

1 **Metagenomics reveals novel microbial signatures of farm exposures in** 2 **house dust**

3 **Ziyue Wang^{1†}, Kathryn R. Dalton^{2†}, Mikyeong Lee², Christine G. Parks², Laura E. Beane**
4 **Freeman³, Qiyun Zhu⁴, Antonio González⁵, Rob Knight^{5,6,7,8}, Shanshan Zhao¹, Alison A**
5 **Motsinger-Reif^{1#}, Stephanie J. London^{2#*}**

6 ¹ Biostatistics and Computational Biology Branch, National Institute of Environmental Health
7 Sciences, National Institutes of Health, Durham, NC, USA

8 ² Genomics and the Environment in Respiratory and Allergic Health Group, Epidemiology Branch,
9 National Institute of Environmental Health Sciences, National Institutes of Health, Durham, NC,
10 USA

11 ³ Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and
12 Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

13 ⁴ School of Life Sciences, Biodesign Center for Fundamental and Applied Microbiomics, Arizona
14 State University, Tempe, AZ, USA

15 ⁵ Department of Pediatrics, University of California San Diego, La Jolla, CA, USA

16 ⁶ Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA

17 ⁷ Department of Bioengineering, University of California San Diego, La Jolla, CA, USA

18 ⁸ Department of Computer Science and Engineering, University of California San Diego, La Jolla,
19 CA, USA

20 [†] Equal contribution co-first authors

21 [#] Co-senior authors

22 *** Correspondence:**
23 Stephanie J. London
24 london2@niehs.nih.gov

25 **Keywords: indoor microbiome, home dust microbiota, whole genome sequencing, farming**
26 **environmental exposures, Agricultural Health Study, environmental microbiology.**

27 **Abstract**

28 Indoor home dust microbial communities, important contributors to human health outcomes, are
29 shaped by environmental factors, including farm-related exposures. Detection and characterization of
30 microbiota are influenced by sequencing methodology; however, it is unknown if advanced
31 metagenomic whole genome shotgun sequencing (WGS) can detect novel associations between
32 environmental exposures and the indoor built-environment dust microbiome, compared to
33 conventional 16S rRNA amplicon sequencing (16S). This study aimed to better depict indoor dust
34 microbial communities using WGS to investigate novel associations with environmental risk factors

35 from the homes of 781 farmers and farm spouses enrolled in the Agricultural Lung Health Study. We
36 examined various farm-related exposures, including living on a farm, crop versus animal production,
37 and type of animal production, as well as non-farm exposures, including home cleanliness and indoor
38 pets. We assessed the association of the exposures on within-sample alpha diversity and between-
39 sample beta diversity, and the differential abundance of specific microbes by exposure. Results were
40 compared to previous findings using 16S. We found most farm exposures were significantly
41 positively associated with both alpha and beta diversity. Many microbes exhibited differential
42 abundance related to farm exposures, mainly in the phyla *Actinobacteria*, *Bacteroidetes*, *Firmicutes*,
43 and *Proteobacteria*. The identification of novel differential taxa associated with farming at the genera
44 level, including *Rhodococcus*, *Bifidobacterium*, *Corynebacterium*, and *Pseudomonas*, was a benefit
45 of WGS compared to 16S. Our findings indicate that characterization of dust microbiota, an
46 important component of the indoor environment relevant to human health, is heavily influenced by
47 sequencing techniques. WGS is a powerful tool to survey the microbial community that provides
48 novel insights on the impact of environmental exposures on indoor dust microbiota, and should be an
49 important consideration in designing future studies in environmental health.

50 **1 Introduction**

51 Humans spend 90% of their lives indoors (1), with much of this time spent in the home, where they
52 both contribute to and are exposed to environmental microbiota. Home dust microbiota are
53 commonly captured by vacuuming living spaces, including bedrooms. Exposure to bacterial and
54 fungal communities inside the home has been associated with allergic, atopic, and respiratory
55 conditions in children and adults (2-5). These associations could reflect the direct impacts of
56 environmental microbial exposure on inhabitants' health, as well as through indirect effects of dust
57 microbiota on the human gut, skin, oral, and respiratory microbiomes (6-8). Housing characteristics
58 and other environmental exposures have been shown to influence indoor microbial communities,
59 including farm-related exposures (8-11). Living in or near a farm environment entails unique
60 microbial exposures and subsequent health concerns. Farm exposures have been associated with
61 altered microbial composition in home dust, which in turn have been associated with allergic
62 outcomes in adults and children (4, 12-14). Identifying environmental factors that influence home
63 dust microbiota is a critical first step in determining exposure pathways relevant to health outcomes.

64 The emergence and optimization of high-throughput sequencing have enabled new approaches to
65 assessing the composition of bacterial communities present in home dust samples, which have a
66 complex matrix and low microbial biomass compared to host-associated microbiome samples such as
67 stool. 16S rRNA amplicon sequencing (16S) is a traditional next-generation technique in which all
68 amplified products are sequenced from a single gene (i.e., the 16S rRNA gene). The technique is
69 limited, however, because annotation is based on putative associations of the 16S rRNA gene with
70 bacterial taxa defined computationally as operational taxonomic units (OTUs). Thus, specific
71 bacterial entities are not directly sequenced, but rather predicted based on OTUs, and consequently
72 have more uncertainty at the lower taxonomy ranks of genus and species (15-18). Metagenomic
73 whole genome shotgun sequencing (WGS), in which random fragments of the genome are
74 sequenced, is an alternative approach and offers a major advantage in that taxa can be more
75 accurately defined at the genus/species level (16, 19). However, WGS is more expensive and requires
76 more extensive data processing and analysis (15, 20). Most of the published data on associations of
77 home dust microbiota with environmental exposures or health outcomes have relied on the older 16S
78 methodology.

79 Higher taxonomic classification resolution with WGS provides a more comprehensive description of
80 the microbial community, and may improve the ability to detect novel associations with
81 environmental risk factors, which is important when considering environmental health pathways. In
82 human microbial communities, especially the gut microbiome, WGS generally identifies a larger
83 number of unique phyla and higher overall microbial diversity within samples compared to 16S (16,
84 19-26). However, results are mixed for environmental samples in water and soil (27, 28). At present,
85 no research has evaluated sequencing methodology on microbial community characterization in
86 indoor home dust samples, and how this will impact the upstream associations with farm and non-
87 farm environmental exposures.

88 In the present study, we analyzed samples from 781 participant homes in the Agricultural Lung
89 Health Study (ALHS), a study of farmers and their spouses in North Carolina and Iowa, using
90 advanced WGS methods, and evaluated associations with farm and nonfarm exposures found to be
91 important in previous work based on 16S, in this cohort and others (4, 8, 29). We considered both
92 microbial community diversity levels and specific bacterial taxa, in order to determine whether WGS
93 can provide novel insights into farming environmental exposure pathways, the results of which are
94 relevant to the design of future research integrating environmental health and microbiology.

95 **2 Materials and methods**

96 **2.1 Study population and design**

97 ALHS is a case-control study of adult asthma study nested within the Agricultural Health Study
98 (AHS), a prospective cohort of licensed pesticide applicators, mostly farmers and their spouses,
99 enrolled between 1993 and 1997 (30). ALHS participants were selected from among AHS
100 participants who were either farmers or farm spouses in North Carolina (NC) and Iowa (IA) and
101 completed an AHS telephone follow-up conducted from 2005-2010. ALHS enrolled individuals with
102 asthma diagnosis and current asthma symptoms or medication use along with individuals with
103 symptoms and medication use suggesting likely asthma ($n = 1,223$). The comparison group was a
104 random sample of AHS participants without these criteria ($n = 2,078$). The Supplemental Methods
105 further details study population selection and inclusion criteria. The Institutional Review Board at the
106 National Institute of Environmental Health Sciences approved the study. Written informed consent
107 was obtained from all participants.

108 **2.2 Dust sample and environmental exposure data collection**

109 Of the 3,301 ALHS participants, 2,871 received a home visit and had adequate levels of collected
110 dust from the bedroom (Figure 1), as described in Carnes et al. (31). A trained field technician
111 vacuumed two 1-yd² (0.84-m²) areas—one on participants' sleeping surface and one on the floor next
112 to the bed—for 2 min each with a DUSTREAM Collector (Indoor Biotechnologies Inc.). The
113 samples were divided into aliquots of 50 mg and stored at -20°C until DNA processing.

114 During the home visit, information was obtained on environmental factors, including current (past 12
115 months) farming activities (living on a farm, working with crops, and working with animals), type of
116 animals raised on the farm (beef or dairy cattle, swine, or poultry) and the presence of indoor pets
117 (cats and dogs). Field technicians noted the presence of carpeting in the bedroom and ranked overall
118 home cleanliness on a standardized five-point scale (32). For our analysis, we created a binary
119 variable comprising poor/lower (score of 1 or 2) or good/higher (score of 3–5) home condition. We
120 categorized season of dust collection based on the date of the home visit: March 21–June 20 for

121 spring, June 21–September 20 for summer, September 21–December 20 for fall, and December 21–
122 March 20 for winter.

123 **2.3 DNA extraction**

124 A random selection (n=879, including 333 asthma cases) of dust samples were sent for WGS analysis
125 (Figure 1). DNA extraction is described elsewhere (4). Briefly, DNA was isolated using a MoBio 96
126 well plate PowerSoil DNA extraction kit (QIAGEN Inc.), as recommended by the manufacturer, with
127 the modification of loading 0.3-0.5g per dust sample into each well and incubated in PowerSoil bead
128 solution and C1 buffer at 70°C for 20 min before the beating step to aid in lysis of spores. We
129 quantified using the NanoDrop (A260) (Thermo Fisher Scientific Inc.) and normalized to 5 ng/L
130 DNA.

