Online Data Supplement

EFFECT OF ADVANCED HIV INFECTION ON THE RESPIRATORY MICROBIOME

Homer L. Twigg III[1], Kenneth S. Knox[2], Jin Zhou[2], Kristina A. Crothers[3], David E. Nelson[4], Evelyn Toh[4], Richard B. Day[1], Huaiying Lin[5], Xiang Gao[5], Qunfeng Dong[5], Deming Mi[6], Barry P. Katz[6], Erica Sodergren[7], George M. Weinstock[7]

Supplemental Methods

Participants: Twenty-two HIV-uninfected participants consisting of healthy current smokers and non-smokers were recruited in Indianapolis, Houston, and Atlanta as part of the Lung HIV Microbiome Project to undergo bronchoalveolar lavage and oral wash sampling. The HIV cohort consisted of 30 subjects who underwent bronchoscopy before and at one month, 1 year, and 3 years after starting HAART. This longitudinal cohort was part of a prospective trial to systematically assess risk factors for the immune reconstitution syndrome (IRIS) in HIV-infected subjects starting HAART (Principle Investigator – Ken Knox, M.D.) Inclusion criteria for both cohorts were men and women ages 18 to 80 years, no use of antibiotics or immunosuppressive medications in the past three or six months respectively, and no evidence of a recent acute respiratory process, defined as reported fever, cough, or upper respiratory symptoms in the previous four weeks. Importantly, to increase the likelihood an IRIS would be found, all HIV subjects had to have a CD4 count < 500 cells/mm^3 at baseline. This research was approved by the Institutional Review Board at Indiana University, the University of Arizona, the University of Houston, and Emory University.

BAL collection and processing: Bronchoscopy with bronchoalveolar lavage (BAL) was performed as previously described (1). Recovered lavage fluid was kept on ice and transferred to the lab within one hour of collection for processing. Lavage fluid was passed through a 100 micron mesh to remove debris and then centrifuged at 400 X g for 10 minutes to pellet BAL cells. The acellular supernatant was harvested and stored at -70^{o}C for subsequent batch DNA extraction and 16S rRNA gene sequencing. In this

study we have analyzed acellular BAL fluid (not whole BAL as in the consortium paper) because of our prior work showing that acellular BAL yields a more distinct respiratory microbiome (compared to oral washes) than whole BAL (2).

16S gene sequencing: DNA was extracted from 5 ml of acellular BAL using a DNeasy kit Blood and Tissue kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. All reagents and materials used were certified DNAse/RNase free. Isolated genomic DNA was shipped to The Genome Institute at Washington University where sequencing was performed using Roche 454 FLX Titanium platform as previously described (3, 4). Primers for variable regions 1 through 3 (V1-3) were utilized. The 16S sequences were processed with the Mothur package (v 1.29) (5). Briefly, sequence reads were deconvoluted into individual samples based on perfect match to the barcode sequences. Primers and barcodes were trimmed from each read, and low-quality and chimeric sequences were removed with default Mothur parameters with one minor adjustment (i.e. the trump symbol was not included at the Mothur's *filter.seqs* step due to our observation that it resulted in over-removal of aligned reads). The remaining high-quality 16S sequences from each sample were classified using the RDP Classifier (v2.5) (6) with the default cutoff value of 0.8 from phylum to genus level. Species-level operational taxonomic units (OTUs) were produced by Mothur based on linkage of 97% sequence identity. The OTU sequences were BLASTed (7) against the RDP 162 database release 11.4 (8). Top BLAST hits were used as species-level annotations if the BLAST alignment achieved $\geq$ 95% coverage and $\geq$ 98% sequence identity. Although the annotations are tentative since only top BLAST hits were

considered, the results still provide a reasonable estimation of the closely related species represented by the OTUs.

Data analysis and statistics: Baseline subject characteristics consisting of dichotomous variables were compared using chi-square tests. Numeric variables were compared using two sample t-tests. Skewed variables were logarithmically transformed to approximate a normal distribution and stabilize variances before the t-tests were performed. Sequencing analysis was carried out as follows. To account for the uneven sequencing depth of each sample, samples were normalized using a procedure as previously described (9, 10). Briefly, subsampling of 500 reads with replacement from each sample was performed 10 times using a customized R script (samples less than 500 reads in total were excluded). After filtering out samples with less than 500 total reads there were 20 uninfected, 26 untreated, 22 four-week treated, 25 one-year treated, and 9 three-year treated subjects. The averaged taxonomic abundance for each sample across the 10 subsamples was used for subsequent analyses using customized R scripts. To identify the genera which might be significantly different between uninfected versus baseline, uninfected versus 1 year treated, and uninfected and 3 year treated subjects, we applied the negative binomial (NB). In this model, the taxa counts were used as the response variable. HIV status was used as the explanatory variable. We also explored age, race, and smoking status as confounding factors in the model. Alpha diversity richness (a measure of the number of unique taxa in a sample) was measured using Chao 1 and ACE richness indices, and diversity (a measure of the distribution of taxa in a sample) was assessed using the Shannon

