Using microbial composition within sputum transcriptome data to stratify patients by asthma severity

Daniel J Spakowicz *Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT*
Qing Liu *Section of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine; Yale School of Medicine, New Haven, CT*
Vera Nezgovorova *Section of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine; Yale School of Medicine, New Haven, CT*
Xiting Yan *Section of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine; Yale School of Medicine, New Haven, CT*
George Weinstock *The Jackson Laboratory for Genomic Medicine, Farmington, CT*
Geoffrey Chupp *Section of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine; Yale School of Medicine, New Haven, CT*
Mark Gerstein *Program in Computational Biology and Bioinformatics, Department of Molecular Biophysics and Biochemistry, Department of Computer Science; Yale University, New Haven, CT*

Asthma is a highly heterogeneous disease and many of its clinical manifestations are resistant to treatment. Recent efforts have suggested that some of this heteogeneity may relate to the bacteria, fungi and other organisms that are resident or transiently present in the lungs. Explored here is the relationship between the microbiota observed in sputum and clinical parameters for asthmatic patients with diverse disease presentations. Sputum RNA sequencing was performed on 122 samples and the microbial composition determined. Hierarchical clustering showed slight separation between asthmatics and controls and between moderately severe asthmatics and other asthma severities. Correlations with clinical variables showed signficant relationships between bacteria and fungi for such variables as the number of hospitalizations, lung function spirometry, and the percent of eosinophils in the sputum. In particular, the percent eosinophils was related to fungi in the genus *Candida* and several bacteria. Random forest regression further supported the relationship between *Candida* and the percent eosinophils and to a lesser extent *Campylobacter* and *Saccharomyces*. This demonstrates the need to consider fungi in addition to bacteria in the study of the microbiome associated with asthma.

Keywords: asthma, microbiome, RNAseq, metatranscriptomics

Introduction

Asthma afflicts over 300 million people worldwide and approximately 30 million in the United States. For reasons that are largely unknown, the prevalence of asthma has risen to epidemic proportions over the past five decades [Masoli et al.], resulting in roughly 15 billion dollars in health expenditures in the US each year [Lugogo and Kraft]. While the understanding of disease pathogenesis has increased in recent years, the morbidity related to asthma remains high,

accounting for 10 million school absences each year and limitations to physical activity reported by approximately half of asthma patients [Bousquet et al.]. Efforts to develop better therapeutics are hampered by the heterogeneity of the disease, the source of which remains poorly understood. Recently suggested as a potential source of this heterogeneity is the airway microbiome [Huang and Boushey].

The role of the airway microbiome in the development of disease is being increasingly appreciated. Commensal microbiota have been shown in other contexts to be strong regulators of host immune system development and homeostasis [Round and Mazmanian]. Disturbances in the composition of commensal bacteria can result in imbalanced immune responses and affect an individual's susceptibility to various diseases, including inflammatory (IBD and colon cancer), autoimmune (e.g., celiac disease, arthritis), allergic (e.g., asthma and atopy) and metabolic (e.g., diabetes, obesity, metabolic syndrome) (reviewed in [Shreiner et al.]).

Investigation of the microbiota in the lower respiratory tract is a relatively new field in comparison to the extensive work on the intestinal tract. In fact, the lung was excluded from the original Human Microbiome Project because it was not thought to have a stable resident microbiome [Turnbaugh et al.]. A limited number of reports have investigated the changes in the lung microbiota between healthy, non-smoking and smoking individuals as well as in patients suffering from Cystic Fibrosis (CF), Chronic Obstructive Pulmonary Disease (COPD) or Asthma [Erb-Downward et al., Hilty et al., Huang et al., Morris et al.]. Despite emerging data on airway microbiota, little is known about the role of the lung microbiome in modulating pulmonary mucosal immune responses. The lung microbiota in humans has been observed to include on the order of hundreds of bacterial species per person and exhibits exceptional inter-individual diversity [Zemanick et al.] that relate to the clincal heterogenity of asthma.

Traditional methods for the analysis of airway microbiota involve the amplification of ribosomal RNA (rRNA) gene fragments and then sequencing the mixture of amplicons, however, recently studies have shown that this signal is confounded by non-viable sources. For example, swabbing ATM buttons in different neighborhoods in New York City demonstrated the ability to distinguish neighborhoods by food preferences, such as chicken and fish [Bik et al.]. In addition, the primers used to amplify the rRNA fragments has been shown to bias the results, most strongly in that a single kingdom (typically bacteria) is sampled in each experiment. In contrast,

RNA is more environmentally labile and therefore more likely to be observed only if isolated from intact, metabolically active cells. In addition, deep sequencing of the total RNA present in a sample, so-called meta-transcriptomics, avoids biases introduced by specific primer amplification and enables discovery of organisms from multiple kingdoms.

Here we use RNA sequencing of the sputum of asthmatic patients to identify and quantify the non-human community of organisms. Machine learning approaches are used to correlate microbial taxa with a standard clinical lung function test for asthmatic patients. This work speaks to the phenotypic heterogeneity of asthmatic patients by incorporating differences in their sputum microbiomes and may provide insight into the biological mechanisms that drive those differences.

