The Lung Tissue Microbiome in Chronic Obstructive Pulmonary Disease

Marc A. Sze, Pedro A. Dimitriu, Shizu Hayashi, W. Mark Elliott, John E. McDonough, John V. Gosselink, Joel Cooper, Don D. Sin, William W. Mohn, James C. Hogg

ONLINE DATA SUPPLEMENT

Online Supplement

Lung Tissue Preparation and DNA Extraction

After surgical removal lung tissue was frozen as soon as possible using previously published methods (1-3) and 1.5 x 2 cm (diameter x length) cores of this lung tissue were used. These frozen lung tissue cores were cut in a biosafety cabinet level two and approximately 30mg of tissue was taken from each lung sample for DNA extraction using the Qiagne DNeasy kit (Maryland, USA). The quality and quantity of DNA was assessed by Nanodrop (Delaware, USA) using OD 260/230, OD 260/280, and OD 260 readings (Table S1). Working aliquots of 1part extracted DNA and 5 parts DNase and RNase free water for the DNA from non-smoking and smoking controls and COPD (GOLD 4) and 1 in 10 for cystic fibrosis (CF) patients were stored in a -20°C freezer while the undiluted stock DNA samples were stored in a -80 °C freezer.

Materials

All PCR steps used Qiagen 10X PCR buffer, HotstarTaq DNA Polymerase, and dNTP mix. Primers were from SIGMA (Ontario, Canada). The QPCR assays on the ABI 7900HT Sequencing Detection System (California, USA) used 2x SYBR green PCR master mix (Qiagen). *E.coli* was grown within the lab on LB agar and extracted using the DNeasy Tissue and Blood extraction kit (Qiagen). PCR for both the TRFLP and pyrotag sequencing were performed using Bio-Rad My Cycler Thermal Cycler (California, USA). The *HHaI* restriction enzyme used in the TRFLP was from New England Biolabs (Massachusetts, USA). For the pyrotag sequencing primers, nucleotide sequences were provided by McGill University and Genome Quebec Innovation Centre and oligonucleotides bought from SIGMA. A range from 10.6 – 112.8 ng of total DNA was used in all PCR reactions.

Subjects

Out of the 24 patients that comprised the non-smoking, smoking, and COPD (GOLD 4) group only 9 individuals had a complete medication history and none had information on co-morbidities available. Of these 9 individuals only three were one corticosteroids (1 smoker and 2 COPD (GOLD 4) patients). There was no significant difference in bacteria/1000 human cells between those receiving corticosteroids and those not that did not ($P > 0.05$) (Table S2). None of these 9 patients received antibiotics. Of the 8 CF patients medication data was available for only 6. All 6 patients whose medication data was available were prescribed antibiotics.

QPCR

To quantify total eubacteria, primers specifying the 293bp amplicon of the 16S rRNA gene (4)(Table S3) were applied with PCR cycling conditions that were modified from that reported previously (4) in the following manner: 95˚C for 15 minutes then 40 cycles of 95˚C for 30 seconds , and 63˚C for 1 minute. These 40 cycles were then followed with a denaturing curve sequence. A serial dilution of DNA from *E.coli* was used to generate a standard curve (4). Total *E.coli* DNA, as measured by Nanodrop (Delaware, USA) was used to calculate the number of *E.coli* cells based on the genome size of 4.5 million bp. The standard curve for the 16S rRNA gene assay was $y = -3.1638x + 36.373$ ($R^2 = 0.99$).

The Rpp40 assay cycling conditions were 95˚C for 15 minutes followed by 40 cycles of 95˚C for 30 seconds and 60˚C for 1 minute. The standard curve for the human Rpp40 gene assay was $y = -3.7119x + 34.579$ ($R^2 = 0.99$ for both). As with the 16S rRNA assay a denaturing curve sequence followed. Finally, a correction factor using the formula [(Average number of bacteria per sample – average number of bacteria in the negative control) / (number of Rpp40 copies/sample)] x 1000 = bacterial cells / 1000 human cells was applied. A *Lactobacillus* QPCR assay was performed to validate the signal found in the pyrotag sequencing. The assay used

Lactobacillus/*Lactococcus* specific 16S rDNA primers and cycling conditions that were previously published (5) with a standard curve based on the *Lactobacillus acidophilus* species DNA of $y = -3.5037x + 33.197$ ($R^2 = 0.9974$). The total *Lactobacillus* count was then normalized to the previously obtained total bacterial count for the respective sample to generate a percent abundance of *Lactobacillus*.

TRFLP

PCR protocol

Modifications of TRFLP (5, 6) usedthe restriction enzyme *HHaI* (6) to digest a fluorescently-labeled PCR product of 881bp spanning the hypervariable regions V1-V3 (7). The published PCR conditions (6) were modified to give an initial hot start step at 95˚C for 15 minutes, followed by 40 cycles of 94˚C for 40 seconds, 57˚C for 30 seconds, and 72˚C for 90 seconds. Primer sequences can be found on Table S4.