diversity index and Simpson's index of diversity (1-D).(11). Diversity indices of the HIV-infected (baseline and 1 year treated) and uninfected populations at the OTU or genus level were compared using linear regression models with the diversity indices as the response variable (with or without log transformation) and HIV status as the explanatory variable with and without controlling for age, race, and smoking status as confounding factors. Beta-diversity (a measure of the differentiation of taxa compositions among samples) is a measure of community similarities and was assessed at the genus level using Bray-Curtis dissimilarity (12) computed with the R package *ecodiest* (13) and at the OTU level using UniFrac distances (14) computed with the R package phyloseq (15). Non-metric multidimensional scaling (NMDS), implemented with the R package *vegan* (16) , was applied to visualize the beta diversity among samples. The PERMANOVA test was used to compare UniFrac distances or Bray-Curtis dissimilarities (17, 18) between various populations (untreated HIV-infected subjects, HIV infected subjects with CD4 counts above and below 350 cells/mm^3, and uninfected controls) after adjusting for various combinations or single demographic feature(s) (i.e., age, race, and smoking status). Multivariate dispersion of groups was compared using the *betadisper* command in R *vegan* package by calculating the average UniFrac (weighted, unweighted, normalized, and weighted) distance of group members to the group centroids at the OTU level (19, 20). Principal coordinates analysis (PCoA) of weighted Unifrac distance in 3D plot using beta diversity.py, principal coordinates.py, and make emperor.py commands in QIIME package (21, 22) was performed on the longitudinal OTU abundances for untreated (baseline) samples and 1 month treated, 1

year, and 3 year treated samples after starting therapy. Benjamini-Hochberg corrections were used to adjust for multiple testing.

Supplemental Methods References

1. Twigg HL, Soliman DM, Day RB, Knox KS, Anderson RJ, Wilkes DS, Schnizlein-Bick CT. Lymphocytic alveolitis, bronchoalveolar lavage viral load, and outcome in human immunodeficiency virus infection. *American journal of respiratory and critical care medicine* 1999;159:1439-1444.

2. Twigg HL III ND, Day RB, Gregory RL, Dong Q, Rong R, Knox KS, Crothers K, Sodergren E, Weinstock G. . Comparison of whole and acellular bronchoalveolar lavage to oral wash microbiomes. Should acellular bronchoalveolar lavage be the standard? *Annals of the Am Thorac Soc* 2014;11:S72 - S73.

3. Lozupone C, Cota-Gomez A, Palmer BE, Linderman DJ, Charlson ES, Sodergren E, Mitreva M, Abubucker S, Martin J, Yao G, Campbell TB, Flores SC, Ackerman G, Stombaugh J, Ursell L, Beck JM, Curtis JL, Young VB, Lynch SV, Huang L, Weinstock GM, Knox KS, Twigg H, Morris A, Ghedin E, Bushman FD, Collman RG, Knight R, Fontenot AP, Lung HIVMP. Widespread colonization of the lung by tropheryma whipplei in hiv infection. *American journal of respiratory and critical care medicine* 2013;187:1110-1117.

4. Morris A, Beck JM, Schloss PD, Campbell TB, Crothers K, Curtis JL, Flores SC, Fontenot AP, Ghedin E, Huang L, Jablonski K, Kleerup E, Lynch SV, Sodergren E, Twigg H, Young VB, Bassis CM, Venkataraman A, Schmidt TM, Weinstock GM, Lung HIVMP. Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *American journal of respiratory and critical care medicine* 2013;187:1067-1075.

5. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG,

Van Horn DJ, Weber CF. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* 2009;75:7537-7541.

6. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology* 2007;73:5261-5267.

7. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic acids research* 1997;25:3389-3402.

8. Maidak BL, Olsen GJ, Larsen N, Overbeek R, McCaughey MJ, Woese CR. The ribosomal database project (rdp). *Nucleic acids research* 1996;24:82-85.

