

Nanopore 16S rRNA sequencing reveals alterations in nasopharyngeal microbiome and enrichment of *Mycobacterium* and *Mycoplasma* in patients with COVID 19

Soumendu Mahapatra^{1*}, Rasmita Mishra^{1*}, Punit Prasad^{1**§}, Krushna Chandra Murmu¹, Shifu Aggarwal¹, Manisha Sethi¹, Priyanka Mohapatra¹, Arup Ghosh¹, Rina Yadav¹, Hiren Dodia¹, Shamima Azma Ansari¹, Saikat De¹, Deepak Singh¹, Amol Suryawanshi¹, Rupesh Dash¹, Shantibhushan Senapati¹, Tushar K. Beuria¹, Soma Chattopadhyay¹, Gulam Hussain Syed¹, Rajeeb Swain¹, Sunil K. Raghav¹, Ajay Parida^{1§}

¹Institute of Life Sciences, Bhubaneswar, Odisha, India.

* These authors contributed equally to this work

§ Correspondence: **Ajay Parida, Ph.D.**

Institute of Life Sciences,
Nalco Square, Chandrasekharapur,
Bhubaneswar, Odisha – 751023
Phone: +91-674-2304324
Email: ajayparida@ils.res.in drajayparida@gmail.com

Punit Prasad, Ph.D.

Institute of Life Sciences,
Nalco Square, Chandrasekharapur,
Bhubaneswar, Odisha – 751023
Phone: +91-674-2304319
Email: punit@ils.res.in punit.ils@gov.in

Running title: 16S rRNA sequencing of nasopharyngeal microbiome using Oxford Nanopore™

Keywords: Nasopharyngeal microbiome, Nanopore, COVID-19, 16S rRNA

1 **Abstract**

2 The coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory
3 syndrome corona virus 2 (SARS-CoV-2) is a major global health concern. This virus infects
4 the upper respiratory tract and causes pneumonia-like symptoms. So far, few studies have
5 shown that respiratory infections alter nasopharyngeal (NP) microbiome diversity and enrich
6 opportunistic pathogens. In this study, we have sequenced the 16S rRNA variable regions, V1
7 through V9, extracted from NP samples of control and COVID-19 (symptomatic and
8 asymptomatic) participants using the Oxford Nanopore™ technology. Comprehensive
9 bioinformatics analysis investigating the alpha/beta diversities, non-metric multidimensional
10 scaling, correlation studies, canonical correspondence analysis, linear discriminate analysis,
11 and dysbiosis index analysis revealed control and COVID-19-specific NP microbiomes. We
12 observed significant dysbiosis in COVID-19 NP microbiome with abundance of opportunistic
13 pathogens such as *Cutibacterium*, *Corynebacterium*, *Oerskovia*, and *Cellulomonas* in
14 asymptomatic patients, and of *Streptomyces* and *Mycobacteriaceae* family in symptomatic
15 patients. Furthermore, we observed sharp rise in enrichment of opportunistic pathogens in
16 symptomatic patients, with abundance of *Mycobacteria* and *Mycoplasma*, which strongly
17 correlated with the occurrences of chest pain and fever. Our findings contribute novel insights
18 regarding emergence of opportunistic pathogens in COVID-19 patients and their relationship
19 with symptoms, suggesting their potential role in coinfections.

20
21
22
23

24 **Introduction**

25 The coronavirus disease 2019 (COVID-19) pandemic, a global health threat, is caused by
26 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The symptoms range from
27 fever, throat pain, loss of taste and smell to severe congestion in the chest, drop in oxygen
28 levels, pneumonia, and acute respiratory distress syndrome (1). Furthermore, a significant
29 population worldwide remains asymptomatic, which is considered spreaders of the infection
30 (2). The virus enters the host via the upper respiratory tract (URT) where the spike protein
31 binds to the angiotensin I converting enzyme 2 (ACE2) receptor, an essential step in invading
32 host cells to cause progressive disease (3, 4). Random mutations in the SARS-CoV2 spike
33 protein and receptor-binding domain promote efficient invasion and enhance pathogenicity
34 (5).

35 The nasopharyngeal tract is inhabited by a large number of microbial communities which
36 maintain normal homeostasis (6). Studies have revealed association between microbial
37 communities that influence viral infections of the lung, such as chronic rhinosinusitis, asthma,
38 pneumonia, and cystic fibrosis in the URT (7, 8). URT microbiome dysbiosis may also
39 enhance the opportunistic pathogen population and promote coinfection in the host (9, 10).
40 Reports have shown that nasopharyngeal (NP) swabs in viral transport media can be used to
41 investigate the NP microbial composition in patients with COVID-19 (11, 12). Recent studies
42 have revealed overall compositional changes in the NP microbiota and promotion of
43 opportunistic pathogens such as *Rothia* and *Veillonella* in COVID-19 patients with shortness
44 of breath (11, 13, 14). The secondary infection in patients with COVID-19 is associated with
45 abundance of opportunistic pathogens such as *Moraxella*, *Corynebacterium*, *Haemophilus*,
46 *Stenotrophomonas*, *Acinetobacter*, *Fusobacterium periodonticum*, and *Pseudomonas*

47 *aeruginosa* (15-18). Studies on functional pathways of the NP metagenomics have revealed
48 that the abundance of NP commensal bacteria such as *Gemella morbillorum*, *Gemella*
49 *haemolysans*, and *Leptotrichia hofstadii* was reduced in the respiratory tract of COVID-19
50 patients, indicating the role of distinct functional metabolic pathways in this infection (19, 20).
51 Little is known about the crosstalk between SARS-CoV-2 viral infection and NP microbiota.
52 Moreover, systematic data connecting COVID-19-associated symptoms with microbial
53 composition is lacking. The absence of an animal model makes it difficult to test and validate
54 the role of NP microbiota in SARS-CoV-2 infection. Studies so far have shown differences in
55 the abundance of different opportunistic pathogens in the NP microbiota of patients, which is
56 one of the bottlenecks in this area of research. Hence more studies on the NP microbiome are
57 required for understanding its role in symptomatic and asymptomatic COVID-19 patients and
58 its relation with symptom severity.

59 In this study, we have investigated the alterations in the NP microbial ecosystem of patients
60 with active COVID-19 (n = 46) and compared them with that of healthy individuals (n = 12).
61 We have used the 16S metagenome approach and long-read sequencing (V1–V9) with the
62 Nanopore sequencing method to elucidate the reduction in microbial diversity in patients with
63 COVID-19. The composition of the NP microbiota changed significantly between
64 symptomatic and asymptomatic patients, resulting in enrichment of opportunistic pathogens
65 Interestingly, we found abundance of *Mycoplasma* and *Mycobacterium* at the genus level,
66 which strongly correlated with chest pain and fever in the symptomatic patients.

67 **Materials and Methods**

68 **Ethical approval:** Ethical permission for nasopharyngeal microbiome study and the
69 biorepository was obtained from the Institutional Ethical Committee (IEC)/Institutional

70 Review Board (IRB) of the Institute of Life Sciences [(102/HEC/2020) and (100/HEC/2020)].
71 Approval was also obtained from the Institutional Biosafety Committee (IBSC) (V-122-
72 MISC/2007-08/01/2/2.1) for this study and the biorepository (V-122-MISC/2007-08/01) and
73 from the Review Committee on Genetic Manipulations (RCGM) under Department of
74 Biotechnology, Ministry of Science and Technology.

75 **Sample collection and reverse transcription-polymerase chain reaction (RT-PCR):** In
76 total, 60 NP samples were collected for 16S rDNA amplicon sequencing from the Institute of
77 Life Science (ILS) COVID-19 sample biorepository unit. The COVID-19-positive samples (n
78 = 47) were confirmed by amplifying the genes encoding SARS-CoV-2 nucleocapsid, spike,
79 and ORF1ab/RdRP using either TaqPath™ COVID-19 combo kit (Invitrogen, A47814) or
80 Meril COVID-19 one-step RT-PCR kit (Meril Diagnostics, NCVPCR-02). All samples were
81 collected in the hospital setup prior to the medication. These COVID-19-positive patients were
82 not treated with antibiotics as the patients were not aware of their COVID-19 testing results.
83 The COVID-19 patients were grouped as symptomatic (n = 22) or asymptomatic (n = 25)
84 based on their clinical data. The control samples (n = 13) were negative for SARS-CoV-2
85 virus RNA and none of the subjects from whom the samples were obtained had any flu-like
86 symptoms. All samples were collected in viral transport media (VTM) and stored at -80°C
87 until DNA isolation.

88 **DNA extraction and PCR amplification:** DNA was isolated using the PureLink™
89 microbiome DNA purification kit (Invitrogen, A29790) according to the manufacturer's
90 protocol and eluted in 40 µl elution buffer. The quality and quantity of DNA were determined
91 using the Multiskan™GO spectrophotometer (Thermo Scientific). 16S rDNA amplification,
92 library preparation, and sequencing: V1-V9 variable regions of the 16S rRNA gene were

93 amplified using 130-F (5'-GGCGGATCCAAGGAGGTGTTCCAGCCGC-3') and 139-R (5'-
94 GGCCTCGAGAGAGTTTGGATCCTGGCTCAGG-3') primers. PCR (50 μ l) was set up using
95 total DNA (10 ng) isolated from NP samples, primers (5 nM), and NEB Q5® High-Fidelity 2
96 X master mix (NEB, M0492L) per the manufacturer's protocol. The amplicons (~1.6 kb) were
97 analyzed on 0.8% agarose gel and cleaned using DNA Clean and Concentrator-25 kit (Zymo
98 Research, D4034). The PCR products were quantified using a Qubit 4 fluorometer (Thermo
99 Scientific) using the Qubit® dsDNA BR assay kit (Thermo Scientific, Q32853). Amplicon
100 libraries were generated following the Oxford Nanopore 1D library preparation protocol using
101 the PCR barcoding (96) genomic DNA kit (Oxford Nanopore™, SQK-LSK109). Equimolar
102 amounts of amplicon libraries were pooled and sequenced using the MinION OXFORD
103 NANOPORE™ device at the ILS DNA sequencing facility.

104 **Microbiome data processing:** RAW fast5 files were generated using the MinKNOW™ tool
105 for individual samples. Base calling was performed using the Guppy base-caller and fastq files
106 were generated. FastQC of each sample was performed using the Babraham fastqc suite
107 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), followed by trimming of low
108 quality reads using nanoflit. Operational taxonomic units (OTUs) were generated using
109 Kraken2 (<https://ccb.jhu.edu/software/kraken2/index.shtml>) (21) and the unclassified reads
110 were filtered for downstream analysis using the 'phyloseq' 'R' package to generate combined
111 OTUs for all the samples and metadata (Supplemental Table 1). Read counts for mitochondria
112 and chloroplast were discarded. Normalization and differential OTU abundance were
113 determined between control, and symptomatic and asymptomatic subjects using the DESeq2
114 function (cutoff of p-value ≤ 0.05). The accession ID in NCBI is PRJNA774098.