131 **2.4 Metagenomic whole genome shotgun sequencing and preprocessing**

132 The University of California San Diego IGM Genomics Center performed library preparation,
133 multiplexing, and whole genome shotgun sequencing using standard techniques (33). Extracted DNA
134 was quantified via Qubit™ dsDNA HS Assay (ThermoFisher Scientific). The library size was
135 selected for fragments between 300 and 700 bp using the Sage Science PippinHT and sequenced as a
136 paired-end 150-cycle run using an Illumina HiSeq2500 v2 in Rapid Run mode.

137 We performed several quality control steps, which are summarized in Supplementary Figure S1. We
138 first trimmed low-quality reads, duplicates, and adapters based on FastQC results (v0.11.5) (34). We
139 then identified and removed reads not from microbial genomes, as potential contaminant host
140 genomic sources (human, PhiX, cow, pig, chicken, turkey, horse, goat, sheep, dog, cat, and dust mite
141 genomes) (Supplementary Table S1) using Bowtie2 (35) and KneadData (v0.7.10) (36). We further
142 assessed the taxonomic classification of sequences using Kraken2 (v2.1.1) (37) and obtained accurate
143 estimations of abundance using Bracken (v2.5.0) (38) with pre-compiled data comprising RefSeq
144 genomes for bacteria, archaea, eukaryotes, fungi, viruses, and plasmids and NCBI taxonomy
145 information. Supplementary Tables S2 and S3 summarize the overall read sequence statistics and
146 proportion of host genome contaminants across samples. Additionally, we accounted for the potential
147 introduction of contaminant DNA sequences during sample collection or laboratory processing by
148 incorporating negative ‘blank’ sequencing controls of sterile water, with contaminants identified and
149 removed with the decontam R package (v1.10.0) (39). A total of 168 taxa were filtered out
150 (Supplementary Table S4). Because dust samples have low microbial biomass (fewer microbes), we
151 performed two sequencing runs, each with separate quality control processes, and then performed
152 abundance pooling across the two runs. At the sample level, we excluded low-quality samples
153 defined by sequencing depths less than 1000 (Supplementary Figure S2). Rare taxa were filtered out
154 if they did not appear in at least 10 samples (Supplementary Figure S2). This quality control pipeline
155 left 781 samples and 6,528 taxa for downstream analysis. A taxonomy chart was created that
156 assigned all taxa to a taxonomic classification across the seven phylogenetic levels - kingdom,
157 phylum, class, order, family, genus, and species. The Supplemental Methods provides details of the
158 bioinformatic procedures.

159 **2.5 Statistical analysis**

160 We performed all statistical analyses and visualization in R (v4.0.3) (40). We rarefied data to the
161 minimum library size (1,003) across all samples before calculating alpha and beta diversities using
162 the phyloseq R package (v1.34.0) (41). We considered both non-farming exposures, including state

163 of residence, sex, presence of indoor pets, home condition, and season of dust collection, and farming
164 exposures in the past 12 months, including living on a farm, crop farming, and animal farming. All
165 exposures were treated as binary variables. For season of dust collection, we compared one season to
166 all other seasons combined. We included asthma as a covariate in all models due to the nested case-
167 control design.

168 To evaluate intra-group alpha diversity and its association with farming and non-farming exposures
169 we used the Shannon index, exponentially transformed for normality, as the outcome in linear
170 models. We first fitted a baseline univariable regression model for each exposure to identify
171 exposures associated with alpha diversity. We also considered whether associations differed by state
172 of residence (IA or NC) by using product terms. Our final multivariable model included any exposure
173 with significant association to alpha diversity from the baseline univariable model, along with any
174 significant product terms for the individual interactions of each exposure with state of residence.
175 Detailed analytical formula were described in Supplemental methods (SM3). We set $p < 0.05$ as the
176 statistical significance threshold for all analyses.

177 To explore beta diversity, we calculated unweighted and weighted UniFrac distance metrics. We
178 conducted permutational multivariate analysis of variance (PERMANOVA) analysis to test the
179 differences in microbial community structure across exposure levels using the *adonis* method in the
180 R *vegan* package (v2.5.7) (42, 43). We used the R^2 value to quantify the percentage of variance
181 explained. We did similar analysis as alpha diversity to evaluate differences in associations by state.
182 We conducted non-metric multidimensional scaling (NMDS) analysis to visualize the separation
183 between samples by exposure levels in a two-dimensional space using the *phyloseq* (v1.34.0) (41)
184 and R *ggplot2* (v3.3.6) (44) packages.

185 To identify differentially abundant taxa for each exposure, we used analysis of composition of
186 microbiomes with bias correction (ANCOM-BC, v1.0.5) models (45), which is based on a linear
187 regression framework on the log transformed taxa counts, with exposures as dependent variables and
188 sampling fraction as an offset term. To account for variation in sequencing depth, we performed
189 normalization by estimating the sampling fraction using the ANCOM-BC built-in algorithm. We
190 tested taxa at the OTU level and summarized the results by genus and phylum rank. We also
191 calculated the log₂ fold-difference which is the ratio of the mean abundance after normalized by
192 ANCOM-BC across exposure levels. We controlled the false discovery rate (FDR) at 0.05 with the
193 Benjamini-Hochberg (BH) method (46). We determined a taxon to be significantly differentially
194 abundant if it had both $p < 0.05$ after FDR correction and had log₂ fold-difference larger than 1 or
195 smaller than -1. We performed sensitivity analyses to evaluate differences in associations by state of
196 residence.

197 Lee et al. (47) analyzed samples for the same population with 16S rRNA amplicon sequencing. To
198 examine differences of house dust microbial profile between these two methods, we compared the
199 taxonomic chart from our WGS data to the previous 16S data to determine the number of unique and
200 overlapping microbial organisms, at the phyla rank, detected by each sequencing method. We note
201 how common or rare the uniquely identified phyla were based on the frequency of assigned taxa and
202 the relative abundance across samples. In addition, we evaluated the differences between alpha
203 diversities (richness and Shannon index) generated by the two sequencing methods by calculating the
204 Spearman's correlation coefficient.

205 3 Results

206 3.1 Summary statistics for the study population and metagenomics characteristics

207 Table 1 summarizes the demographic characteristics and environmental exposures of the study
208 population. Iowa residents accounted for 68% of samples; North Carolina for 32% (247). Sixty
209 percent of participants were male. Indoor pets (dogs or cats) were present in 43% of homes. Most
210 homes (78%) were in good/higher cleanliness, and nearly all had carpeted floors (93%). Overall, 83%
211 of participants lived on a farm, 56% farmed crops in the past 12 months, and 51% worked with farm
212 animals in the past 12 months. Of the 401 (51%) participants who reported animal farming, 281
213 worked with beef cattle, 48 worked with dairy cattle, 120 worked with hogs, and 90 worked with
214 poultry. Overall, 31% of dust samples were obtained in summer, 25% in spring, 20% in fall, and 23%
215 in winter. Current asthma was present in 296 (37.9%) participants and the overall mean age of
216 participants was 62 years (standard deviation 11).

217 After filtering out samples with low sequencing depth and filtering out rare taxa, 781 samples and
218 6,528 taxa remained for downstream analysis with 183,025,561 reads across all samples. At the
219 Kingdom phylogenetic level, 5,661 taxa were assigned to Bacteria, 156 to Archaea, 96 to Eukaryota,
220 and 615 to viruses, with an average of 2,247 ($\pm 1,226$) taxa per sample ($n=781$). Figure 2 outlines the
221 phylum composition across all samples. Among the 59 phyla identified from WGS, 16 had relative
222 abundance greater than 1% in at least one sample (Figure 2, Supplementary Table S5). Phyla
223 *Firmicutes*, *Proteobacteria*, *Actinobacteria*, and *Bacteroidetes* were the most prominent among home
224 dust microbial communities. At lower taxonomy rank, 1789 unique genera were identified, where 36
225 had relative abundance greater than 10% in at least one sample. The five most abundant genera were
226 *Mycobacterium*, *Serratia*, *Toxoplasma*, *Lactobacillus*, and *Alcaligenes* (Supplementary Table S6).

227 3.2 House dust microbial community diversity analysis

228 Figure 3 shows the association between alpha diversity and each exposure. The presence of indoor
229 pets and farming status (living on a farm, crop farming, animal farming with beef cattle, hogs, and
230 poultry) were positively associated with alpha diversity, while good/higher home cleanliness was
231 negatively associated with alpha diversity ($p < 0.050$). State of residence had a suggestive significant
232 association with alpha diversity with $p = 0.057$. In our multivariable primary model including all
233 statistically significant exposures and all significant interaction terms with state of residence, living
234 on a farm and animal farming remained significantly positively related to alpha diversity
235 (Supplementary Table S7).

236 For beta-diversity, PERMANOVA analysis revealed significant differences in beta diversity for all
237 demographic characteristics and exposure levels based on unweighted UniFrac distance although the
238 percent variance explained by the exposure groups (R^2 values) were small (Supplementary Figure
239 S3). Current farming accounted for relatively greater explained microbial diversity variance (0.5%
240 for crop farming and 0.7% for animal farming) compared to other farm and nonfarm exposures
241 (Figure 4a, 4b). The differences in the microbial composition of home dust samples by state of
242 residence explained around 1% of the variance of bacterial communities ($p = 0.001$) (Figure 4c). The
243 results with weighted UniFrac distance were similar to unweighted metric (Supplementary Figure
244 S4).

