Microbial communities as predictors of outcomes in industrial and clinical applications

by

Allison Rose Perrotta

B.S., Biochemistry
University of New Hampshire, 2007

Submitted to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Environmental Microbiology
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2017

© Massachusetts Institute of Technology 2017. All rights reserved.

Signature redacted
Author ...........
Department of Civil and Environmental Engineering
May 19 2017

Signature redacted
Certified by...
Eric J. Alm
Professor
Thesis Supervisor

Signature redacted
Accepted by ...........
Jesse Kroll
Chair of Committee on Graduate Students
Department of Civil and Environmental Engineering
Microbial communities as predictors of outcomes in industrial and clinical applications

by
Allison Rose Perrotta

Submitted to the Department of Civil and Environmental Engineering
On May 19th 2016, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Environmental Microbiology

Abstract

Environmental and host associated microbial communities provide an extensive reservoir of genetic and functional diversity. This diversity represents a wealth of potential for applications in many fields. To harness this potential for engineering applications, the impact of temporal dynamics need to be better understood. Yet most of the data we have are in the form of static surveys of diversity. In this thesis, I use 16S rRNA sequencing analysis to measure community composition across time series to predict outcomes for three applications: bioreactor function; a non-invasive diagnostic of endometriosis; and commercial chicken rearing. I identify bacteria that exhibit distinct temporal dynamics within each application, and discuss the implications of those dynamics in the context of each application. Despite the diverse communities covered in this work, temporal dynamics emerge as a common theme that can impact these engineering applications which rely on stable and predictable community performance.

Thesis Supervisor: Eric J. Alm
Title: Professor
Acknowledgements

Many people contributed to the work and process involved in my PhD, including:

- My advisor Eric Alm, whom I thank for his guidance, humor, unending support, creativity, friendship, and advocacy

- My thesis committee members Jeff Gore and Martin Polz

- My outside collaborators, without whom this work would not have been possible

- The MIT BioMicro Center staff, particularly Stuart Levine, for giving me opportunities to explore and pursue my professional goals.

- The Parsons community, for all of the inspiring interdisciplinary discussions and helping to make MIT feel like a home

- The entire Alm Lab, for helping to build and maintain a collaborative and inspiring community to work in

- My collaborators and mentors in the Alm Lab, especially Sarah Preheim, Ilana Brito, Mathilde Poyet, Sean Gibbons, Sean Kearney, and Sarah Spencer

- My husband Daniel Blackler, for his ceaseless patience
## Contents

1. Introduction ................................................. 5

2. Inoculum composition determines microbial community and function in an anaerobic sequential batch reactor ............................................. 11

3. The vaginal microbiome is predictive of endometriosis severity in a Brazilian cohort .................................................. 31

4. Fecal microbial composition is predictive of boiler chicken feed conversion performance at the production breeder level ................................... 55

5. Conclusions .................................................. 85

A. Appendix A: Supplementary Information for Chapter 2 .................................................. 90

B. Appendix B: Supplementary Information for Chapter 3 .................................................. 97

C. Appendix C: Supplementary Information for Chapter 4 .................................................. 106
Chapter 1:

Introduction

Microbial communities provide an extensive reservoir of genetic and metabolic diversity that are hugely important to all ecosystems on earth. One of the most common ways that microbial communities are studied is by 16S rRNA surveys. In the early 1990s microbiologists began to take advantage of the sequence variation found within the vertically inherited 16S rRNA gene to phylogenetically classify and estimate the diversity of bacteria from environmental samples (1-3). 16S rRNA sequence analysis proved to be a powerful tool that allowed the observation of uncultured environmental bacteria without needing to mimic convoluted environmental conditions in the laboratory. The high levels of sequence conservation that occurs in particular regions of the gene also allows for the development of new molecular biology based techniques to improve and build upon previous approaches (1,4). 

In the last decade, the development of next-generation sequencing (NGS) technologies capable of generating high-resolution data has revolutionized many fields, and environmental microbiology has not been an exception. The combination of traditional 16S rRNA techniques and NGS has allowed us to observe the diversity and relative abundances of complex and rare bacterial communities with higher throughput and specificity. Additionally, because of the volume of data generated with NGS, these techniques have allowed us to perform rigorous computational analyses and modeling to derive and test ecological and microbiological theories that were previously extremely difficult, if not impossible, to perform using low-throughput or culture based techniques. The knowledge gained by investigating environmental bacterial communities using NGS techniques has moved beyond ecological theory and into specific applications focused on utilizing and
manipulating bacterial communities for a vast array of applications including bioremediation, agriculture, and the treatment of human diseases (5–8).

While much work on the diversity and composition of environmental and host associated bacterial communities has been performed, with illuminating and advantageous results; a great amount of this work has focused on individual samplings. Recent work focusing on sample collections over time series has revealed the necessity in considering the temporal variation and dynamics of these communities in a given environment or system (9–12). These temporal factors appear to be inherent to many bacterial communities, and provide challenges but also opportunities for applications seeking to exploit the functions of these communities. In this thesis I present three applications that aim at utilizing bacterial community composition over time to predict application outcomes. In each case 16S rRNA analysis, NSG techniques, and time series were utilized.

1.1 Addressing the role of distinct environmental inocula in the composition and production profile of a bioreactor

In the second chapter of this thesis, the role that inocula composition plays in product generation of a sequential batch reactor is addressed. Microbial fermentation is often used to recover resources from wastewater. A common assumption made in these processes is that high community diversity within the inoculum source is sufficient for the selection of a functional microbial community. However, variability of the fermentation product profile is a persistent challenge in the field of resource recovery.

In the course of this thesis we investigated the impact of inoculum on bioreactor microbial community structure and function over time. Sequential batch reactors in a controlled laboratory setting were inoculated with three complex microbial inocula. Chemical and microbial composition was monitored by HPLC and 16S rRNA amplicon analysis, respectively, over the course of the experiment. Within the sequential batch reactors, the temporal dynamics of the microbial community and chemical profiles were reproducible, distinct to initial inoculum, and not specific to the diversity of the initial inoculum. Our
results suggest that the composition of the original inoculum predictably contributes to bioreactor community structure and function.

1.2 Using host associated microbial communities to predict host disease and phenotype

In the third and fourth chapters of this thesis, the ability to predict disease and host growth rate from the composition of host associated microbial communities, commonly referred to as the microbiome, is addressed. Data was collected over a time course for both projects to assess the impact that temporal dynamics have upon the power of those predictions.

1.2.1 Predicting disease severity from the vaginal microbiome

Firstly, in chapter three the case of a commonly occurring human reproductive disorder is investigated. Endometriosis, the misplaced ectopic growth of endometrial tissues, is a complex multi-factorial disorder effecting 10-15% of women worldwide. The pathogenesis of this disorder remains uncertain, but there is evidence that genetic, immune, and environmental factors contribute. Invasive diagnosis, performed by laparoscopy, serves as an additional clinical challenge.

In the course of this thesis, we sought to explore the potential of utilizing the human microbiome as a less-invasive diagnostic tool for endometriosis. 16S rRNA analysis was performed on rectal and vaginal samples collected from subjects with or without endometriosis. Previous studies have observed dynamic shifts and variation in both the vaginal and gut microbiomes over short time scales (10,12). To investigate the temporal dynamics in the microbiome of this patient population, and how this might inform differences in disease state and severity, samples were collected during the follicular and menses phases of the menstrual cycle.

Temporal dynamics greatly affected the microbial composition observed in this study, particularly in the vaginal microbiome. We observed flux and stability of the vaginal microbiome that differed between individuals and across time points. These temporal
dynamics generated variability that impacted our ability to classify disease state and severity. By independently analyzing the time points, we were able to build a machine learning based classifier for disease severity during menses. This work helps to emphasize the great importance of understanding the temporal dynamics of a system and the large impact these dynamics can have on translating the microbiome to a clinical application.

1.2.2 Predicting feed conversion performance in broiler chickens from the fecal microbiome

In chapter four of this thesis, the potential of using the fecal microbiome as an indication of chicken growth rate in pedigree breeder lineages is investigated. The gut microbiome has a substantial impact on the nutrient availability for its host. For this reason, much work has been dedicated to identifying bacterial taxa that have the potential to augment the growth rate performance of agricultural animals such as chickens. However, much of the work to identify probiotic targets for poultry research has been conducted in adult chickens. Studies experimentally testing the effect of probiotics on growth rate have shown initial improvements during early stages of growth that do not persist at later stages of development (13).

In the course of this thesis we performed 16S rRNA analysis on the gut microbiome of adult chickens and on a time series of maturing chickens through out the rearing process. We built a machine learning classification model that can predict feed conversion performance from the gut microbiome in adult chickens. We then investigated the abundance dynamics of predictive bacterial taxa from our classification model across the developmental time series. We found that the gut microbiome of maturing chickens is dominated by previously observed OTUs that are not unique to an individual after 20 days of age. Additionally, we found that OTUs predictive of high performance are differentially represented in the guts of young chickens. This work suggests that investigating microbial community dynamics over host development could help future studies to more informatively select early-life probiotic targets.
1.3 Temporal dynamics are informative in utilizing microbial communities for translational applications

The work performed in this thesis demonstrates the impact that temporal dynamics can have on harnessing microbial communities for industrial and clinical applications. In the final chapter of this thesis I discuss the limitations of this work as well as suggest future work to further develop and understand the presented results.

References


Chapter 2

Inoculum composition determines microbial community and function in an anaerobic sequential batch reactor

The contents of this chapter were published as: Perrotta, Allison R., Rajkumari Kumaraswamy, Juan R. Bas das-Oyanedel, Eric J. Alm, Jorge Rodríguez. “Inoculum composition determines microbial community and function in an anaerobic sequential batch reactor.” *PLoS one* (2017): 12(2), e0171369.

Figures and tables are provided within the text. Supplemental information for this chapter is provided in Appendix A.
Abstract

The sustainable recovery of resources from wastewater streams can provide many social and environmental benefits. A common strategy to recover valuable resources from wastewater is to harness the products of fermentation by complex microbial communities. In these fermentation bioreactors high microbial community diversity within the inoculum source is commonly assumed as sufficient for the selection of a functional microbial community. However, variability of the product profile obtained from these bioreactors is a persistent challenge in this field. In an attempt to address this variability, the impact of inoculum on the microbial community structure and function within the bioreactor was evaluated using controlled laboratory experiments. In the course of this work, sequential batch reactors were inoculated with three complex microbial inocula and the chemical and microbial compositions were monitored by HPLC and 16S rRNA amplicon analysis, respectively. Microbial community dynamics and chemical profiles were found to be distinct to initial inoculate and highly reproducible. Additionally we found that the generation of a complex volatile fatty acid profile was not specific to the diversity of the initial microbial inoculum. Our results suggest that the composition of the original inoculum predictably contributes to bioreactor community structure and function.

2.1 Introduction

Research in wastewater treatment, traditionally focused on pollutant removal, has expanded to include sustainable recovery of resources such as biofuels and organic acids from waste streams. Valuable products such as methane, hydrogen, solvents and bio-plastics can be generated from wastewater streams by microbial fermentation (1–6). However, the large-scale deployment of bioreactors for resource recovery purposes is limited due to persistent variability of the chemical profiles produced (7). Most research focusing on product variability from microbial fermentations focuses on how abiotic operational conditions of bioreactors affect production and stability. In parallel, many environmental microbiologists focus on the role of community assembly in defining complex microbial community structure and function (7–13). Yet these two complementary fields show minimal overlap in the literature. There are other examples
in the literature in which the role that the initial microbial inoculum plays in bioreactor production has been investigated. But, these works have focused on other product profiles, specifically methane production, and have infrequent sampling of the microbial community (13, 14).

Previous studies have identified a correlation between high diversity and functional redundancy within complex microbial communities. This work has lead to an emphasis in inoculum diversity over composition (15-17). However, many of the studies demonstrating this correlation focused upon broad ecosystem functions such as respiration and biomass and not upon the production of a desired chemical profile (15-17). Additionally, microbial successional dynamics have been shown to play a large role in initiating bioreactors and can greatly affect community predictability, diversity, and the complexity of the product profile (9,10,13,17).