115 **In-depth microbiome data analysis:**

116 Diversity analysis: Alpha diversity was assessed using the Shannon diversity index and
117 Simpson Diversity index. Statistical significance was estimated using the Wilcoxon rank sum
118 test. The beta diversity significance among groups was examined with PERMANOVA (p-
119 value 0.001). Ordination analysis was performed by PCoA, NMDS and CCA. R packages used
120 are ‘microbiome’, ‘Vegan’, ‘ade4’, ‘ggpubr’ for analysis and ‘ggplot2’ for visualization.

121 Dysbiosis index: Microbiome dysbiosis in each sample was calculated based on Bray-Curtis
122 distances. All samples were subjected to PCoA using Bray-Curtis distances. Next, the centroid
123 (median) of the control subjects was calculated along PCoA axes. The dysbiosis score for each
124 sample was calculated as a Euclidian distance between its position in the PCoA space and
125 control centroid ($DI(X, HC) = \sqrt{(X_i - HC_i)^2 + (X_j - HC_j)^2}$ | (DI: Dysbiosis Index, X: Samples, HC:
126 Control Centroid). Their significance was assessed using Wilcoxon and Kruskal-Wallis test
127 (22).

128 Sample correlation: Correlation matrix between samples and OTUs for each taxonomic level
129 (phylum, order, family, and genus) from differential OTUs was obtained using Spearman’s
130 correlation method and it was visualized as a heat map. Correlation coefficients for each
131 sample correlation pair and each classification level and density plot were plotted with mean
132 and median. The Kolmogorov test (KS) was used to determine the significance in sample
133 groups (control, asymptomatic, and symptomatic).

134 Linear discriminant analysis (LDA) effect size (LEfSe) analysis: The LEfSe was calculated
135 using the online Galaxy web application with the Huttenhower lab’s tool
136 (<https://huttenhower.sph.harvard.edu/galaxy/>). LDA effect size was calculated using the
137 Kruskal-Wallis sum rank test (alpha = 0.05) and it detected differential abundant features at
138 genus and species level within three sample groups. The taxonomic-level significance was

139 then tested using the pairwise Wilcoxon rank-sum tests ($\alpha = 0.05$). Finally, the effect size
140 of each differentially abundant feature was estimated using LDA. One-against-all sample
141 groups were compared and a linear discriminant analysis score greater than 3.6 was set as the
142 threshold; all-against-all sample groups were compared and a linear discriminant analysis
143 score greater than 2.0 was set as the threshold. Cladogram was used for identification of taxa
144 at different levels of the taxonomic hierarchy between sample groups (LDA score > 2).

145 Network analysis: Network was constructed using weighted correlation network analysis or
146 weighted gene co-expression network analysis (WGCNA). Briefly, pairwise Spearman
147 correlation between OTUs (which was generated from LefSe analysis) was calculated using
148 the WGCNA function. Network metrics such as betweenness, closeness, Eigen centrality, and
149 PageRank centrality of the resulting network were calculated and visualized using ‘Gephi’,
150 (<https://gephi.org/>) (23).

151

152 **Results**

153 **Study design and subject attributes**

154 The role of the microbiome in viral infections is an emerging field. We collected NP samples
155 from COVID-19 patients between 11th May 2020 and 10th October 2020 to study alterations
156 in the NP microbiome. The schematic representation of the nasal microbiome study with 16S
157 rDNA amplicon sequencing is shown in Figure 1A. In total, 60 NP samples subjects (infected,
158 $n = 47$ and control, $n = 13$ subjects, positive and negative for SARS-CoV2 RT-qPCR test
159 respectively) were obtained from the Institute of Life Sciences biorepository. Out of 47 SARS-
160 CoV-2-positive subjects, 25 were asymptomatic and 22 were symptomatic with mild
161 symptoms (Figure 1B). In total, 179,59,691 reads were generated. Two samples with low read

162 counts (1 from control and other from symptomatic category) were excluded and the final
163 study was performed with 58 subjects, including the control (C) [n = 12 (21%)], asymptomatic
164 [IA, infected asymptomatic; n = 25 (43%)], and symptomatic [IS, infected symptomatic; n =
165 21 (36%)]. The details of the participants considered for this study are described in Table 1.
166 Differential OTUs (n = 795, $p \leq 0.05$) were obtained from a total of 3482 OTUs using the
167 `deseq2` function by comparing with control NP subjects. For downstream analysis differential,
168 795 OTUs were considered. We used the t-distributed stochastic neighbor embedding (t-SNE)
169 dimension reduction method to obtain the overall distribution of NP samples with 795 OTUs
170 (Figure 1C). We found that the control and SARS-CoV-2-infected subjects showed distinct
171 segregation of OTUs in the NP microbiome, while asymptomatic and symptomatic subjects
172 showed modest separation. This indicated that the abundance of 795 differential OTUs
173 potentially determines the compositional distribution patterns.

174 **NP microbiome diversity was significantly altered in COVID-19 patients**

175 Distinct distribution of OTUs from control and infected patients prompted us to compare the
176 evenness and richness of bacterial community compositions using Shannon and Simpson
177 alpha indices. The Shannon and Simpson alpha microbial diversity indices between control
178 and SARS-CoV-2-infected participants differed significantly ($p\text{-value} \leq 0.05$) in pairwise
179 Wilcoxon rank test (Shannon $p\text{-value} = 3.0 \times 10^{-4}$ and Simpson $p\text{-value} = 3.3 \times 10^{-3}$) (Figure 2A,
180 B). Although the alpha diversity indices for samples from symptomatic and asymptomatic
181 patients compared to control subjects were found to be significantly reduced, no difference
182 was observed between symptomatic and asymptomatic samples (Figure 2C, D). Furthermore,
183 we used a linear regression model to establish the association between total OTU read counts
184 for each sample and Shannon/Simpson alpha diversity indices. We found negative correlation

185 for both Shannon (IA - $R = -0.35$, $R^2 = 0.44$, $p = 0.083$; IS - $R = -0.54$, $R^2 = 0.48$, $p = 0.012$)
186 and Simpson (IA - $R = -0.58$, $R^2 = 0.68$, $p = 0.0028$; IS - $R = -0.77$, $R^2 = 0.63$, $p = 7.7 \times 10^{-5}$)
187 alpha diversity indices with 95% confidence intervals with total OTU counts (Figure 2E, F).
188 To further understand the microbial composition dissimilarity within the samples, we analyzed
189 beta diversity using principal coordinate analysis (PCoA) and applied both unweighted
190 (microbial richness) and weighted (microbial richness and abundance) unifrac distance
191 methods. The first two components of PCoA showed 60.3% and 80.1% variance for the
192 unweighted and weighted unifrac method. The overall difference in microbial population
193 showed two different clusters of control and SARS-CoV-2-infected patients (IA and IS) in the
194 unifrac weighted method, while the unifrac unweighted method showed more clear
195 segregation between symptomatic and asymptomatic samples (Figure 2G). We assessed the
196 significance of beta diversity to calculate unifrac distance matrix (PERMANOVA test with
197 999 permutations) for both unweighted and weighted methods and found that the three sample
198 groups (C, IA, and IS) differed significantly ($P = 0.001$) with 18% variance explained ($R^2 =$
199 0.18842).

200 **NP microbiome dysbiosis in COVID-19 patients**

201 Alterations in the microbial diversity prompted us to determine microbial dysbiosis index (DI)
202 (alterations in the microbial community) across the three groups (C, IA, and IS). We
203 performed PCoA using the Bray Curtis distance matrix and found that NP microbiota was
204 significantly altered ($p = 0.001$) with 61% variation in distances explained ($R^2 = 0.6136$)
205 assessed by ADONIS test. Next, we calculated the Euclidean distance from the centroid for
206 samples from control (median = 0.3404), asymptomatic (median = 0.1881) and symptomatic
207 (median = 0.1511) individuals and calculated the DI (Supplementary figure 1B). The overall

208 observed DI was significant (Kruskal-Wallis test, $p = 1.317E-07$) across all the groups.
209 Pairwise comparison showed significant dysbiosis between control vs symptomatic ($p =$
210 $5.6E-09$) and control vs. asymptomatic ($p = 1.1E-09$) groups; however, dysbiosis between
211 asymptomatic and symptomatic ($p = 0.016$) pair was not highly significant (Figure 2H). We
212 also observed highly significant dysbiosis ($p = 2.2E-12$) between the control and infected
213 group (Supplemental Figure 1A, 1C). This showed that compared to that in the control
214 subjects, the NP microbial community is severely altered in both symptomatic and
215 asymptomatic COVID-19 patients.

216 **Distinct microbial composition and abundance at phylum and family levels in patients** 217 **suffering from SARS-CoV2 infection**

218 The alpha and beta diversities, and DI showed that the NP microbiome was significantly
219 altered in COVID-19 patients. Next, we aimed to identify the microbial communities that were
220 altered at the phylum and family levels in three sample groups. We found 795 differential
221 OTUs, out of which, 12 phyla, 65 orders, 126 families, and 240 genera were present in all
222 three groups (C, IA, and IS) (Supplemental Table 1). The 12 phyla and their significance is
223 shown in Table 2. The most significant bacteria in phylum level were Actinobacteria ($p =$
224 $9.96E-07$) and Proteobacteria ($p = 9.61E-07$), including 9 other phyla assessed using the
225 Kruskal-Wallis test. The abundance of phyla Firmicutes ($p = 4.65E-02$) and Actinobacteria (p
226 $= 9.96E-07$) were significantly higher in the SARS-CoV-2-infected groups (symptomatic and
227 asymptomatic). In contrast, Bacteroidetes ($p = 1.48E-06$) and Proteobacteria ($p = 6.56E-07$)
228 were highly abundant in the control group (non-infected) (Supplemental Figure 2A).
229 Furthermore, we analyzed relative abundance of top 30 families and found enrichment of
230 *Mycobacteriaceae*, *Propionibacteriaceae*, and *Streptomyetaceae* (Supplemental Figure 2B).

231 These families contain opportunistic pathogens in both symptomatic and asymptomatic
232 COVID-19 patients, while these families are absent in control subjects. Top families and their
233 significance is shown in Table 3.