9. Hawlena H, Rynkiewicz E, Toh E, Alfred A, Durden LA, Hastriter MW, Nelson DE, Rong R, Munro D, Dong Q, Fuqua C, Clay K. The arthropod, but not the vertebrate host or its environment, dictates bacterial community composition of fleas and ticks. *The ISME journal* 2013;7:221-223.

10. Zhou M, Rong R, Munro D, Zhu C, Gao X, Zhang Q, Dong Q. Investigation of the effect of type 2 diabetes mellitus on subgingival plaque microbiota by high-throughput 16s rdna pyrosequencing. *PloS one* 2013;8:e61516.

11. Faith DP. Phylogenetic pattern and the quantification of organismal biodiversity. *Philos Trans R Soc Lond B Biol Sci* 1994;345:45-58.

12. Lozupone C, Knight R. Unifrac: A new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* 2005;71:8228-8235.

13. Goslee SC, Urban, D. L. The ecodist package for dissimilarity-based analysis of ecological data. *Journa of Statistical Software* 2007;22:1-19.

14. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. Unifrac: An effective distance metric for microbial community comparison. *The ISME journal* 2011;5:169-172.

15. McMurdie PJ, Holmes S. Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PloS one* 2013;8:e61217.

16. Oksanen j, Blanchet, F. G, Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Wagner, H. Vegan: Community ecology package. R package version 2.3-0. 2015.

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

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

19. Anderson MJ, Walsh, D. C. I. Permanova, anosim, and the mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs* 2013;83:557-574.

20. Anderson MJ. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* 2006;62:245-253.

21. Kuczynski J, Stombaugh J, Walters WA, Gonzalez A, Caporaso JG, Knight R. Using qiime to analyze 16s rrna gene sequences from microbial communities. *Current protocols in microbiology* 2012;Chapter 1:Unit 1E 5.

22. Vazquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. Emperor: A tool for visualizing high-throughput microbial community data. *GigaScience* 2013;2:16.

Supplemental Figure Legends

1. Comparison of alpha diversity in acellular BAL fluid at the genera level between uninfected (black), untreated HIV-infected subjects (red), and HIV-infected treated with HAART for one year (purple). Richness (A and B) and diversity (C and D) are significantly greater in the uninfected population compared to a treatment naïve HIV-infected population. However, differences between the HIV infected population and uninfected controls is less after 1 year of treatment.
2. Comparison of beta diversity in BAL at the OTU level between uninfected control subjects from different sites (A) and all untreated HIV-infected subjects compared to HIV-infected subjects (B) using UniFrac Principal Coordinate Analysis. While there are differences between uninfected subjects from different sites, significant differences remain between the entire uninfected and HIV cohorts even when considering the acquisition site of uninfected controls [green = Indiana ($p = 0.001$), blue = Houston ($p = 0.002$), black = Emory University in Atlanta ($p = 0.009$)].
3. Comparison of beta diversity in BAL at the genus level between uninfected and untreated HIV-infected subjects using Bray-Curtis dissimilarity visualized by NMDS. **A.** The HIV-infected population (red) was significantly different compared to BAL from uninfected subjects (black). $p = 0.001$ using Bray-Curtis dissimilarity in the PERMANOVA test. **B.** Subdividing HIV-infected subjects into those with CD4 counts above (green) or below (blue) 350 cells/mm^3 did not affect the results.