Pyrotag sequencing

The conditions of the first round of the nested PCR were 95˚C for 15 minutes followed by 40 cycles of 94˚C for 40 seconds, 57˚C for 30 seconds, and 72˚C for 90 seconds. This was followed by the second round of PCR utilizing primers that amplified the 550bp sequence spanning the hypervariable regions V1-V3 (7) as described previously (8) with conditions (9) modified to 95˚C for 15 minutes followed by 40 cycles of 94˚C for 40 seconds, 61˚C for 40 seconds and 72˚C for 60 seconds. Primer design can be found on Table S5

Initial pyrotoag sequencing without first round PCR (non-nested) yielded enough amplicons for the CF samples only. Thus a first round and second round PCR (nested PCR) were initiated. Before the analysis of different sample groups were undertaken a comparison of the non-nested cystic fibrosis sequencing versus the nested PCR cystic fibrosis sequencing was completed. The first round nested consisted of the eight unique CF samples while the nested consisted of six of these eight samples and two samples that were duplicates of these six. This was to assess whether or not there was a significant variation between repeated samples in the nested PCR. The two samples from the eight original samples that were excluded were chosen at random.

There was no significant difference (Figure E1) between the cystic fibrosis non-nested samples and the cystic fibrosis nested samples with respect to clustering using a principle coordinate analysis ($P > 0.05$). There was also no significant difference between the duplicate samples and the original or nested samples ($P>0.05$).

There were a total of 686,280 reads from the pooled pyrotag sequencing run. The total reads/sample for the non-smoking and smoking control and GOLD 4 ranged from 3181 – 17295; the CF group, 3265 – 64099; and the negative control group, 3134 – 10570. After sequence cleanup the total pooled pyrotag reads were 319,961. The reads/sample for the non-smoking and smoking control and GOLD 4 groups ranged from 3006 – 16784; the CF group 3260 – 64008; and the negative control group 3062 – 10568.

Pyrotag sequencing pipeline

Sequence analysis

Sequence processing was performed with mothur commands (10) except where indicated. The pipeline consisted of the following steps: (i) Reads having a quality score <25, ambiguous bases, and homopolymers >**6** bp were discarded. (ii) Tags containing the regions targeted by the bacterial primers were extracted from the high-quality reads with the software tool *V-Xtractor,* which implements Hidden Markov Models to locate specific hypervariable regions in 16S rRNA sequence collections (11). Because regions V1-V3 and V1-V2 produced comparable taxonomic profiles (based on classifying a subset of the sequences with the online Ribosomal Database Project *Classifier* tool), we selected the latter segment for all further analyses. (iii) Chimeric sequences were identified and removed with *chimera.uchime* using the “reference=self” option. (iv) Pyrotags were taxonomically classified, using a bootstrap support threshold of 60%, with the naive Bayesian algorithm implemented in the *classify.seqs* command. The sequence collection against which we compared our reads was compiled from the 16S rRNA Greengenes database by trimming entries to regions V1-V3 with V-Xtractor. (v) Pyrosequences were strictly de-replicated, sorted first by abundance and then by length, and binned into operational taxonomic units (OTUs) with an improved version of the algorithm described by (12) which clusters sequences having less than *k* Levenshtein or edit distance values (13). In the implementation of this algorithm, the metric determines the number of deletions, insertions, or substitutions required to transform a *de novo* selected seed sequence into a query sequence. By sorting unique reads according to abundance, we ensured that clusters were seeded with sequences that are sequentially less likely to be error-inducing, minimizing the chance of recruiting reads into spurious OTUs (14). We selected a *k* of 7, resulting in OTUs delimited at a ~97% similarity threshold based on an average read length of 252 bp. Our method also [1] avoids the pitfalls associated with clustering algorithms that rely on multiple sequence alignments, which become unreliable when comparing taxa that span a broad divergence range (15); and [2] excludes homopolymeric counts, a major cause of errors in pyrosequencing data (16), in pair-wise similarity calculations. (vi) OTUs were assigned to a taxon, using the *classify.otu* routine, if at least 50% of the reads within the OTU shared the taxonomic string delineated by classify.seqs.

Statistical analysis

To track changes in the composition of bacterial communities, we calculated pairwise Bray-Curtis (d_{BC}) dissimilarities after square-root-transforming of relative OTU abundances. Bacterial community patterns were visualized with principal coordinates analyses (PCoA) of dissimilarity matrices. The effect of sample groups —healthy smokers, healthy non-smokers, cystic fibrosis, COPD, and negative controls— on bacterial community composition was investigated with a permutational (non-parametric) multivariate analysis of variance (PERMANOVA; (17, 18)) on the basis of Bray-Curtis distances. The statistical significance of the one-way models was evaluated with unrestricted permutations of raw data (*n* = 9,999).

To determine the strength of the association between OTUs and lung disease state, we applied a diagnostic species analysis to all disease state combinations as implemented in Gingko v. 1.7 (28). Rather than using the original IndVal approach, which would aid in identifying the group of samples (state) to which a species (OTU) is maximally associated (19), we chose the biserial correlation coefficient (r_{pb}) as an index of association. This strategy allowed us to determine the degree of preference of an OTU for a given type compared to the remaining types (20). The significance of each indicator value was tested with 9,999 permutations. A false-discovery rate correction (*qvality*; (21)) was applied to account for multiple comparisons

QPCR & Pyrotag Reproducibility

Multiple QPCR of a COPD (GOLD 4) extracted DNA sample (different 30mg tissue piece) yielded similar numbers of total bacteria each time (43.2, 29.1, and 23.8 bacterial cells) and these values were not significantly different than the average of the COPD (GOLD 4) sample group (P > 0.05). For the pyrotag sequencing two CF samples that went through the nested PCR were sequenced twice with good agreement between the two sequencing repeats (less than a 2% difference between relative abundance of the major phylad identified).