234 **Taxonomic classifications based on OTU abundance showed sample group segregation**
235 **at the genus level**

236 To further our understanding regarding the 795 differentially abundant OTUs, we used the
237 NMDS approach at phylum, order, family, and genus levels for C, IA, and IS sample groups
238 using the Bray-Curtis distance matrix. Statistical significance using ANOSIM for phylum (R
239 $= 0.262$, $p = 1.7E-03$), order ($R = 0.322$, $p = 3E-04$), family ($R = 0.3461$, $p = 3E-04$), and
240 genus ($R = 0.3507$, $p = 3E-04$) showed gradual increase in R -value for genus. This indicated
241 that as we go lower in the taxonomic classification, the variance in the OTUs provides better
242 sample segregation. The differential OTUs present at the genus level in three sample groups
243 have high level of dissimilarity (35%) with $R = 0.3507$ and show clear sample segregation
244 (Figure 3A). To further validate the NMDS findings and identify the NP OTU differences
245 between C, IA, and IS sample groups, we used sample correlation (Spearman matrix)
246 (Supplemental Figure 3A-D). The sample correlation matrix clearly showed distinction among
247 C, IA, and IS with respect to taxonomic classification (Figure 3B). To further reconcile the
248 distinct sample segregation at higher to lower taxonomic level based on OTU abundances, we
249 plotted density histogram of correlation coefficient values (obtained in sample correlation).
250 The mean and median value of each density plot revealed lack of difference between the C,
251 IA and IS groups at the phylum level. Furthermore, subtle differences were observed at the
252 order and family level. However, at the genus level, we found comprehensible differences
253 between C (mean = $7.95E-01$; median = $6.39E-01$), IA (mean = $5.65E-01$; median = $8.33E-$

254 01) and IS (mean = 6.51E-01; median = 7.01E-01) (Table 4) (Figure 3B). To evaluate the
255 statistical significance of densities based on sample segregation, we calculated cumulative
256 distribution distance (D) and significance between C, IA, and IS groups using the
257 Kolmogorov–Smirnov (KS) test for each taxonomic rank (Table 5). We observed that
258 compared to that at other taxonomic levels, all the comparisons were highly significant at the
259 genus level. Based on the ‘D’ value comparison the samples were well distributed in C vs. IA
260 (D = 5.94E-01; p-value < 2.2E-16), C vs. IS (D = 5.06E-01; p = 3.308E-14), and IA vs. IS (D
261 = 2.28E-01; p = < 2.2E-16) at the genus level. The overall sample distribution differences
262 were highly enriched at the genus level than between these three groups. Although the ‘D’
263 value between IA and IS groups was less but the distribution pattern showed significant
264 differences between them. The above observance enables us to consider the genus level OTUs
265 (n = 240) for downstream analysis.

266 **Cluster-specific OTUs at genus level identified unique sample-specific OTUs**

267 To gain insight regarding how the bacterial genera were segmented among three groups of
268 samples, we performed genus level OTU correlation (n = 240) and calculated the correlation
269 coefficient (Spearman correlation), followed by unsupervised hierarchical clustering (Figure
270 3C). We identified 5 distinct clusters, C1 (n = 23), C2 (n = 109), C3 (n = 59), C4 (n = 33) and
271 C5 (n = 16), with variable number of OTUs (Supplemental Table 2). The heat map
272 corresponding to each cluster is shown in Supplemental Figure 4. Cluster-wise OTUs relative
273 abundance density maps were constructed, which distinguished OTUs that were enriched in
274 IA/IS (C1, C3, C4, and C5) and in control samples (C2) (Figure 3D). Some of the enriched
275 cluster-specific OTUs in IA/IS are *Mycobacterium* (1763), *Mycolicibacterium* (1766),
276 *Mycobacteroides* (1774), *Halothiobacillus* (927), *Flavobacterium* (986), *Bifidobacterium*

277 (1695), *Streptomyces* (1884), *Rothia* (2047), and *Mycoplasma* (2100). C2, a control-specific
278 cluster, contains OTUs such as *Thermomicrobium* (500), *Kingella* (502), *Enterobacter* (547),
279 *Bacteroides* (821), and *Prevotella* (840). Thus, this analysis shows the distinction in genus-
280 specific OTUs for both SARS-CoV-2-infected and control subjects. Next, we performed CCA
281 on each of the clusters (C1 to C5) to eliminate sample heterogeneity and enhance the
282 stringency of our analysis pipeline (Figure 3E). We considered the first two components of
283 CCA that explained cumulative variance for the clusters. Cluster C4 explained the highest
284 cumulative variance of 94.8%, while cluster C3 showed the lowest variance of 12.9% (Figure
285 3F, Supplemental Figure 5A-E). CCA showed efficient sample clustering, which is
286 reminiscent of the density plot (Figure 3D-E).

287 **LefSe analysis identified unique OTUs at genus level in COVID-19 patients**

288 The CCA analysis prompted us to select clusters with maximum variance explained.
289 Therefore, we considered all the clusters with $\geq 30\%$ variance, which includes all the clusters
290 except C3, for LDA. OTUs ($n = 181$) were extracted from clusters (C1, C2, C4, C5) and plotted
291 in a heat map with their abundance (Figure 4A). Different genera could be clearly
292 distinguished between C, IA, and IS sample groups. Next, we performed LefSe to distinguish
293 the most significant microbiomes from C, IA and IS groups. In a one-against-all comparison
294 (C with IA and IS), we got 40 genera in a control group, 34 genera in the symptomatic group,
295 and 4 genera in an asymptomatic group (LDA score $[\log_{10}] > 3.6$). The genera obtained from
296 one-against-all are highlighted in heatmap (Figure 4A and Supplemental Figure 6A). Relative
297 abundance of each OTU obtained from C, IA, and IS groups are shown in stack plots with
298 clear segregation in OTUs for individual samples (Supplemental Figure 6B-D). The DI
299 calculated from these genera showed high dysbiosis between control and SARS-CoV-2-

300 infected patients (Supplemental Figure 6E-F). We further increased the LefSe stringency by
301 using all-against-all (each sample group compared with each other) comparisons and
302 constructed a cladogram and a bar plot (Figure 4B-C). All the genera obtained from LefSe
303 (One against all and all against all) with their LDA scores and comparison are listed in
304 Supplemental Table 3 and Table 6. We obtained 12 significantly enriched genera of
305 *Gallibacterium*, *Orientia*, *Acidocella*, and *Citrobacter* in control samples (LDA score
306 $[\log_{10}] > 2.0$), *Mycoplasma*, *Streptosporangium*, *Mycobacterium*, *Mycolicibacterium*,
307 *Mycolicibacillus*, and *Mycobacteroides* in symptomatic samples, and *Oerskovia* and
308 *Cellulosimicrobium* in asymptomatic samples (Figure 4C). The histogram showing the relative
309 abundance of the 12 genera for each C, IA, and IS sample group clearly distinguishes each
310 sample type (Figure 4D). Finally, we used weighted correlation network analysis to construct
311 a network (Spearman correlation) with 12 genera identified using the LDA analysis. The
312 network creates two distinct modules, one for control groups and another for both symptomatic
313 and asymptomatic groups. We obtained strong correlation within the genera of C, IA, and IS
314 sample groups (Table 7). However, the correlation between C vs. IA was extremely weak and
315 correlation was not obtained for C vs. IS groups. The network analysis suggested that the NP
316 microbiota of the control group was clearly distinct from that of the asymptomatic and
317 symptomatic groups. The DI of the 12 genera showed the highest significance between C vs.
318 IS ($p = 4.7E-05$), while significant dysbiosis was not observed between IA and IS groups
319 (Figure 4F). Overall, our analysis confirms the significance of the genera identified and their
320 associations with symptomatic and asymptomatic COVID-19 patients.

321 **Distinct correlation of OTUs with clinical symptoms in COVID-19 patients**

322 To evaluate the accuracy of LDA classification that identified eight bacterial genera in the IA
323 and IS sample group, we tested the ROC (receiver operating characteristics) – AUC (area
324 under the curve) score. We obtained a value of 0.8 with 95% confidence interval for true
325 positive classification, showing 80% sensitivity and specificity of data obtained from LDA
326 analysis (Figure 5A). Next, we used the Spearman correlation matrix to identify the
327 association of symptoms with the genera. Interestingly, chest pain showed high positive
328 correlation with *Mycoplasma*, *Mycobacterium*, *Mycolicibacterium*, *Mycolicibacillus*, and
329 *Mycobacteroides* which were related to IS group, and weak correlation with *Oerskovia* and
330 *Cellulosimicrobium*, which were associated with the IA group. *Mycoplasma*, however,
331 showed a strong correlation with both chest pain (0.4446) and fever (0.4214) (Figure 5B).
332 ROC-AUC analysis for chest pain and fever showed 0.90 and 0.79 scores, respectively, with
333 eight bacterial genera (Figure 5C-D). We extended our study at the species level for the 12
334 genera found in LDA analysis and observed that 54 species were represented in a heat map
335 for C, IA, and IS group (Supplemental Figure 7A). Several known opportunistic pathogens
336 such as *Mycobacterium tuberculosis*, *Mycobacterium avium*, and *Mycoplasma pneumonia*
337 were highly abundant in the SARS-CoV-2-infected patients. The significance of the 54
338 bacterial species was assessed using the Kruskal-Wallis test and the top 30 significant species
339 were plotted in the bubble plot (Supplemental Figure 7B). In sum, we established the
340 association of pathogenic microbes with COVID-19 disease and showed susceptibility to
341 alterations in the NP microbiome in case of infection in SARS-COV-2. We also identified the
342 compositional difference in NP microbiota between symptomatic and asymptomatic group.

343 **Discussion**

344 Scientists worldwide are trying to understand the pathophysiology of SARS-CoV-2 infection
345 and the associated alterations in the host, including those in the microbiome. As SARS-CoV-
346 2 infection initiates in the upper respiratory tract, we investigated the alterations in the NP
347 microbiota of COVID-19 patients. We amplified the 16rRNA gene of variable regions (V1-
348 V9) and performed long-read sequencing using Oxford Nanopore technology. Subsequently,
349 we have used multiple bioinformatics approaches to cross-validate our data sets at various
350 levels and identify the most significant bacterial population in the NP microbiome of COVID-
351 19 patients. We found significant changes in abundance, diversity, and DI of SARS-CoV-2-
352 infected patients compared to those of the control. The IA and IS groups also showed overall
353 significant alterations in microbiota composition. We found abundance of opportunistic
354 pathogens such as *Mycoplasma* and *Mycobacterium* in symptomatic patients, which correlated
355 strongly with patient symptoms such as chest pain and fever. Insights into species level
356 abundance revealed the presence of *Mycoplasma pneumoniae*, *Mycobacterium tuberculosis*,
357 *Mycobacterium avium*, and *Mycolicibacterium sp.* in the SARS-CoV-2-infected patients. To
358 the best of our knowledge, this is the first comprehensive study to report abundance of
359 opportunistic pathogens such as *Mycoplasma pneumoniae* and *Mycobacterium tuberculosis*
360 based on the complete sequence of the 16S rRNA variable regions in patients with SARS-
361 CoV-2 infection.