245 3.3 Differential abundance analysis of individual taxa

246 There were 372 unique taxa belonging to 175 genera within 16 unique phyla, that were differentially
247 abundant in relation to at least one exposure (Supplementary Table S8, Supplementary Table S9,
248 Supplementary Figure S5). Animal farming and living on a farm were associated with more

249 differentially abundant taxa than non-farming exposures. Figure 5 includes volcano plots of
250 differentially abundant taxa related to the presence of indoor pets, living on a farm, crop farming, and
251 animal farming in the past 12 months, color coded by phylum. The top 10 taxa based on FDR values
252 are labeled by their genus rank. Working with hogs was identified with the greatest number of
253 differentially abundant taxa compared with other types of farming animals (Figure 5a, Supplementary
254 Figure S5).

255 Living on a farm was associated with differential abundance of 101 taxa (increased abundance for
256 100 taxa and decreased abundance for one taxon in genus *Dickeya*), which were mainly in phylum
257 *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria* (Figure 5b). Among the top 10 taxa,
258 two were in genus *Bifidobacterium*. The 26 differentially abundant taxa all had increased abundance
259 related to crop farming were mainly in phyla *Actinobacteria*, *Firmicutes*, and *Proteobacteria* (Figure
260 5c). The most significant taxa were genus *Methanobrevibacter* and *Jeotgalibaca*. Animal farming
261 was associated with increased abundance for 191 taxa and decreased abundance for one taxon in
262 phylum *Firmicutes* (Figure 5d). Genera *Methanobrevibacter*, *Jeotgalibaca*, *Corynebacterium*,
263 *Chryseobacterium*, *Glutamicibacter*, *Pseudomonas*, and *Rhodococcus* were among the top 10 taxa.
264 Forty-nine taxa were differentially abundant for the presence of indoor pets, mostly in phylum
265 *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Fusobacteria*, and *Proteobacteria* (Figure 5e). The taxa
266 with the smallest FDR value were genus *Frederiksenia* and *Poerphyromonas*. Only a few
267 differentially abundant taxa belonging to phylum *Proteobacteria* were related to the season of dust
268 collection (Supplementary Table S8, Supplementary Figure S5).

269 Many differentially abundant taxa were shared among exposures, but there were some taxa uniquely
270 related to individual farming exposures (Figure 6, Supplementary Table S9). In particular, there were
271 103 taxa assigned to 67 genera within 7 phyla (*Proteobacteria*, *Actinobacteria*, *Bacteroidetes*,
272 *Euryarchaeota*, *Firmicutes*, *Tenericutes*, *Chloroflexi*) specific to animal farming. For crop farming, 2
273 taxa were unique – *Tatumella citrea* in phylum *Proteobacteria* and *Fusarium graminearum* in
274 phylum *Ascomycota* (Supplementary Table S9). There were only 4 taxa (*Bacillus* [*Firmicute* phyla],
275 *Campylobacter* [*Proteobacteria*], *Streptomyces* [*Actinobacteria*], and *Acholeplasma* [*Tenericutes*])
276 that were identified to be associated with both animal farming and crop farming (Supplementary
277 Table S9). In terms of specific type of farm animals, 89 taxa were unique to hogs, including
278 *Clostridium*, *Campylobacter*, *Pseudomonas*, and *Streptococcus suis*, 14 unique to poultry, including
279 *Enterococcus*, *Brucella*, and *Escherichia* genera, 5 unique to dairy cattle, including *Mycoplasma* and
280 *Acinetobacter*, and 26 unique to beef cattle, including *Corynebacterium* and *Bacillus* (Supplementary
281 Table S9). Several taxa were identified in multiple types of farming animals: 15 taxa were shared for
282 hogs, beef cattle and dairy cattle, only one taxon (*Carnobacterium sp._CPI*) were common among
283 hogs, poultry, and beef cattle, and 24 taxa including *Methanobrevibacterium* was related to either
284 cattle type (Figure 6, Supplementary Table S9).

285 As for non-farming exposures, 44 taxa were uniquely differentially abundant for presence of indoor
286 pets, including animal-related *Staphylococcus* species *pseudintermedius* and *felis*. Additionally, 4
287 taxa were unique to home condition, 16 unique to carpeting, and 3 unique to spring dust collection
288 (Supplementary Table S9).

289 3.4 Sensitivity Analysis by State of Residence

290 For interaction effects by state of residence with either alpha or beta diversity, only sex, home
291 condition, crop farming, general animal farming, beef cattle farming, and spring dust collection had
292 significant interactions, but most effect sizes were minimal (Supplementary Table S10,

293 Supplementary Table S11, Supplementary Table S12). Therefore, we did not carry interaction
294 products into the differential abundance analysis. When stratifying by state of residence, several
295 exposures, including the presence of indoor pets, living on a farm, and general animal farming, were
296 significantly associated with either alpha or beta diversity in Iowa, where about 2/3 of participants
297 resided but not in North Carolina which has a much smaller sample size (Supplementary Table S13,
298 Supplementary Table S14). Fourteen phyla were consistent for both states with differentially
299 abundant taxa by at least one exposure (Supplementary Table S8, Supplementary Figure S6,
300 Supplementary Figure S7).

301 **3.5 Additional findings with WGS from 16S rRNA sequencing results**

302 WGS data identified many more taxa and phyla than 16S rRNA. The 6,526 taxa identified by WGS
303 data were assigned to 59 phyla, compared to 1,346 taxa from 18 phyla for 16S. The three phyla with
304 the largest proportion of taxa assignment (most frequent) for WGS results (*Proteobacteria*,
305 *Actinobacteria*, *Firmicutes*) were identical for 16S results. Among the 18 phyla identified from 16S
306 sequencing, 17 were present in the WGS results (Figure 7, Supplementary Table S5). 47 phyla were
307 uniquely identified by WGS, of which the most frequent phyla were *Uroviricota* with 518 (7.9%)
308 taxa assigned, *Ascomycota* with 51 (0.8%) taxa assigned, *Spirochaetes* with 38 (0.6%) taxa assigned,
309 *Cossaviricota* with 35 (0.5%) taxa assigned, and *Apicomplexa* with 25 (0.4%) taxa assigned
310 (Supplementary Table S5). Additionally, many of the unique phyla in WGS were not rare, including
311 *Apicomplexa* with average relative abundance across all samples at 3%, and *Ascomycota*,
312 *Cossaviricota*, *Basidiomycota*, *Nucleocytoviricota*, and *Uroviricota* at 2% each (Supplementary
313 Table S5). When examining differences in the alpha diversity of results from WGS and 16S
314 sequencing, Spearman's correlation coefficient for richness ($\rho=0.413$, $p<2.2e-16$) and the Shannon
315 index ($\rho=0.355$, $p<2.2e-16$) were moderate.

316 Because more microbial organisms were detected by WGS, we observed additional associations with
317 farming exposures compared to 16S data presented by Lee et al. (39). Notably, a unique phylum
318 (*Ascomycota*) detected only by WGS was significantly associated with crop farming. One of phyla
319 identified by both WGS and 16S (*Tenericutes*) had differentially abundant taxa based on animal
320 farming using WGS not with 16S (Supplementary Table S5, Supplementary Table S8). In addition,
321 WGS provided the ability to assign taxa to genus taxonomic levels, including the 175 genera with
322 differential abundance taxa related to at least one exposure (Supplementary Table S8), compared to
323 16S results at the phyla and family level. Of 175 genera, 16 had relative abundance greater than 10%
324 in at least one sample including *Lactobacillus*, *Staphylococcus*, and *Bacillus* (Supplementary Table
325 S6, Supplementary Table S8).

326 **4 Discussion**

327 In this study, we evaluated the associations between farming exposures and house dust microbiota
328 using the whole genome shotgun sequencing method in a US agricultural population. Our results
329 indicate that both indoor microbial diversity and composition in homes differ in relation to current
330 farming exposures; living on a farm, and crop and animal farming were associated with increased
331 within-sample microbial diversity levels and altered microbial composition. Expanding on our
332 previous findings performed with 16S rRNA gene amplicon sequencing, we identified four times
333 more unique microbial taxa. The improved detection of unique taxa with WGS enabled us to detect
334 novel associations between farm exposures and increased abundance of specific microbes including
335 *Rhodococcus*, *Bifidobacterium*, *Corynebacterium*, and *Pseudomonas*. Enhanced identification of

336 factors that impact the indoor microbiome can improve understanding of environmental exposure
337 pathways relevant to human health.

338 A unique aspect of this study was the use of the whole genome shotgun sequencing technique,
339 compared to many previous home dust microbiome studies that use the 16S rRNA amplicon
340 sequencing technique (4, 12). This work is the first reported to use WGS to evaluate farm exposures
341 in home dust microbiota. WGS has the advantage of sequencing the entire microbial genome, versus
342 just a single gene, which can more accurately assign taxonomic classifications (48). In this study, the
343 use of WGS identified more unique microbial phyla – 42 phyla were found only using WGS,
344 including both common and rare taxa, versus only one phylum using the 16S technique. Detection of
345 a greater number of unique phyla from WGS compared to 16S enables better characterization of the
346 mixed, complex microbial composition of indoor dust in homes. Consequently, we observed novel
347 environmental exposure associations with the newly detected microbial outcomes from this more
348 comprehensive WGS method. Expanded taxonomic detection and depiction, as well as the
349 development of updated, robust bioinformatic and statistical tools for metagenomic data (49), will
350 then have downstream effects on the interpretation of association to environmental exposures.