We sought to test how initial microbial inocula derived from different sources affect microbial community structure and ecosystem function with a focus on producing a complex and even profile of volatile fatty acids (VFAs). Identical sequential batch reactors (SBRs) were inoculated with three naturally occurring sources for microbial inocula: camel manure (Camel), mangrove sediment (Mangrove), and wastewater treatment sludge (Sludge). Experiments were conducted in triplicate and abiotic operational conditions favored the fermentative production of VFAs; specifically, dark anaerobic fermentation where methanogenesis is inhibited via low pH (6). Glucose was the sole carbon source provided throughout the experiment. Reactors underwent sequential batch cycles every 48 hours to ensure a steady supply of glucose as well as limit product inhibition and contamination (19). Chemical production and microbial composition of each reactor were tracked during a 14-day fermentation period.

2.2 Materials and Methods

2.2.1 Inocula and fermentation

Camel manure from Al Ain camel market, mangrove intertidal sediment (less than 30 cm in depth) from Abu Dhabi coast (24° 31’45.5” N 54° 33’21”E), and anaerobic sludge...
from Al Mafra wastewater treatment plant (Abu Dhabi Sewerage Services Company, Al Dhafrah, Abu Dhabi, UAE) served as starting inocula. These three sources were chosen as they represent three distinct environments and thus would provide very distinct yet highly diverse microbial communities. All sample collections were obtained from locations that are open to the public and did not involve any endangered or protected species.

Inocula were stored at 4 °C for less than 5 days prior to the start of experiments (S1 Text). Fermentations were carried out at 37 °C in 150 mL serum bottles with a working volume of 60 mL. Anaerobic serum bottles were loaded with media (see below for composition) and respective inoculum. The initial biomass concentration for all the three inocula was 10 g/l, measured as dry weight matter.

Sterile media consisted of the following components: glucose 5 g/L, phosphate buffer Na₂HPO₄·2H₂O 0.2 g/L and KH₂PO₄ 2.5 g/L diluted in basal anaerobic media (20), NH₄Cl 1 g/L, NaCl 0.1 g/L, MgCl₂·6H₂O 0.1 mg/L, CaCl₂·2H₂O 1.2 g/L, (NH₄)₆Mo₇O₂₄·4H₂O 0.05 g/L, CoCl₂·6H₂O 0.05 g/L, FeCl₂·4H₂O 2 g/L, MnCl₂·4H₂O 0.05 g/L, NiCl₂·6H₂O 0.1 g/L, Na₂SeO₃·5H₂O 0.1 g/L, H₃BO₃ 0.05 g/L, CuCl₂·H₂O 0.04 g/L, ZnCl₂ 0.05 g/L, AlCl₃ 0.05 g/L, EDTA 0.5 g/L. Media pH was adjusted to 5.5 with HCl 1 M to suppress methanogenic activity. The glucose solution was autoclaved separately from the mineral media.

Reactors were inoculated by mixing inocula (Camel, Mangrove, or Sludge), mineral media and glucose solution. After the initial inoculation step, the glucose present in the media was the only carbon source provided. Following inoculation, each bottle was crimped using sterile rubber stoppers, and flushed with pure N₂ for 2 minutes, in order to achieve anaerobic conditions. Flushing was performed using sterile filters, 0.45 μm pore size (Corning®, Corning, NY, USA), and syringes (BD®, San Jose, CA, USA).

All reactors were incubated without agitation in a temperature-controlled incubator, at 35 °C for a total of 14 days. Three replicate reactors were utilized per inoculum source, which amounted to nine reactors being simultaneously incubated. Gas and liquid sampling, followed by SBR cycling (Fig 1), were performed every 48 hours. A total of 72
samples were collected consisting of baseline (cycle 0) and seven subsequent 48-hour time points (cycles 1 - 7). Each SBR cycle involved centrifuging the microcosms for 20 minutes at 4,000 g, removal of supernatant and addition of fresh media (Fig 1). After SBR cycling, the bottles were again crimped and flushed with N₂. All inoculation, liquid sampling, and media addition/removal steps were performed in a UV sterilized laminar flow chamber. Liquid sampling was performed using sterile needles and consisted of collecting 3 ml from the broth for DNA extraction and liquid chemical analysis. Rubber stoppers were cleaned with 70% ethanol prior to liquid and gas sampling.

Fig 1. Sequential Batch Reactor (SBR) cycling was performed. SBR cycles involved centrifuging microcosms, for 20 minutes at 4,000 g, removal of the supernatant, and addition of fresh media. A total of 7 SBR cycles were performed per reactor.

2.2.2 Chemical Analysis

Gas production was measured by liquid volume displacement. The liquid consisted of a saturated solution of NaCl (3 M) at pH 2 (21). Chemical production of VFAs was detected by UV (210 nm) and quantified using an Agilent 1260 Infinity HPLC, Hi-Plex H 300 x 7.7 mm column. The column temperature was 65 °C, with an eluent flow rate of 0.6 mL/min, the eluent consisted of 5 mM H₂SO₄. Glucose and ethanol were detected by refractive index detection at an optical unit temperature of 35 °C.
2.2.3 DNA extraction and amplicon based Illumina sequencing of 16S rRNA genes

DNA was extracted from the initial inocula and the bioreactor microbial communities using the Ultra Clean Soil DNA isolation kit (MOBIO Laboratories, Carlsbad, CA, USA) according to the manufacturer's procedure. Paired-end Illumina sequencing libraries were constructed using a two-step PCR approach targeting 16S rRNA genes previously described by Preheim et al. (22).

All paired-end libraries were multiplexed into one lane and sequenced with paired end 150 bases on each end on the Illumina MiSeq platform at the Biomicro Center (MIT, Cambridge, MA).

2.2.4 Sequencing data processing

Raw data were quality filtered using QIIME (version 1.3.0) (23). Fastq files from the forward and reverse reads were processed using the split_library_fastq.py program of QIIME. After quality filtering, a total of 4,825,857 sequence reads were generated with an average of 68,930 per sample (±61,056, std. dev., n=72).

Sequences were truncated when more than one base in a row dropped below a Phred quality score of 14, corresponding to a probability of error around 3.98%. Only sequences at least 99 bps long after quality filtering was retained. After quality-filtering, 16S rRNA universal primer sequences were removed from all reads, Operational Taxonomic Units (OTUs) were assigned using distribution-based clustering with default parameters, and singletons were removed using methods described in Preheim, et al. (22) (S2 Text).

2.2.5 OTU analysis

After initial processing for quality, a total of 2,736 OTUs were identified and their relative abundances were assessed. 83 OTUs were identified as having a relative
abundance of at least 1 % across all samples; the majority of these being present only in the original inocula (cycle zero). 11 OTUs were identified with a significant abundance, greater than or equal to 1 %, in samples collected during cycles 1 to 7. The taxonomic classification of all OTUs was determined using the Ribosomal Database Project (RDP) with a support cut-off of 0.5 (24) and were confirmed using BLAST (see above).

The analysis package QIIME was used to filter the OTU table by input inoculate and remove samples with less than a thousand read counts from further processing (25) (S3 Text). Two samples, M2.1 (Mangrove, second replicate first cycle) S1.6 (Sludge, first replicate sixth cycle), were removed due to low read count. As a result, a total of 70 samples were considered for downstream analyses.

Alpha (Shannon diversity) and beta (Jensen-Shannon distance) diversities (26) were calculated using pysurvey (S4 Text). The effective number of species, of order 1, represented in each sample was calculated by taking the inverse of the natural log of the calculated Shannon entropy.

Phylogenetic trees were produced for organization and visualization purposes (Fig 1) using unique OTU representative sequences. Prior to building a phylogenetic tree, QIIME was used to align OTU unique sequences to full-length 16S reference sequences in the greengenes database (version gg_13.5) (27). The aligned sequences were trimmed of gaps using trimAL (version 1.2) (28) and a constraint file created from the RDP taxonomic classifications was generated using a custom Python script (version 2.7.10). FastTree (29) was then used to build the phylogenetic tree (S5 Text). Abundance heat maps associated with the phylogenetic trees were generated using Interactive Tree of Life (30).

2.2.6 Statistical analysis
To assess the significance of differences in chemistry and microbial composition as well as that of replicate variance between the inoculum sources, multivariate analyses were performed using the vegan package in R (S6 Text). Distance matrices for these analyses were calculated using Euclidean and Jensen-Shannon distance (26) for chemistry and
microbial data, respectively. Due to the lack of independence in measurements across cycles, these analyses were performed within and not across each cycle. Significance was defined as a p-value less than or equal to 0.05 for both forms of analysis.

2.3 Results and Discussion

2.3.1 Environmental selection contributes to community structure

In the course of this work we sought to compare the impact of initial inoculum upon microbial community structure and ecosystem function within a bioreactor. Our ecosystem function of interest focused on producing a predictable and stable profile of VFAs. It was important to assess if there was selection acting upon the microbial communities due to the abiotic operational conditions of the bioreactors over the course of the experiment. We hypothesized that the shared and consistent abiotic conditions of the bioreactors, for example pH and retention time, should serve as shared selective pressures and lead to similarities across all communities. Thus we expected the common abiotic pressures in our system to result in selection of similar OTUs, decreased divergence over time, and reproducibility across replicates.

Abiotic selection pressures had consistent effects upon the microbial communities and confirm the reproducibility of our experiments. An enrichment of common OTUs was observed in all bioreactors (Fig 2 and S1 Fig). An overall decrease in pairwise divergence between communities was also observed. As seen in Fig 3A, pairwise comparisons of all microcosms consistently decreased in divergence over time, as measured by Jensen-Shannon distance (26). The phylogenetic structure of replicates was also reproducible as evidenced by low divergence between replicates over time (Fig 3B). These conclusions are supported by recent work by Vanwonterghem et al. (13) in which microbial community structure within a methanogenic anaerobic digester was shown to be predictable and reproducible over an extended period of time.
Fig 2. Microbial communities are reproducible and distinct to the initial microbial inoculum. Relative abundances of OTUs are represented as log relative abundances. Only OTUs that were present at equal or greater than 1% across all time points have been included. OTUs are organized vertically by phylogenetic relationship and horizontally by time point within a replicate, cycle 0 to cycle 7.
Unsurprisingly, the observed enrichment represented by the most abundant OTUs, consisted of species known to produce hydrogen and VFA products in both mono- and mixed-culture bioreactors. More specifically, all bioreactors were enriched for *Clostridium pasteurianum, Clostridium acetobutyricum, Escherichia coli*, as well as other OTUs classified within the *Clostridium* and *Enterobacter* genera (Fig 4A) (1,19,31–33). All enriched community members, with the exception of the *C. pasteurianum*, were present at very low abundances in at least two out of three replicates at cycle zero. *C. pasteurianum* was only detectible in one Camel replicate at time point zero. The abundance of these OTUs at cycle zero indicates that these community members were present at low relative abundances in the inoculum and were enriched for in our system.
Fig 4. Dynamics of enriched OTUs and reactor chemical profiles are distinct to initial microbial inoculum. We illustrate changes in average relative abundances of enriched OTUs over time (A) and the changes in average product yields in terms of electron equivalents over glucose consumed from provided feed (B) for the three inocula sources Camel, Mangrove, and Sludge reactors, respectively.

These results suggest that the total diversity available for selection is extremely large, and underestimated even by deep 16S rRNA gene amplicon sequencing (S1 Fig). However, even in the light of the similar microbial enrichment observed in all bioreactors there were differences in the evenness and diversity of the microbial communities that were distinct to the microbial inoculum (Fig 4). This implies that even if diversity is large and shaped by abiotic selection, it may still fall short of the Bass-Becking "everything is everywhere" concept (34,35) in which the effects of initial inoculum on final community structure is negligible. Thus we tested the relative importance of inoculum on community structure.

2.3.2 Microbial inoculum governs community structure and dynamics

While broad commonalities between bioreactor communities were observed they were overwhelmed by the effect of inoculum. At the community level, each inoculum generated a phylogenetic structure that was strikingly distinct and highly reproducible...
The dynamics of individual abundant OTUs over time were highly dependent on the starting inoculum supplied (Fig 4A) and highly reproducible across replicates with in each of the inoculum sources (S3 Fig). Within each cycle, the microbial community in the bioreactors was significantly different between each inoculum group and the variance between replicates was insignificantly different, with the exception of cycle 6, as determined by multivariate analysis (S1 Table).