362 Respiratory infections alter the NP microbiota, which reduces the diversity of the NP microbial
363 ecosystem and promotes the growth of opportunistic pathogens (24). At the phylum level,
364 Proteobacteria, Firmicutes, and Actinobacteria were detected in all NP samples. However, the
365 abundance of Firmicutes and Actinobacteria was significantly higher in both symptomatic and
366 asymptomatic groups. Our results are in partial agreement with those of Ventero et al., who

367 found the abundance of Firmicutes, Bacteroidota, Proteobacteria, and Actinobacteria in the
368 NP samples of COVID-19 patients (13). Only few studies have shown either no alterations or
369 significant changes in the microbiome composition of the nasopharynx during COVID-19
370 infection. Maio et al. and Braun et al. did not find any significant alterations in NP microbial
371 composition (12, 25). However, other studies showed prevalence of opportunistic pathogens
372 such as *Staphylococcus*, *Anelloviridae*, *Pseudomonas*, *Haemophilus*, *Stenotrophomonas*,
373 *Redondoviridae*, and *Pseudomonas aeruginosa* in COVID-19 patients (11, 13, 15-18).
374 Compared to earlier reports, our study also revealed overall changes in the composition of the
375 NP microbial community, reduction in bacterial diversity due to COVID-19 infection and the
376 presence of opportunistic pathogens such as *Mycoplasma* and *Mycobacterium* in COVID-19
377 patient cohort.

378 Most of the NP microbial studies amplify short 16S rRNA gene using the Illumina platform,
379 which is more accurate but is limited by taxonomic resolution owing to sequencing of shorter
380 reads (26) and sequencing of the specific variable region. The taxonomic resolution can be
381 improved to genus, species, and even at the strain level by sequencing the V1–V9 (~ 1600 bp)
382 variable regions of the 16S rRNA gene (26, 27). In this study, we have used the Oxford
383 Nanopore™ long read sequenced platform and sequenced V1 to V9 (~1.6 kb) of 16S variable
384 regions and successfully obtained taxonomic resolution to genus and some extent species
385 level. This has provided us immense advantage of determining the abundance of opportunistic
386 pathogens in the NP of the COVID-19 patients. Until now, only Mostafa et al. has used
387 metagenomics for COVID-19 NP samples using Oxford Nanopore technology. They have
388 sequenced both RNA and DNA from the NP samples without any PCR amplification. They
389 not only identified the SARS-CoV-2 virus in the samples but also potential pathogens that

390 may lead to co-infections (18). Our study is the first to use 16S amplification of ~1.6 Kb
391 variable regions to identify the bacterial community associated with infected and control NP.
392 However, 16S rRNA gene amplification may introduce PCR biases, however, more subjects
393 and a robust analysis pipeline may dilute these biases. This study is the first comprehensive
394 study from Odisha cohort and second from India. Gupta et al. used the Illumina platform for
395 16S amplicon sequencing and found enrichment of several opportunistic pathogens (17).
396 Interestingly, *Mycoplasma*, *Mycolicobacterium*, and *Mycobacterium* were not present in
397 their list. This could be due to the analysis pipeline or region-specific differences.
398 Nevertheless, the identification of opportunistic pathogens and their increase in abundance in
399 COVID-19 patients is one of the important aspects of this study.

400 Our comprehensive bioinformatics analysis with sample and OTU correlation analysis
401 distinguished COVID-19 infected and control samples at the genera level. Furthermore, LDA
402 analysis identified a significantly high abundance of *Mycobacterium* and *Mycoplasma* in
403 symptomatic patients, which correlated well with the occurrence of fever and chest pain.
404 Significantly high relative abundance of members of family *Mycobacteriaceae* in the
405 symptomatic COVID-19 group indicates the presence of both pathogenic and non-pathogenic
406 bacteria. Further dissection into genus level revealed the presence of several key genera,
407 namely, *Mycobacterium*, *Mycolicobacterium*, *Mycolicibacillus*, and *Mycobacteroides*.
408 *Mycobacterium* genera are well associated with several pulmonary diseases, for example,
409 *Mycobacterium tuberculosis* is responsible for tuberculosis in humans and is associated with
410 pulmonary infection (28), while *Mycobacterium avium* is highly associated with lung disease
411 (29). *Mycolicobacterium* and *Mycolicibacillus* are generally non-pathogenic but some species
412 have been associated with pathogenicity in humans and were isolated from hospitalized

413 patients (30). *Mycobacteroides* are potentially associated with soft tissue infections and
414 *Mycobacteroides abscessus* is a known pulmonary pathogen (30-32). Members of genus
415 *Mycoplasma* is a well-recognized pathogen, *Mycoplasma pneumoniae* being responsible for
416 pneumonia and other respiratory infections in humans (33, 34). *Cellulosimicrobium* and
417 *Oerskovia* were detected in asymptomatic COVID-19 patients. Few species of
418 *Cellulosimicrobium* are pathogenic, although their pathogenicity was not clear under normal
419 conditions because those species were isolated from hospitalized patients with acute renal
420 failure (35, 36). Results of our and other reports have proved the association of opportunistic
421 pathogens with alterations in the diversity of the microbial communities in symptomatic and
422 asymptomatic COVID-19 patients. This study establishes a new set of opportunistic pathogens
423 in the context of NP microbiome in COVID-19 infected patients. Moreover, this study clearly
424 distinguishes between the NP microbial composition of symptomatic and asymptomatic
425 groups using LefSe with AUC-ROC validation. Thus, we believe that SARS-CoV-2 virulence
426 may promote the growth of opportunistic pathogens and may lead to coinfection or secondary
427 infection in COVID-19 patients.

428 Our study has certain limitations. The subject size is limited and a larger cohort would have
429 strengthened our findings. The clinical manifestations are limited, and therefore, the larger
430 picture is difficult to interpret. Future studies should include NP samples of vaccinated,
431 asymptomatic, and hospitalized COVID-19 patients with detailed pathophysiology.
432 Furthermore, blood biochemistry and metabolite studies from the serum would boost
433 conclusions regarding functional aspects of the NP microbiome.

434 **Author Contribution**

435 P.P. and A.P. conceptualized the study and secured funding. P.P and S.M. initiated the work,
436 directed overall workflow, interpreted data, and troubleshoot the experiments. R.M. did most
437 of the bioinformatics analysis and S.M., K.C.M. and A.G. helped in bioinformatics analysis
438 and troubleshooting. S.M., S.A., M.S., P.M., R.Y., H.D., S.A.A., S.D., and D.S., helped with
439 the preprocessing of the samples in Biosafety level 3 (BSL3) facility and nucleic acid
440 extractions. S.M., S.A., M.S., P.M., and R.Y. were involved in amplicon library preparations.
441 R.S. provided samples from the Biorepository. A.S., R.D., S.S., T.K.B., S.C., G.H.S., R.S.,
442 S.K.R., P.P., and A.P. coordinated with COVID-19 sampling and testing at BSL3. P.P., S.M.,
443 and R.M wrote the manuscript.

444 **Conflict of interest**

445 The authors declare no competing commercial or financial interests in relation to this work.

446 **Acknowledgments**

447 We acknowledge the institute's core funding from the Department of Biotechnology (DBT),
448 Government of India. This work was also supported by the ILS flagship project
449 (BT/ILS/Flagship/2019) from DBT, Ramalingaswami Re-entry fellowship (BT/RLF/Re-
450 entry/25/2015), and SERB core research grant (CRG/2018/002052). We also acknowledge
451 Biorepository, BSL-3, and BSL-2 laboratories, qPCR, and DNA-sequencing institutional
452 central core facilities. R.M., S.M., and K.C.M received their fellowships from
453 Ramalingaswami, ILS Flagship, and SERB core research grant, respectively. We thank all the
454 volunteers who provided samples for research purposes.

455

456

457

458 **References**

- 459 1. He Y, Wang J, Li F, Shi Y. Main Clinical Features of COVID-19 and Potential
460 Prognostic and Therapeutic Value of the Microbiota in SARS-CoV-2 Infections. *Front*
461 *Microbiol.* 2020;11:1302.
- 462 2. Khatiwada S, Subedi A. Lung microbiome and coronavirus disease 2019 (COVID-19):
463 Possible link and implications. *Hum Microb J.* 2020;17:100073.
- 464 3. Zou X, Chen K, Zou J, Han P, Hao J, Han Z. Single-cell RNA-seq data analysis on the
465 receptor ACE2 expression reveals the potential risk of different human organs vulnerable to
466 2019-nCoV infection. *Front Med.* 2020;14(2):185-92.
- 467 4. Hou YJ, Okuda K, Edwards CE, Martinez DR, Asakura T, Dinnon KH, 3rd, et al. SARS-
468 CoV-2 Reverse Genetics Reveals a Variable Infection Gradient in the Respiratory Tract. *Cell.*
469 2020;182(2):429-46 e14.
- 470 5. Mlcochova P, Kemp SA, Dhar MS, Papa G, Meng B, Ferreira I, et al. SARS-CoV-2
471 B.1.617.2 Delta variant replication and immune evasion. *Nature.* 2021.
- 472 6. Belkaid Y, Harrison OJ. Homeostatic Immunity and the Microbiota. *Immunity.*
473 2017;46(4):562-76.
- 474 7. Fazlollahi M, Lee TD, Andrade J, Oguntuyo K, Chun Y, Grishina G, et al. The nasal
475 microbiome in asthma. *J Allergy Clin Immunol.* 2018;142(3):834-43 e2.
- 476 8. de Steenhuijsen Piters WA, Sanders EA, Bogaert D. The role of the local microbial
477 ecosystem in respiratory health and disease. *Philos Trans R Soc Lond B Biol Sci.*
478 2015;370(1675).
- 479 9. Kumpitsch C, Koskinen K, Schopf V, Moissl-Eichinger C. The microbiome of the upper
480 respiratory tract in health and disease. *BMC Biol.* 2019;17(1):87.