References

1. Gosselink JV, Hayashi S, Chau E, Cooper J, Elliott WM, Hogg JC. Evaluation of small sample cdna amplification for microdissected airway expression profiling in COPD. *COPD* 2007;4:91-105.

2. Gosselink JV, Hayashi S, Elliott WM, Xing L, Chan B, Yang L, Wright C, Sin DD, Pare PD, Pierce JA, Pierce RA, Patterson A, Cooper J, Hogg JC. Differential expression of tissue repair genes in the pathogenesis of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2010;181:1329-1335.

3. McDonough JE, Yuan R, Suzuki M, Seyednejad N, Elliott WM, Sanchez PG, Wright AC, Gefter WB, Litzky L, Coxson HO, Pare PD, Sin DD, Pierce RA, Woods JC, McWilliams AM, Mayo JR, Lam SC, Cooper JD, Hogg JC. The relationship between small airway obstruction and emphysema in chronic obstructive pulmonary disease. *N Engl J Med* 2011;365:1567-1575.

4. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, Davies J, Ervine A, Poulter L, Pachter L, Moffatt MF, Cookson WOC. Disordered microbial communities in asthmatic airways. *PLOS One* 2010;5:1-9.

5. Ferreira RBR, Gill N, Willing BP, Antunes LCM, Russell S, Croxen MA, Finlay BB. The intestinal microbiota plays a role in *salmonella*-induced colitis independent of pathogen colonization. *PLOS One* 2011;6:1-11.

6. Liu W-T, Marsh TL, Cheng H, Forney LJ. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl and Environ Microbiol* 1997;63:4516-4522.

7. Neefs J-M, Van de Peer Y, Hendriks L, De Watcher R. Compiliation of small ribosomal subunit RNA sequences. *Nucleic Acids Res* 1990;18Suppl:2237-2317.

8. Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, Schmidt LA, Young VB, Toews GB, Curtis JL, Sundaram B, Martinez FJ, Huffnagle GB. Analysis of the lung microbiome in the "Healthy" Smoker and in COPD. *PLOS One* 2011;6:1-12.

9. Bailey MT, Dowd SE, Parry NMA, Galley JD, Schauer DB, Lyte M. Stressor exposure disrupts commensal microbial populations in the intestines and leads to increased colonization by citrobacter rodentium. *Infect Immun* 2010;78:1509-1519.

10. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres BB, Thanllinger GG, Horn DJ, Weber CF. Introducing Mothur: Open-source, platform-independent, community-supported software for describing and comparing mircobial communities. *Appl Environ Microbiol* 2009;75:7537-7541.

11. Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH. V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *J Microbiol Methods* 2010;83:250-253.

12. Stoeck T, behnke A, Christen R, Amaral-Zettler L, Rodriguez-Mora MJ, Chistoserdov A, Orsi W, Edgcomb VP. Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol* 2009;7:72.

13. Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. *Sov Phys Doklady* 1966;13;451-461

14. Huse SM, Welch DM, Morrison HG, Sogin ML. Ironing out the wrinkles in the rare bioshpere through improved OTU clustering. *Environ Microbiol* 2010;12:1889-1898.

15. Schwarz R, Seibel PN, Rahmann S, Schoen C, Huenerberg M, Müller-Reible C, Dandekar T, Karchin R, Schultz J, Müller T. Detecting species-site dependencies in large multiple sequence alignments. *Nucleic Acids Rese* 2009;37:5959-5968.

16. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. Wrinkles in the rare bioshpere: Pyrosequencing errors can lead to artificial inlfation of diversity estimates. *Environ Microbiol* 2010;12:118-123.

17. Anderson MJ. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 2001;26:32-46.

18. McArdle BH, Anderson MJ. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* 2001;82:290-297.

19. Dufrene M, Legendre P. Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs* 1997;67:345-366.

20. De Cáceres M, Legendre P, Moretti M. Improving indicator species analysis by combining groups of sites. *Oikos* 2010;119:1674-1684.

21. Käll L, Storey JD, Noble WS. Qvality: Non-parametric estimation of q-values and posterior error probabilities. *Bioinformatics* 2009;25:964-966.

Table E1: Nanodrop results for extracted DNA from each sample.

	Sample ID	ng/µL	OD 260	OD 260/280	OD 260/230
Non-Smoking Controls	1977	37	0.75	1.81	1.33
	3037	134	2.58	1.87	1.94
	3262	32	0.64	1.84	1.26
	3480	75	1.50	1.88	1.77
	5909	90	1.81	1.92	1.90
	6376	34	0.68	2.03	1.20
	6788	47	0.96	1.91	1.38
	7180	38	0.76	1.92	1.20
Smoking Controls	2014	122	2.18	1.88	1.64
	2431	13	0.25	1.85	0.76
	5771	18	0.37	1.87	1.04
	5882	94	1.91	1.91	1.62
	6043	37	0.74	2.02	1.38
	6077	34	0.68	1.96	1.41
	6651	74	1.48	1.92	1.52
	6894	72	1.43	1.89	1.59
COPD (GOLD 4)	6965	117	2.34	1.97	2.06
	6967	84	1.68	1.93	1.99
	6968	74	1.49	1.97	1.87
	6969	31	0.61	2.18	1.66
	6971	81	1.63	1.96	1.85
	7013	81	1.64	1.95	1.79
	7014	122	2.42	1.94	1.70
	7015	15	0.32	1.98	1.02
Cystic Fibrosis	2877	157	3.07	1.91	2.33
	5723	201	4.01	1.99	2.30
	5894	333	6.63	1.97	2.33
	5901	127	2.55	1.92	2.31
	5915	152	3.07	1.91	2.33
	5928	173	3.42	1.89	2.26
	5938	221	4.45	1.94	2.33
	6058	204	4.02	1.94	2.38

Table E2: Comparison of bacterial load and corticosteroid use

	Corticosteroid Positive (n=3)	Corticosteroid Negative (n=6)
Bacteria Cells/1000 Human Cells	12.9 ± 12.7	47.5 ± 53.3

Table E3: QPCR primer sequences.