In Mangrove bioreactors, successional dynamics were observed and the final community contained OTUs belonging to *E. coli*, *C. pasteurianum*, *Enterobacteriaceae*, *Clostridium*, *Bacteroides*, and *Viellonellaceae*. Within this final community no single OTU made up more than 27% average relative abundance. Camel bioreactors underwent very little community succession, and the final community was dominated by two OTUs, *E. coli* and *C. acetobutylicum*, that were present at 47.2% and 50.8% average relative abundance respectively. In the Sludge bioreactors microbial succession was observed, but a single OTU belonging to the *Enterobacteriaceae* genus dominated the final cycles at 79.5% average relative abundance.

Many large environmental surveys have demonstrated that both biotic and abiotic processes shape community assembly in deterministic and stochastic ways and that their relative importance can vary greatly based on environmental pressures (36–38). Our results indicate that inoculum source deterministically contributes to shaping community structure within an anaerobic SBR system.

### 2.3.3 Microbial inoculum governs chemical composition

In addition to having community profiles that are reproducibly distinct, each inoculum was reproducibly predictive of bioreactor production over the two-week period of the experiment. Much like the observed community dynamics, chemical profiles were distinct to starting inoculum (Fig 4B). Within each cycle, the chemical profile of the bioreactors was significantly different between each inoculum group and the variance between replicates was insignificantly different, as determined by multivariate analysis (S2 Table). The results of these experiments suggest that the functional capabilities of these communities arise in a deterministic fashion that is dependent upon starting
Mangrove reactors maintained a high level of glucose conversion and a complex product profile throughout the experiment. These reactors exhibited a short adaptation period, evident due to a high production of ethanol (35% of all the products synthesized) during the initial cycles, followed by production of the desired products acetic, butyric, and propionic acids. The chemical profile of Mangrove reactors shifted to butyric acid production, up to 50%, in later cycles. Acetic and propionic acid as well as minor lactic acid production are observed throughout all cycles (Fig 4B).

In contrast to the Mangrove reactors, Camel and Sludge reactors demonstrated low glucose conversion over the course of the experiment. Glucose conversion in Camel microcosms dropped by almost half after the first cycle. By the final time point, ethanol made up more than 40% of all chemical products, and up to 79% of the glucose remained unconverted. This chemical profile indicates a loss of the desired function (Fig 4B). Sludge microcosms also failed to produce the desired fermentation profile in this experiment. In Sludge reactors, 30% of the provided glucose remained unconverted in the first cycle and up to 80% remained unconverted in later stages. The main products synthesized were ethanol and acetic acid (Fig 4B).

While the reasons for the difference in production between inoculum remain unclear, differences in the production potential of the community that are distinct from initial inoculum diversity could play a role. The initial Mangrove inoculum did not have the highest diversity of the three inoculum used (Fig 5). However, the community selected for in Mangrove reactors maintained a higher diversity and evenness throughout the experiment (Fig 5) and produced a complex chemical profile (Fig 4B). Sludge and Camel, demonstrating the highest and lowest initial diversity respectively, both consistently decreased in diversity throughout the experiment and did not produce a stable chemical profile. High diversity within a community is highly correlated with functional redundancy, when multiple community members are capable of the same metabolic process (39). Diversity is also correlated with functional complementation, when community members differ in metabolic capability (16). Both functional redundancy and complementation promote stability in a complex community (16,17,40). Functional
complementation is especially important if, as in our case, a complex range of products is desired (41). Our results indicate that initial inoculum diversity is not sufficient to predict the diversity and production potential of the microbial community that arises from that inoculum within an anaerobic fermentation bioreactor. This contrasts the idea that high inoculum diversity alone is sufficient for ecosystem function (15) as well as the classical idea of given high diversity, “everything is everywhere and the environment selects” (34,35). While inoculum diversity may outweigh composition for broad ecosystem functions such as respiration and biomass (15) our results suggest that under the conditions tested, inoculum composition plays a significant role in specific ecosystem functions.

![Graph showing diversity over time for different bioreactors.](image)

**Fig 5. Initial inoculum alpha diversity is not correlated with reactor diversity.** The microbial inocula used vary in diversity. The diversity of the microbial communities present in the reactors decreases over the course of the experiment. Diversity is presented as the effective number of species over time for Camel (dark blue), Mangrove (purple), and Sludge (red) bioreactors.

The decreased production observed in the Camel and Sludge reactors indicate that key species required for multispecies metabolic pathways may be missing. Bacterial species belonging to the class Clostridia have been shown to be important for the production of VFAs via carboxylate (1) and CoA pathways (32). Compared to the other inoculum types, Mangrove reactors had higher overall abundances and diversity (Fig 6 and S1 Table) of OTUs classified as Clostridia. It is important to note that the initial Mangrove inoculum, observed as the community profile at cycle zero, did not contain the highest abundance...
of OTUs classified as Clostridia (S1 Table). This indicates that the Mangrove inoculum may contain the highest number of Clostridia that are resistant to the experimental selection pressures and hints at the large impact strain level differences could have in production capability.

Fig 6. Dynamics of abundant Clostridia species are distinct to starting inoculum. We illustrate changes in average relative abundances of Clostridia OTUs over time in Camel (A), Mangrove (B), and Sludge (C) bioreactors. Abundance is represented as average percent relative abundance. OTUs were filtered for equal or greater than 0.1% relative abundance across all time points, and are labeled with the lowest taxonomic classification identified, genus level unless otherwise noted (f, family; o, order; c, class).

Our results indicate that inoculum source deterministically contributes in shaping community structure and specific ecosystem function. Inoculum-specific outcomes suggest the influence of additional factors such as species-species interactions or variable metabolic gene content (42,43). While the role of strain level differences cannot be tested with 16S rRNA gene data alone, previous studies have observed differences associated with metabolism and complex ecological interactions when comparing bacterial strains from distinct environments (42,43). It is also important to note that it is conceivable the Camel and Sludge bioreactor communities are experiencing a slower adaptive period compared to the Mangrove community, or that they are experiencing senescence. Further studies will need be conducted for extended periods of time and incorporate whole-shotgun metagenomics or transcriptomics analysis in order to confirm the growth stage of these communities and the role
individual bacterial groups play in the observed differences in reactor function.

2.4 Acknowledgements

We thank the MIT BioMicro Center for assistance with sequencing. A.R.P., J.R.B-O. and R.K. were funded under the Masdar Institute/MIT joint research program grant 11WAMA1. The National Institute of Environmental Health Sciences of the NIH under award P30-ES002109 funded sequencing efforts through the MIT BioMicro Center.

2.5 Data Availability

The raw sequencing data and associated metadata from this study was submitted to the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under the Bioproject accession number PRJNA327400 with corresponding Biosample accession numbers SAMN05371773-SAMN05371844.

References


34. Baas-Becking LGM. Geobiologie; of inleiding tot de milieukunde. WP Van Stockum


Chapter 3
The vaginal microbiome is predictive of endometriosis severity in a Brazilian cohort

The contents of this chapter contribute to a publication in preparation by the following authors: Allison R. Perrotta, Giuliano Borrelli, Sabri Sanadani, Linda G. Griffith, Eric J. Alm, Mauricio S. Abrao

Figures and tables are provided within the text. Supplemental information for this chapter is provided in Appendix B.
Abstract

Endometriosis is a complex multi-factorial disorder that is one of the most commonly occurring gynecological disorders in the world. The pathogenesis of this disorder remains uncertain; however, there is evidence that genetic, immune, and environmental factors contribute to the development of the disease. Diagnosis serves as an additional clinical challenge as laparoscopy is currently the most effective diagnostic. In this study, we sought to explore the potential of utilizing the human microbiome as a less-invasive diagnostic tool for endometriosis. We performed 16S rRNA analysis on rectal and vaginal samples collected during the follicular and menses phases of the menstrual cycle from patients with and without endometriosis. We built a machine learning classification model capable of predicting severity of disease from the vaginal microbiota during menses. The distribution of vaginal community state types (CSTs) observed in this study reflects those previously reported in women of European descent. As far as we know, this is the first study to describe the vaginal microbiomes and distribution of CSTs in women with endometriosis and in a South American cohort. This work helps to inform future studies of the challenges and potentials of using the microbiome, particularly the vaginal microbiome, to develop alternative diagnostics for endometriosis.

3.1 Introduction

Endometriosis is a multi-factorial and heterogeneous gynecological disorder defined as the presence and adherence of endometrial cells and tissues outside of the uterine cavity, primarily to tissues in the peritoneal cavity such as the pelvic peritoneum, ovaries, and recto-vaginal septum [1,2]. This unregulated misplacement of endodermal tissues results in reoccurring inflammation and lesions at the sites of adherence. Endometriosis is one of the most commonly occurring gynecological disorders and affects 10-15% of reproductive age women and 30% of infertile women worldwide [2–6]. Common symptoms of this disorder can include severe abdominal pain, bleeding, infertility, pain during vaginal intercourse, constipation, and in some cases even bowel stenosis and obstruction [2].
Endometriosis is also associated with high morbidity and an increased rate of ovarian cancer [6,7]. There is no known cure for endometriosis and current treatments focus on relief of symptoms, in most cases a combination of anti-inflammatory, hormonal regulation, and surgical therapies are utilized [8].

The pathogenesis of endometriosis is unknown. There are three main theories on the origin of this disorder, the most commonly supported of these is the implementation theory, in which flow-back of endometrial tissue through the fallopian tubes results in the presence of endometrial cells in the peritoneal cavity [9–12]. The epidemiology of this disorder is complex and many genetic, environmental, immunological, and hormonal factors are believed to be involved in endometriosis [13]. Endometriosis has been shown to aggregate in families and has been connected with polymorphisms in multiple genes [14–21]. This disorder is also associated with increased and chronic inflammatory activity, and multiple studies have observed altered inflammatory signaling and function in women suffering from endometriosis [22–29].

In addition to the severity of symptoms and unknown pathogenesis, there is an additional challenge in diagnosing this disorder. It is important that diagnosis occurs as early in the life of the patient as possible due to the morbidity, risk of ovarian cancer, and development of ectopic lesions associated with endometriosis. However, surgery involving laparoscopy is the only sure diagnostic currently used. Large lesions can be imaged using ultrasound and magnetic resonance imaging, but these methods will not assist in identifying small lesions or adhesions that occur during early development of the disease [8,30].

Given the connections between endometriosis and gastrointestinal and immunological factors, we explored using the human microbiome as a less-invasive diagnostic tool. We performed a 16S rRNA analysis on rectal and vaginal samples collected during the menses and follicular phases of the menstrual cycle from patients with and without endometriosis. We built a machine-learning based model that can accurately predict the stage of disease (early verses late) from the vaginal microbiota composition during menses. In this study, the distribution of vaginal community state types (CSTs) observed in subjects reflects the distributions observed previously in women of European descent [31]. As far as we know,
this is the first study to describe the human microbiome and the distribution of vaginal CSTs in women with endometriosis as well as the first study to investigate the vaginal microbiome in a South American cohort. This work helps to inform future studies of the challenges and potentials of using the microbiome, particularly the vaginal microbiome, as a potential diagnostic tool for endometriosis.

3.2 Materials and Methods

3.2.1 Subjects and sample collection

Rectal and vaginal samples were collected during the menses and follicular phases of 64 patient’s menstrual cycles. A total of 40 patients with diagnosed endometriosis and 24 patients without endometriosis were sampled. Five patients were removed from the study due to uncertain disease state (patients 18, 56, 69, 41, and 54). The Internal Review Boards of the University of Sao Paulo and the Massachusetts Institute of Technology approved this protocol. Written informed consent was obtained from all participants.

All samples were collected using Sterile Catch-All sample collection swabs (Fisher Scientific, Pittsburgh, PA, USA). Rectal samples were collected by swabbing the rectal tissue at a depth of 3cm. Rectal swabs were utilized instead of fecal collections in this study so we could better sync the timing of collection between body sites to the timing of the menstrual cycle. The prevalence of constipation and pain during defecation that is common in subjects with endometriosis also made rectal swab collection appealing for work with this patient cohort. Rectal swabs have been previously shown to be an effective way of alternatively sampling the intestinal microbiota [32].