Group	Control	Asthma		
Asthma Severity		MILD	MODERATE	SEVERE
Num Indiv	16	21	41	44
Mean Age (St. Dev)	44.6 (15.2)	44.3 (17.2)	50.1 (16.6)	46 (12.8)
Perc Female	62.5	76.2	80.5	70.5
BMI	26.7	24.9	28.9	33.2
BDR	3.5	5.0	8.8	12.8
ACT Score	19.4	20.0	17.5	12.7
Median Percent Eosinophils	1.7	5.1	7.5	9.5
Perc White	88	76	83	45
Perc Black	6	14	10	32
Percent Other	0	10	7	23

Table 1: Patient characteristics

Methods

Sample collection and sequencing

Sputum induction was performed with hypertonic saline, the mucus plugs dissected away from saliva, the cellular fraction separated and the RNA purified as described previously. Briefly, RNA was purified using the All-in-One purification kit (Norgen Biotek) and its integrity assayed by Agilent bioanalyzer (Agilent Technologies, Santa Clara, CA). Ten ng RNA was amplified using random primers and the WT-Ovation Pico RNA amplification System (NuGen, San Carlos, CA). Samples were sequenced using an Illumina HiSeq 4000 with 2x125 bp reads, with an average of 40 million reads per sample.

RNAseq processing by exceRpt

An adapted version of the software package exceRpt [**?**, ???], was used to process and conservatively search for exogenous sequences within RNA-seq data. Briefly, RNA-seq reads are subjected to quality-assessment using the FastQC software v.0.10.1 [**?**, ???] both prior to and following 3′ adapter clipping. Adapters were removed using FastX v.0.0.13 [**?**, ???]. Identical reads were counted and collapsed to a single entry and reads containing N's are removed. Clipped, collapsed reads are mapped directly to the human reference genome and pre-miRNA sequences using STAR [Dobin et al.]. Reads that did not align are mapped against a ribosomal reference library of bacteria, fungi and archaea, compiled by Ribosome Database Project [Cole et al.]. Remaining reads are aligned to genomes of other organisms including bacteria, fungi, plants and viruses, retreived from GenBank [Benson et al.].

Microbial abundance counting and normalization

Reads mapping to taxa were normalized to the total sequencing output for each sample and presented as the number reads mapping to each taxon per million sequencer reads. Reads that mapped to multiple reference genomes were assigned to the phylogenetic tree node shared by all genomes to which the read mapped. For example, if a read mapped equally-well to two species in the genus Bacterioides, the read would be assigned to the genus node.

Results and Discussion

The RNA isolated from sputum samples from 122 patients were sequenced with a median of 47.5 million reads per sample. A median of 60% of the reads aligned to the human reference genome and 50% to annotated transcripts (Figure 1, green bars), which is consistent with other RNA sequencing efforts on samples of this type. A median of 0.7% of the input reads aligned to exogenous sources, with some samples containing as much as 28.1% exogenous reads. A large portion of the reads remained unmapped to any references (median = 27.8%, min = 5.6, max = 90)(Figure 1, grey bars).

A median of 87.4-thousand reads aligned to exogenous ribosomal RNA references, or 0.196%. Of these reads 8.01% aligned uniquely to 1746 species; the remaining alignments scored equally