Target	Primer Sequence	Primer Name
16S rRNA gene	5’- GCAGGCCTAACACATGCAAGTC-3’	63F
	5’- CTGCTGCCTCCCGTAGGAGT-3’	355R
Rpp40 gene	5’- CGTAAGCAAGTTTAGTGAATACCTGAA-3’	Forward
	5’- GCACAGCTTCCATCTTACTCAATC-3’	Reverse

Table E4: TRFLP primer sequences.

Target	Primer Sequence	Primer Name
16S rRNA gene	5’ - AGAGTTTGATCMTGGCTCAG-3’	8F
	5’ – CCGTCAATTCMTTTGAGTTT-3’	926R

Table E5: Pyrotag PCR primer sequences

Target	Primer Sequence	Primer Name
16S rRNA gene (1st Round PCR)	5’- AGAGTTTGATCMTGGCTCAG-3’	8F
	5’- CCGTCAATTCMTTTGAGTTT-3’	926R
16S rRNA gene (2nd Round PCR)	5’-X…-N…-AGAGTTTGATCMTGGCTCAG-3’	8F
	5’-X…- GWATTACCGCGGCKGCTG-3’	519R

X = sequence used to attach amplicons to bead sequencing system, N = unique identifer nucleotide sequence assigned to each sample

Table E6: The Top 150 aligned OTU's based on the total abundance of all sequence reads obtained after sequence cleanup.

OTU	Total Abundance	Taxon
1	73597	Bordetella hinzii
2	33723	Flavobacteriaceae unclassified
3	26356	Aquabacterium unclassified
4	15054	Pseudomonas unclassified
5	10880	Acidovorax unclassified
6	13626	Burkholderiales unclassified
7	10315	Diaphorobacter unclassified
8	9535	Pseudomonas geniculata
9	7017	Aggregatibacter aphrophilus
10	6319	Prevotella oris
11	5036	Novosphingobium unclassified
12	4623	Burkholderia unclassified
13	6115	Brevundimonas mediterranea
14	4260	Streptococcus constellatus
15	3935	Corynebacterium unclassified
16	3595	Lactobacillus unclassified
17	3580	Burkholderia multivorans
18	3393	Propionibacterium acnes
19	4142	Pedobacter unclassified
20	3954	Caulobacter leidyia
21	3255	Sphingobacteriales unclassified
22	2065	Allobaculum unclassified
23	1963	Flavobacterium unclassified
24	1861	mle1-12 unclassified
25	1741	Bacteroidales unclassified
26	1603	Rhodocyclaceae unclassified
27	2206	Actinomyces unclassified
28	1505	Lactobacillus unclassified
29	1401	Arcicella unclassified
30	1887	Methylobacterium populi
31	1290	Sphingobium unclassified
32	1205	Bacteroides acidifaciens
33	1095	Psychrobacter unclassified
34	1025	Roseateles unclassified
35	989	Staphylococcus aureus
36	990	Candidatus Rhodoluna unclassified
37	946	Bacteroidales unclassified
38	911	Rhodocyclaceae unclassified

39	847	Gemmata unclassified
40	839	Sphingomonas unclassified
41	803	Clostridium unclassified
42	779	Acidovorax caeni
43	771	Burkholderiales unclassified
44	710	Acinetobacter unclassified
45	706	Sphingomonadaceae unclassified
46	653	Bradyrhizobium unclassified
47	604	Brevundimonas diminuta
48	604	Streptococcus pseudopneumoniae
49	682	Sphingobacteriales unclassified
50	559	Lachnospiraceae unclassified
51	556	Pedobacter unclassified
52	520	Ochrobactrum unclassified
53	533	Chromatiales unclassified
54	532	Prevotella unclassified
55	525	Chryseobacterium unclassified
56	491	Erythrobacteraceae unlcassified
57	490	Rhodoplanes unclassified
58	479	Bacteroidales unclassified
59	467	Blastococcus unclassified
60	450	Allobaculum sp ID4
61	447	Bacteroidales unclassified
62	436	Candidatus_Odyssella unclassified
63	419	Acinetobacter unclassified
64	408	Pseudomonas mendocina
65	407	ACK-M1 unclassified
66	404	Corynebacterium unclassified
67	406	Mycobacterium unclassified
68	409	Agrococcus jenensis
69	397	Sphingomonas azotifigens
70	394	Methylobacterium adhaesivum
71	393	mle1-12 unclassified
72	394	Sphingobacteriales unclassified
73	391	Alcaligenaceae unclassified
74	376	Enterococcus cecorum
75	368	Polaromonas unclassified
76	353	Prevotella unclassified
77	346	Hylemonella unclassified
78	344	Xanthomonas unclassified
79	310	Prevotella melaninogenica