Vaginal samples were collected in the mid-vagina at the site of the vaginal introits. No speculums or lubricants were used prior to collection. Immediately after collection all swabs were immersed in the lysis buffer of Mobio PowerBead tubes (MoBio Laboratories Inc., Carlsbad, CA, USA) and held against the sides of the tube three times for 20 seconds.
each to ensure transfer of biological material. All tubes were then transferred on wet ice to a clinical lab in less than two hours and then stored at -80°C until extraction. Vaginal exams were performed after all vaginal collections to ensure patients did not have any active vaginal infections.

### 3.2.2 DNA extraction and 16S rRNA amplicon library preparation for Illumina sequencing

All samples were extracted using the MoBio Powersoil extraction kit (MoBio Laboratories Inc., Carlsbad, CA, USA). Additional proteinase K (Qiagen, Valencia, CA, USA) and 65°C heat treatments were added to the extraction protocol for improved lysis of gram positive bacteria. All other steps were in accordance with the manufacturer’s instructions. Paired-end Illumina sequencing libraries were constructed using a two-step 16S rRNA gene PCR amplicon approach described previously by Preheim et al. [33]. All paired-end libraries were multiplexed into one lane and sequenced on the Illumina MiSeq platform at the Biomicro Center (MIT, Cambridge, MA). Libraries were sequenced with 2x250 bases.

### 3.2.3 Sequencing data processing and OTU analysis

From the sequencing data, forward and reverse reads were merged and filtered by estimated error per read using USEARCH (version 8) (S1 Text). After error filtering, further processing and clustering of Operational Taxonomic Units (OTUs) was performed using an in house pipeline (S1 Text). OTUs were taxonomically classified using the Ribosomal Database Project with a confidence cut off of 0.5 [34].

Samples with less than 500 total reads and OTUs present in less than a total of three subjects were filtered out using custom python scripts and were not considered in downstream analyses (Python version 3.5.2). A total of 1400 OTUs and 161 samples representing 50 patients were retained and included in downstream analyses. All analyses
were performed on an OTU table of normalized fractions generated by dividing the count of each OTU by the total number of reads present for a sample. Log transformation of the normalized OTU table was performed by adding a scalar to all values in the table. The scalar used, -13.07, was calculated as one half multiplied by one divided by the maximum total sample read count in the un-normalized OTU table. Alpha (Shannon diversity) and beta (Jensen-Shannon divergence) diversities were calculated using Qiime and pysurvey, respectively (S1 Text) [35,36].

3.2.4 Vaginal community state type assignment

To determine if the five CSTs and eight sub-CSTs observed in other vaginal microbiome studies were present in this study, CST assignment was determined using data derived from previous studies. CST assignment was not determined using hierarchical clustering of samples within our study alone. Species level classification is required for discriminating between many of the Lactobacillus dominated CST groups. In order to perform the CST assignment, an environment specific species level classification table was generated using MCclassifier (S3 Text). After the species table was generated, a support vector machine-learning model, previously trained on data derived from multiple vaginal microbiome studies, was utilized to predict the CST of each sample (see acknowledgements). This was performed for only the sequencing reads generated from the vaginal samples and was only utilized for assigning a CST to each sample.

3.2.5 Identification of predictive features in the microbiome and statistical analysis

Correlations of 16S rRNA amplicon data with clinical metadata (disease state, disease severity) were conducted using a random forest based machine-learning analysis [37]. All OTU abundance data were analyzed with a log transformation and all metadata categories were analyzed without transformation. The parameters utilized can be found in
supplemental methods (S2 Text). The accuracy of the classification models was determined using area under the curve (AUC) analysis. Contributions of specific OTUs to the classification were ranked using mean decrease accuracy.

Significance of OTU abundance between discrete categories was calculated using the Mann-Whitney U-test (scipy, Python). Significance tests were performed for OTUs with a non-log transformed relative abundance of at least $10^{-6}$. P-values generated from significance testing were used to calculate q-values using a Benjamini-Hochberg false discovery rate correction (statsmodels, Python). Negative and positive correlations between OTU abundance and disease state were calculated using a two-tailed T-test for independent samples (scipy, Python). To test the significance of categorical count data, specifically when testing the significance of CST distributions between groups of patients, a two-sided Fisher exact test for count data was used in R (version 3.3.1). The testing method used is specified through out the text and figure text accordingly.

3.2.6 Phylogenetic analysis

Phylogenetic trees were produced to investigate the phylogenetic relationship between the top predictive OTU identified in our classification model and other closely related representatives of the same genus. Representative sequences were downloaded from the SILVA database [38] and filtered for uncultured and unassigned bacteria. All sequences from the SILVA database were trimmed for the 16S rRNA v4 region by mapping the predictive OTU to the reference sequences and then subsequently trimming the references at positions corresponding to 450 and 770 nt using custom python scripts (Python version 3.5.2). A reference sequence from a *Eubacterium infirmum*, strain F0142, was used as an out-group. Prior to building a phylogenetic tree, muscle was used in QIIME to align the sequences [39]. The aligned sequences were trimmed of gaps using trimAL (version 1.2) [40] and the phylogenetic tree was built using FastTree [41][S5 Text]. The tree was then visualized and pruned using the Interactive Tree of Life tool [42].
3.3 Results and Discussion

3.3.1 No clear disease signal was revealed by comparative analysis

Neither rectal nor vaginal samples grouped by patient disease status (Fig 1). Grouping by disease status was conducted by hierarchical linkage clustering based upon the community composition and OTU abundance of samples within a sample type and time point. This lack of clustering likely reflects the high inter-subject variability of the intestinal and vaginal microbiomes [31,43,44]. It is important to note that while there was no sample grouping by disease status, the vaginal samples did group by CST (S1 Fig).
Figure 1. Rectal and vaginal microbial communities do not cluster by disease state. We illustrate the hierarchical linkage clustering of samples based upon the community composition and OTU abundance for patient rectal and vaginal samples collected during the follicular and menses time points. Patient disease status is indicated as endometriosis or control group, blue and black respectively. OTU abundances are displayed as log relative abundance.
Alpha diversity of rectal and vaginal samples, measured as Shannon Diversity, also did not differ significantly between endometriosis and control subjects (S2 Fig). The pairwise beta diversity of rectal samples in endometriosis subjects was significantly lower than that of control subjects during follicular and menses phases (S3 Fig). These results indicate that there is less variability in the gut microbiome of endometriosis subjects than control subjects even though they do not overtly cluster by disease state. The microbiome of endometriosis subjects was not more variable over time compared to control subjects as indicated by calculating the pairwise beta diversity within a subject between the two time points (S3 Fig).

Overall, the results of these comparative analyses indicate that a bacterial community biomarker for endometriosis could be elusive and subtle in the face of the inherent person-to-person variability of these communities. The confounding factors contributing to this person to person variability in the microbiome is currently unknown, but host genetics, immune control, diet, sexual activity and hygiene have all been indicated [43–48].

3.3.2 Distribution of vaginal Community State Types significantly differs between follicular and menses stages of the menstrual cycle.

The CST distributions of endometriosis and control subjects during the follicular phase were consistent with those observed in non-pregnant reproductive age women of European descent, and did not differ significantly between subject groups for either time point [31,45,46,49] (Fig 2). It is important to note that our subject cohort included white, mixed race, and black subjects, 61%, 22%, and 16.9% of the cohort, respectively (STable 1). Endometriosis is more common in Caucasian and Asian populations and CST distribution has been suggested to vary by race [2,27,42]. Because most of the subjects in this study are white, it is possible that the CST distributions we observed are driven by this subject bias.
However, this is the first time, that we know of, that the distribution of CSTs has been investigated in a South American cohort.

In this study, two *Lactobacillus* dominated CSTs, II and V, dominated by high abundances of *Lactobacillus gasseri* and *L. jensenii*, respectively [31], were lost during the menses phase in both patient groups (Fig 2). This difference between time points was significant (p=0.021, Fisher’s exact test). The loss of CST II and V during menses could be due to a lack of power and too few women being sampled to observe these CSTs in both time points. However, an alternative hypothesis could be the drop in estrogen levels that occurs during the menses phase in reproductive age women [44]. During cycle phases associated with higher estrogen levels, increased vaginal secretions and thickness of the epithelium lead to an accumulation of glycogen that potentially encourages the growth of *Lactobacillus* species [50,51]. In support of this theory is the fact that during pregnancy, when estrogen levels increase dramatically, women have vaginal microbiomes that are more stable over time, tend to have higher abundances of *Lactobacillus* species, and CSTs that are not dominated by *Lactobacilli* are rare [49,52]. Pregnant women are also more likely to have vaginal communities dominated by *L. crispatus* and *L. iners*, corresponding to CST I and III, respectively. Additionally, dynamics of estrogen levels and glycogen content that span a woman’s lifetime have been reported to correlate with abundances of *Lactobacilli* in the vaginal bacterial community [51,53,54]. Given these previous observations and the results of this study, it is possible that *L. gasseri* and *L. jensenii*, corresponding to CSTs II and V, respectively, are not as resilient to lower estrogen levels as other species of *Lactobacilli*. 
Figure 2. The distribution of vaginal community state types differs between follicular and menses phases of the menstrual cycle. Distribution of vaginal community state types observed in endometriosis and control subjects during follicular and menses phases. The number of subjects included is 27, 18, 21, and 16; top left, right, bottom left, and right, respectively.

3.3.3 CST flux and stability between follicular and menses time points is observed in both endometriosis and control cohorts

The CST assignment of some subjects changes between the follicular and menses time points, while other subjects remain in the same CST (Fig 3). In this study, there were no significant differences in transitions between endometriosis and control groups as calculated by Fisher’s exact test. However, the observation that there is more flux into the subCST IV-A during menses, particularly in endometriosis subjects, is very interesting as this subCST, which is not dominated by Lactobacillus nor Gardnerella species, has recently been correlated with higher states of inflammation and increased risk of HIV infection [55]. Menses is the highest inflammation state in the menstrual cycle. Menses is of additional interest when studying endometriosis, as patients experience the most pain during menses,
and also because non-menstruating women have a significantly lower frequency of endometriosis [2].

Figure 3. Flux and stability of vaginal sub-community state types of endometriosis and control subjects across follicular and menses phases. We illustrate change or lack of change in assigned sub-community state type across the follicular and menses phases for (A) endometriosis and (B) control subjects. An arrow base represents patient state during the follicular time point and the head represents the state during menses time point. Loops around a state represent patients that remain in the same state for both time points. Line weight signifies the number of patients that each arrow represents, with the thinnest representing a single patient and the thickest representing three or two patients in A and B, respectively.

The transitional dynamics observed in this study are similar to those seen in other studies. Gajer and colleagues showed that over a period of 15 weeks the dynamics of the vaginal microbiome vary by person with some women regularly transitioning between CSTs, while others remain stable over time. These transitions often occur during menses and do not occur with equal likelihood [44]. In agreement with the dynamics observed by Gajer et. al. we see no transitions between CSTs I and II in our study. As mentioned above, no subjects in this study transitioned into CST II or V. Gajer and colleagues did not observe this result;
however, transitions to CST II were rare in their study. Differences in transition patterns between our study and previous vaginal time series studies could be due to many factors including power limitations or the timing of sampling as transitional states have been shown to rarely persist for long periods of time [44].

### 3.3.4 Vaginal community composition during menses accurately predicts disease severity

Given the differences in beta diversity between endometriosis and control groups in the rectal microbiome (S3 Fig), and the CST distribution differences between follicular and menses time points (Fig 2); we asked whether there were any microbial indicators associated with disease severity. One way to approach this question is to computationally classify clinical meta data from the microbial composition [33]. We built machine-learning models based on the rectal or vaginal microbiota composition during the follicular and menses time points. Using this model, we were able to accurately predict early (N = 12, stages 1 and 2) verses late stage endometriosis (N = 8, stages 3 and 4) from the vaginal bacterial community of endometriosis patients during menses (Fig 4). The predictive power of this model was significant with a p-value of 0.019, calculated using model derived confusion matrices and Fisher’s exact test (S4 Fig). We were unable to accurately predict disease state from any of the sample types or time points (S4 Fig). We were also unable to accurately predict either stage of disease from control subjects (S4 Fig). While this latter point is worrisome, it is possible that this inability to classify disease verses control could be due to high variability in the control samples (S3 Fig). Our control cohort did not consist of healthy subjects but instead subjects suffering from other non-endometriosis related gynecological symptoms, which could be a contributing factor to our difficulty in accurately classifying this group.
**Figure 4. Microbiota composition predicts stage of disease.** Here we illustrate (A) accuracy, as determined by AUC analysis of ROC curves, of the model built to predict early versus late stage endometriosis from the composition of the vaginal microbial community during the menses time point. (B) The top contributing OTU features for this model, ranked by Mean Decreased Accuracy. Taxonomic level for each OTU displayed is genus except in the case of a single OTU where f indicates the family level. Correlations of OTU abundance with stage of disease, calculated using a two-sided T-Test, are indicated as positive (blue, T statistic $> 0$ and p-value $< 0.05$), negative (red, T statistic $< 0$ and p-value $< 0.05$), and no correlation (grey, p-value $> 0.05$).