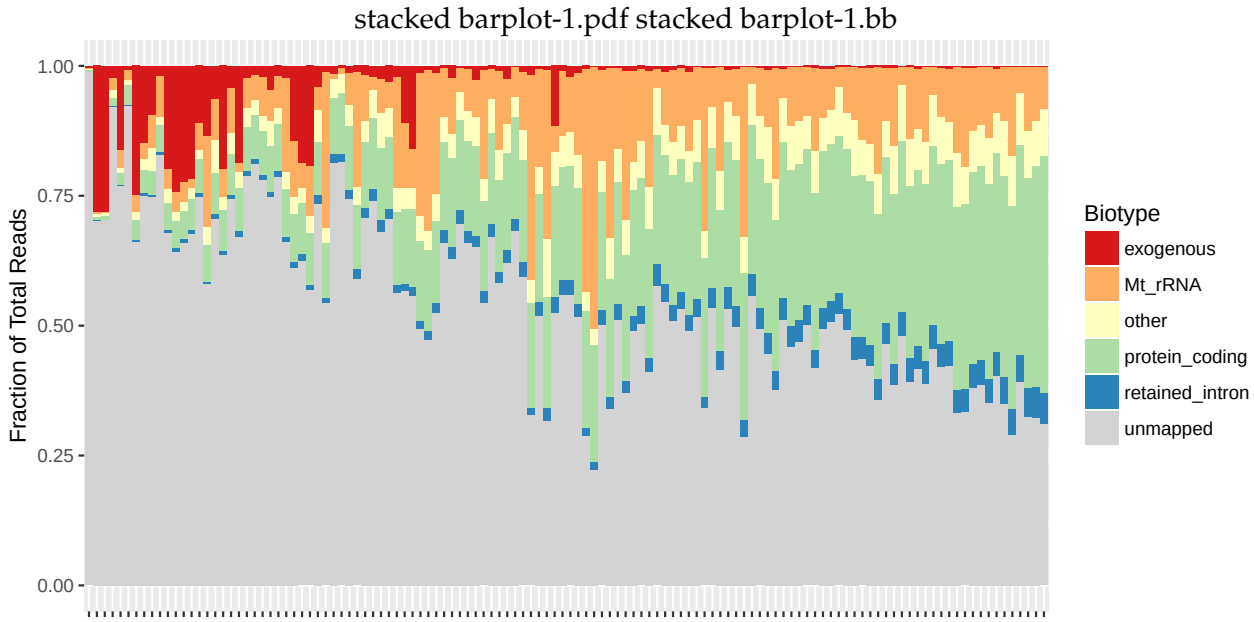

Figure 1: Alignment summary for the RNAseq of control and asthmatic sputum. There is a large intra-sample variability in the percent of reads aligning to each biotype.

well to multiple references and were assigned to the nearest parent node representing the top scoring matches.

The alpha diversity of these samples was not signficantly different between the different asthma severity groups (Figure 2), as defined by the amount of fluticasone or equivalent per day to control symptoms (mild = $> 200\mu g$, moderate = $200 - 800\mu g$, severe = $> 800\mu g$). The Fisher's alpha, Shannon and Simpson diversity metrics showed slightly higher diversity in mild asthmatics relative to other groups, though with a wide distrubution of values for both asthmatics and controls. Other studies have observed lower alpha diversity in asthmatics (n = 6) relative to controls (n = 8) using transcriptomics of nasal cavity swabs from children and adolescents [Castro-Nallar et al.]. It is possible that the differences observed here are due to the wider distribution of patient ages, sampling of the sputum rather than the nasal cavity, or larger number of samples. Notably, more fungi were observed in our study than by Castro-Nallar et al.

The dominant phyla observed in the samples was Proteobacteria, followed by Firmicutes and Bacteroidetes (Figure 3). The abundance of Proteobacteria is in contrast to observations from the gut where Bacterioides predominate [Turnbaugh et al.]. Also notable was the presence of two phyla of fungi among the eight most abundant overall, though clearly in lower abundance than

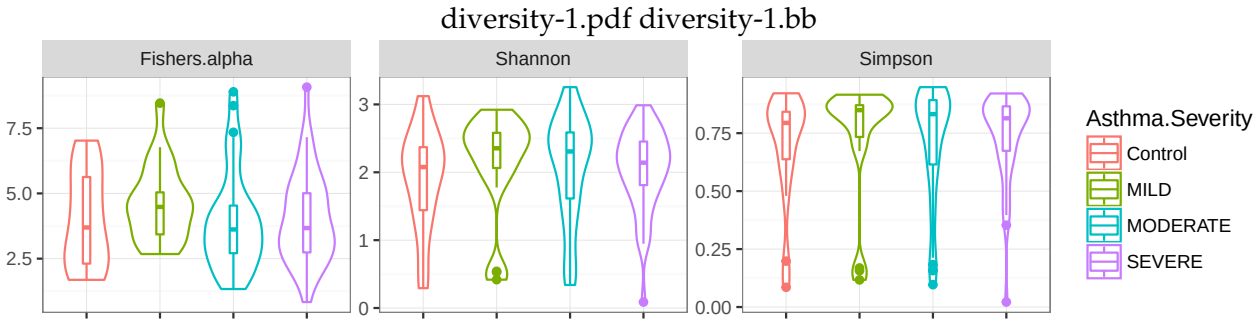

Figure 2: Alpha diversity metrics by asthma severity. All asthmatics have a lower fisher's alpha than the control group. The severe asthma group has a higher shannon and simpson diversity than the others.

many of the bacterial phyla.

Though the asthma severity categories were not significantly different in their alpha diversities, significant category enrichment when clustering by the beta-deversity metric Bray-Curtis distance was observed (dengrogram cut height 0.7, fisher's exact test p-value = 0.02)(Figure 7). In particular, one of the three major clusters observed was significantly depleted in control samples (permutation test p-value = 0.0085) and significantly enriched in moderate asthmatics (permutation test p-value = 0.0058). This group has moderate levels of both Proteobacteria and Firmucutes, but the highest Bacteroidetes, Spirochaetes and Fusobacteria levels in the cohort. However, this cluster was not signficantly different from the other clusters in any continuous clincal variables after accounting for multiple hypothesis testing. We therefore sought to identify if specific taxa were significantly correlated with the clinical parameters by regression approaches.

Correlation with clinical variables

Microbial ribosomal RNA abundances at all taxonomic levels were correlated with continuous clinical variables. After controlling for the effects of age, body mass index and gender 14 (40% of total) clinical variables were significantly associated with one or more of 66 1% of total) exogenous taxa (FDR < 0.05)(Figure 4). This included the total signal for exogenous sequences, which was strongly positively associated with the total pack years of smoking for the patients. Interestingly, none of the individual taxa were associated with the total pack years of smoking, perhaps suggesting an overall effect of smoking as increasing the total microbial load without selecting for a subset of the organisms. Alternative explanations include that the chronic inflammation in