351 Consistent with findings using 16S, our data with WGS found that numerous bacteria were
352 associated with environmental exposures across various phyla. At the phyla level, *Actinobacteria*,
353 *Bacteroidetes*, *Firmicutes*, and *Proteobacteria* were positively associated with farm exposures,
354 including living on a farm and crop and animal farming. These trends are similar to our findings
355 using 16S, which found *Firmicutes* and *Proteobacteria* to be associated with farm exposures. In
356 previous research, these phyla have been associated with various health conditions, such as asthma,
357 atopy, and cardiometabolic outcomes (50-52). However, our 16S findings found that crop farming
358 was associated with significant decreased abundance of taxa in 16 of the 19 phyla (4), compared to
359 using WGS, where all 26 of our significantly associated taxa had an increased abundance with crop
360 farming. Complementary studies evaluating home dust in Germany and Finland (12) and classroom
361 dust in China (53) have found positive associations between nearby farm exposure and increased
362 abundance of *Proteobacteria* (also known as *Alphaproteobacteria*) and *Actinobacteria*.

363 WGS enables improved classification of microbial taxa at lower taxonomic levels, including the
364 identification of genera that are differentially abundant by environmental exposures. Using WGS, we
365 ascertained genera that were associated with our farming exposures, including *Rhodococcus*,
366 *Bifidobacterium*, *Corynebacterium*, and *Pseudomonas*. *Rhodococcus* and *Corynebacterium*, gram-
367 positive bacteria, and *Pseudomonas*, a gram-negative bacterium, are found commonly in
368 environmental sources (54-56). Certain strains of each can be pathogenic in immunocompromised
369 individuals (54-56), and their abundance has been shown to be elevated in dust from children with
370 asthma and atopy (57). *Pseudomonas* was also found to be increased using WGS in classroom dust
371 samples in rural regions near farms compared to suburban areas in China (53). Interestingly,
372 *Rhodococcus*, *Pseudomonas*, and *Methylobacterium* (another microbe positively associated with
373 farm exposures in our data) have been previously identified in agricultural settings, where they can
374 be bioremediation agents and degrade certain pesticides (58). *Bifidobacterium* is ubiquitous in the
375 human and animal gastrointestinal tract and is associated with positive gut homeostasis, inhibition of
376 pathogen colonization, and modulation of the local and systemic immune system (59, 60). We
377 observed that *Methanobrevibacter* and *Jeotgalibaca*, both previously associated with cattle rumen
378 and manure fermentation (61, 62), were increased with crop and animal farming, and unique to dairy
379 and beef cattle farming, which is consistent with previous studies evaluating farm exposures in
380 human microbial communities (12, 63, 64). Two taxa unique to crop farming, *Tatumella citrea* and
381 *Fusarium graminearum*, are pathogens associated with grain production (65, 66). Reassuringly, we

382 noted an increased abundance of microbes specific to farm and companion animals associated with
383 concurrent exposure to those animals, such as *Streptococcus suis* with hog farming exposure (67) and
384 *Staphylococcus pseudintermedius* and *felis* with dog and cat exposure (68, 69).

385 Our findings suggest that the home dust microbial diversity levels differ between participants
386 exposed to farming activities, as well as pets, both for alpha and beta diversity levels. Overall, the
387 findings from this study were generally similar to those performed previously using 16S (4). For
388 microbial composition beta diversity, we found distinct microbial community structure based on farm
389 and non-farm exposures, which was significant for all explored variables, similar to results from 16S.
390 The coefficient-of-determination R-squared (R²) statistic was greater using 16S, which supports the
391 hypothesis that WGS resulted in more diverse microbial community identification with greater
392 heterogeneity, so the same exposure would account for less of the variability. Both WGS and 16S
393 findings had low R² explained variance, consistent with previous research (70). Both analyses
394 showed positive associations between alpha diversity and crop and animal farming. Living on a farm
395 was a significant factor using WGS but not 16S. In addition, there were differences based on the type
396 of animal production, with hog production having a positive association using WGS but not 16S, and
397 dairy cattle production having a positive association using 16S but not WGS (although there was a
398 positive trend).

399 The differences in associations between exposures and Shannon alpha diversity in the WGS
400 compared to our previous 16S data are to be expected given differences between the methods and
401 batch effects when comparing two different methods run three years apart in different laboratories.
402 Alpha diversity was slightly higher in WGS than 16S samples with moderate correlation (Spearman's
403 rho=0.36); unsurprisingly, as a greater number of unique microbes were identified with WGS and is
404 similar to previous research on environmental samples (19). The discrepancies in measurements and
405 effect sizes between WGS and 16S can lead to altered interpretations regarding risk factors for
406 dysbiosis in home dust microbial composition and highlights the importance of how the processing of
407 microbiome samples can impact downstream analyzes.

408 The positive associations with farm exposures and alpha diversity reinforce trends observed in other
409 literature (3, 10, 12, 14, 53), in addition to our prior 16S analyses (4). In a study of 203 homes in
410 Finland and Germany, homes located on farms had significantly higher indoor microbial richness and
411 diversity compared to rural non-farm home indoor dust, which was associated with decreased asthma
412 risk in child inhabitants (12). Amin et.al. reported that airborne bacterial diversity was more abundant
413 in farmer's indoor environment than in suburban homes (10). Using WGS, a study in Shanxi
414 Providence, China, found higher microbial diversity in schools in rural area near farms compared to
415 urban non-farm schools (53).

416 A limitation of this work is that we only have a single dust sample per household, collected in the
417 bedroom. Thus, we assume the sample reflects the normal home condition. To the extent that
418 microbial composition differs across the household (11), this may not be true. However, people
419 spend about a third of their time in the bedroom, making this a logical single sampling location. This
420 limitation would be expected to lead to nondifferential misclassification of exposure and a bias
421 toward the null. Our work benefits from an advanced next-generation technique, whole genome
422 shotgun sequencing, to explore the impact of detailed farm exposures on the indoor microbiome in a
423 large sample size compared to previous studies. The improved detection from WGS across novel
424 phyla at the genus level adds insights on factors influencing the built environment microbiota, which
425 plays a key component on host microbiome composition and subsequent health outcomes. Future
426 investigations on the functional capabilities of the dust microbiota, such as presence of antibiotic

427 resistance genes, can help better understand human health and disease etiology caused by
428 environmental exposures.

429 **5 Conclusions**

430 We evaluated a comprehensive set of factors related to farming to determine their influence on home
431 dust microbiome assessed using state of the art whole genome shotgun sequencing. The increased
432 identification by WGS of microbial entities led to detection of associations missed using older 16S
433 technology. Identifying significant predictors of indoor built environmental microbiota is an
434 important element in understanding environmental exposure health pathways. The use of advanced
435 whole genome shotgun sequencing techniques produced novel insights into these health pathways
436 and may be considered an optimal metagenomic method for future environmental health studies.

437 **6 Conflict of Interest**

438 The authors declare that the research was conducted in the absence of any commercial or financial
439 relationships that could be construed as a potential conflict of interest.

440 **7 Author Contributions**

441 SL and ML were responsible for study design and data acquisition. CP and LF initiated ALHS study
442 and were responsible for the sample collection. ZW designed and performed all bioinformatics and
443 statistical analysis, with SZ, AM, KD, SL and ML providing analytical input. QZ, AG and RK
444 planned shotgun metagenomics sequencing and prepared raw sequences data. ZW and KD
445 formulated the research ideas and drafted the manuscript. All authors contributed to the interpretation
446 of results and editing of the manuscript.

447 **8 Funding**

448 This work was supported by the Intramural Research Program of the National Institutes of Health
449 (NIH), the National Institute of Environmental Health Sciences (NIEHS) (Z01-ES049030 and Z01-
450 ES102385), the National Cancer Institute (Z01-CP010119B), and by American Recovery and
451 Reinvestment Act funds.

452 **9 Ethics**

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

455 **10 Acknowledgments**

456 This work was supported by the Intramural Research Program of the National Institutes of Health
457 (NIH), the National Institute of Environmental Health Sciences (NIEHS) (Z01-ES049030 and Z01-
458 ES102385), the National Cancer Institute (Z01-CP010119B), and by American Recovery and
459 Reinvestment Act funds. The Center for Microbiome Innovation at the University of California San
460 Diego provided support by generating sequencing data. We appreciate all the study participants for
461 their contribution to this research. We thank Drs. F. Day of NIEHS for expert computational
462 assistance and J. Hoppin (North Carolina State University, Raleigh, NC) for her important
463 contribution to the Agricultural Lung Health Study during her tenure at NIEHS. We thank Dr. Gail

464 Ackermann of UCSD for assistance with metadata curation, and Dr. Greg Humphrey for laboratory
465 processing.

466 **11 Supplementary Material**

467 See Supplemental Materials Table of Contents document for a list of the supplementary methods,
468 tables, and figures referenced.

469 **12 Data Availability Statement**

470 The raw sequencing data presented in this study are stored in online platform Qiita. Contact the
471 corresponding author for permission for the raw sequence data and metadata.