A Firmicutes in the genus *Anaerococcus* was the feature contributing the most to our classification of early versus late disease stage (Fig 4), and the relative abundance of this OTU significantly differed by stage using the Mann-Whitney U-Test (Fig 5). However, there was no significant difference in abundance for this OTU between the early or late disease and control subject groups (q=0.28 and 0.35 respectively). Late disease and control or early subject groups had large effect sizes, calculated as Cohen's d (d=-1.03 and -1.9, respectively), while the effect size of early disease and control subject groups was small (d=0.33) [52]. The large effect sizes of the late disease group indicate that while the abundance of the *Anaerococcus* OTU may not significantly differ from control it does suggest that there are differences. *Anaerococci* are strict anaerobes and are commensals...
frequently found in the vaginal microbiome, particularly in women with non-*Lactobacillus* dominated CSTs. This predictive *Anaerococcus* OTU was distantly related to cultured representatives of this genera in the SILVA database [38], but grouped into a clade including *A. murdochii* and *A. degenerii* (Fig 6).

**Figure 5. The relative abundance of the top predictive feature differs between early and late endometriosis patients.** We illustrate the abundance in vaginal samples during the menses time point of the most predictive feature from our machine learning model, an OTU classified as belonging to the genera *Anaerococcus*. Log relative abundances for this OTU are shown for control, early stage endometriosis, and late stage endometriosis subjects. Q-values shown were calculated using the Mann-Whitney U-Test with an FDR correction as described in the materials and methods.
Figure 6. The phylogenetic relationship between the predictive *Anaerococcus* OTU and other cultured representatives of this genus. We illustrate the phylogenetic relationship between the *Anaerococcus* OTU identified in this study as predictive of disease severity (highlighted with a grey box) and *Anaerococcus* representatives in the SILVA database. Four major clades are indicated: (red) a clade including *A. hydrogenalis* and *A. vaginalis* strains; (blue) *A. lactolyticus* strains; (purple) the predictive *Anaerococcus* OTU, *A. degenerii*; and (green) *A. prevotii*.
murdochi, and A. degenerii strains; and (green) A. *tetraadius*, and A. *prevotii* strains. The OTU indicated in the disease severity classifier is highlighted with a grey box.

### 3.4 Conclusions

Endometriosis is a complex multi-factorial disorder associated with high morbidity and decreased quality of life for women who suffer from it. Due to high symptom heterogeneity, unknown pathogenesis, and invasive diagnostic techniques this disorder is very challenging to diagnose and treat.

In this work we have described the rectal and vaginal microbiomes and the distribution of vaginal CSTs in women with endometriosis and in a South American for, as far as we know, the first time. We identified an *Anaerococcus* OTU that differs significantly in abundance between early and late endometriosis patients during menses and that highly contributes to an early verses late classifier. It is encouraging for future diagnostic development that distinctions between stages of disease severity were identified during menses as this is when women with endometriosis experience the most severe symptoms; and also because non-menstruating women have a significantly lower occurrence of the disorder [2]. A future extension of this work would be to ascertain how generalizable and useful the results outlined in this study maybe for the development of a diagnostic; for example, by assessing the consistency of these results across larger and demographically diverse subject cohorts. The work outlined in this thesis helps to inform future studies of the challenges and potentials of using the vaginal microbiome as a potential diagnostic for endometriosis.

### 3.5 Acknowledgements

Library preparation and sequencing efforts through the MIT BioMicro Center were funded by the National Institute of Environmental Health Sciences of the NIH under award P30-ES002109. The authors would also like to acknowledge the support of Marta Privato for
providing invaluable organizational and communication support over the course of this research; and also Pawel Gajer and Jacque Ravel for providing access to McClassifier and the trained support vector machine learning model utilized for CST assignment in this study.

References

9. Sampson JA. Metastatic or embolic endometriosis, due to the menstrual
dissemination of endometrial tissue into the venous circulation. Am J Pathol.
American Society for Investigative Pathology; 1927;3(2):93.


21. GONÇALVES-FILHO RP, Brandes A, Christofolini DM, Lerner TG, Bianco B, Barbosa CP.


42. Anahtar MN, Byrne EH, Fichorova RN, Kwon DS, Anahtar MN, Byrne EH, et al. Cervicovaginal Bacteria Are a Major Modulator of Host Inflammatory Responses in the Female Genital Article Cervicovaginal Bacteria Are a Major Modulator of Host Inflammatory Responses in the Female Genital Tract. Immunity [Internet]. Elsevier Inc.; 2015;42(5):965–76. Available from: http://dx.doi.org/10.1016/j.immuni.2015.04.019


Chapter 4

Fecal microbial composition is predictive of boiler chicken feed conversion performance at the production breeder level

The contents of this chapter contribute to a publication in preparation by the following authors: Sandra Díaz-Sánchez*, Allison Perrotta*, Isaac Rockafellow, Eric J Alm, Rachel Hawken, and Irene Hanning

* These authors contributed equally to this work

Figures and tables are provided within the text. Supplemental information for this chapter is provided in Appendix C.
Abstract

The gut microbiome has a substantial impact on the nutrient availability for its host. For this reason, much work has been dedicated to identifying members of the microbial gut community that have potential to augment the growth rate performance of agricultural animals such as chickens. However, much of the work performed in poultry has been conducted in production level adult chickens. In this study we investigated how the gut microbiome differs between high and low performing chickens within the context of a pedigree breeder model. The gut microbiome of adult chickens and a time series of maturing chickens throughout the rearing process for two pedigree lineages was measured using 16S rRNA amplicon analysis. We identified compositional differences in the gut community between high and low performing birds and between genetic lines. We built a machine-learning based model that can predict feed conversion performance from the gut microbial community. Our results also suggest that the gut microbiome of maturing chickens is stable after 20 days of age and that OTUs predictive of feed conversion performance have abundances that are dynamic during host development and are differentially represented in the guts of young chickens.

4.1 Introduction

The microbiome of the intestinal tract has a substantial impact on nutrient availability, which in turn affects the ability of an animal to harvest energy from food. For this reason, the gut microbiome also has a substantial impact on the growth rate of agriculturally important animals. Correlations between microbiome profiles and industry metrics of growth rate have been previously described in production level chicken lines (1–3). However, correlations between performance metrics and the gut microbiome has not been investigated in breeder pedigree lineages. Additionally, the abundance dynamics of bacterial taxa correlated with performance over a time series of host development has not been previously studied. Understanding how the gut microbial community varies at the highest level of poultry production, the level of pedigree lineages, and throughout host
development has potential to improve production costs as well as future work seeking to manipulate the chicken gut microbiome (4,5).

In poultry production, pedigree lineages are bred for industry performance factors like feed conversion, weight gain and egg yield. High performing pedigree lines are then crossed to generate production breeders. The offspring of these production breeders are the products that are sold at retail, including meat and eggs (6). To ensure only individuals with desired traits are utilized as breeders, chickens are evaluated or graded for performance factors during specific developmental stages. Introducing microbial grading through sequencing and rigorous analysis of fecal samples could provide predictive information for the future performance of a given chicken. Additionally, manipulation of the gut microbiota early in life could help select for desired traits independent of genetic lineage.

In this work, we used a breeder pedigree model to correlate growth rates and feed efficiency with the gut microbiome. Two distinct pedigree lines were used as previous studies have suggested that the gut microbiome varies across production lineages of chickens (5,7). We identified high-level differences in the gut community between high and low performing birds and birds from different genetic lineages. We used a machine learning model to identify microbial biomarkers that predict feed conversion performance. To investigate how identified biomarkers vary over the course of host development, the gut microbiome of chicks was monitored during maturation from hatch to final selection after performance grading. Our results suggest that the microbiome profile of high and low performing chickens, and the OTUs predictive of that performance, differ across pedigree lineages. In this study we find that the gut microbiome of maturing chickens is dominated by previously observed OTUs that are not unique to an individual after 20 days of age. We also find that OTUs predictive of high performance are differentially represented in the guts of young chickens.

4.2 Materials and Methods
4.2.1 Ethics statement

All poultry were raised on Cobb-Vantress, Inc. farms under rearing conditions and protocols adjusted for company specific protocols. All sampling was conducted by Cobb-Vantress employees.

4.2.2 Poultry housing and sampling

For this study two lines of pedigree breeder broilers were chosen and labeled Line A and Line B. A total of 3 independent experiments were conducted: 1) maturation of the microbiome; 2) feed conversion performance grading; and 3) weight gain performance grading. For each experiment a separate group of 200 chickens was used. Lines A and B and birds for each sub-project were housed separately.

In the maturation experiments, all chickens were wing tagged so that the same chicken could be sampled consecutively throughout the experiments. Line A was sampled at day of hatch, then 2, 3, 5, 6 and 16 weeks of age. Line B was sampled at day of hatch, then 2, 3, 6, 9, and 16 weeks of age. Differences in ages of sample collection are consistent with industry recognized developmental time points for the two pedigree lineages used. Dietary changes, corresponding to common industrial rearing protocols of decreasing dietary protein over time, occurred prior to sample collection at each time point. For the maturation experiment, a total of 576 samples were collected, 48 of which were sequenced reflecting four samples at six different ages for both lines used.

For weight gain and feed conversion experiments, weight gain metric is defined as the weight gained by an individual bird over time; and feed conversion is a calculation of feed consumed divided by weight. For Line A chickens, weight gain and feed conversion were measured at 5 and 6 weeks of age, respectively. For Line B, weights and feed conversion were measured at 6 and 9 weeks of age, respectively. Differences in grading and sample collection age are consistent with industry protocols for the two genetic lines used in this study. Samples were obtained after the weights and feed conversion data were collected. A
total of 20 samples from each line were collected for each of the grading metric experiments. Specifically, 10 samples from high and 10 samples from low performing chickens for both performance metrics.

All fecal samples were collected by placing sterile baking paper under individual chickens. On occasion, chickens were gently squeezed to expel feces onto the baking paper. Once a fresh fecal sample was available, it was placed into a sterile 50 mL centrifuge tube and stored on ice. Samples were frozen at -20°C until the DNA could be extracted. Cloacal and fecal samples have been demonstrated to reflect different bacterial populations (8,9). Fecal samples were utilized in this study due to concerns that cloacal swabs may not collect enough material for sufficient analysis and the need to not sacrifice birds in the maturation sub-project so that individuals could be tracked over time.

4.2.3 DNA extraction, amplicon preparation, and sequencing

The DNA from all fecal samples was extracted in an individual manner. Fecal samples were weighed (100 mg) and DNA was extracted using a bead-beating procedure. DNA extracted was processed with the QIAmp®DNA Stool Mini Kit (Qiagen, Valencia, CA) according to the manufacturer’s guidelines to remove PCR fecal inhibitors. Extracted DNA was quantified by spectrophotometry using a QuBit fluorometer (Life Technologies, Grand Island, NY) prior to storage at -80°C.