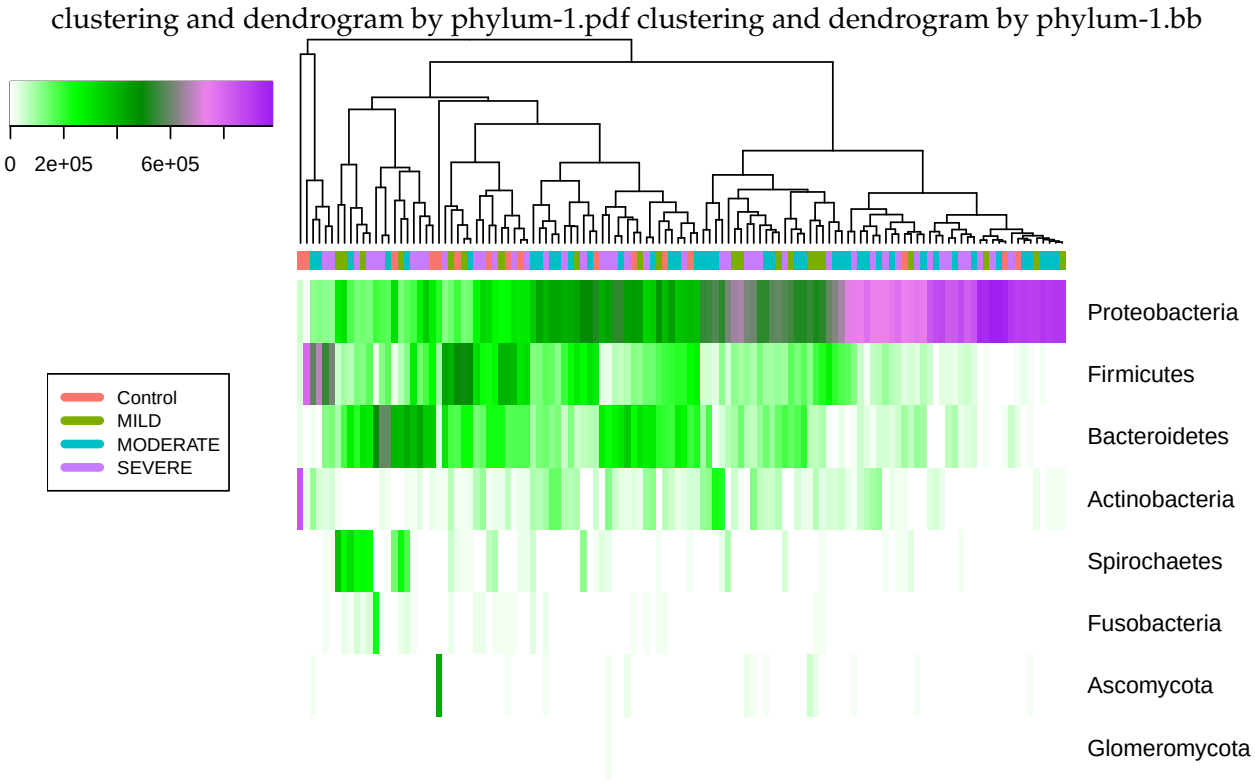

Figure 3: Heatmap of the ten most abundant phyla observed in the data. The amount of Proteobacteria observed strongly influences the hierarchical clustering of the samples.

asthmatic lungs has phenotypic overlap with the inflammation caused by smoking, leading the environment to be similar between smoking and non-smoking asthmatics [Larsen et al.].