- 481 10. Yildiz S, Mazel-Sanchez B, Kandasamy M, Manicassamy B, Schmolke M. Influenza A
482 virus infection impacts systemic microbiota dynamics and causes quantitative enteric
483 dysbiosis. *Microbiome*. 2018;6(1):9.
- 484 11. Engen PA, Naqib A, Jennings C, Green SJ, Landay A, Keshavarzian A, et al.
485 Nasopharyngeal Microbiota in SARS-CoV-2 Positive and Negative Patients. *Biol Proced*
486 *Online*. 2021;23(1):10.
- 487 12. De Maio F, Posteraro B, Ponziani FR, Cattani P, Gasbarrini A, Sanguinetti M.
488 Nasopharyngeal Microbiota Profiling of SARS-CoV-2 Infected Patients. *Biol Proced Online*.
489 2020;22:18.
- 490 13. Ventero MP, Cuadrat RRC, Vidal I, Andrade BGN, Molina-Pardines C, Haro-Moreno
491 JM, et al. Nasopharyngeal Microbial Communities of Patients Infected With SARS-CoV-2
492 That Developed COVID-19. *Front Microbiol*. 2021;12:637430.
- 493 14. Feehan AK, Rose R, Nolan DJ, Spitz AM, Graubics K, Colwell RR, et al.
494 Nasopharyngeal Microbiome Community Composition and Structure Is Associated with
495 Severity of COVID-19 Disease and Breathing Treatment. *Applied Microbiology*.
496 2021;1(2):177-88.
- 497 15. Rhoades NS, Pinski AN, Monsibais AN, Jankeel A, Doratt BM, Cinco IR, et al. Acute
498 SARS-CoV-2 infection is associated with an increased abundance of bacterial pathogens,
499 including *Pseudomonas aeruginosa* in the nose. *Cell Rep*. 2021;36(9):109637.
- 500 16. Nardelli C, Gentile I, Setaro M, Di Domenico C, Pinchera B, Buonomo AR, et al.
501 Nasopharyngeal Microbiome Signature in COVID-19 Positive Patients: Can We Definitively
502 Get a Role to *Fusobacterium periodonticum*? *Front Cell Infect Microbiol*. 2021;11:625581.

- 503 17. Gupta A, Karyakarte R, Joshi S, Das R, Jani K, Shouche Y, et al. Nasopharyngeal
504 microbiome reveals the prevalence of opportunistic pathogens in SARS-CoV-2 infected
505 individuals and their association with host types. *Microbes Infect.* 2021:104880.
- 506 18. Mostafa HH, Fissel JA, Fanelli B, Bergman Y, Gniazdowski V, Dadlani M, et al.
507 Metagenomic Next-Generation Sequencing of Nasopharyngeal Specimens Collected from
508 Confirmed and Suspect COVID-19 Patients. *mBio.* 2020;11(6):1-13.
- 509 19. Liu J, Liu S, Zhang Z, Lee X, Wu W, Huang Z, et al. Association between the
510 nasopharyngeal microbiome and metabolome in patients with COVID-19. *Synth Syst*
511 *Biotechnol.* 2021;6(3):135-43.
- 512 20. Haiminen N, Utro F, Seabolt E, Parida L. Functional profiling of COVID-19 respiratory
513 tract microbiomes. *Sci Rep.* 2021;11(1):6433.
- 514 21. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome*
515 *Biol.* 2019;20(1):257.
- 516 22. Lloyd-Price J, Arze C, Ananthkrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW,
517 et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature.*
518 2019;569(7758):655-62.
- 519 23. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and
520 Manipulating Networks. *Proceedings of the International AAAI Conference on Web and*
521 *Social Media.* 2009;3(1):361-2.
- 522 24. Santacroce L, Charitos IA, Ballini A, Inchingolo F, Luperto P, De Nitto E, et al. The
523 Human Respiratory System and its Microbiome at a Glimpse. *Biology (Basel).* 2020;9(10).

- 524 25. Braun T, Halevi S, Hadar R, Efroni G, Glick Saar E, Keller N, et al. SARS-CoV-2 does
525 not have a strong effect on the nasopharyngeal microbial composition. *Sci Rep.*
526 2021;11(1):8922.
- 527 26. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al.
528 Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.
529 *Nat Commun.* 2019;10(1):5029.
- 530 27. Kaul D, Rathnasinghe R, Ferres M, Tan GS, Barrera A, Pickett BE, et al. Microbiome
531 disturbance and resilience dynamics of the upper respiratory tract during influenza A virus
532 infection. *Nat Commun.* 2020;11(1):2537.
- 533 28. Peto HM, Pratt RH, Harrington TA, LoBue PA, Armstrong LR. Epidemiology of
534 extrapulmonary tuberculosis in the United States, 1993-2006. *Clin Infect Dis.*
535 2009;49(9):1350-7.
- 536 29. Hwang JA, Kim S, Jo KW, Shim TS. Natural history of *Mycobacterium avium* complex
537 lung disease in untreated patients with stable course. *Eur Respir J.* 2017;49(3):1600537.
- 538 30. Gupta RS, Lo B, Son J. Phylogenomics and Comparative Genomic Studies Robustly
539 Support Division of the Genus *Mycobacterium* into an Emended Genus *Mycobacterium* and
540 Four Novel Genera. *Front Microbiol.* 2018;9:67.
- 541 31. Batchelder HR, Story-Roller E, Lloyd EP, Kaushik A, Bigelow KM, Maggioncalda EC,
542 et al. Development of a penem antibiotic against *Mycobacteroides abscessus*. *Commun Biol.*
543 2020;3(1):741.
- 544 32. Tortoli E. Microbiological features and clinical relevance of new species of the genus
545 *Mycobacterium*. *Clin Microbiol Rev.* 2014;27(4):727-52.

- 546 33. Beeton ML, Zhang XS, Uldum SA, Bebear C, Dumke R, Gullsby K, et al. Mycoplasma
547 pneumoniae infections, 11 countries in Europe and Israel, 2011 to 2016. Euro Surveill.
548 2020;25(2).
- 549 34. Foy HM. Infections caused by Mycoplasma pneumoniae and possible carrier state in
550 different populations of patients. Clin Infect Dis. 1993;17 Suppl 1:S37-46.
- 551 35. Sharma A, Gilbert JA, Lal R. (Meta)genomic insights into the pathogenome of
552 Cellulosimicrobium cellulans. Sci Rep. 2016;6:25527.
- 553 36. Delpont J, Wakabayashi AT, Anantha RV, Lannigan R, John M, McCormick JK.
554 Cellulosimicrobium cellulans isolated from a patient with acute renal failure. JMM Case
555 Reports. 2014;1(2):e000976.

556 **Figure Legends:**

557 **Figure 1: Schema of nasopharyngeal sample processing, 16S sequencing, and OTU-**
558 **based sample distribution.** (A) Flow chart showing nasopharyngeal sample processing for
559 DNA extraction, amplicon library preparation, Oxford Nanopore™ sequencing, and
560 bioinformatics analysis pipeline. (B) Pie chart showing nasopharyngeal samples (controls,
561 symptomatic, and asymptomatic) used in this study. (C) t-SNE plot showing the OTU-based
562 sample distribution and ordination points for control, symptomatic, and asymptomatic
563 samples.

564

565 **Figure 2: Alpha/beta diversities and dysbiosis index in COVID-19-positive and negative**
566 **nasopharyngeal sample.** (A-B) Alpha diversity index (Shannon/Simpson) between control
567 and COVID-19-infected samples (pairwise Wilcoxon rank-sum test $p = \leq 0.05$). (C-D) Same
568 analysis as above where the COVID-19-infected samples are classified as asymptomatic and

569 symptomatic compared to the control group. (E-F) Linear regression model showing the
570 association between total OTU count and Shannon/Simpson diversity index for each sample;
571 the shaded grey region represents 95% confidence intervals of two groups, symptomatic and
572 asymptomatic, with correlation (Spearman) regression line [Shannon: $R = -0.35$
573 (asymptomatic), $R = -0.54$ (symptomatic) and Simpson: $R = -0.58$ (asymptomatic), $R = -0.77$
574 (symptomatic)]. (G) Principal coordinate analysis (PCoA) showing beta diversity in
575 asymptomatic, symptomatic, and control sample groups based on unfrac
576 (weighted/unweighted) distance ($p = 0.001$, PERMANOVA). (H) Violin plot showing
577 dysbiosis indexes of samples from control, asymptomatic, and symptomatic participants
578 (pairwise Wilcoxon rank-sum test $p = \leq 0.05$).

579

580 **Figure 3: Taxonomic classification of bacterial communities using non-metric**
581 **multidimensional scaling (NMDS), correlation, and canonical correspondence analysis**
582 **(CCA). (A) NMDS ordination of Bray-Curtis distance matrix based on all samples and**
583 **bacterial communities of each taxonomy level (phylum, order, family, and genus) (ANOSIM**
584 **$p = < 0.05$). (B) The density plot representing the Spearman correlation coefficient at each**
585 **taxonomy level (phylum, order, family, and genus); dotted line indicates the mean value of**
586 **each sample group (Kolmogorov–Smirnov (KS) Test $p \leq 0.05$). (C) Heat map of Spearman**
587 **correlation for genus level with sample correlation (lower) and OTU correlation (upper). Five**
588 **clusters (C1, C2, C3, C4, and C5) were generated using unsupervised hierarchical clustering**
589 **from the OTU correlation plot. (D) Sample-wise OTU density plot for each cluster (C1, C2,**
590 **C3, C4, and C5) showing relative abundance. (E-F) CCA plot of microbial community**
591 **composition for each cluster and bar plot representing cumulative variation percentage from**

592 two components [C1 (92.4%), C2 (80.5%), C3 (12.9%), C4 (94.8%), and C5 (39.4%)]. Dotted
593 line shows 30% variance cut-off for downstream analysis.