The V4 hypervariable region of the 16S rRNA gene was amplified from extracted DNA in a 25 μl reaction containing 1x Five Primer Hot Master Mix (5 PRIME, Fisher Scientific Company, LLC, Pittsburgh, PA) and 0.2 μM of each primer, PCR grade water (MoBio Laboratories, Inc., Carlsbad, CA) and 50 ng DNA template. The primers and sample barcodes used were described by Carporaso et al. (10,11). The PCR conditions used were 94°C for 3 min; 35 cycles of 94°C for 45 sec, 50°C 60 sec and 72°C 90 sec; followed by 72 °C 10 min. Three amplification reactions were conducted for each sample (25 μL per PCR reaction). After amplification, which included incorporating barcodes, triplicate PCR reactions were combined into a single volume resulting in a total of 75 μL for each sample.
The PCR product was evaluated on a 1.5% agarose gel to confirm the target band size of 350 bp. Amplicon quality and concentrations were measured using a QuBit fluorometer (Life Technologies). Samples were then combined in equal concentrations (240 ng/sample) and cleaned using the MoBio UltraClean PCR Clean Up Kit following the manufacturer’s instructions (MoBio Laboratories, Inc., Carlsbad, CA). Samples were checked for quality and concentration using the Bioanalyzer™ 2100 and Agilent High Sensitivity DNA kit (Agilent Technologies, Palo Alto, CA, USA). All samples were sequenced using the Illumina MiSeq platform following the procedures described by Caporaso et al., (10,11).

4.2.4 Sequencing data processing and OTU analysis

From the sequencing data, forward and reverse reads were merged and filtered by estimated error per read using USEARCH (version 8) (SText 1). After error filtering, the data further processing and clustering of Operational Taxonomic Units (OTUs) was performed using an in house pipeline (SText 1). OTUs were taxonomically classified using the Ribosomal Database Project (RDP) with a confidence cut off of 0.5 (12).

Samples with less than 500 total reads and OTUs present in less than a total of three subjects were filtered out using custom python scripts (Python version 3.5.2) and were not considered in down stream analysis. A total of 1025 OTUs and 121 samples were retained and included in downstream analysis. All analyses were performed on an OTU table of normalized fractions generated by dividing the count of each OTU by the total number of reads present for a sample. Log transformation of the normalized OTU table was performed by adding a scalar to all values in the table. The scalar used, -13.77, was calculated as one half multiplied by one divided by the maximum total sample read count in the un-normalized OTU table. Alpha (Shannon diversity) and beta (Jensen-Shannon divergence) diversities (13) were calculated using Qiime (11) and pysurvey, respectively (SText 1).

4.2.5 Identification of predictive features in the microbiome
To confirm that differences between groups classified as high and low performance in weight gain or feed conversion were statistically significant, a One-way analysis of variance (ANOVA) was used. To identify genomic biomarkers differing between high and low performance groups, Linear discriminant analysis effect size (LEfSe) was performed (14). No cut-off values were assigned to the binned and classified sequences so that all taxa were analyzed regardless of percent abundance in a specific sample. Taxonomic group comparisons were conducted at the Phyla, Class, Order, and Family levels. Analyses were performed separately for each sub-project and genetic line. When possible, analyses were also conducted without respect to Line. Output of these analyses included a linear discriminant analysis (LDA) model of the differential features ranked by effect size. All parameters were default. Briefly, the non-parametric factorial Kruskal-Wallis sum rank test was used to detect overabundant taxa in each class. The Wilcoxon rank-sum test was then used to determine if the overabundant taxa were consistent among microbiomes within a group defined by the Metadata. Finally, an LDA was used to estimate the effect size of each overabundant taxa with respect to metadata characteristics.

All OTUs in the normalized OTU table with an average relative abundance greater than $1 \times 10^{-6}$ were subjected to univariate analysis. Specifically, a Mann-Whitney U-test (scipy, Python) was performed between individual OTU relative abundances in high and low performing groups. P-values calculated in significance testing were then corrected for multiple hypotheses using a Benjamini-Hochberg false discovery rate correction (statsmodels, Python) to generate corresponding q-values. OTUs that were identified as discriminating using linear discriminant analysis (LDA) were then filtered for those with q-values greater than 0.2.

Correlations with metadata (age, weight, feed conversion) treatment were conducted using a machine learning approach (15). All OTU abundance data were analyzed with a log transformation. All metadata categories were analyzed without transformation. All parameters utilized can be found in supplemental methods (SText 2). Model accuracy was determined by calculating the area under the curve (AUC) of the corresponding receiver operating characteristic curve. Model accuracy was also determined by calculating the significance of the true and predicted values of the corresponding confusion matrix using a
Fischer Exact Test. Top predictive OTUs contributing to the models were ranked using mean decrease accuracy (MDA). Correlations of OTU abundances and continuous metadata were determined using Kendall-Tau correlation coefficients.

Phylogenetic trees were produced for organization and visualization purposes using unique OTU representative sequences. Prior to building a phylogenetic tree, QIIME was used to align OTU unique sequences to full-length 16S reference sequences in the greengenes database (version gg_13_8) (16). The aligned sequences were trimmed of gaps using trimAL (version 1.2) (17) and a constraint file created from the RDP taxonomic classifications was generated using a custom Python script. FastTree (18) was then used to build the phylogenetic tree (SText 3). Abundance heat maps associated with the phylogenetic trees were generated using the Interactive Tree of Life tool (19).

4.3 Results and Discussion

4.3.1 The microbiome is dynamic and changes during host development

Shifts in the microbial community, observed as relative abundances at the phyla level, occur throughout the rearing process (Figure 1). At hatch the gut microbial communities of both lines are dominated by Proteobacteria. Over the course of development, both lineages show a switch to a community dominated by Firmicutes as well as increasing abundances of Actinobacteria, and Bacteroidetes. These results are consistent with previous studies (20,21). Proteobacteria are aerobic bacteria that have been shown to be the primary colonizers of the chicken intestinal tract (20). As the host ages histological development such as crypt formation, mucus production, and villi elongation combines with the reduction of carbon sources by the microbial community to shift the intestinal environment to an anaerobic profile of Firmicutes and Bacteroidetes (20,21). It is possible that dietary changes, as described in the materials and methods, also affect the microbial community and contribute to the community dynamics observed in this study.
Figure 1. The microbiome is dynamic and differs between genetic lines during host development. Here we illustrate the percent relative abundance of bacterial phyla present in the gut microbiome across a maturation time series. Bird age is indicated as days since hatch.

While phylum-level trends are similar between Lines A and B, there are notable distinctions. Specifically, the two lines differ in the abundances of Bacteroidetes in later developmental stages (Figure 1). In Line A chickens, Bacteroidetes are consistently abundant during the post feed-selection and final selection time points. In Line B chickens, only one chicken has a substantial abundance of Bacteroidetes during the final selection time point. These differences in phylum abundance between the genetic lineages may be partly due to cage effects, and differences in age during sampling as noted in the materials and methods section.
4.3.2 A majority of OTUs in the gut microbiome persist over host development

During early phases of growth, primary colonization and increasing diversity is followed by a community that is dominated by OTUs that persist between time points and are not unique to an individual chicken. Specifically, between hatch and 2 weeks of age, occurrences of new OTUs are more likely to be shared between at least two chickens, than they are to be unique to a single chicken (Figure 2). While new OTUs are observed throughout the time series, at 20 days of age the fraction of previously observed OTUs, that are observed in at least two different birds, is greater than 65% for both genetic lines. Additionally, after 20 days less than 20% of the total observed OTUs are unique to a single bird. From these observations we can infer that after 20 days of age the chicken gut microbiome is highly stable and consistent across chickens of the same genetic lineage. These results are consistent with observations of successional dynamics followed by community stability in the gut microbiome during the course of human and mouse developmental maturation (22–25).
Figure 2. During host development a majority of the OTUs are persistent and not unique. OTUs are defined as “Seen before” if they were observed in at least one previous time point, and “New” if they have never been observed in a previous time point. OTUs are defined as “unique” if they have been observed in only one chicken, and “shared” if they are observed in at least 2 chickens. Trends in OTU persistence as shown for Line A (A) and Line B (B) chickens.

4.3.3 Phylum level differences between high and low grading differ by genetic line and grading metric

A comparison of phyla abundances between high and low performing chickens shows high-level differences between genetic lines and grading metrics (Figure 3). In Line A chickens there were no significant differences between high and low performing chickens at the phylum level for either metric. However, a single OTU classified as Chloroplast has a significantly higher abundance in low performing chickens using both the weight gain and feed conversion metrics (SFigure 2). Chloroplast reads are often observed in human and animal gut 16S rRNA studies. These reads are derived from off-target amplification of food related plant material and are often removed prior to downstream analysis (26-28).

However, there are studies where the abundance of Chloroplast reads have been utilized as a proxy for abundance of plant dietary content (29,30) or low bacterial biomass (31). The significantly higher abundance of Chloroplast in poorly performing Line A chickens could indicate a higher consumption of feed with low increases in body mass or an overall lower bacterial biomass in the guts of these chickens, and therefore lower metabolic capability and nutritional benefit for the host.
Figure 3. There are phylum level differences between genetic lines and high or low performance graded adult chickens. We illustrate the log relative abundances of phyla observed in Line A and B chickens graded using the weight gain and feed conversion metrics. Only phyla with a non-log transformed mean abundance of at least 1% are included. Phyla that significantly differ between high and low chickens are indicated with a darker tinted background. Significance is defined as a p-value less than or equal to 0.05 using an uncorrected Mann-Whitney U-test.

Line B chickens differed significantly in the abundance of Bacteroidetes and Firmicutes, between high and low weight gain graded chickens (Figure 3). Line B chickens graded as high or low using feed conversion, also differed moderately in the abundance of Firmicutes, although this difference was only marginally significant (p=0.054).

Given the phylum-level differences observed between genetic lines and grading metrics, it is important to note that with the exception of Line B chickens graded using the weight gain metric, there are no significant differences in alpha diversity between high and low graded chickens (SFigure 3). This may indicate that weight and food conversion performance are dictated by specific bacterial groups in the community and are not a function of diversity as indicated in some human weight gain and obesity studies (32,33).
4.3.4 OTUs discriminating between high and low performance are specific for genetic line and grading metric.

Using LDA and univariate analysis we identified features in the gut microbiome that significantly differ in abundance between high and low performing Line B birds graded by weight gain and Line A birds graded by feed conversion (14) (Fig 4). When comparing high and low performing weight gain groups in Line B birds, OTUs belonging to the clade Campylobacteraceae were enriched in the high performing group (Fig 4A). In the chicken intestinal tract, Campylobacter is a commensal that typically colonizes the chicken intestinal tract around 2 and 3 weeks of age when the microbiome is more stable (34,35). Campylobacter is thought to promote the production of SCFAs by serving as a hydrogen sink (36). For this reason, higher concentrations of these bacteria could lead to an increased production of organic acids that can be used as an energy source for the host (37). It is important to note that in this study we are limited to the genus level of taxonomic resolution. The genus Campylobacter contains nonpathogenic and pathogenic organisms and pathogenesis can not be determined from genus level 16S rRNA gene classification alone.

![LDA Score (log 10)](image)

Figure 4. LDA and univariate analysis identify taxa that discriminate between high and low performing chickens. Histograms of the LDA scores show differentially abundant features in high (blue) and low (red) (A) Line B chickens graded with weight gain; and (B)
Line A chickens graded with the feed conversion metric. Only taxonomic groups representing features identified as discriminating in LDA analysis that also had a q-value less than or equal to 0.2 are included. Taxonomic levels are displayed as c, o, f, and g indicating class, order, family, and genus levels respectively.

Low weight gain performing Line B chickens were enriched for many OTUs in the Phylum Firmicutes. Specifically, OTUs classified as Bacilli, Lactobacillales, Burkholderiaceae, and Enterococcaceae (Figure 4A). There are confounding reports on the beneficial effects of these bacteria on nutrient availability and host body weight. For example, poultry and pig models report that Lactobacillales were associated with decreased weight gain and the production of bile salt hydrolase (38–40). Yet, Lactobacilli are a widely used bacteria in commercial probiotics for chickens and have been reported to increase weight gain (41). However the positive effect on performance is very strain-specific (42,43). It is noteworthy that improved weight gain is typically during early stages of early growth and probiotic and non-probiotic treated groups do not significantly differ in weight at later stages (44).

OTUs classified as Pelomonas and Cupriavidus were enriched in high feed converting birds when comparing high and low feed conversion groups Line A (Figure 4B). While these OTUs passed our q-value cut-off (see Materials and Methods), it is important to note that these were the only two to pass that cut-off when comparing high and low feed conversion groups in Line A chickens. We have maintained these OTUs in our findings, because they were also identified in our machine learning model described below. The identification of Cupriavidus as being enriched in high feed converting chickens could actually be derived from feed and not the gut community, similar to Chloroplast as described above. This is because different species of this genus are provided as dry cells in poultry feed as a protein source (45,46).