80	314	Microbacterium aurum
81	304	Chryseobacterium unclassified
82	311	Sphingobium unclassified
83	298	Comamonas unclassified
84	296	Capnocytophaga sputigena
85	289	Actinomycetospora unclassified
86	288	Ralstonia unclassified
87	258	Pseudochrobactrum unclassified
88	323	Burkholderiales unclassified
89	383	Bosea vestrisii
90	254	Methylobacterium unclassified
91	251	Lactobacillus iners
92	248	Oxalobacteraceae unclassified
93	235	Micrococcus antarcticus
94	233	Rhizobiales unclassified
95	228	Pseudomonas stutzeri
96	227	Microbacteriaceae unclassified
97	223	Hydrogenophilus unclassified
98	276	Burkholderiales unclassified
99	218	Staphylococcus unclassified
100	217	Flavobacterium unclassified
101	216	Bacteroides unclassified
102	213	Paracoccus marcusii
103	209	Chryseobacterium unclassified
104	209	Lachnospiraceae unclassified
105	205	Corynebacterium durum
106	203	Beijerinckiaceae unclassified
107	202	Bacillus unclassified
108	200	Novosphingobium unclassified
109	262	Pelomonas puraquae
110	198	Parabacteroides distasonis
111	198	Kocuria palustris
112	195	Parvimonas micra
113	189	Bacteroidales unclassified
114	186	Rhodoplanes unclassified
115	184	Pseudomonas unclassified
116	183	Rhodopseudomonas unclassified
117	181	Streptococcus unclassified
118	184	Flavobacteriaceae unclassified
119	177	Rhodospirillaceae unclassified
120	175	Lachnospiraceae unclassified

121	171	Gordonia polyisoprenivorans
122	172	Sporosarcina unclassified
123	175	Xanthomonadaceae unclassified
124	170	Rhodocyclales unclassified
125	173	Pedobacter unclassified
126	167	Neisseria unclassified
127	148	Pseudomonas unclassified
128	167	Patulibacteraceae unclassified
129	168	Nocardioides unclassified
130	155	Lachnospiraceae unclassified
131	150	Staphylococcus epidermidis
132	151	Bacteroidales unclassified
133	149	Janibacter limosus
134	147	Clostridium unclassified
135	147	Spirosoma unclassified
136	145	Methylophilus unclassified
137	136	Lactobacillus reuteri
138	135	Oxalobacteraceae unclassified
139	135	Roseomonas unclassified
140	131	Dermacoccus unclassified
141	131	Prevotella unclassified
142	131	Clostridium unclassified
143	129	Thermicanus unclassified
144	131	Nocardioides alkalitolerans
145	125	Mucispirillum unclassified
146	119	Clostridium unclassified
147	113	Clostridium unclassified
148	111	Agrobacterium unclassified
149	111	Stramenopiles unclassified
150	108	Lachnospiraceae unclassified

Table E7: Top 52 bacterial species based on indicator species analysis with P value < 0.05. Listed in order of lowest P-value for each sample group.

Sample Group	Species Name	Biserial Correlation Coefficient	P-value	Q-value
COPD (GOLD 4)	Lactobacillus Unclassified	0.69101	0.0003	0.011299
COPD (GOLD 4)	Burkholderia Unclassified	0.6489	0.0004	0.014123
COPD (GOLD 4)	Lactobacillus Unclassified	0.68462	0.0007	0.023262
COPD (GOLD 4)	Lactobacillus Unclassified	0.6078	0.0008	0.025108
COPD (GOLD 4)	Burkholderiales Unclassified	0.61993	0.0024	0.07136
COPD (GOLD 4)	Bacteroidales Unclassified	0.52069	0.0104	0.234045
COPD (GOLD 4)	Allobaculum sp ID4	0.49339	0.0116	0.234045
COPD (GOLD 4)	Allobaculum Unclassified	0.52738	0.0141	0.274675
COPD (GOLD 4)	Lactobacillus reuteri	0.51136	0.0156	0.291883
COPD (GOLD 4)	Bacteroides acidifaciens	0.51366	0.0169	0.291883
COPD (GOLD 4)	Burkholderia Unclassified	0.52097	0.0178	0.291883
COPD (GOLD 4)	Acinetobacter Unclassified	0.38883	0.0182	0.291883
COPD (GOLD 4)	Burkholderiales Unclassified	0.49329	0.021	0.312201
COPD (GOLD 4)	Bacteroidales Unclassified	0.48038	0.0359	0.405884
COPD (GOLD 4)	Allobaculum Unclassified	0.47403	0.0364	0.405884
COPD (GOLD 4)	Lactobacillus Unclassified	0.43042	0.0393	0.405884
COPD (GOLD 4)	Lactobacillus Unclassified	0.47295	0.0393	0.405884
COPD (GOLD 4)	Burkholderia heleia	0.48038	0.0393	0.405884
COPD (GOLD 4)	Bacteroides acidifaciens	0.48038	0.0399	0.405884
COPD (GOLD 4)	Bacteroides acidifaciens	0.48038	0.0399	0.405884
COPD (GOLD 4)	Bacteroides acidifaciens	0.48038	0.0399	0.405884
COPD (GOLD 4)	Burkholderiales Unclassified	0.48038	0.0404	0.405884
COPD (GOLD 4)	Caulobacter Unclassified	0.37953	0.0406	0.405884
COPD (GOLD 4)	Staphylococcus hominis	0.47295	0.0406	0.405884
COPD (GOLD 4)	Oscillospira Unclassified	0.39413	0.0407	0.405884
COPD (GOLD 4)	Caulobacteraceae Unclassified	0.45266	0.0411	0.405884
COPD (GOLD 4)	Bacteroidales Unclassified	0.47295	0.0411	0.405884
COPD (GOLD 4)	Lactobacillus Unclassified	0.47295	0.0423	0.405884
Smoking Control	Aquabacterium Unclassified	0.50939	0.0115	0.234045
Smoking Control	Rhodocyclaceae Unclassified	0.43033	0.0424	0.405884
Smoking Control	Acidovorax caeni	0.52223	0.0496	0.424557
Non-Smoking Control	Comamonadaceae Unclassified	0.6083	0.0032	0.09039
Non-Smoking Control	Comamonadaceae Unclassified	0.55322	0.0038	0.102226
Non-Smoking Control	Brevundimonas diminuta	0.53428	0.0103	0.234045
Non-Smoking Control	Diaphorobacter Unclassified	0.49996	0.0114	0.234045
Non-Smoking Control	Comamonadaceae Unclassified	0.51732	0.0172	0.291883
Non-Smoking Control	Hylemonella Unclassified	0.45193	0.0174	0.291883
Non-Smoking Control	Acidovorax Unclassified	0.46682	0.0202	0.308424