594

595 **Figure 4: Linear discriminant analysis effect size (LefSe) analysis reveals distinct genus-**

596 **level OTUs in control, asymptomatic and symptomatic.** (A) Heat map showing genus level

597 OTU (n = 181) abundance distribution from four clusters (C1, C2, C4, and C5) identified from

598 CCA analysis in control, asymptomatic and symptomatic samples. The OTUs marked on

599 either side of the heat map were obtained from one-against-all and all-against-all comparison

600 in LDA analysis (B) The cladogram shows the output of the LefSe (LDA score >2.0), which

601 identifies taxonomic differences between sample groups. Each circle represents a bacterial

602 taxon, and each ring of taxonomy level starting with kingdom in the innermost circle is

603 followed by phylum, class, order, family, and genus in the outermost circle. The different color

604 intensities indicate the different taxonomy levels and the diameter of each circle is

605 proportional to the taxon's abundance and correlates with the LDA score. (C) The histogram

606 of the LDA scores (score >2.0 and all-against-all) was computed for differentially abundant

607 taxa between sample groups. The effect size of specific taxa in the particular group at the

608 genus level. (D) Histogram of the all LefSe-specific taxa (*Mycoplasma*, *Streptosporangium*,

609 *Citrobacter*, *Acidocella*, *Mycolicibacterium*, *Mycolicibacillus*, *Mycobacterium*,

610 *Mycobacteroides*, *Orientia*, *Gallibacterium*, *Cellulosimicrobium*, and *Oerskovia*) showing

611 relative abundance across sample groups. Solid and dotted lines show median and mean

612 relative abundance respectively. (E) Weighted correlation network analysis (WGCNA) was

613 used for network construction and plotted using Gephi. Each node of the network represents

614 the individual bacterial genera with their respective abundance size and the edges represent

615 correlation strength with edge weight by thickness. The pie chart within each node represents
616 abundance for each genus. The dotted line shows two distinct modules (control and infected)
617 created in the network analysis. (F) Violin plot showing the dysbiosis indexes of LefSe sample
618 groups (pairwise Wilcoxon rank-sum test p-value < 0.05).

619

620 **Figure 5: Area under the curve-receiver operating characteristic (AUC-ROC) validation**
621 **and correlation of genera with the symptoms of COVID-19 subjects.** (A) ROC curve for
622 LDA classified symptomatic and asymptomatic group. The AUC w 0.80 with a 95%
623 confidence interval (CI). (B) Correlation between bacteria at genus level and clinical
624 symptoms of patients. (C-D) ROC curve for chest pain and fever in the symptomatic and
625 asymptomatic group. The AUCs were 0.904 (chest pain) and 0.793 (fever) with a 95%
626 confidence interval (CI).

627

628

629

630

631

632

633

634

635

636

637

638 **Table 1: Details of samples included in this study**

	Control (n=12)	Asymptomatic (n=25)	Symptomatic (n=21)
Sex			
Male	5 (41.66%)	18 (72%)	19 (90.47%)
Female	7 (58.33%)	7 (28%)	2 (9.52%)
Age (years)	31 (median)	26 (median)	32 (median)
Symptoms			
Dry Cough	NA	NA	7 (33.3%)
Fever	NA	NA	17 (80.95%)
Tiredness	NA	NA	8 (38.09%)
Sore throat	NA	NA	11(52.38%)
Body pain	NA	NA	13(61.9%)
Chest pain	NA	NA	9(42.85%)
Fever + Body pain	NA	NA	4(19.04%)
Fever + multiple symptoms*	NA	NA	16(76.19%)
Fever + chest pain	NA	NA	8(38.09%)
Loss of smell/taste + multiple symptoms* without fever	NA	NA	2(9.52%)

639 *Multiple symptoms refer to having more than one symptoms from symptoms list.

640 **Table 2: Phylum based on relative abundance and their respective values**

Phylum	Mean	1st Quartile	Median	3rd Quartile	p_value	BH_FDR
Proteobacteria	1.51E-01	3.79E-02	4.53E-02	7.29E-02	6.56E-07	6.56E-07
Fusobacteria	2.25E-03	1.31E-03	1.72E-03	2.71E-03	9.00E-07	9.00E-07
Actinobacteria	6.83E-01	7.09E-01	7.62E-01	7.99E-01	9.96E-07	9.96E-07
Bacteroidetes	8.77E-03	5.71E-03	6.66E-03	1.00E-02	1.48E-06	1.48E-06
Tenericutes	9.37E-04	1.35E-04	5.60E-04	1.21E-03	2.21E-06	2.21E-06
Chloroflexi	2.25E-03	6.47E-04	9.84E-04	1.46E-03	6.54E-06	6.54E-06
Chlamydiae	5.93E-04	2.82E-04	4.98E-04	7.76E-04	7.27E-06	7.27E-06
Fibrobacteres	5.65E-04	2.86E-04	4.05E-04	7.10E-04	4.77E-05	4.77E-05
Thermodesulfobacteria	2.62E-02	2.09E-02	2.84E-02	3.24E-02	8.13E-04	8.13E-04
Aquificae	2.31E-04	1.19E-04	2.30E-04	2.93E-04	1.59E-03	1.59E-03
Firmicutes	1.23E-01	1.19E-01	1.28E-01	1.44E-01	4.65E-02	4.65E-02
Chlorobi	1.02E-03	6.81E-04	9.32E-04	1.27E-03	8.58E-01	8.58E-01

641

642

643

Table 3: Top 30 family based on relative abundance and their respective values

Family	Mean	1st quartile	Median	3rd quartile	p_value	BH_FDR
<i>Acetobacteraceae</i>	5.08E-02	5.00E-02	5.12E-02	5.28E-02	9.14E-07	9.14E-07
<i>Actinomycetaceae</i>	2.23E-02	2.10E-02	2.18E-02	2.32E-02	8.18E-03	8.18E-03
<i>Aeromonadaceae</i>	2.96E-02	2.54E-02	3.07E-02	3.38E-02	8.18E-03	8.18E-03
<i>Alcaligenaceae</i>	8.14E-02	8.50E-02	8.71E-02	9.43E-02	1.41E-06	1.41E-06
<i>Bifidobacteriaceae</i>	7.69E-02	4.59E-02	6.52E-02	1.04E-01	2.99E-03	2.99E-03
<i>Brevibacteriaceae</i>	4.37E-02	2.72E-02	3.82E-02	5.98E-02	6.35E-04	6.35E-04
<i>Cellulomonadaceae</i>	6.75E-02	4.60E-02	6.29E-02	8.52E-02	8.64E-05	8.64E-05
<i>Chromobacteriaceae</i>	2.30E-02	2.27E-02	2.30E-02	2.32E-02	9.37E-07	9.37E-07
<i>Corynebacteriaceae</i>	5.49E-02	4.18E-02	5.05E-02	6.52E-02	2.36E-05	2.36E-05
<i>Enterobacteriaceae</i>	1.26E-01	3.61E-02	1.65E-01	2.17E-01	1.73E-06	1.73E-06
<i>Erwiniaceae</i>	3.19E-02	2.97E-02	3.10E-02	3.46E-02	2.01E-06	2.01E-06
<i>Erysipelotrichaceae</i>	2.26E-02	2.21E-02	2.26E-02	2.30E-02	1.69E-06	1.69E-06
<i>Eubacteriaceae</i>	2.99E-02	2.58E-02	2.84E-02	3.28E-02	1.85E-06	1.85E-06
<i>Lactobacillaceae</i>	2.12E-02	2.05E-02	2.10E-02	2.18E-02	5.39E-03	5.39E-03
<i>Micrococcaceae</i>	3.52E-02	2.65E-02	3.37E-02	4.36E-02	3.49E-03	3.49E-03
<i>Micromonosporaceae</i>	2.41E-02	2.25E-02	2.43E-02	2.53E-02	2.65E-06	2.65E-06
<i>Morganellaceae</i>	5.01E-02	4.51E-02	5.30E-02	5.93E-02	1.90E-05	1.90E-05
<i>Mycobacteriaceae</i>	2.19E-01	1.85E-01	2.42E-01	2.69E-01	7.93E-07	7.93E-07
<i>Neisseriaceae</i>	5.24E-02	5.08E-02	5.23E-02	5.43E-02	1.29E-06	1.29E-06
<i>Nocardiaceae</i>	3.15E-02	2.46E-02	3.41E-02	3.63E-02	8.90E-07	8.90E-07
<i>Pasteurellaceae</i>	4.73E-02	3.34E-02	5.14E-02	5.75E-02	5.89E-05	5.89E-05
<i>Pectobacteriaceae</i>	2.17E-02	2.05E-02	2.15E-02	2.27E-02	2.09E-06	2.09E-06
<i>Peptostreptococcaceae</i>	3.01E-02	2.56E-02	2.82E-02	3.23E-02	1.89E-06	1.89E-06
<i>Propionibacteriaceae</i>	1.17E-01	9.00E-02	1.20E-01	1.33E-01	8.90E-07	8.90E-07
<i>Pseudonocardiaceae</i>	2.94E-02	2.70E-02	2.98E-02	3.27E-02	1.44E-06	1.44E-06

645

646

Table 4: Mean and median value of density plots.

Taxonomic rank	Statistic value	Control	Asymptomatic	Symptomatic
Phylum	Mean	8.96E-01	8.68E-01	8.90E-01
	Median	9.30E-01	8.95E-01	9.16E-01
Order	Mean	8.58E-01	7.70E-01	8.36E-01
	Median	8.78E-01	8.21E-01	8.93E-01
Family	Mean	8.22E-01	6.86E-01	7.60E-01
	Median	8.63E-01	7.46E-01	8.06E-01
Genus	Mean	7.95E-01	5.65E-01	6.51E-01
	Median	6.39E-01	8.33E-01	7.01E-01

647 **Table 5: Result of Kolmogorov–Smirnov (KS) test between the densities of each**
 648 **taxonomic rank.**

Taxonomic rank	Control vs Asymptomatic	Control vs Symptomatic	Asymptomatic vs Symptomatic
Phylum	D = 1.88e-01	D = 1.07e-01	D = 1.01e-01
	p-value = 3.03e-02	p-value = 0.4834	p-value = 9.78e-04
Order	D = 3.18e-01	D = 9.97e-02	D = 2.30e-01
	p-value = 1.237e-05	p-value = 0.5706	p-value < 2.2e-16
Family	D = 4.14e-01	D = 2.93e-01	D = 2.25e-01
	p-value = 2.706e-09	p-value = 4.75e-05	p-value < 2.2e-16
Genus	D = 5.94e-01	D = 5.06e-01	D = 2.28e-01
	p-value < 2.2e-16	p-value = 3.308e-14	p-value < 2.2e-16

649

650 **Table 6: Linear discriminate analysis (LDA) score for all-against-all analysis**

651

Genus	highest mean among all the classes	Samples	LDA score (log 10)	pvalue
<i>Oerskovia</i>	3.83	Asymptomatic	3.44	1.96E-02
<i>Cellulosimicrobium</i>	3.85	Asymptomatic	3.5	1.73E-02
<i>Gallibacterium</i>	3.84	Control	3.36	8.46E-04
<i>Orientia</i>	3.88	Control	3.46	3.07E-04
<i>Acidocella</i>	4.07	Control	3.81	4.68E-07
<i>Citrobacter</i>	4.08	Control	3.89	2.82E-07
<i>Mycobacteroides</i>	3.88	Symptomatic	3.61	3.19E-07
<i>Mycolicibacillus</i>	3.98	Symptomatic	3.67	3.19E-07
<i>Mycolicibacterium</i>	3.93	Symptomatic	3.71	2.55E-07
<i>Mycobacterium</i>	3.9	Symptomatic	3.79	2.55E-07
<i>Streptosporangium</i>	4.16	Symptomatic	3.83	1.65E-04
<i>Mycoplasma</i>	4.26	Symptomatic	3.95	1.64E-06