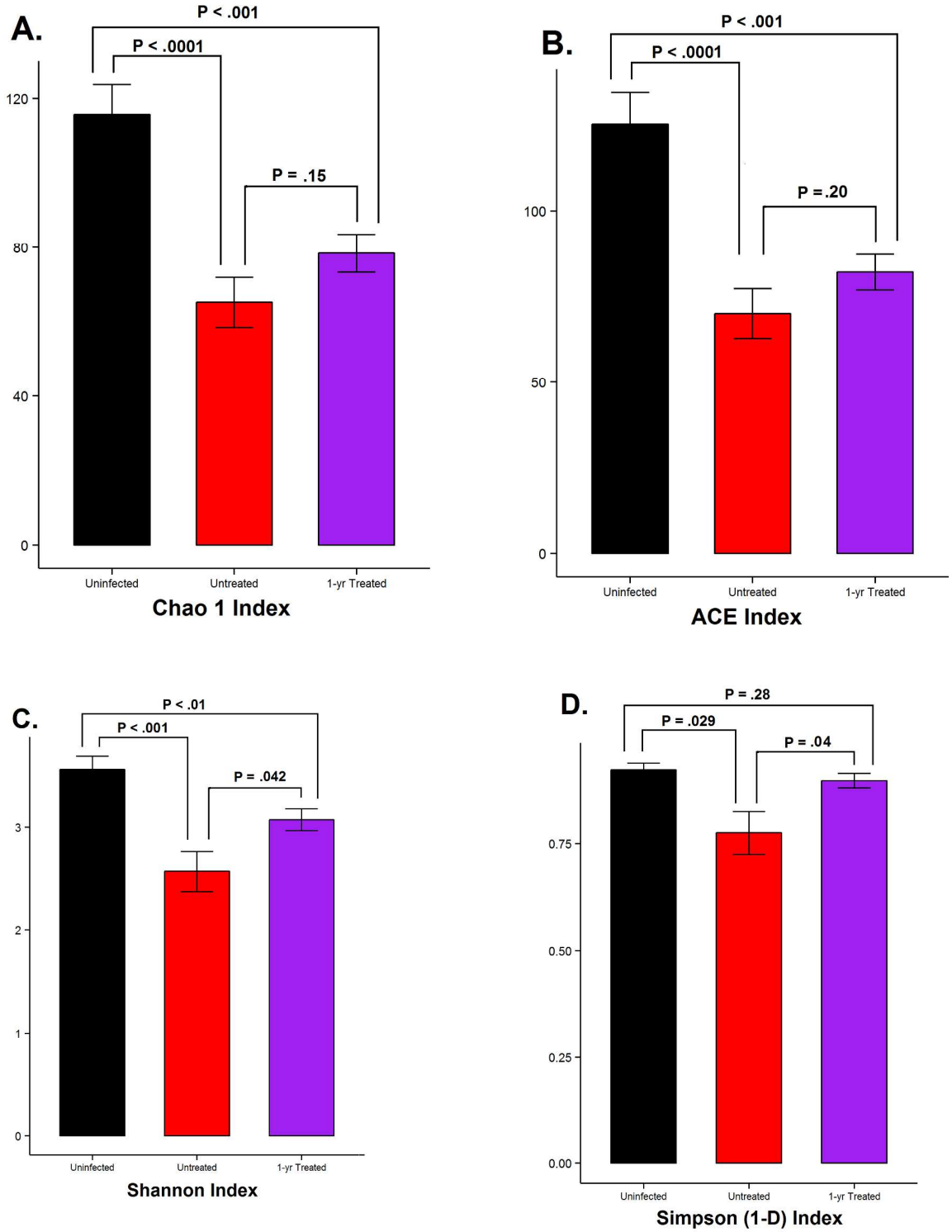
4. Comparison of beta diversity in BAL at the genus level between uninfected and treated HIV-infected subjects using Bray-Curtis dissimilarity visualized by NMDS. The top panel represents HIV-infected subjects who have been treated for one year, the bottom panel three years. The HIV-infected population (red) remains significantly different compared to BAL from uninfected subjects (black) even after 1 year on therapy. This difference is slightly less after three years of HAART.
5. Comparison of differences in beta diversity between uninfected controls versus untreated HIV-infected subjects and uninfected controls versus HIV subjects treated for one year. Only subjects with both baseline and 1 year treated samples were used in this analysis (n=22). 100 pair-wise weighted Unifrac distances from one group (Uninfected vs. Untreated) were randomly selected and compared to their counterparts in the other group (Uninfected vs. 1-yr Treated). Unifrac distances between uninfected subjects and HIV-infected subjects declined slightly but significantly with therapy ($p < 0.001$ using paired T test).
6. Comparison of oral wash (OR - light blue), acellular BAL (BA - dark blue), and environmental controls consisting of a bronchoscope prewash (PW - gold) in normal subjects who made up the control population for this manuscript. Oral wash, acellular BAL, and environmental controls are different. Furthermore, the environmental controls were dominated by Dyella and Streptophyta, two organisms found exceedingly rarely (< 0.001%) in BAL samples.