of corrected correlation-1.pdf of corrected correlation-1.bb

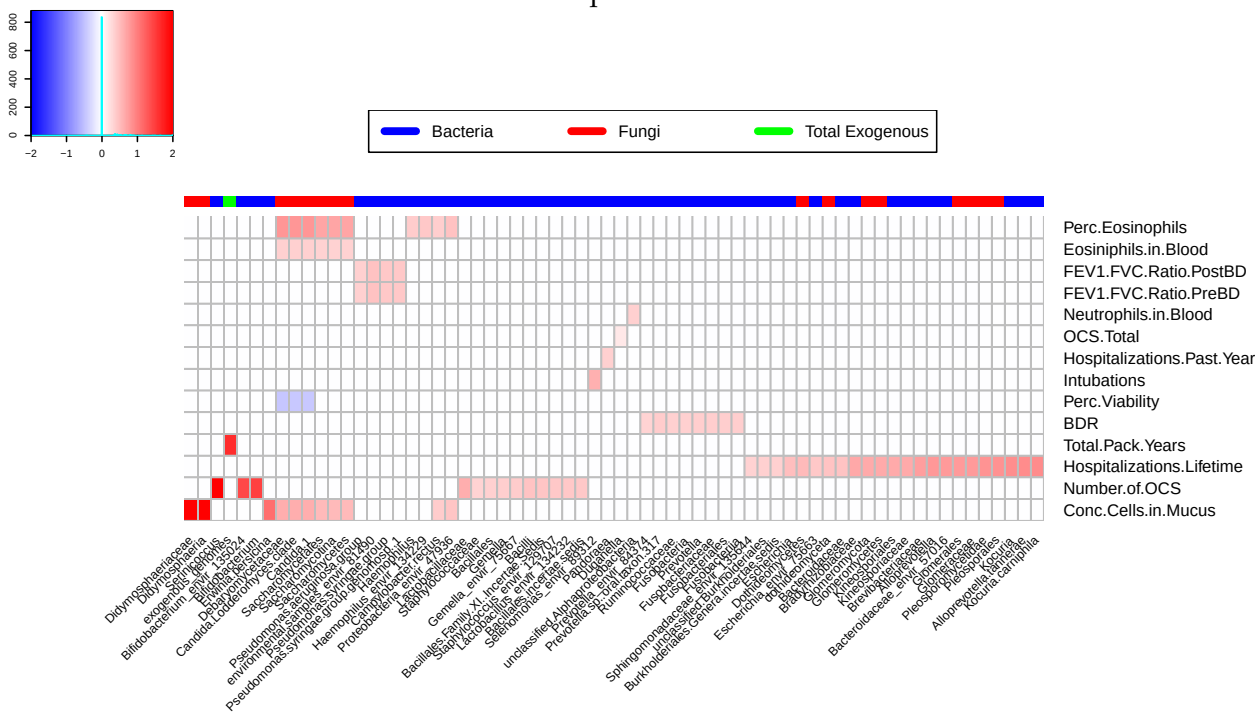

Figure 4: Heatmap of the regression coefficients for each taxon corrected for the effects of age, gender and BMI.

Microbial taxa were associated with healthier metrics as well as with potentially pathogenic roles. In the case of spirometry lung function metrics, the ratio of Forced Expiratory Volume in one second (FEV1) to the Forced Vital Capacity (FVC) was positively associated with members of the genus *Pseudomonas*. This suggests that *Pseudomonas* is associated with a reduction in obstructive defects to the airway, which was true both before and after treatment with bronchodilators. However, *Pseudomonas* was not associated with the response to bronchodilators (BDR); rather three bacterial groups were: the family Ruminococcaceae, the genus Fusobacteria and its family and order, as well as the genus *Prevotella*. This result agrees with previously reported observation that *Prevotella* does not promote Toll-like receptor 2-independent lung inflammation, whereas members of the phylum Proteobacteria did, including *Haeomophilus* and *Moraxella* [Larsen et al.].

There were far greater numbers of taxa correlated with negative health effects. Interestingly, roughly one third of the significant correlations were with fungi, highlighting the importance of

analyzing more than the 16S of bacteria. In particular, the number of hospitalizations that the patient has experienced correlated with both fungi and bacteria in roughly equal proportions. Proteobacteria taxa such as *Escherichia coli* were observed, as well as the fungal orders of Glomeraleas and Pleosporales. Glomerales is an order of arbuscular mycorrhizal fungi not known to be associated with humans. The order Pleosporales contains a known human pathogen but has not been associated with the lungs or asthma.

Fungal and bacterial taxa were also correlated with the concentration of cells in the mucus as well as the percent of eosinophils in both the sputum and the blood. *Haemophilus*, which has been reported to increase inflammation, was positively correlated with the percent of eosinophils in the sputum but not in the blood, nor the overall concentration of cells in the mucus. However, the fungal genus *Candida* was associated with all three. Pulmonary candidiasis has long been associated with allergic bronchial asthma and inflammation [Masur et al.].

Model for the percent eosinophils

To further explore the association of exogenous microbes with the percent of eosinophils found in the sputum we used a machine learning approach. A random forest model was applied to the 150 genera with the most variance in the dataset. In the context of this large number of genera, *Candida* is shown to have the greatest influence in the model (Figure 5). The next most influential genera were *Campylobacter* and another yeast genus, *Saccharomyces*. *Campylobacter* has been associated with chronic diseases of various types including asthma [Doorduyn et al.], while *Saccharomyces* has been shown to be protective against the development of asthma-like symptoms in mice [Fonseca et al.]. Moreover, Fonseca *et al.* observed the protective effect of *Saccharomyces* to be mediated in part through decreased airway eosinophils. In the present study, the contrary is observed, in that each partial dependence plots for each of those taxa has an overall positive slope (Figure 8). However, the particular strain used in the mouse model study, *Saccharomyces cerevisiae* UFMG A-905, could not be unambiguously identified in this study, in that reads aligned to several *S. cerevisiae* genomes equally well.

One of the benefits of analyzing the bulk sputum by RNAseq, in addition to being able to survey both the bacteria and fungi, is the ability to simultaneously view the human transcriptome signal. Future work will analyze the human reads to determine if particular pathways are

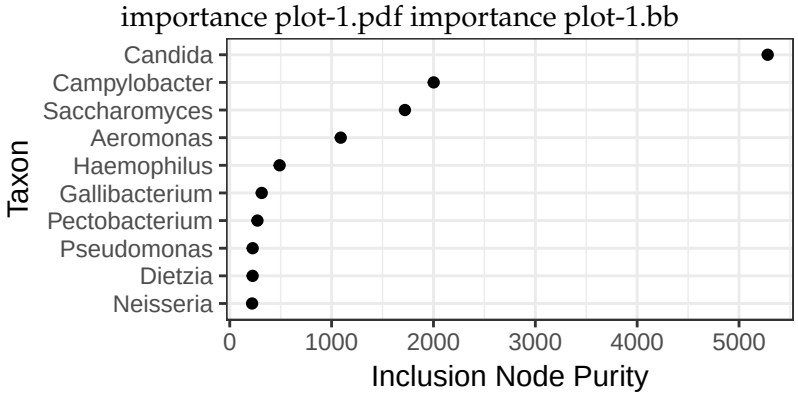

Figure 5: Variable importance plot for the random forest model of the hundred genera with the most variance. Candida and Campylobacter have the strongest effects on node purity.

associated with the microbial taxa observed. For example, are the same patterns relating clinical and exogenous sequences observable in the human transcriptome signal, such as in inflammation response pathways? This has the potential to speak directly to the mechanisms by which the microbial taxa are having an effect, and perhaps shed light on the mechanisms and role of microbes in asthma heterogenity.