652

653

654

655

Table 7: WGCNA network data table

ID	Genus	OTUs	Module	absolute abundance	Degree	Page ranks	Eigen centrality	Modularity class	Clustering	Triangles
1	<i>Acidocella</i>	525	M1	1362	5	0.0673	0.4963	0	1	10
2	<i>Citrobacter</i>	546	M1	3463	5	0.0673	0.4963	0	1	10
3	<i>Gallibacterium</i>	750	M1	285	5	0.0673	0.4963	0	1	10
4	<i>Orientia</i>	784	M1	263	5	0.0673	0.4963	0	1	10
5	<i>Cellulosimicrobium</i>	1710	M2	11832	10	0.1202	1	1	0.5555	25
6	<i>Oerskovia</i>	1713	M2	13210	10	0.1202	1	1	0.5555	25
7	<i>Mycobacterium</i>	1763	M2	139017	7	0.0852	0.8208	1	0.9047	19
8	<i>Mycolicibacterium</i>	1766	M2	81430	7	0.0852	0.8208	1	0.9047	19
9	<i>Mycobacteroides</i>	1774	M2	7172	7	0.0852	0.8208	1	0.9047	19
10	<i>Mycolicibacillus</i>	1798	M2	2749	7	0.0852	0.8208	1	0.9047	19
11	<i>Streptosporangium</i>	2002	M2	163	5	0.0642	0.5732	1	1	10
12	<i>Mycoplasma</i>	2100	M2	475	7	0.0852	0.8208	1	0.9047	19