Non-Smoking Control	Comamonadaceae Unclassified	0.45968	0.0246	0.356344
Non-Smoking Control	Flavobacterium Unclassified	0.4581	0.0374	0.405884
Non-Smoking Control	Burkholderiales Unclassified	0.4504	0.0386	0.405884
Non-Smoking Control	Flavobacteriaceae Unclassified	0.42184	0.0433	0.405884
Non-Smoking Control	Bradyrhizobiaceae Unclassified	0.52223	0.045	0.405884
Non-Smoking Control	Rhizobiales Unclassified	0.38754	0.0453	0.405884
Non-Smoking Control	Rhodopseudomonas Unclassified	0.36609	0.0457	0.405884
Non-Smoking Control	Burkholderia unclassified	0.51404	0.0466	0.405884
Non-Smoking Control	Acidovorax Unclassified	0.40674	0.0467	0.405884
Cystic Fibrosis	Pseudomonas Unclassified	0.54776	0.0064	0.164345
Cystic Fibrosis	Pseudomonas Unclassified	0.52855	0.0083	0.203868
Cystic Fibrosis	Alcaligenaceae Unclassified	0.54825	0.0186	0.291883
Cystic Fibrosis	Alcaligenaceae Unclassified	0.47414	0.0314	0.405884
Cystic Fibrosis	Pseudomonadaceae Unclassified	0.4765	0.0382	0.405884