It is important to note that LDA also identified discriminating OTUs in the cases of Line A weight gain and Lin B feed conversion chickens. However, none of those identified OTUs passed our univariate analysis.
4.3.5 Microbiota composition predicts feed conversion performance

Given the distinctions in microbiota composition between high and low performing birds observed at the phylum level and in the LDA analyses, we asked whether these differences were robust enough to enable computational prediction of performance within a genetic line and grading method. We built a random forest machine-learning model based on the microbiota composition and tried to predict high or low weight gain or feed conversion performance within each genetic line. Using this model, we were able to accurately predict if chickens were from high or low feed conversion groups in Lines A and B, with AUCs equal to 0.85 (p=0.04, Fisher Exact test) and 0.76 (p=0.003, Fisher Exact test), respectively (Figure 5 A and C, STable 1).
Figure 5. Microbiota composition predicts feed conversion grading. This figure depicts the accuracy, as determined by AUC analysis, of the models built to predict high verses low feed conversion grading from the microbiome composition of Line A and B chickens (A, C, respectively). The top contributing OTU features of those models are shown for Line A and B in B and D, respectively. Taxonomic level for each OTU is displayed is genus unless otherwise noted. Phyla, class, order, and family levels are indicated as p, c, o, and f, respectively. Correlations of OTU abundance and continuous feed conversion scores, calculated using Kendall-Tau correlation, are indicated as high (teal, Tau coefficient < -0.3), low (red, Tau coefficient > 0.3), and no correlation (grey, Tau coefficient < absolute value of 0.3). Because lower feed conversion scores reflect desirable performance, an OTU with a
negative correlation was deemed as contributing to high performance and an OTU with a positive correlation was deemed as contributing to low performance. * Indicates OTUs shared between B and C.

The classification models built for the prediction of high or low weight gain performance were less effective than those for feed conversion. For Line A chickens the classification was less accurate than that for feed conversion, AUC = 0.67, and the predictive power was insignificant (p=0.17) (Stable 1). The classification model built for Line B chickens was equally accurate to that of feed conversion, AUC = 0.76, but the predictive power was insignificant (p=0.17) (Stable 1).

The finding that the feed conversion classifier was more accurate than the weight gain classifier is consistent with a history of difficulties in predicting weight gain and obesity from gut microbial composition (47). Nonetheless, there is substantial evidence connecting the gut microbiota and nutrient availability which would be reflected in the feed conversion data (36,48–51).

### 4.3.6 OTUs that predict performance are distinct to genetic lines

OTUs contributing to the predictive models of feed conversion performance differed between the two genetic lines. Only two OTUs, an OTU classified as Clostridiales and one as Lactobacillus, are shared between the top OTUs for Line A and Line B feed conversion classifiers (Figure 5 B and D). For Line A chickens a majority of the model predictive OTUs have abundances that correlate with high feed conversion performance (Figure 5B). However, for Line B chickens many of the predictive OTUs correlate with low performance or do not have a strong correlation with performance at all (Figure 5D). Regardless of these differences, there are commonalities at various taxonomic levels. Many Firmicutes OTUs such as those classified as Lactobacillus, Lachnospiraceae, and Clostridiales are highly represented as predictive of feed conversion in both genetic lines (Figure 5 B and D). These results are consistent with previous studies that have indicated a correlation between
Clostridia, *Lactobacillus*, and Lachnospiracea and improved or diminished performance grading (1,2,5).

As mentioned above, a univariate analysis comparing OTU abundances between high and low feed conversion graded chickens resulted in only two OTUs maintaining a q-value less than or equal to 0.2 for both Lines A and B. This indicates that while each of the OTUs identified in our model do not individually have a strong enough correlation between abundance and feed conversion to be considered significantly different across groups, as a community they are predictive of feed conversion performance.

4.3.7 OTUs that predict performance are dynamic across a maturation time series

To investigate the presence and abundance of our model predictive OTUs in developing chickens, we next assessed the relative abundances of these OTUs within a maturation time series (Figure 6). Notably, multiple *Lactobacillus* OTUs were present in the group of OTUs contributing to low feed conversion performance (Figure 6B). This is consistent with previous studies that have shown a correlation between *Lactobacillus* species and poor feed conversion in chickens as well as mouse studies that have shown that *Lactobacillus* can inhibit weight gain even when a high fat diet is consumed (1,2,52). These results call into question the policy of using Lactobacillus bacteria as probiotics, which is commonly done in the poultry industry (53). The popularity of *Lactobacillus spp.* as a popular probiotic in the poultry industry is possibly based on protection from gastrointestinal pathogens and zoonoses including *Salmonella* and not upon desired phenotypes alone (54).
Figure 6. Top predictive OTUs correlated feed conversion performance are dynamic during host development. Relative abundance of top predictive OTUs that correlate with high performance (A) and low performance (B) are shown across a maturation time series in Line A and B chickens. Time points 1-6 in figure correspond to development stages: 1) Hatch, 2) Brooding-1, 3) Brooding-2, 4) Grading, 5) Post Feed Selection, 6) Final selection.

An OTU classified as *Faecalibacterium* is correlated with poor feed conversion in this study. Stanley et al. (2) found that a *Faecalibacterium prausnitzii* OTU was correlated with high
feed conversion and this genus is commonly considered a butyrate producing group (55,56). This difference of Faecalibacterium being associated with desired or undesirable feed conversion reflects the complexity of microbial-host metabolic interactions and the need for complex community manipulation to generate a reliable phenotype in the host. These differences between studies could also be due to functional differences at the species or strain level, as has been previously argued for the differential effects of Lactobacillus species on chicken performance factors (2).

While OTUs classified as belonging to the phylum Firmicutes were highly represented in OTU groups correlated with both the high and low performance (Figure 6); some important distinctions can be made when potential OTU metabolism and its affect on host nutrition is considered. The bacterial taxa that ferment carbohydrates to produce specific short chain fatty acids (SCFA) are polyphyletic and can not be determined by 16S rRNA analysis alone (55,57), but much research has focused on the bacterial groups capable of producing specific SCFAs, especially butyrate and propionate (4,55,57,58).

Many of the OTUs correlated with desired feed conversion belong to bacterial groups known to contain butyrate producing species including Enterococcus, Eubacterium, Lachnospiraceae, and Subdoligranulum (55,56). Butyrate serves as a local energy source for colonocytes and is believed to have anti-inflammatory effects, encourage epithelial cell proliferation, and serve a potential role in weight control (48,56,58). Multiple studies have correlated butyrate production with desired performance in chickens (2,4). However, if successful colonization of young chickens is required to reliably manipulate the gut community, then differences in the abundances of butyrate producing bacteria during the maturation time series could help inform future work. For example, the OTUs classified as Enterococcus, Subdoligranulum, Pelomonas, Cupriavidus, and Burkholderia in Figure 6A are present in at least one chicken during hatch and persist through out the time series. This indicates that they are naturally present in young chickens and are therefore may be more likely to colonize if introduced as probiotics or synthetic communities early in life. In contrast, the OTUs classified as Eubacterium and Lachnospiraceae in Figure 6A are absent or very sparse in young chickens and may be less likely to colonize and persist and may be less useful as an early-life probiotic.
Many of the OTUs correlated with poor feed conversion have been identified as acetate and propionate producers in other studies. Specifically OTUs classified as *Lactobacillus spp.*, *Blautia spp.*, as well as some *Lachnospiracea spp.* in Figure 6B (4,49,56,59). Previous studies have found evidence for high levels of propionate and acetate production in the chicken gut (4,36). Propionate can contribute to weight control and is thought to do so through the enhancement of satiety (56,58). Acetate production has been linked to weight control independent of satiety (55,58). High acetate production would generate higher levels of ATP for the acetogenic bacteria but allows for little NADH to be recycled and proves less nutritious for the host than other SCFAs because of its lower thermodynamic payoff (36,60). Jackson et. al. (61) also found that the extent of butyrate degradation in a system is negatively correlated with acetate concentrations.

In light of the possible correlation between high propionate and acetate production and undesirable feed conversion phenotypes, it is important to keep in mind that both propionate and acetate play important roles in healthy gut metabolism. Propionate and acetate serve important roles in the liver and other peripheral tissues (48). High acetate concentrations have been shown to favor the growth of beneficial bacterial communities, including *Bifidobacterium* (62). Additionally, acetogenesis can play an important role in the gut community by providing a hydrogen sink and serving as a potential precursor for the synthesis of butyrate (48,57). It is also worth noting that there can be variability in complex SCFA metabolism, and therefore the nutrients available to the host; and that this variability is highly dependent upon initial carbohydrate metabolism steps and the presence of different syntrophic hydrogen-providing and hydrogen-using pairs available within the community as a whole (61).

### 4.4 Conclusions

In this work, we have identified bacterial groups that are predictive of feed conversion in breeder pedigree lineages that represent the highest production level in the poultry industry. Additionally, we have identified predictive OTUs that are present during early life
stages and persist into adulthood allowing for a more informed selection of candidates for future probiotic work. Considering the significant impact that the microbiome has on weight gain and feed conversion, assessment of these biomarkers through sequencing and rigorous analysis of fecal samples could provide predictive information for the future performance of a given chicken thereby enhancing traditional performance grading processes. Alternatively, manipulation of the gut microbiota early in life could be used to encourage desired traits independent of genetic lineage has the potential to render grade selection between pedigree producers and production breeders obsolete.

Successfully and reliably manipulating the gut community for a given phenotype is not an easy task. The connections between bacterial metabolism, its affects on host phenotype, and how these metabolic regimes change with host development are highly complex. In the context of the phenotypes discussed in this study, a chicken may have high weight gain, but low feed conversion performance due to high feed consumption. Likewise, a chicken may be very lean but have high feed conversion performance. Thus, the data collected in these experiments indicate that future probiotic design should take into account that some probiotics may positively impact one factor while negatively impacting another.

Additionally, in the course of this work we have shown that one factor contributing to this difficulty is the successional dynamics observed during the maturation (Figure 1 and 6). Because the microbiome is dynamic, we suggest that future work focused on identifying or experimentally testing probiotic bacterial groups that are predictive of desired phenotype, investigate the dynamics of those bacteria in young chickens and not just adults. As bacteria that are present in young chickens and persistent into adulthood could reflect unknown colonization factors (4) that could be important to the success of the treatment.

References


43. Stanley D, Geier MS, Denman SE, Haring VR, Crowley TM, Hughes RJ, et al.


61. Jackson BE, Mcinerney MJ. Anaerobic microbial metabolism can proceed close to

Chapter 5

Conclusions

In this thesis we utilized 16S rRNA analysis, NGS sequencing, and time series data to investigate utilizing environmental and host associated bacterial communities as predictors in industrial and clinical applications. Three specific applications were investigated, including: the role of inoculum in bioreactor function; the potential for a microbiome-based diagnostic of endometriosis; and using the microbiome to predict chicken flock performance in a commercial setting. In each application the abundance dynamics over time played a pivotal role in the reliability and impact of the prediction. In this chapter the main conclusions, limitations, and potential for future work specific to and common across each application will be discussed.

5.1 Chapter 2 Conclusions

In the second chapter, we found that the microbial community composition and function in sequential batch reactors over time was distinct to the starting inoculum used. We also found that the generation of a complex volatile fatty acid production profile was not specific to the diversity of the starting inoculum. These results suggest that the composition of the inoculum predictably contributes to bioreactor community structure and function.

While these results are informative, the mechanisms that made one bioreactor more productive than another was not explored. Additional factors such as species-species interactions or variable metabolic gene content across strains could contribute (1,2). It is also important to note that it is conceivable that the two less productive reactors in this work, those inoculated with camel dung and wastewater sludge, are not less functional communities, but instead may be experiencing a slower adaptive period or are experiencing senescence. Since this work was performed with 16S rRNA analysis, strain level, functional, and expression differences between the bioreactor microbial communities could not be addressed. Additionally, this experiment was conducted for a total of 14 days,
which may not be enough time to fully account for a slower adaptive period in the less productive bioreactors.