472 **References**

- 473 1. U.S. Environmental Protection Agency. Report to Congress on indoor air quality. 1989.
474 Contract No.: EPA/400/1-89/001C.
- 475 2. Ege MJ, Mayer M, Normand AC, Genuneit J, Cookson WO, Braun-Fahrländer C, et al.
476 Exposure to environmental microorganisms and childhood asthma. *N Engl J Med*. 2011;364(8):701-
477 9.
- 478 3. Stein MM, Hrusch CL, Gozdz J, Igartua C, Pivniouk V, Murray SE, et al. Innate Immunity
479 and Asthma Risk in Amish and Hutterite Farm Children. *New England Journal of Medicine*.
480 2016;375(5):411-21.
- 481 4. Lee MK, Carnes MU, Butz N, Azcarate-Peril MA, Richards M, Umbach DM, et al.
482 Exposures Related to House Dust Microbiota in a U.S. Farming Population. *Environ Health Perspect*.
483 2018;126(6).
- 484 5. Dannemiller KC, Gent JF, Leaderer BP, Peccia J. Indoor microbial communities: Influence
485 on asthma severity in atopic and nonatopic children. *Journal of Allergy and Clinical Immunology*.
486 2016;138(1):76-83.e1.
- 487 6. Gupta S, Hjelmsø MH, Lehtimäki J, Li X, Mortensen MS, Russel J, et al. Environmental
488 shaping of the bacterial and fungal community in infant bed dust and correlations with the airway
489 microbiota. *Microbiome*. 2020;8(1):115.
- 490 7. Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, et al.
491 Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science*.
492 2014;345(6200):1048-52.
- 493 8. Dannemiller KC, Gent JF, Leaderer BP, Peccia J. Influence of housing characteristics on
494 bacterial and fungal communities in homes of asthmatic children. *Indoor Air*. 2016;26(2):179-92.
- 495 9. Panthee B, Gyawali S, Panthee P, Techato K. Environmental and Human Microbiome for
496 Health. *Life*. 2022;12(3):456.
- 497 10. Amin H, Santl-Temkiv T, Cramer C, Vestergaard DV, Holst GJ, Elholm G, et al. Cow
498 Farmers' Homes Host More Diverse Airborne Bacterial Communities Than Pig Farmers' Homes and
499 Suburban Homes. *Front Microbiol*. 2022;13:883991.
- 500 11. Zhou JC, Wang YF, Zhu D, Zhu YG. Deciphering the distribution of microbial communities
501 and potential pathogens in the household dust. *Sci Total Environ*. 2023:162250.
- 502 12. Kirjavainen PV, Karvonen AM, Adams RI, Täubel M, Roponen M, Tuoresmäki P, et al.
503 Farm-like indoor microbiota in non-farm homes protects children from asthma development. *Nature*
504 *Medicine*. 2019;25(7):1089-95.

- 505 13. Lee MK, Wyss AB, Carnes MU, Richards M, Parks CG, Beane Freeman LE, et al. House
506 dust microbiota in relation to adult asthma and atopy in a US farming population. *Journal of Allergy*
507 *and Clinical Immunology*. 2021;147(3):910-20.
- 508 14. Birzele LT, Depner M, Ege MJ, Engel M, Kublik S, Bernau C, et al. Environmental and
509 mucosal microbiota and their role in childhood asthma. *Allergy*. 2017;72(1):109-19.
- 510 15. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic
511 classification and assembly. *Briefings in Bioinformatics*. 2019;20(4):1125-36.
- 512 16. Laudadio I, Fulci V, Palone F, Stronati L, Cucchiara S, Carissimi C. Quantitative Assessment
513 of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut
514 Microbiome. *OMICS*. 2018;22(4):248-54.
- 515 17. Campanaro S, Treu L, Kougias PG, Zhu X, Angelidaki I. Taxonomy of anaerobic digestion
516 microbiome reveals biases associated with the applied high throughput sequencing strategies.
517 *Scientific Reports*. 2018;8(1).
- 518 18. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 16S rRNA gene sequencing of
519 mock microbial populations- impact of DNA extraction method, primer choice and sequencing
520 platform. *BMC Microbiology*. 2016;16(1).
- 521 19. Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho LFM, et al. Large-scale
522 differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci*
523 *Rep*. 2017;7(1):6589.
- 524 20. Durazzi F, Sala C, Castellani G, Manfreda G, Remondini D, De Cesare A. Comparison
525 between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut
526 microbiota. *Sci Rep*. 2021;11(1):3030.
- 527 21. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome:
528 Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res*
529 *Commun*. 2016;469(4):967-77.
- 530 22. Clooney AG, Fouhy F, Sleator RD, O' Driscoll A, Stanton C, Cotter PD, et al. Comparing
531 Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. *PLOS*
532 *ONE*. 2016;11(2):e0148028.
- 533 23. Tedersoo L, Anslan S, Bahram M, Pölme S, Riit T, Liiv I, et al. Shotgun metagenomes and
534 multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of
535 fungi. *MycKeys*. 2015;10:1-43.
- 536 24. Chan TF, Ji KM, Yim AK, Liu XY, Zhou JW, Li RQ, et al. The draft genome, transcriptome,
537 and microbiome of *Dermatophagoides farinae* reveal a broad spectrum of dust mite allergens. *J*
538 *Allergy Clin Immunol*. 2015;135(2):539-48.
- 539 25. Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmiento H, et al.
540 Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore
541 diversity and structure of microbial communities. *Environmental Microbiology*. 2014;16(9):2659-71.
- 542 26. Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM. Microbial Community Analysis with
543 Ribosomal Gene Fragments from Shotgun Metagenomes. *Appl Environ Microbiol*. 2016;82(1):157-
544 66.
- 545 27. Poretsky R, Rodriguez RL, Luo C, Tsementzi D, Konstantinidis KT. Strengths and limitations
546 of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS*
547 *One*. 2014;9(4):e93827.
- 548 28. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, et al. Cross-biome
549 metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad*
550 *Sci U S A*. 2012;109(52):21390-5.
- 551 29. Sitarik AR, Havstad S, Levin AM, Lynch SV, Fujimura KE, Ownby DR, et al. Dog
552 introduction alters the home dust microbiota. *Indoor Air*. 2018;28(4):539-47.

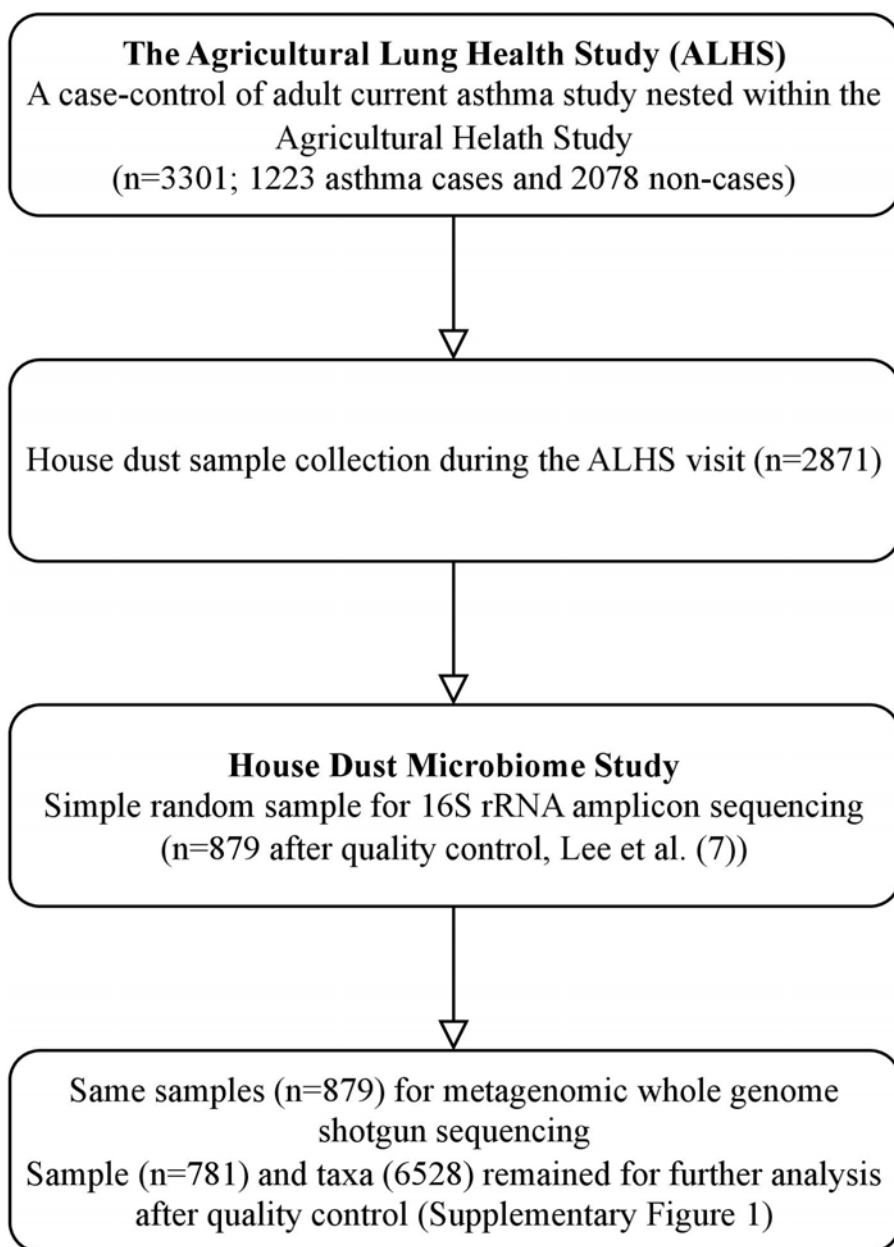
- 553 30. Alavanja MC, Sandler DP, McMaster SB, Zahm SH, McDonnell CJ, Lynch CF, et al. The
554 Agricultural Health Study. *Environ Health Perspect.* 1996;104(4):362-9.
- 555 31. Carnes MU, Hoppin JA, Metwali N, Wyss AB, Hankinson JL, O'Connell EL, et al. House
556 Dust Endotoxin Levels Are Associated with Adult Asthma in a U.S. Farming Population. *Ann Am
557 Thorac Soc.* 2017;14(3):324-31.
- 558 32. Arbes SJ, Jr., Cohn RD, Yin M, Muilenberg ML, Burge HA, Friedman W, et al. House dust
559 mite allergen in US beds: results from the First National Survey of Lead and Allergens in Housing. *J
560 Allergy Clin Immunol.* 2003;111(2):408-14.
- 561 33. Sanders JG, Nurk S, Salido RA, Minich J, Xu ZZ, Zhu Q, et al. Optimizing sequencing
562 protocols for leaderboard metagenomics by combining long and short reads. *Genome Biol.*
563 2019;20(1):226.
- 564 34. Babraham Institute. *FastQC.* 2010.
- 565 35. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.*
566 2012;9(4):357-9.
- 567 36. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating
568 taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3.
569 *Elife.* 2021;10.
- 570 37. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol.*
571 2019;20(1):257.
- 572 38. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in
573 metagenomics data. *PeerJ Computer Science.* 2017;3:e104.
- 574 39. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical
575 identification and removal of contaminant sequences in marker-gene and metagenomics data.
576 *Microbiome.* 2018;6(1):226.
- 577 40. R Core Team. *R: a language and environment for statistical computing.* Vienna, Austria: R
578 Foundation for Statistical Computing; 2020.
- 579 41. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and
580 graphics of microbiome census data. *PLoS One.* 2013;8(4):e61217.
- 581 42. Anderson MJ. Permutational multivariate analysis of variance (PERMANOVA). Wiley
582 statsref: statistics reference online. 2014:1-15.
- 583 43. Oksanen J, Simpson GL, Blanche FG, Kindt R, Legendre P, Minchin PR, et al. Community
584 ecology package. R package version. 2013;2(0):321-6.
- 585 44. Wickham H. *ggplot2: Elegant Graphics for Data Analysis:* Springer; 2016.
- 586 45. Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. *Nat
587 Commun.* 2020;11(1):3514.
- 588 46. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful
589 approach to multiple testing. 1995. p. 289-300.
- 590 47. Lee MK, Carnes MU, Butz N, Azcarate-Peril MA, Richards M, Umbach DM, et al.
591 Exposures Related to House Dust Microbiota in a U.S. Farming Population. *Environ Health Perspect.*
592 2018;126(6):067001.
- 593 48. Rausch P, Rühlemann M, Hermes BM, Doms S, Dagan T, Dierking K, et al. Comparative
594 analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of
595 animal metaorganisms. *Microbiome.* 2019;7(1).
- 596 49. Berg G, Rybakova D, Fischer D, Cernava T, Verges MC, Charles T, et al. Microbiome
597 definition re-visited: old concepts and new challenges. *Microbiome.* 2020;8(1):103.
- 598 50. Lynch SV, Wood RA, Boushey H, Bacharier LB, Bloomberg GR, Kattan M, et al. Effects of
599 early-life exposure to allergens and bacteria on recurrent wheeze and atopy in urban children. *Journal
600 of Allergy and Clinical Immunology.* 2014;134(3):593-601.e12.