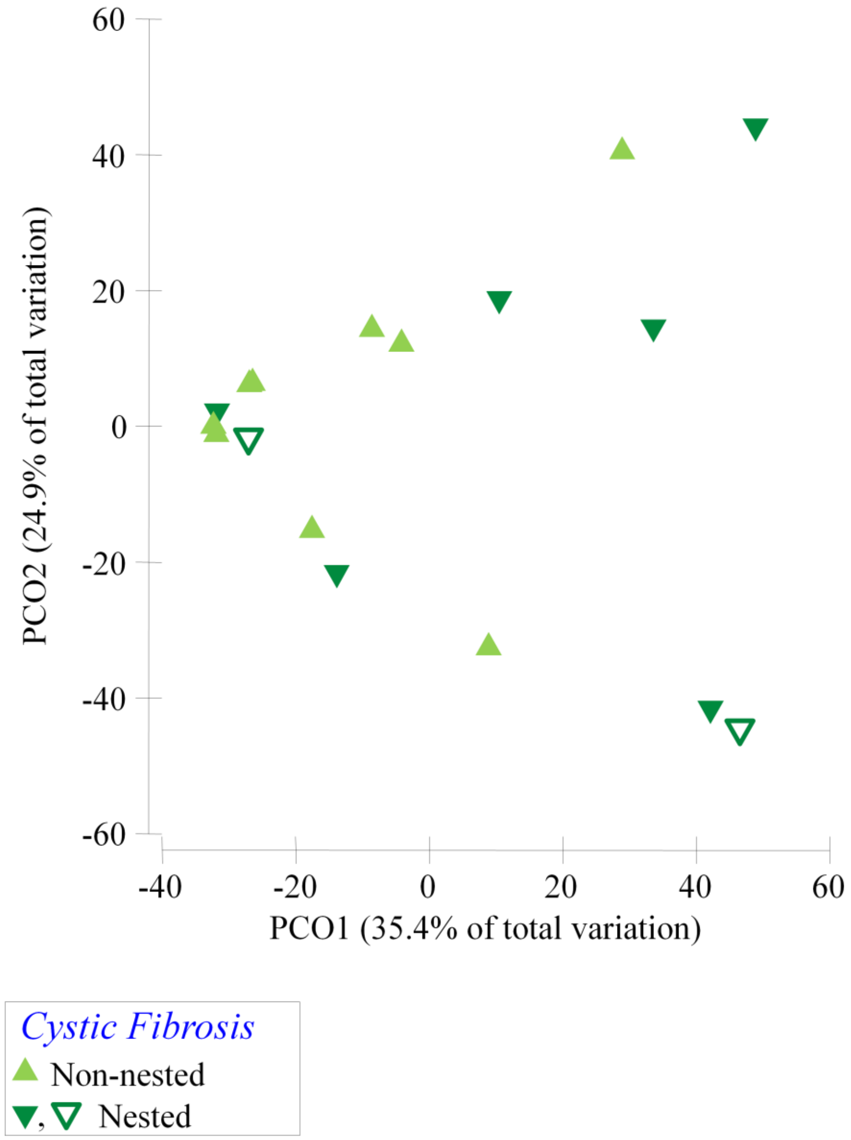

Figure E1: Principle co-ordinate analysis of the nested and non-nested sequencing of the cystic fibrosis samples. The nested samples consisted of 6 unique samples and 2 technical replicates (duplicates of two of the 6 unique samples). Dark green upside-down triangles represents the duplicate of the nested sample closest to it.

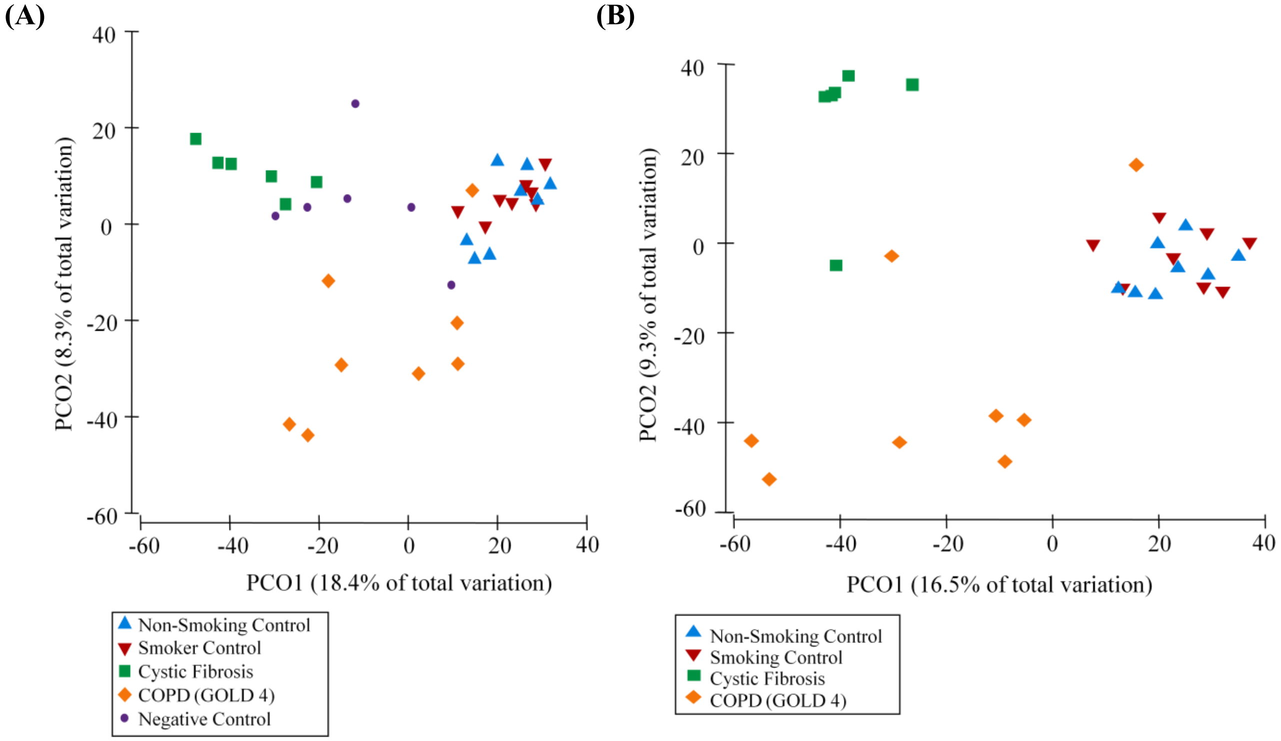

Figure E2: Principle coordinate analysis of pyrotag sequencing results both without subtraction of the OTUs that shared three or greater percent similarity to reads from the negative controls (A) and with subtraction of the negative controls (B). Lung sample groups: n=8, negative controls and cystic fibrosis: n=6.