Figure 1

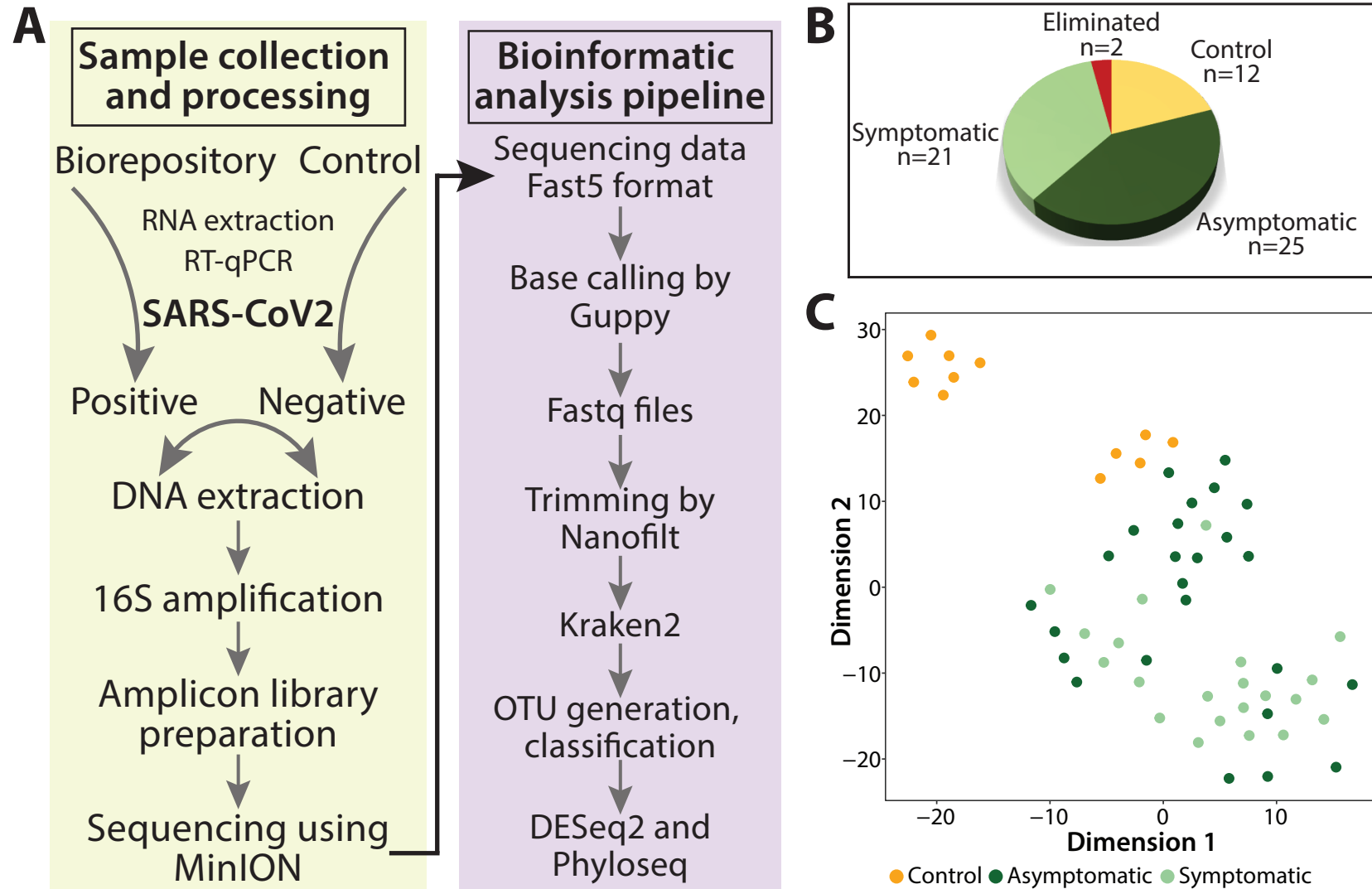


Figure 2

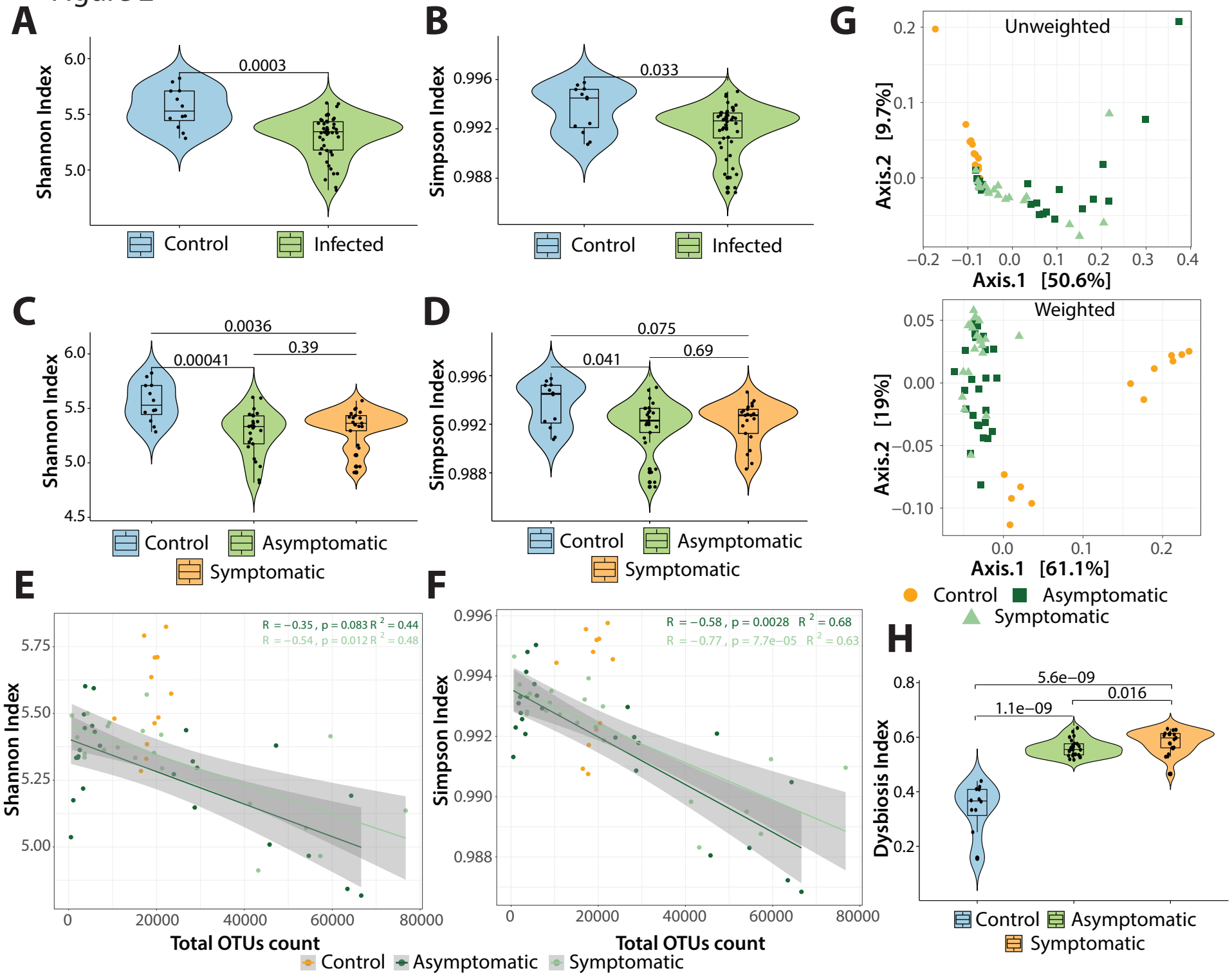


Figure 3

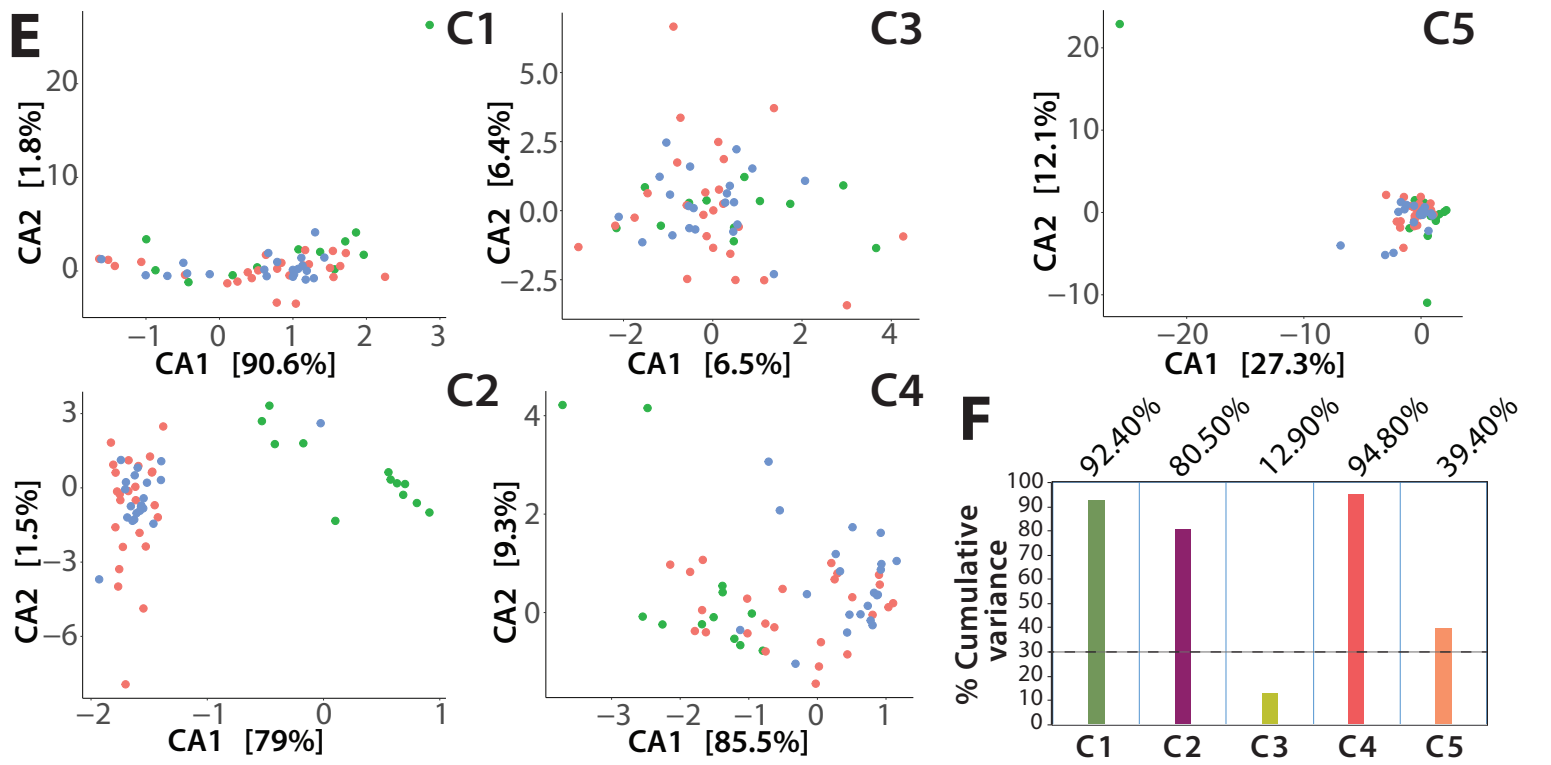
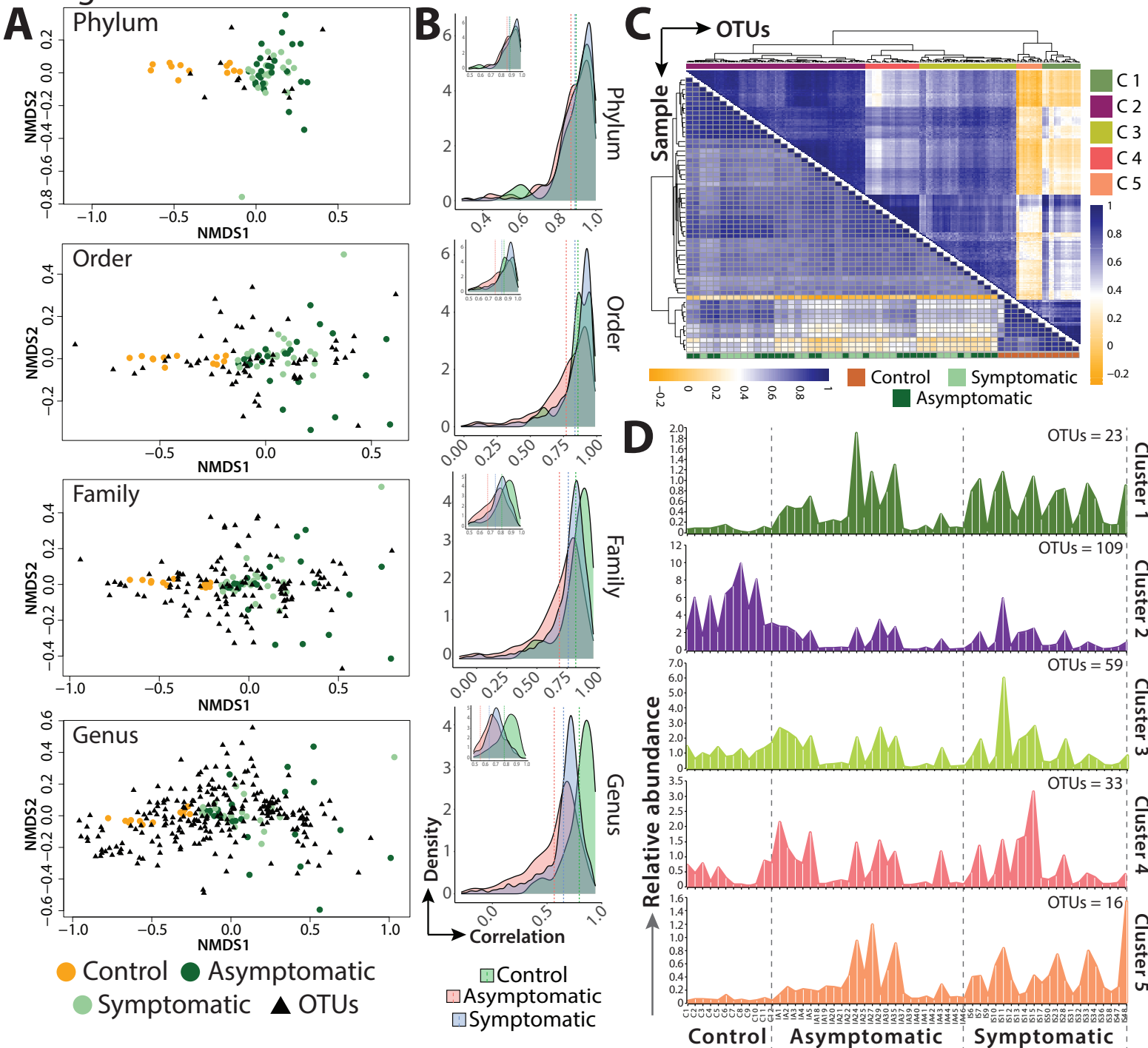


Figure 4

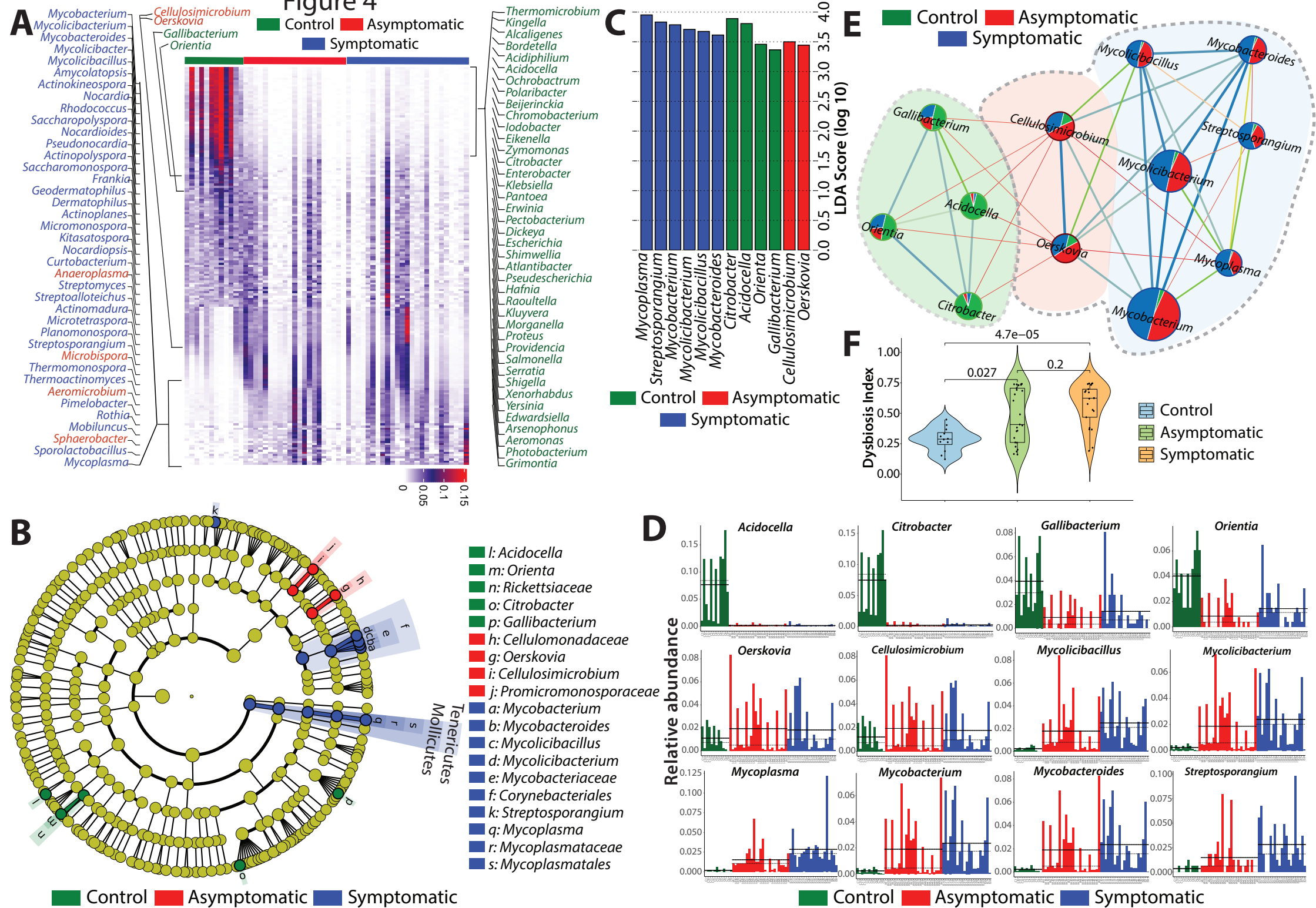


Figure 5

