.
35. Shigdel R, Johannessen A, Lin H, Peddada S, Gomez Real F, Ringel-Kulka T, Svanes C, Bertelsen RJ. Oral bacterial composition associated with lung function and lung inflammation in a community-based Norwegian population. *Respir Res*. 2023;24:183.
36. Krzysciak W, Pluskwa KK, Jurczak A, Koscielniak D. The pathogenicity of the *Streptococcus* genus. *Eur J Clin Microbiol Infect Dis*. 2013;32:1361–76.
37. Murphy TF, Brauer AL, Grant BJ, Sethi S. *Moraxella catarrhalis* in chronic obstructive pulmonary disease: burden of disease and immune response. *Am J Respir Crit Care Med*. 2005;172:195–9.
38. Sethi S, Evans N, Grant BJ, Murphy TF. New strains of bacteria and exacerbations of chronic obstructive pulmonary disease. *N Engl J Med*. 2002;347:465–71.
39. McGregor K, Chang BJ, Mee BJ, Riley TV. *Moraxella catarrhalis*: clinical significance, antimicrobial susceptibility and BRO beta-lactamases. *Eur J Clin Microbiol Infect Dis*. 1998;17:219–34.
40. Mortaz E, Adcock IM, Folkerts G, Barnes PJ, Paul Vos A, Garssen J. Probiotics in the management of lung diseases. *Mediators Inflamm*. 2013;2013:751068.
41. Abuqwidder J, Altamimi M, Mauriello G. *Limosilactobacillus reuteri* in health and disease. *Microorganisms*. 2022;10(3):522.
42. Ballesteros S, Virseda I, Escobar H, Suarez L, Baquero F. *Stenotrophomonas maltophilia* in cystic fibrosis patients. *Eur J Clin Microbiol Infect Dis*. 1995;14:728–9.
43. Ryan RP, Monchy S, Cardinale M, Taghavi S, Crossman L, Avison MB, Berg G, van der Lelie D, Dow JM. The versatility and adaptation of bacteria from the genus *Stenotrophomonas*. *Nat Rev Microbiol*. 2009;7:514–25.
44. Gehring U, Heinrich J, Jacob B, Richter K, Fahlbusch B, Schlenvoigt G, Bischof W, Wichmann HE, Indoor F, Genetics in Asthma Study G. Respiratory symptoms in relation to indoor exposure to mite and cat allergens and endotoxins. *Indoor Factors and Genetics in Asthma (INGA) Study Group*. *Eur Respir J*. 2001;18:555–63.
45. Chen Y, Fu X, Ou Z, Li J, Lin S, Wu Y, Wang X, Deng Y, Sun Y. Environmental determinants and demographic influences on global urban

microbiomes, antimicrobial resistance and pathogenicity. *NPJ Biofilms Microbiomes*. 2023;9:94.

46. Sun Y, Meng Y, Ou Z, Li Y, Zhang M, Chen Y, Zhang Z, Chen X, Mu P, Norback D, et al. Indoor microbiome, air pollutants and asthma, rhinitis and eczema in preschool children - a repeated cross-sectional study. *Environ Int*. 2022;161:107137.
47. Amin H, Santl-Temkiv T, Cramer C, Vestergaard DV, Holst GJ, Elholm G, Finster K, Bertelsen RJ, Schlunssen V, Sigsgaard T, Marshall IPG. Cow farmers' homes host more diverse airborne bacterial communities than pig farmers' homes and suburban homes. *Front Microbiol*. 2022;13:883991.
48. Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of microbiome data in the presence of excess zeros. *Front Microbiol*. 2017;8:2114.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.