Supplemental Table 1: Baseline pulmonary function in HIV-infected subjects (mean ± SD).			
Measurement	**Smoking Status at Baseline**		
Diffusion Capacity	**Combined (n=28)**	**Smoker (n=16)**	**Nonsmoker (n=12)**
DsbHb Actual	24.2 (7.0)	22.5 (7.3)	26.5 (6.1)
DsbHb %Predicted	76.0 (17.9)	70.3 (16.2)	83.8 (17.8)
D/VAsbHb Actual	4.1 (1.0)	3.7 (0.8)	4.6 (1.0)
D/VAsbHb %Predicted	73.5 (23.0)	65.9 (22.7)	83.8 (20.1)
Spirometry	**Combined (n=30)**	**Smoker (n=16)**	**Nonsmoker (n=14)**
FEV1 (L)	3.6 (0.9)	3.6 (1.0)	3.7 (0.7)
FEV1 (L) %Predicted	95.2 (18.0)	92.6 (11.5)	98.1 (23.4)
FVC (L)	4.8 (1.1)	4.7 (1.4)	4.8 (0.7)
FVC (L) %Predicted	100.7 (16.3)	98.4 (13.3)	103.4 (19.3)
FEF 25-75 (L/s)	3.5 (1.4)	3.2 (1.0)	3.9 (1.7)
FeF 25-75 (L/s) %Predicted	83.9 (30.2)	84.7 (17.8)	83.1 (40.8)

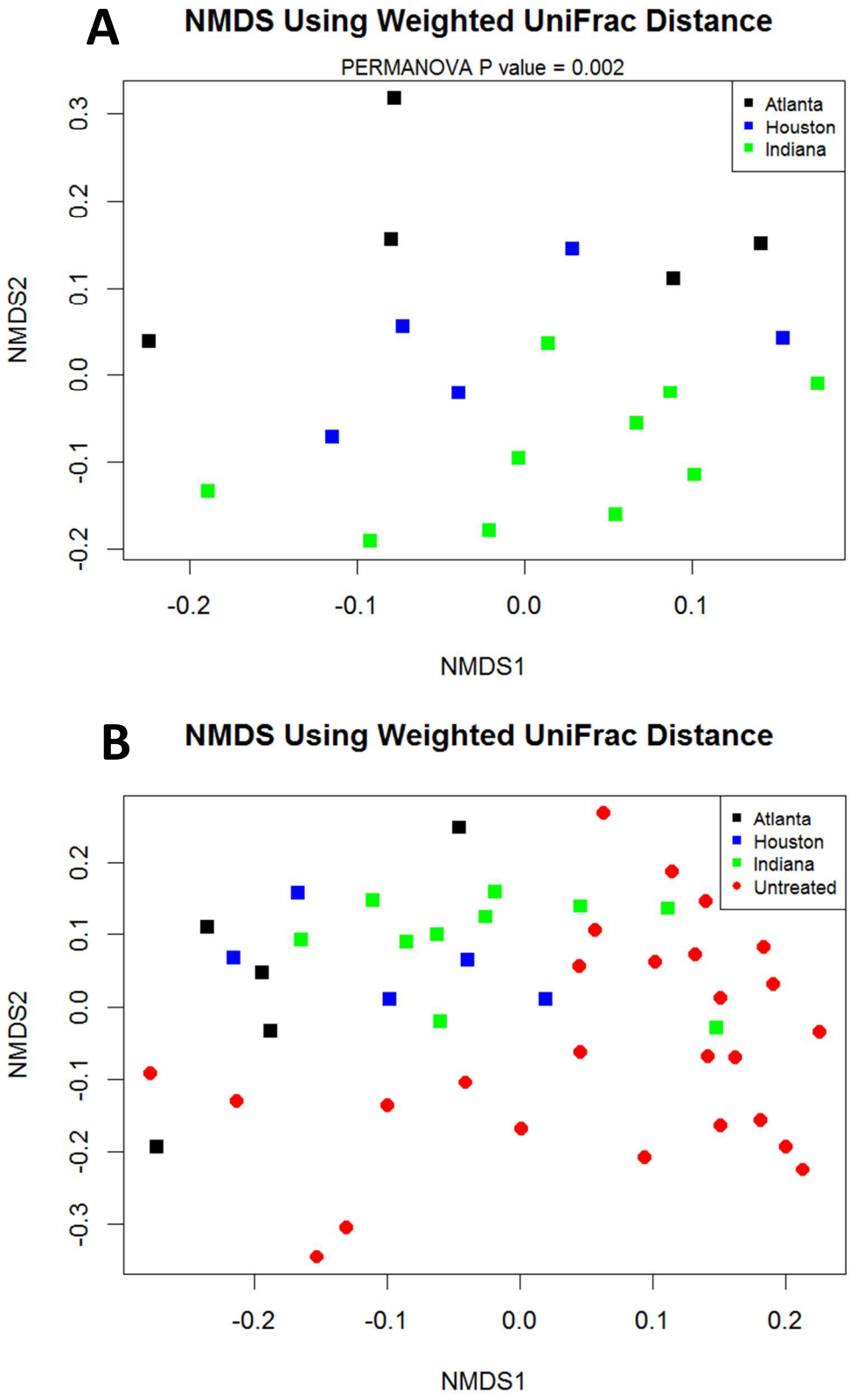
Supplemental Table 2: 16s sequence yields in uninfected and HIV-infected groups.				
	Uninfected	**Baseline HIV**	**1 year treated HIV**	**3 year treated HIV**
High Quality Sequences	4646±3095	8068±9781	3218±2920	3006±4069
% bacteria classified	86%	95%	94%	95%