Acknowledgements

This work was supported by a National Library of Medicine fellowship to DS (5T15LM007056-28) and an NHLBI grant to GC (1R01HL118346-01). The authors would like to thank the support of the Yale High Performance Computing services (Grace, Ruddle, Farnam).

Supplemental Figures

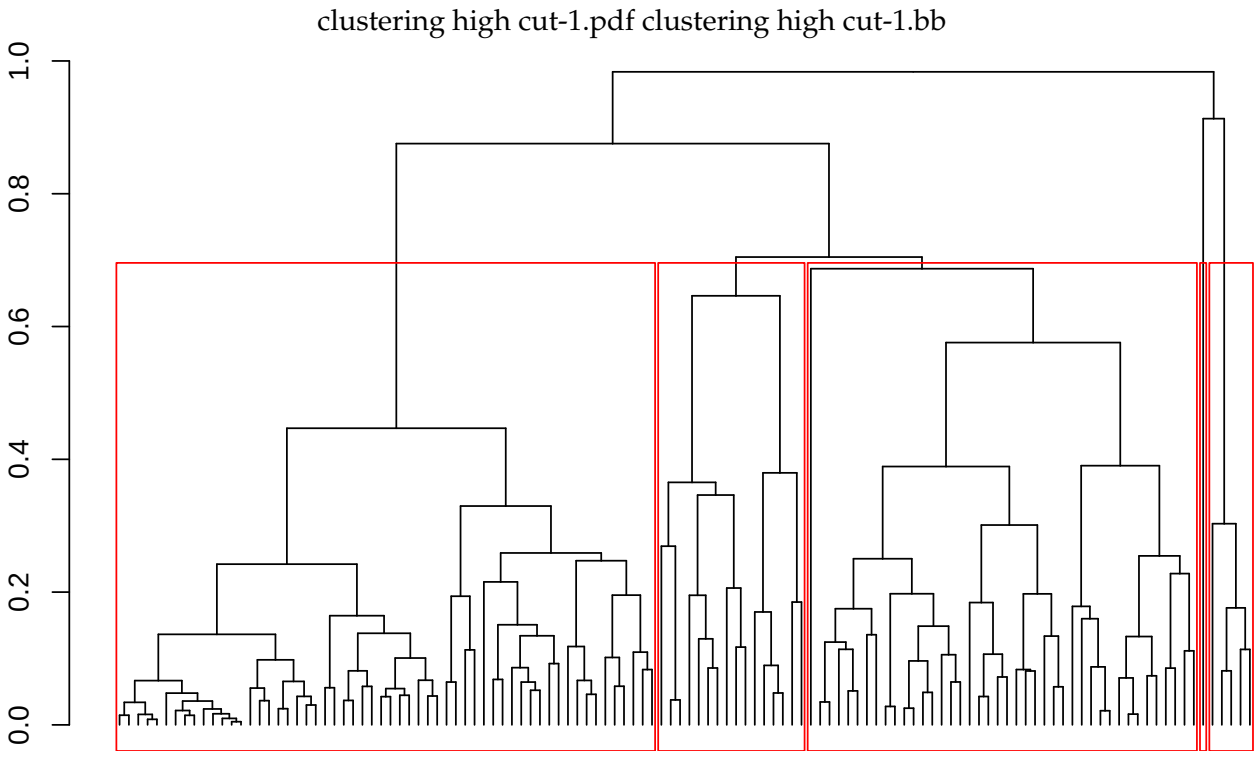

Figure 6: Hierarchical clustering by Bray-Curtis distance, with five clusters boxed (cut height 0.7). Cluster two (center) is significantly depleted of control samples (p-value = 0.0085) and significantly enriched in moderately severe asthmatics (p-value = 0.0058) by permutation test