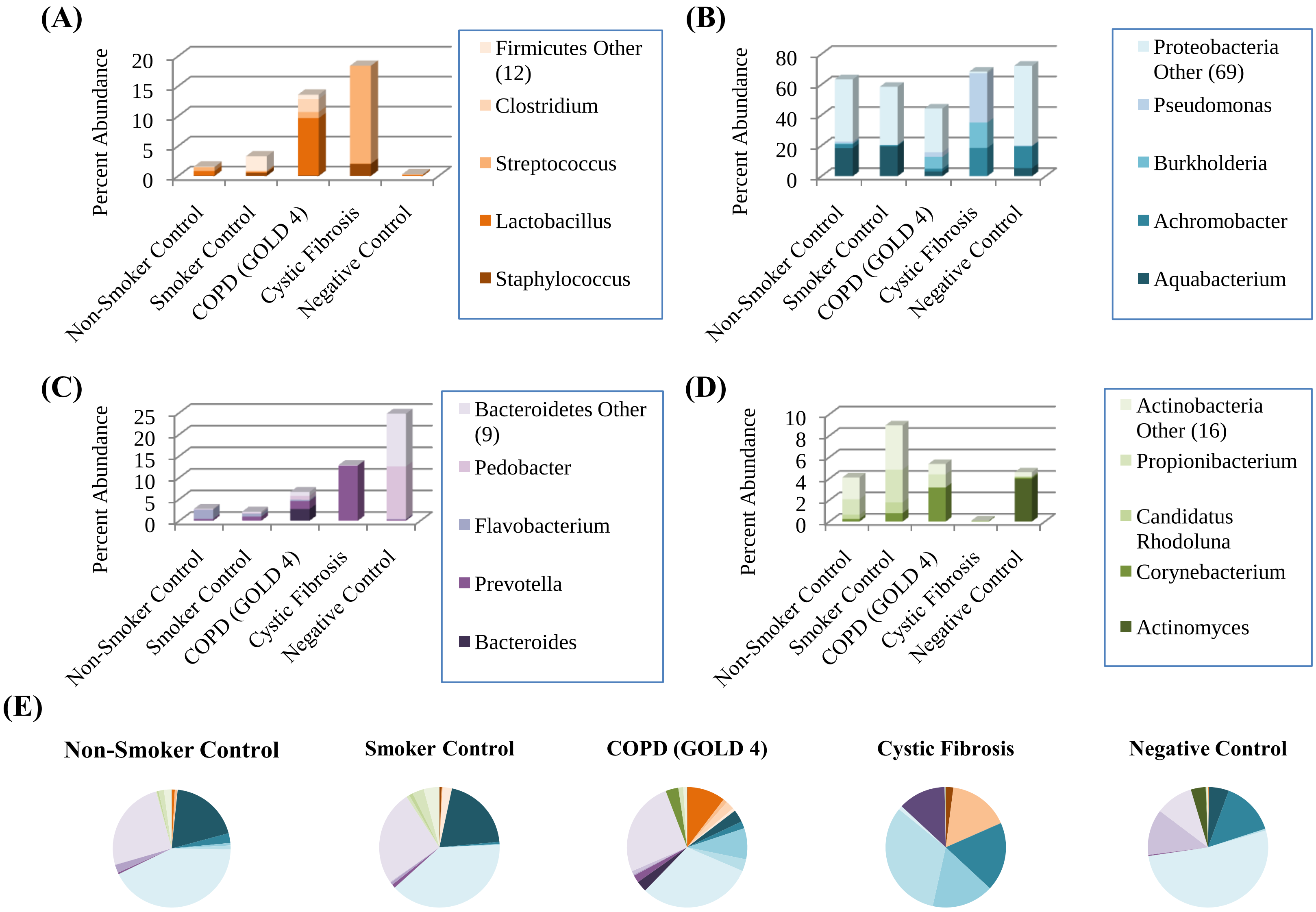

Figure E3: The percent abundance in each sample group of the top 5 genera of the four the major bacterial phyla. A.) The Firmicute phylum. B.) The Proteobacteria phylum. C.) The Bacteroidetes phylum. D.) The Actinobacteria phylum. E.) Pie chart showing overall distribution of the top 5 genera for all phyla in each sample group with colours of each genus corresponding to those used in A, B, C, and D.

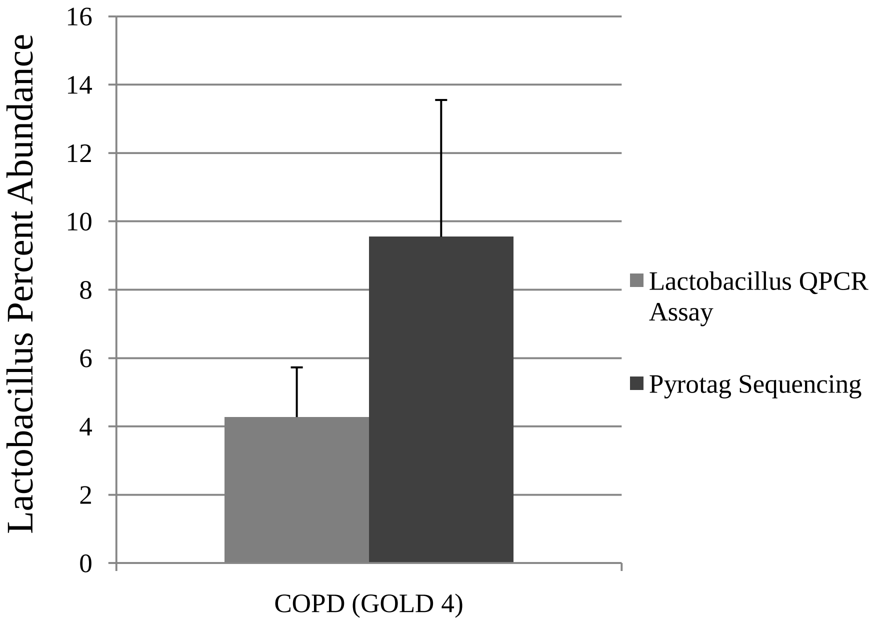

Figure E4: Comparison between pyrotag sequencing and QPCR assay results for *Lactobacillus* in the COPD (GOLD 4) group. There was no significant difference between the two different measurements (P>0.05).