Supplemental Table 3: 15 most abundant genera in uninfected, HIV-infected, and 1 year treated HIV infected subjects.					
Uninfected		**Untreated HIV-infected**		**1 year treated**	
Organism	**%**	**Organism**	**%**	**Organism**	**%**
Flavobacterium	13.2%	Streptococcus	27.9%	Prevotella	21.9%
Streptococcus	11.9%	Prevotella	13.9%	Streptococcus	21.5%
Prevotella	9.8%	Tropheryma	9.4%	Veillonella	8.1%
Lactobacillus	5.0%	Lactobacillus	5.6%	Actinomyces	7.9%
Veillonella	3.9%	Veillonella	5.5%	Porphyromonas	5.7%
Fusobacterium	3.6%	Actinomyces	3.8%	Fusobacterium	3.8%
Porphyromonas	3.3%	Propionobacterium	3.0%	Tropheryma	3.4%
Staphylococcus	2.0%	Fusobacterium	2.8%	Gamella	1.7%
Actinomyces	2.0%	Porphyromonas	1.9%	Neisseria	1.6%
Tropheryma	1.9%	Gamella	1.9%	Lactobacillus	1.5%
Corynebacterium	1.9%	Neisseria	1.5%	Rothia	1.4%
Neisseria	1.6%	Burkholderia	1.4%	Treponema	1.2%
Treponema	1.4%	Capnocytophaga	1.3%	Granulicatella	1.2%
Gamella	1.3%	Campylobacter	1.3%	Megasphaera	1.2%
Propionibacterium	1.0%	Rothia	1.1%	Sneathia	1.1%