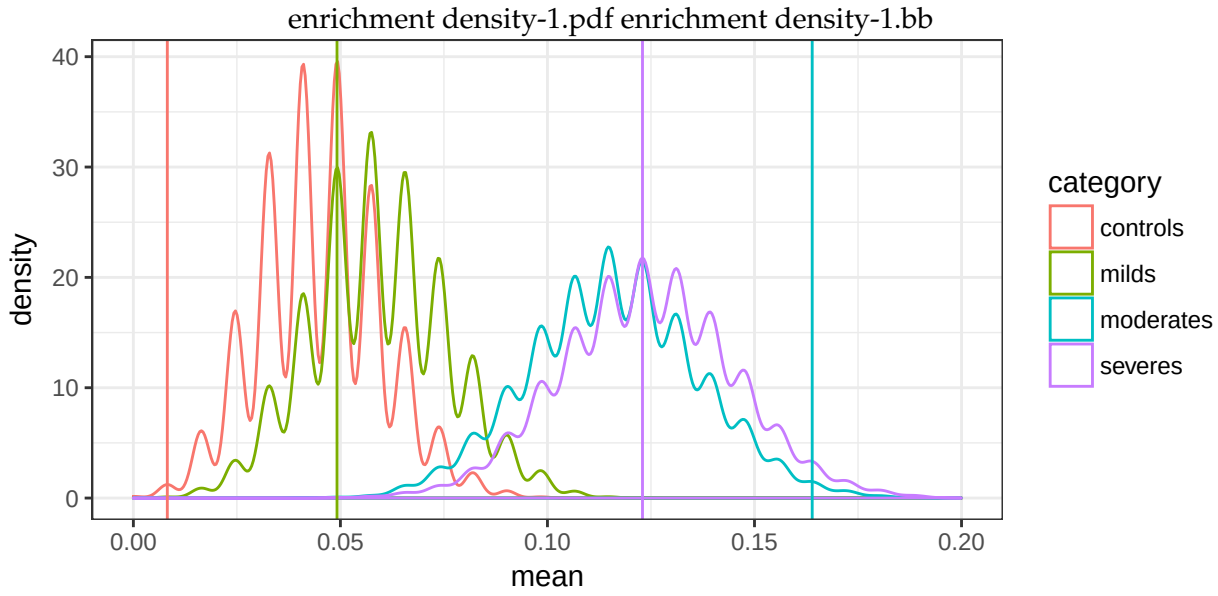

Figure 7: Permutation test results for 10K random shufflings of asthma severity labels, testing for membership in cluster two. Vertical lines show the fraction observed in the hierarchical clustering results, colored to match their permutation test densities. Control samples are significantly depleted and moderately severe asthmatics are significantly enriched.

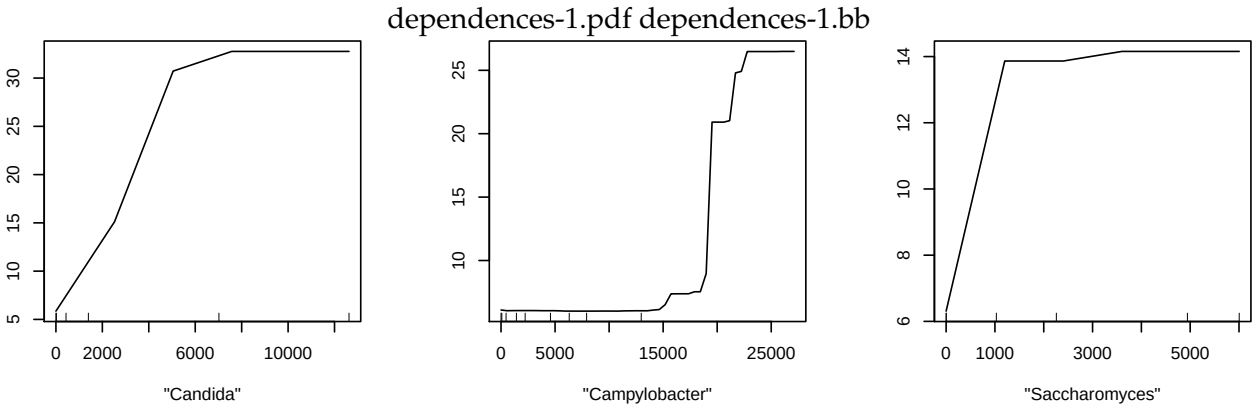

Figure 8: Partial dependences of the three most influential taxa in the random forest model.

References

Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. GenBank. 41:D36–42. ISSN 1362-4962. doi: 10.1093/nar/gks1195.

Holly M. Bik, Julia M. Maritz, Albert Luong, Hakdong Shin, Maria Gloria Dominguez-Bello, and Jane M. Carlton. Microbial community patterns associated with automated teller machine keypads in new york city. 1(6):e00226–16. ISSN 2379-5042. doi: 10.1128/mSphere.00226-16. URL http://msphere.asm.org/content/1/6/e00226-16.

Jean Bousquet, Philippe J. Bousquet, Philippe Godard, and Jean-Pierre Daures. The public health implications of asthma. 83(7):548–554. URL http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S0042-96862005000700016.

Eduardo Castro-Nallar, Ying Shen, Robert J. Freishtat, Marcos Pérez-Losada, Solaiappan Manimaran, Gang Liu, W. Evan Johnson, and Keith A. Crandall. Integrating microbial and host transcriptomics to characterize asthma-associated microbial communities. 8:50. ISSN 1755-8794. doi: 10.1186/s12920-015-0121-1. URL http://dx.doi.org/10.1186/s12920-015-0121-1.

James R. Cole, Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. Ribosomal database project: data and tools for high throughput rRNA analysis. 42:D633–D642. ISSN 0305-1048. doi: 10.1093/nar/gkt1244. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965039/.

Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. page bts635. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bts635. URL http://bioinformatics.oxfordjournals.org/content/early/2012/10/25/bioinformatics.bts635.

Y. Doorduyn, W. Van Pelt, C.ăL.ăE. Siezen, F. Van der Horst, Y.ăT.ăH.ăP. Van Durnhoven, B. Hoebee, and R. Janssen. Novel insight in the association between salmonellosis or campylobacteriosis and chronic illness, and the role of host genetics in susceptibility to these diseases. 136(9): 1225–1234. ISSN 0950-2688. doi: 10.1017/S095026880700996X. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2870923/.

John R. Erb-Downward, Deborah L. Thompson, Meilan K. Han, Christine M. Freeman, Lisa McCloskey, Lindsay A. Schmidt, Vincent B. Young, Galen B. Toews, Jeffrey L. Curtis, Baskaran Sundaram, Fernando J. Martinez, and Gary B. Huffnagle. Analysis of the lung microbiome in the healthy smoker and in COPD. 6(2):e16384. ISSN 1932-6203. doi: 10.1371/journal.pone.0016384. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0016384.

Vanessa M. B. Fonseca, Thamires M. S. Milani, Rafael Prado, Vania L. D. Bonato, Simone G. Ramos, Flaviano S. Martins, Elcio O. Vianna, and Marcos de C. Borges. Oral administration of saccharomyces cerevisiae UFMG a-905 prevents allergic asthma in mice. ISSN 1440-1843. doi: 10.1111/resp.12990.

Markus Hilty, Conor Burke, Helder Pedro, Paul Cardenas, Andy Bush, Cara Bossley, Jane Davies, Aaron Ervine, Len Poulter, Lior Pachter, Miriam F. Moffatt, and William O. C. Cookson. Disordered microbial communities in asthmatic airways. 5(1):e8578. ISSN 1932-6203. doi: 10.1371/journal.pone.0008578. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0008578.

Yvonne J. Huang and Homer A. Boushey. The microbiome in asthma. 135(1):25–30. ISSN 0091-6749. doi: 10.1016/j.jaci.2014.11.011. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287960/.

Yvonne J. Huang, Snehal Nariya, Jeffrey M. Harris, Susan V. Lynch, David F. Choy, Joseph R. Arron, and Homer Boushey. The airway microbiome in patients with severe asthma: Associations with disease features and severity. 136(4):874–884. ISSN 1097-6825. doi: 10.1016/j.jaci.2015.05.044.

Jeppe M Larsen, Hanieh S Musavian, Tariq M Butt, Camilla Ingvorsen, Anna H Thysen, and Susanne Brix. Chronic obstructive pulmonary disease and asthma-associated proteobacteria, but not commensal prevotella spp., promote toll-like receptor 2-independent lung inflammation and pathology. 144(2):333–342. ISSN 0019-2805. doi: 10.1111/imm.12376. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4298427/.

Njira L. Lugogo and Monica Kraft. Epidemiology of asthma. 27(1):1–15. ISSN 0272-5231, 1557-8216. doi: 10.1016/j.ccm.2005.10.006. URL http://www.chestmed.theclinics.com/article/S0272-5231(05)00121-8/abstract.

Matthew Masoli, Denise Fabian, Shaun Holt, Richard Beasley, and Global Initiative for Asthma (GINA) Program. The global burden of asthma: executive summary of the GINA dissemination committee report. 59(5):469–478. ISSN 0105-4538. doi: 10.1111/j.1398-9995.2004.00526.x.

Henry Masur, P. Peter Rosen, and Donald Armstrong. Pulmonary disease caused by candida species. 63(6):914–925. ISSN 0002-9343. doi: 10.1016/0002-9343(77)90546-0. URL http://www.sciencedirect.com/science/article/pii/0002934377905460.

Alison Morris, James M. Beck, Patrick D. Schloss, Thomas B. Campbell, Kristina Crothers, Jeffrey L. Curtis, Sonia C. Flores, Andrew P. Fontenot, Elodie Ghedin, Laurence Huang, Kathleen Jablonski, Eric Kleerup, Susan V. Lynch, Erica Sodergren, Homer Twigg, Vincent B. Young, Christine M. Bassis, Arvind Venkataraman, Thomas M. Schmidt, George M. Weinstock, and Lung HIV Microbiome Project. Comparison of the respiratory microbiome in healthy nonsmokers and smokers. 187(10):1067–1075. ISSN 1535-4970. doi: 10.1164/rccm.201210-1913OC.

June L. Round and Sarkis K. Mazmanian. The gut microbiome shapes intestinal immune responses during health and disease. 9(5):313–323. ISSN 1474-1733. doi: 10.1038/nri2515. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4095778/.

Andrew B. Shreiner, John Y. Kao, and Vincent B. Young. The gut microbiome in health and in disease. 31(1):69–75. ISSN 0267-1379. doi: 10.1097/MOG.0000000000000139. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4290017/.

Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. 449(7164):804. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3709439/.

Edith T. Zemanick, Scott D. Sagel, and Jonathan K. Harris. The airway microbiome in cystic fibrosis and implications for treatment. 23(3):319–324. ISSN 1531-698X. doi: 10.1097/MOP.0b013e32834604f2.