- 601 51. Abrahamsson TR, Jakobsson HE, Andersson AF, Björkstén B, Engstrand L, Jenmalm MC.
602 Low diversity of the gut microbiota in infants with atopic eczema. *Journal of Allergy and Clinical*
603 *Immunology*. 2012;129(2):434-40.e2.
- 604 52. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Human gut microbes associated with obesity.
605 *Nature*. 2006;444(7122):1022-3.
- 606 53. Fu X, Ou Z, Zhang M, Meng Y, Li Y, Wen J, et al. Indoor bacterial, fungal and viral species
607 and functional genes in urban and rural schools in Shanxi Province, China—association with asthma,
608 rhinitis and rhinoconjunctivitis in high school students. *Microbiome*. 2021;9(1).
- 609 54. Weinstock DM, Brown AE. *Rhodococcus equi*: An Emerging Pathogen. *Clinical Infectious*
610 *Diseases*. 2002;34(10):1379-85.
- 611 55. Wong KY, Chan, Y.C. and Wong, C.Y., . *Corynebacterium striatum* as an emerging
612 pathogen. *Corynebacterium striatum* as an emerging pathogen. 2010;76(4):371-2.
- 613 56. De Bentzmann SaP, P. The *Pseudomonas aeruginosa* opportunistic pathogen and human
614 infections. *Environmental microbiology*. 2011;13(7):1655-65.
- 615 57. Valkonen M, Täubel M, Pekkanen J, Tischer C, Rintala H, Zock J-P, et al. Microbial
616 characteristics in homes of asthmatic and non-asthmatic adults in the ECRHS cohort. *Indoor Air*.
617 2018;28(1):16-27.
- 618 58. Pujar NK, Premakshi HG, Ganeshkar MP, Kamanavalli CM. Biodegradation of Pesticides
619 Used in Agriculture by Soil Microorganisms. *Enzymes for Pollutant Degradation: Springer Nature*
620 *Singapore*; 2022. p. 213-35.
- 621 59. Fiocchi A, Burks W, Bahna SL, Bielory L, Boyle RJ, Cocco R, et al. Clinical Use of
622 Probiotics in Pediatric Allergy (cuppa): A World Allergy Organization Position Paper. *World*
623 *Allergy Organization Journal*. 2012;5(11):148-67.
- 624 60. Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI. Human nutrition, the gut
625 microbiome and the immune system. *Nature*. 2011;474(7351):327-36.
- 626 61. Skillman LC, Evans PN, Strompl C, Joblin KN. 16S rDNA directed PCR primers and
627 detection of methanogens in the bovine rumen. *Letters in Applied Microbiology*. 2006;42(3):222-8.
- 628 62. Hatti-Kaul R, Chen L, Dishisha T, Enshasy HE. Lactic acid bacteria: from starter cultures to
629 producers of chemicals. *FEMS Microbiology Letters*. 2018;365(20).
- 630 63. Kraemer JG, Aebi S, Hilty M, Oppliger A. Nasal microbiota composition dynamics after
631 occupational change in animal farmers suggest major shifts. *Sci Total Environ*. 2021;782:146842.
- 632 64. Shukla SK, Ye Z, Sandberg S, Reyes I, Fritsche TR, Keifer M. The nasal microbiota of dairy
633 farmers is more complex than oral microbiota, reflects occupational exposure, and provides
634 competition for staphylococci. *PLOS ONE*. 2017;12(8):e0183898.
- 635 65. Bull CT, De Boer SH, Denny TP, Firrao G, Fischer-Le Saux M, Saddler GS, et al. List of
636 New Names of Plant Pathogenic Bacteria. *Journal of Plant Pathology*. 2012;94(1):21-7.
- 637 66. Goswami RS, Kistler HC. Heading for disaster: *Fusarium graminearum* on cereal crops.
638 *Molecular Plant Pathology*. 2004;5(6):515-25.
- 639 67. Staats JJ, Feder I, Okwumabua O, Chengappa MM. *Streptococcus suis*: Past and Present.
640 *Veterinary Research Communications*. 1997;21(6):381-407.
- 641 68. Bannoehr J, Guardabassi L. *Staphylococcus pseudintermedius* in the dog: taxonomy,
642 diagnostics, ecology, epidemiology and pathogenicity. *Veterinary Dermatology*. 2012;23(4):253-e52.
- 643 69. Sepich-Poore GD, Zitvogel L, Straussman R, Hasty J, Wargo JA, Knight R. The microbiome
644 and human cancer. *Science*. 2021;371(6536).
- 645 70. Dunn RR, Fierer N, Henley JB, Leff JW, Menninger HL. Home Life: Factors Structuring the
646 Bacterial Diversity Found within and between Homes. *PLoS ONE*. 2013;8(5):e64133.
- 647
648
649

650 **Table 1. Characteristics of Study Population.** ^ϕ percentage based on full cohort versus within each
 651 state. Exposures that were different by state of residence using Pearson’s chi-squared test (p<0.05):
 652 Dogs, Living on a Farm, Crop Farming, Animal Farming, Working with Beef Cattle, Working with
 653 Dairy Cattle, Working with Hogs.
 654