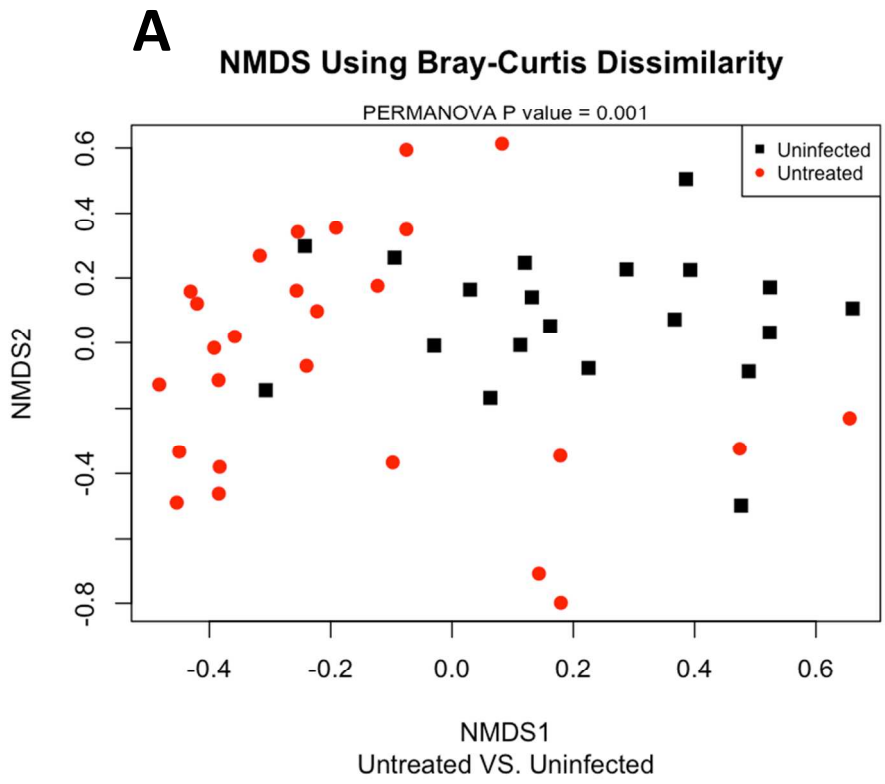
Supplemental Figure 1

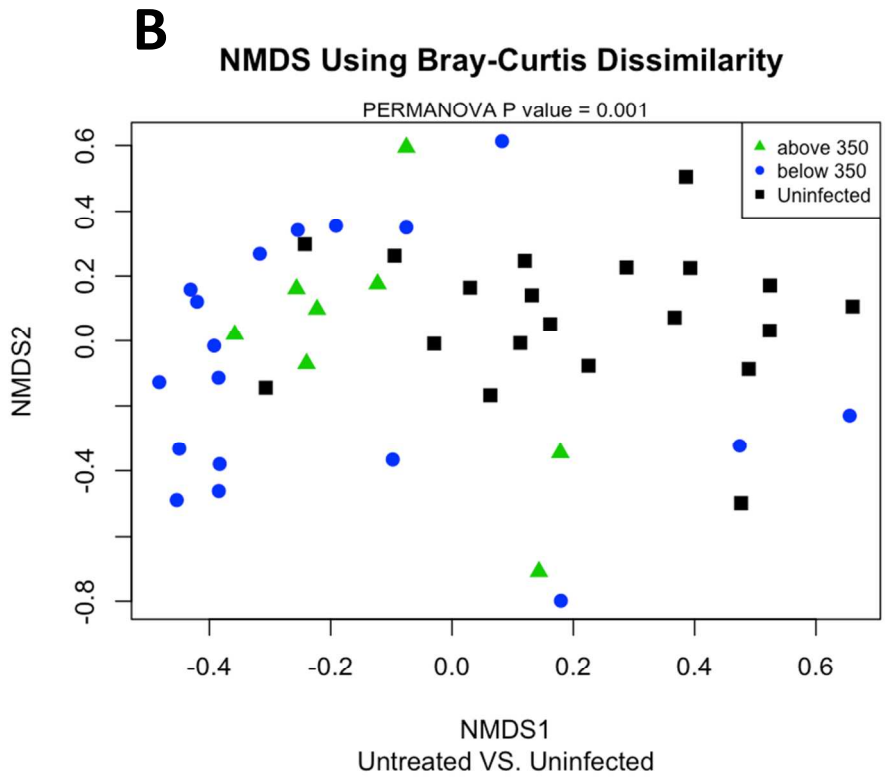


Supplemental Figure 2

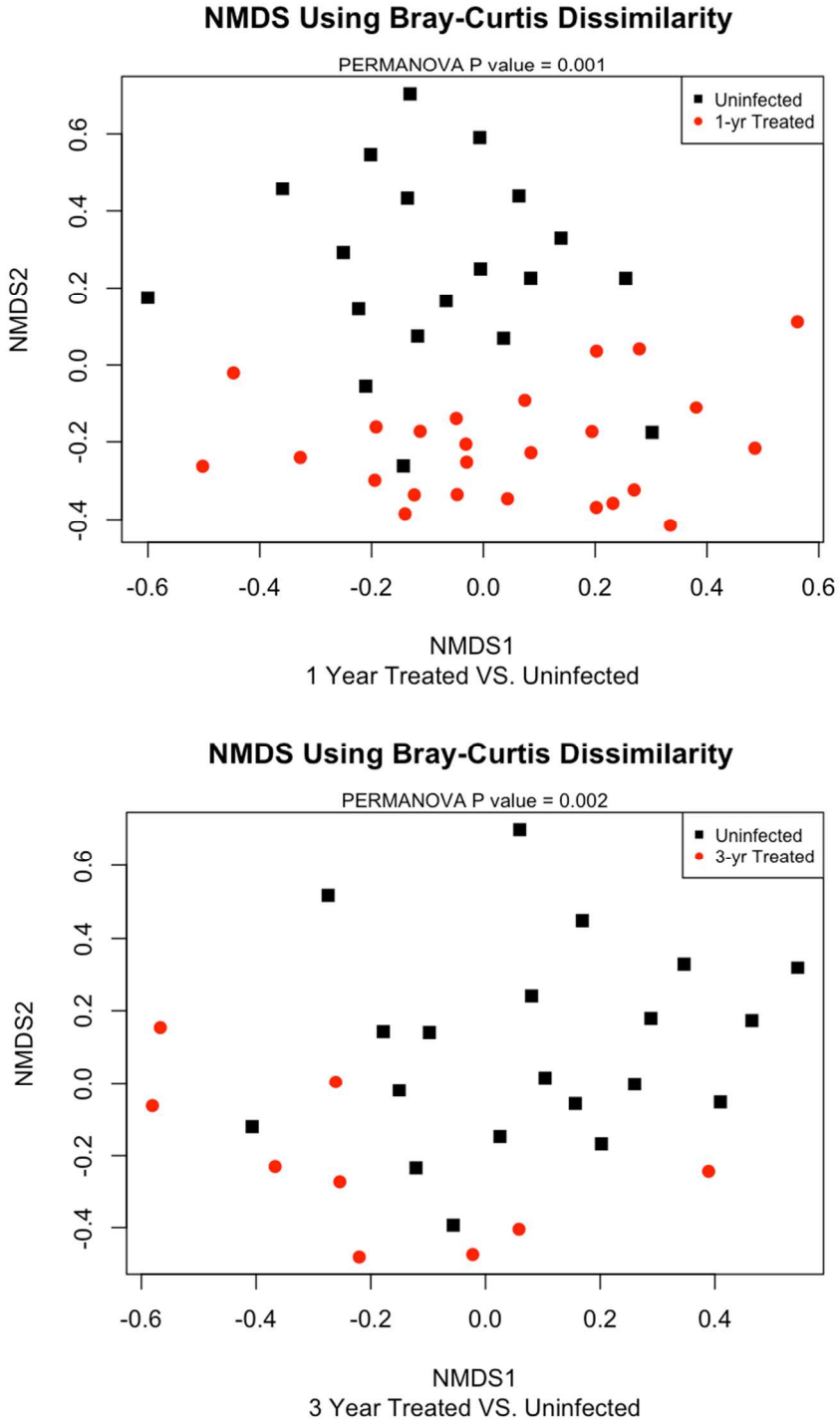


Supplemental Figure 3

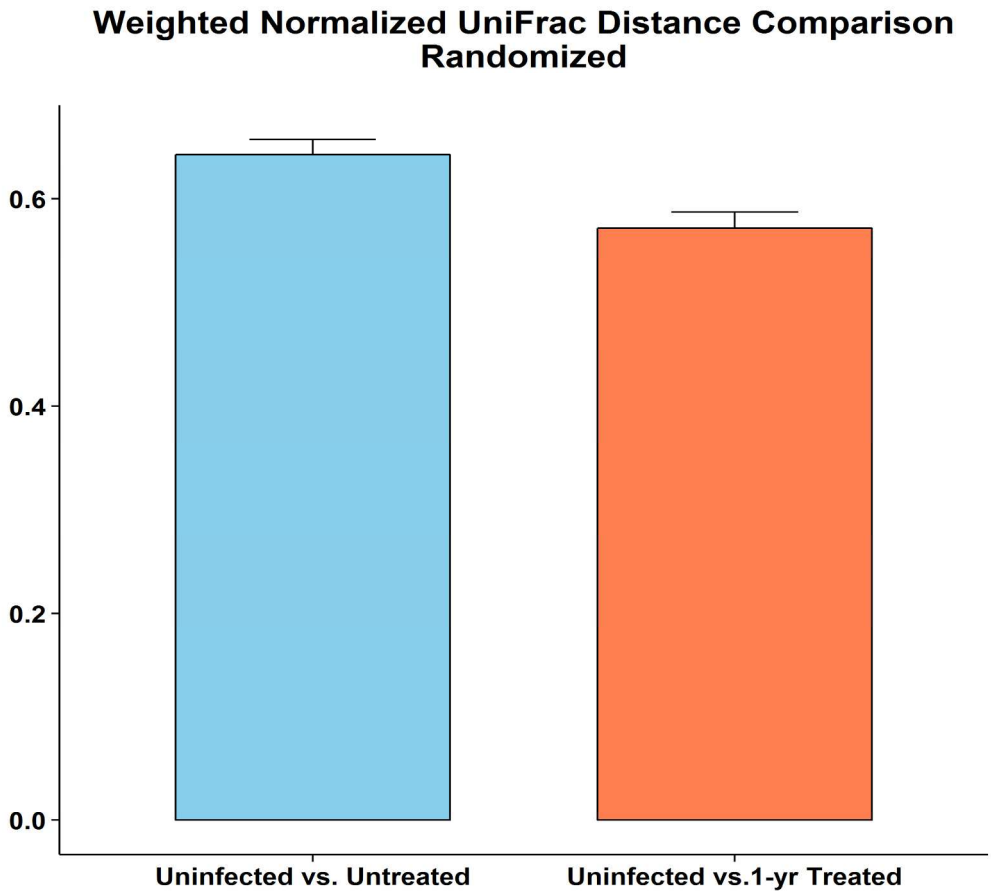




Supplemental Figure 4



Supplemental Figure 5



Supplemental Figure 6