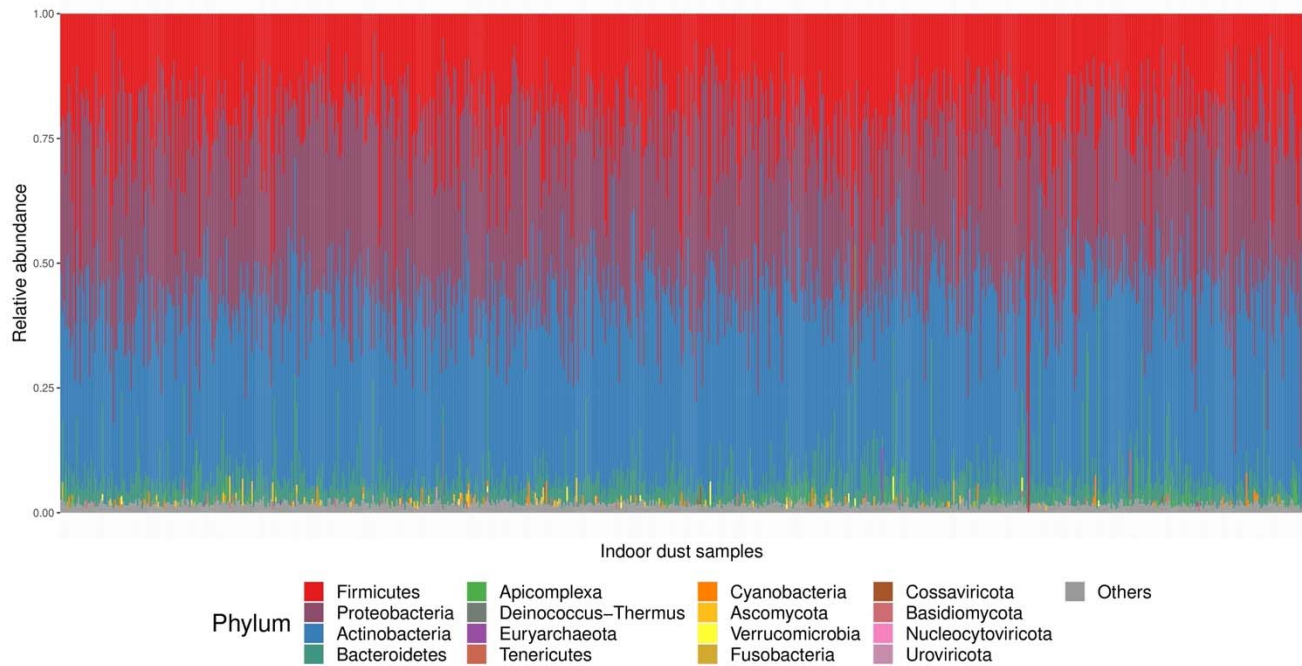
Category	Exposure	Total, N (% ^ϕ)	NC, N (%)	IA, N (%)
Total		781	247 (31.6) ^ϕ	534 (68.4) ^ϕ
Demography	Male sex	469 (60.1)	140 (56.7)	329 (61.6)
	Age in years, Mean (SD)	62 (11)	63 (11)	61 (11)
Presence of Indoor Pets	Dogs or cats	338 (43.3)	118 (47.8)	220 (41.2)
	Dogs	248 (31.8)	95 (38.5)	153 (28.7)
	Cats	165 (21.1)	49 (19.8)	116 (21.7)
Home Condition	Home condition, higher category	607 (77.8)	183 (74.4)	424 (79.4)
	Carpeting, carpeted surface	727 (93.3)	223 (90.7)	504 (94.6)
Current Farming Status	Living on a farm	651 (83.4)	194 (78.5)	457 (85.6)
	Crop farming	437 (55.9)	85 (34.4)	352 (65.9)
	Animal farming	401 (51.3)	98 (39.7)	303 (56.7)
	Working with beef cattle	281 (35.9)	65 (26.3)	216 (40.4)
	Working with dairy cattle	48 (6.1)	7 (2.8)	41 (7.7)
	Working with hogs	120 (15.4)	18 (7.3)	102 (19.1)
	Working with poultry	90 (11.5)	35 (14.2)	55 (10.3)
Season of Dust Collection	Spring	199 (25.5)	68 (27.5)	131 (24.5)
	Summer	245 (31.4)	69 (27.9)	176 (33)
	Fall	159 (20.4)	46 (18.6)	113 (21.2)
	Winter	178 (22.8)	64 (25.9)	114 (21.3)
Current Asthma Status, Case		296 (37.9)	86 (34.8)	210 (39.3)

655



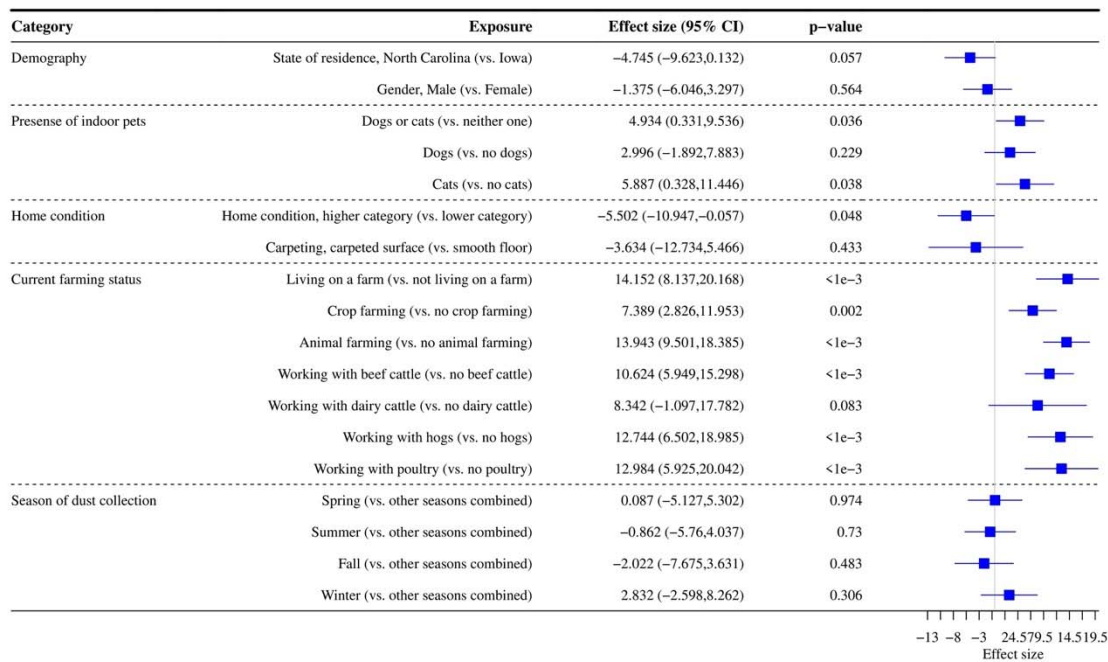
656
657
658
659
660

Figure 1. Workflow of house dust microbiome study in WGS. This workflow includes a summary of sample selection from the Agricultural Lung Health Study (ALHS) (n=3,301) to the house dust microbiome study with 16S (n=879) and WGS sequencing (n=781).



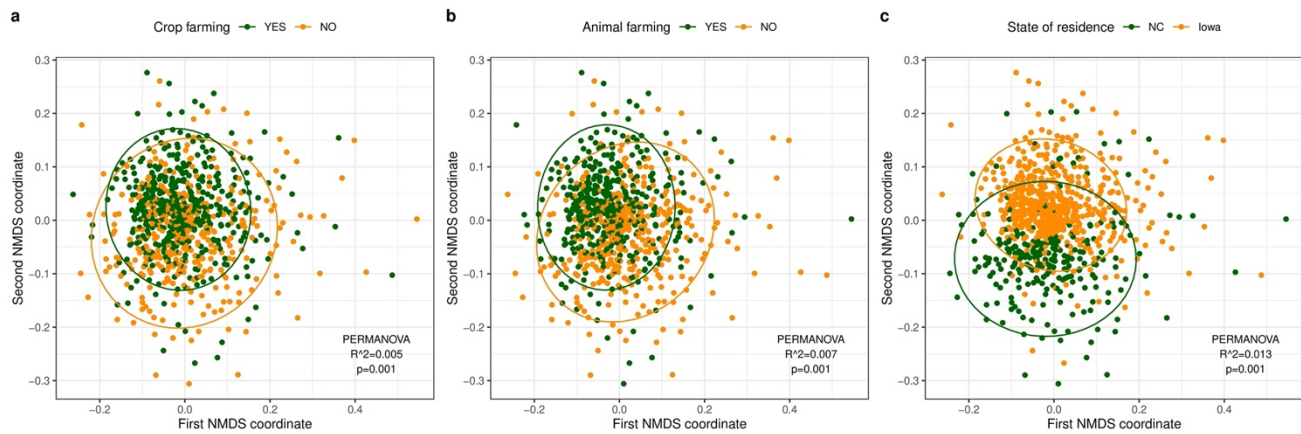
661
662 **Figure 2. Relative abundance at the phylum level across all home dust samples.** The 16 phyla
663 with relative abundance greater than 1% in at least one sample are color-coded according to the
664 legend. All other phyla are represented in grey.

665
666
667
668
669
670
671
672
673

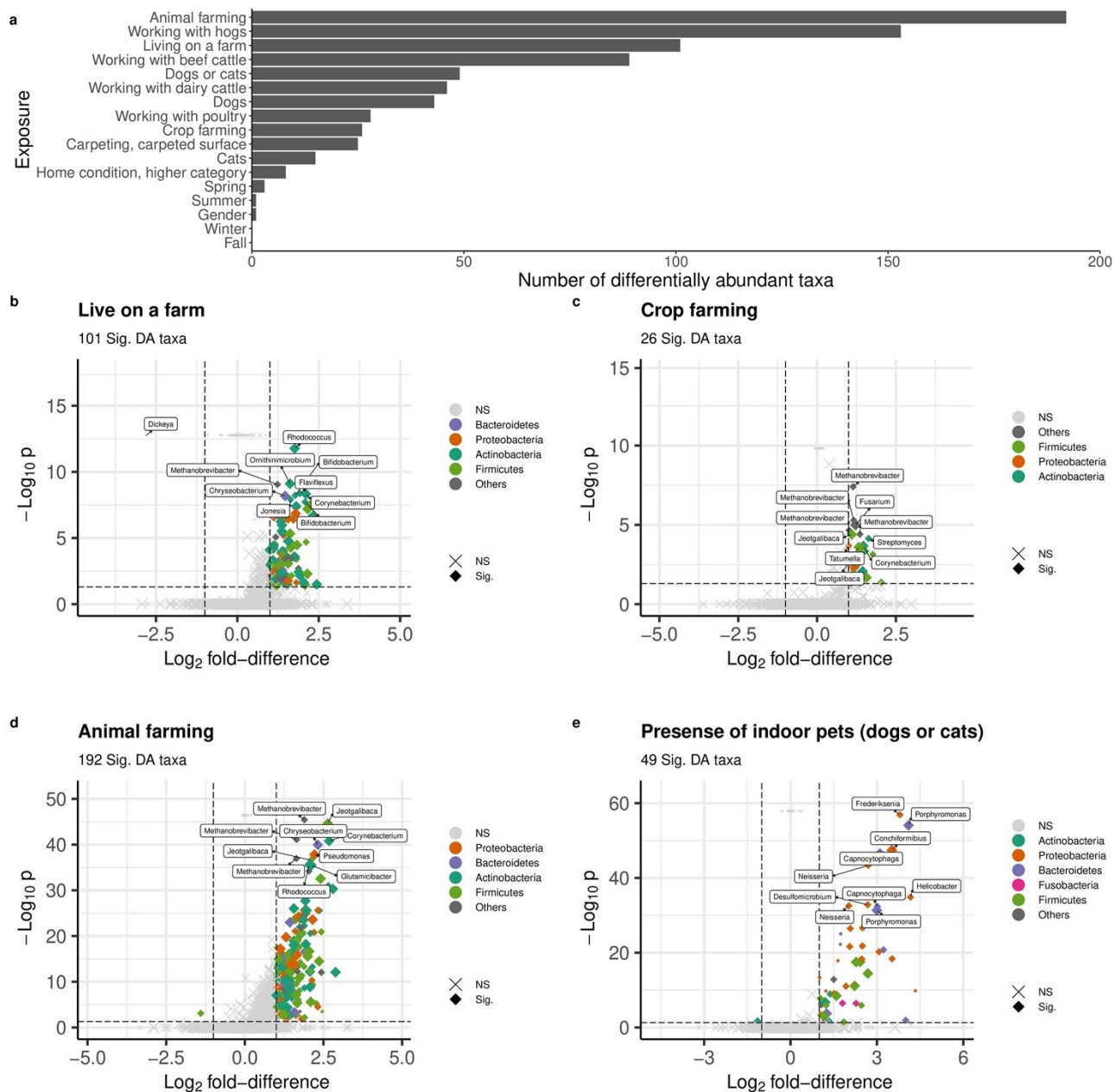


674
 675 **Figure 3. Association between exposures and alpha diversity (Shannon index with exponential**
 676 **transformation).** Data were rarefied to the minimum library size (1,003) across all samples. Effect
 677 size refers to the coefficient from the regression model (difference in alpha diversity for yes versus
 678 no for each exposure). The 95% confidence interval (CI) and p-value for each exposure from the
 679 regression model are reported.

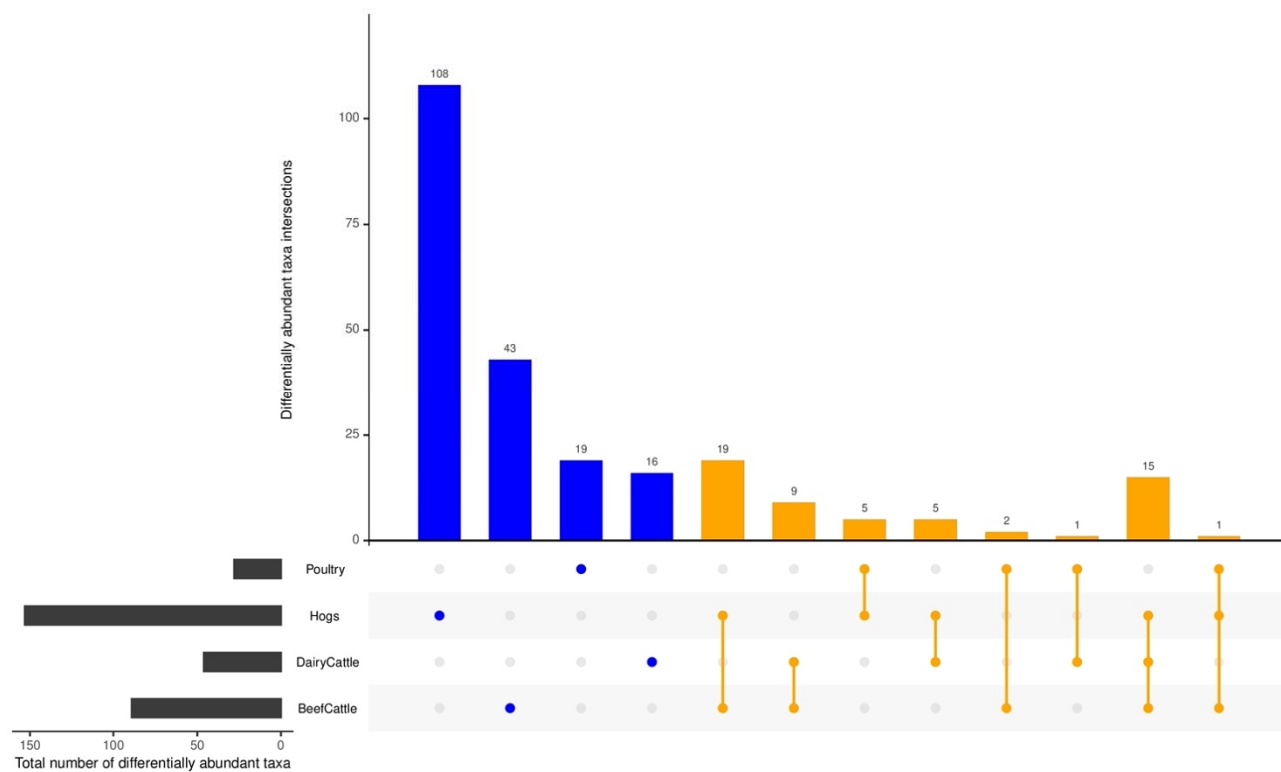
680
 681
 682
 683
 684
 685
 686
 687
 688
 689
 690
 691
 692
 693



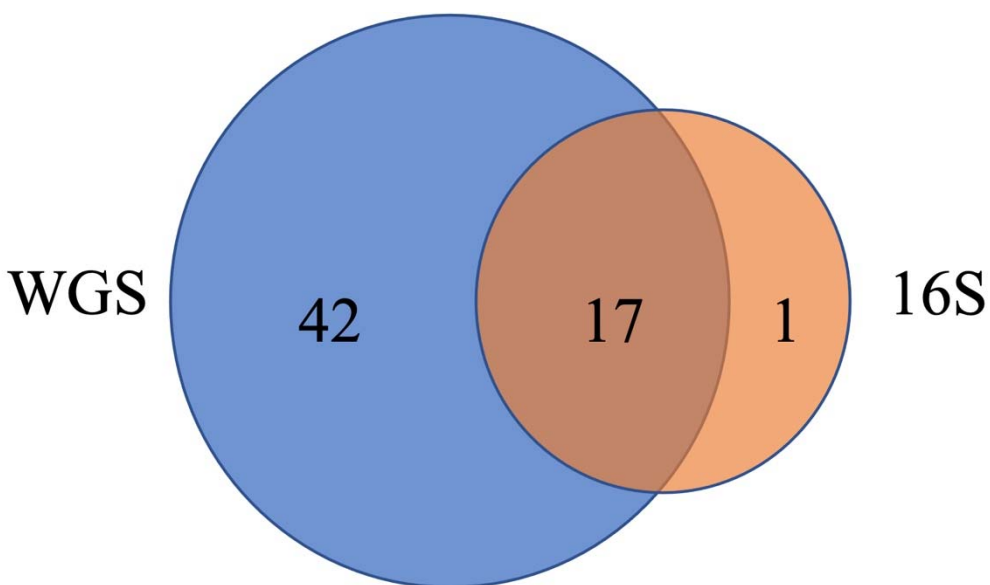
694
695 **Figure 4. Non-metric multidimensional scaling (NMDS) analysis based on unweighted UniFrac**
696 **distances for samples with different exposure levels. (a)** Crop farming (green: with crop farming,
697 yellow: without crop farming). **(b)** Animal farming (green: with animal farming, yellow: without
698 animal farming). **(c)** State of residence (green: North Carolina (NC), yellow: Iowa). The dust
699 microbial community of each sample is represented by a single dot. The ellipse represents the 95%
700 confidence interval for the centroids of each exposure level. R^2 values (percentage of variance
701 explained by an exposure) and p-values from the PERMANOVA analysis are reported.



702
 703 **Figure 5. Differentially abundant (DA) taxa related to individual exposure (FDR<0.05).** (a)
 704 **Number of DA taxa.** (b)-(e) **Volcano plot** for (b) Presence of indoor pets, (c) Living on a farm, (d)
 705 Crop farming, and (e) Animal farming. DA taxa are colored by phylum. The top 10 DA taxa with the
 706 smallest adjusted p-values are labeled by genus. Dot size indicates the medium abundance level for
 707 each taxon. a Benjamini-Hochberg method is used for FDR correction. lfd: log2 fold-difference.
 708 Vertical and horizontal dash lines indicate the threshold of p value after FDR correction and lfd for
 709 filtering DA taxa. Sig: DA taxa with $p < 0.05$ after FDR correction (i.e., $\log_{10} p < 0.5$) and $lfd > 1$ (or
 710 $lfd < -1$); NS: non-DA taxa.



711
 712 **Figure 6. Differentially abundant taxa related to various types of farming animal (FDR<0.05).**
 713 Commonly identified differentially abundant taxa shared by farming animal types were aligned by
 714 lines (orange), while differential taxa unique to farm animal type is identified by a single dot (blue).
 715
 716
 717
 718
 719
 720
 721
 722
 723
 724



725
726 **Figure 7.** Venn diagram of the number of phyla identified in WGS (blue) and 16S (orange). 17 phyla
727 were identified by both methods (Supplementary Table S14).