The research conducted in this thesis chapter establishes a basis for future work to further investigate the impact that starting inoculum has on bioreactor microbial composition and function. Specifically, a future contribution would be to perform metagenomics or transcriptomics to assess the metabolic functions of these distinct bioreactor communities over time. Another exciting direction would be to further investigate how original inoculum affects community stability and resilience to invasion in a bioreactor. This could be performed by introducing an invading bacterial isolate, one not found in any of the original inocula, during the time course and measuring the community and invader abundance dynamics over time.

5.2 Chapter 3 Conclusions

In the third chapter of this thesis, we found that the vaginal microbiome composition can be used to predict severity of the common gynecological disorder endometriosis. The accuracy of the classification of disease severity was specific to the menses time point and reliant upon a single *Anaerococcus* OTU. As far as we know, this work is also the first to describe the distribution of vaginal community state types (CSTs) in a South American cohort. Our results help to inform future studies of the challenges and potentials of using the microbiome, particularly the vaginal microbiome during menses, as an indicator of endometriosis severity.

One of the major challenges faced in performing this research is the subdivision of the subject groups by CST within a time point and also the inter-variability of subjects across time points. These subdivisions within the subject populations were the largest compositional signal in this dataset. Our ability to perform statistical analyses on this dataset was greatly limited by the small patient numbers in each CST group. For these reasons we would suggest that future studies seeking to investigate the vaginal
microbiome as a potential diagnostic of disease to consider CST distributions when calculating the power of their studies.

Future work can build upon the results found in this thesis to reveal how generalizable and useful those results could be in a clinical setting. One interesting extension of the work covered in this thesis would be to assess the consistency of these results across demographically and geographically diverse cohorts, as previous studies have reported that the distribution of vaginal CSTs in a population varies by race and geographic location (3–5). The work performed in this chapter was also under powered for many statistical tests; as such, the reproducibility of our results should be determined using a larger subject cohort. An additional extension of this work would be to investigate how the inclusion of a healthy, and not simply a non-endometriosis, control group affects our ability to predict disease state and severity.

5.3 Chapter 4 Conclusions

In the fourth chapter, we identified compositional differences in the fecal microbiomes of high and low feed conversion chickens and built a classifier for feed conversion performance. We then investigated the abundance dynamics of the top predictive OTUs from the classifier in a maturation time series. We found that OTUs predictive of high feed conversion performance are differentially abundant in the guts of young chickens, a result that could prove informative for future work addressing the use of early-life probiotics.

Some of the limitations of the work discussed in this thesis chapter are the fact that our results are based upon OTU abundance correlations with host phenotypes. The functions of those predictive OTUs and how those functions impact host nutrient availability, specifically short chain fatty acid (SCFA) production, can not be determined by 16S rRNA analysis alone (6,7). To test the correlations of OTU abundance, phenotype, and possible microbial community function, experiments measuring the impact of specific bacterial strains and their affect on performance phenotype and concentrations of SCFA in the gut would need to be conducted. Additionally, future studies can build upon the work
described in this thesis by experimentally testing the impact of bacterial taxa that are predictive of high performance but are differentially represented in the guts of young chickens.

5.4 Common Conclusions

In the course of this thesis we utilized 16S rRNA analysis, NGS sequencing, and time series data to investigate the temporal variation and dynamics of microbial communities in three distinct environments and systems. The environmental and host associated microbial communities covered in this work were used to predict outcomes in specific industrial and clinical applications.

A common limitation of this work is that microbial composition was measured by 16S rRNA analysis. This type of analysis is powerful for investigating the relative abundances and taxonomic diversity of microbial communities, but it can not inform function nor identify mechanisms of interaction with the environment or host. Another challenge of using 16S rRNA analysis is that bacterial sequences are clustered by sequence similarity to form OTUs, which often limits the taxonomic resolution to the genus level and does not allow for species, let alone strain, level identification of bacterial taxa (8–10).

Even with the limitations and challenges discussed above, we have demonstrated the great importance of considering temporal factors in order to informatively and accurately exploit microbial communities for industrial and clinical applications. The work presented in this thesis advances research seeking to utilize naturally occurring microbial communities for specific applications and contributes to the scope of future work in this field.

References


Appendix A
Supplemental Information for Chapter 2

Supplemental Text

S1 Text. Inocula and fermentation methods utilized.
Mangrove, Camel, and Sludge inocula were stored at 4 °C for 43 hours, 24 hours, and 4 days respectively prior to the start of experiments.

S2 Text. Sequence processing methods utilized.
Raw data was quality filtered using QIIME (version 1.3.0) (1). Fastq files were processed using the split_library_fastq.py program of QIIME, truncating sequences when more than one base in a row dropped below Phred quality score 14, corresponding to a probability of error around 3.98% (--last_bad_character N, --max_bad_run_length 1). Only sequences at least 99 bps long after quality filtering was retained (-min_per_read_length 99). All other parameters for this step were default parameters.

Sequences representing OTUs of interest were later reviewed using a minimum length of 150 (-min_per_read_length 150, in Qiime) (1). These longer sequences, 130 and 127 base pairs after primer removal from the forward and reverse read fastq file, respectively; were used for performing a NCBI BLAST (nt, megablast) searches to better taxonomically identify these OTUs below the genera level, the classification level that the RDP classifier is capable of.

S3 Text. OTU table analysis methods utilized.
The filter_samples_from_otu_table.py command in Qiime was used to filter the OTU table by input inoculate and remove samples with less than a thousand read counts from further processing (-n 1000)(1).

S4 Text. Description of alpha and beta diversity calculations. Alpha (Shannon diversity) and beta (Jensen-Shannon distance) diversities were calculated with sample_diversity indices = ['Shannon']) and dist_mat (metric='JS') commands, respectfully, in pysurvey (https://bitbucket.org/yonatanf/pysurvey, https://github.com/swo/pysurvey). The Shannon diversity metric used was shannon entropy and is defined as H = - Sum_i (p_i * log p_i), with p_i being the fraction of OTU_i.

S5 Text. Description of tools and parameters utilized to generate phylogenetic trees.
Aligned sequences used for tree building were trimmed of gaps using trimAL (version 1.2) with a gap threshold of 0.05 so gaps present in 95% or more of the sequences would be removed (-gt 0.05)(2). FastTree (3) was then used to build a phylogenetic tree from the trimmed sequences using nucleotide alignment (-nt), a GTR+CAT model (-gtr), and constraint file created from RDP classifications described above (-constraint). The constraint file was generated using custom a Python script (Python 2.7.10).
S6 Text. Description of statistical analyses utilized. The adonis function of the vegan package (version 2.4-1) in R (version 3.2.3) was utilized to perform permutational multivariate analysis to assess the significance of differences in chemistry and microbial composition between the inoculum sources. The functions permutest and betadisper in the R package vegan (version 2.4-1) were utilized to perform multivariate homogeneity of group dispersion to assess the significance of variance between replicates of the different inoculum sources. 999 permutations were used for both multivariate analyses.

Supplemental Figures and Tables

S1 Fig. Rarefaction plots of species richness. We illustrate the rarefied observed species calculated for each cycle in the Camel (A), Mangrove (B), and Sludge (C) bioreactors.
S2 Fig. Microbial communities are reproducible and distinct to initial inoculum. Abundances of OTUs are represented as percent relative abundance in Camel (A), Mangrove (B), and Sludge (C) reactors. Taxonomic identification represents the lowest classification identified using the GreenGenes database (version gg_13_5) (4); class (c), family (f), genera (g), species (s). Plot was generated using Qiime (1).
S1 Table. Results of permutational multivariate analysis performed on microbial community data. We summarize the results obtained for multivariate analysis of within group dispersion (group dispersion) and between group differences (PERMANOVA) as the p-values obtained from these analyses of the microbial community data for each cycle.

<table>
<thead>
<tr>
<th>Cycle</th>
<th>Group Dispersion p-value</th>
<th>Permanova p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.35</td>
<td>0.005</td>
</tr>
<tr>
<td>1</td>
<td>0.09</td>
<td>0.005</td>
</tr>
<tr>
<td>2</td>
<td>0.14</td>
<td>0.001</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
<td>0.004</td>
</tr>
<tr>
<td>4</td>
<td>0.24</td>
<td>0.005</td>
</tr>
<tr>
<td>5</td>
<td>0.09</td>
<td>0.004</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
<td>0.003</td>
</tr>
<tr>
<td>7</td>
<td>0.35</td>
<td>0.005</td>
</tr>
</tbody>
</table>
S3 Fig. Dynamics of individual OTUs are reproducible across replicates. Abundances are represented as average percent relative abundance of specific OTUs in replicate bioreactors across varying inoculum (replicate 1 = blue, replicate 2 = green, replicate 3 = red). These specific OTUs represent the most abundant OTUs present in bioreactor communities after cycle 1: *Escherichia coli* (A), *Enterobacteriaceae* (B), *Clostridium acetobutylicum* (C), Gammaproteobacteria (D), *Clostridium pasteurianum* (E), *Bacteroides* (F), *Prevotella* (G), *Clostridium* (H), *Vibrio fortis* (I), Viellonellaceae (J), *Bifidobacterium animalis* (K). Points for samples representing Mangrove bioreactor replicate two, cycle one and Sludge bioreactor replicate one, cycle six have been excluded.

<table>
<thead>
<tr>
<th>Cycle</th>
<th>Group Dispersion p-value</th>
<th>Permanova p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>0.62</td>
<td>0.005</td>
</tr>
<tr>
<td>2</td>
<td>0.62</td>
<td>0.013</td>
</tr>
<tr>
<td>3</td>
<td>0.42</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>0.54</td>
<td>0.003</td>
</tr>
<tr>
<td>5</td>
<td>0.47</td>
<td>0.003</td>
</tr>
<tr>
<td>6</td>
<td>0.15</td>
<td>0.001</td>
</tr>
<tr>
<td>7</td>
<td>0.1</td>
<td>0.005</td>
</tr>
</tbody>
</table>

S2 Table. Results of permutational multivariant analysis performed on chemical profile data. We summarize the results obtained for multivariate analysis of within group dispersion (group dispersion) and between group differences (PERMANOVA) as the p-values obtained from these analyses of the chemical profile data for each cycle. P-values for cycles zero are not included, as chemical data was not collected at the baseline time point.

<table>
<thead>
<tr>
<th>Cycle</th>
<th>Camel Average</th>
<th>Camel St. Dev.</th>
<th>Mangrove Average</th>
<th>Mangrove St. Dev.</th>
<th>Sludge Average</th>
<th>Sludge St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.01</td>
<td>0.36</td>
<td>0.02</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.03</td>
<td>0.01</td>
<td>0.37</td>
<td>0.1</td>
<td>0.2</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>0.07</td>
<td>0.02</td>
<td>0.76</td>
<td>0.04</td>
<td>0.1</td>
<td>0.08</td>
</tr>
<tr>
<td>4</td>
<td>0.32</td>
<td>0.04</td>
<td>0.73</td>
<td>0.09</td>
<td>0.16</td>
<td>0.14</td>
</tr>
<tr>
<td>5</td>
<td>0.47</td>
<td>0.04</td>
<td>0.61</td>
<td>0.18</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>6</td>
<td>0.53</td>
<td>0.08</td>
<td>0.47</td>
<td>0.12</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>0.52</td>
<td>0.1</td>
<td>0.34</td>
<td>0.09</td>
<td>0.01</td>
<td>0</td>
</tr>
</tbody>
</table>

S3 Table. Abundance of Clostridia OTUs in initial inoculum is not predictive of Clostridia abundance in reactor community. Averages represent relative abundance of
OTUs belonging to class Clostridia. St. Dev. column represents the standard deviation of the relative abundances across replicate reactors, n=3.

References


Appendix B

Supplemental Information for Chapter 3

Supplemental Methods

S1 Text. Sequencing data processing and OTU analysis
From the sequencing, the forward and reverse read FASTQ files were merged and the estimated error per read was determined using the usearch (version 8) command -fastq_mergepairs with options -eetabbeout, -fastq_ascii 33 (1). Merged reads were then filtered for an estimated error rate of 1 using the usearch8 command -fastq_filter (-fastq_maxee 1). After initial processing a total of 8,633,270 sequence reads were generated with an average of 53,622.79 reads per sample (±67,755.45, n=161).

Filtered reads were then input to an in house pipeline for further processing and clustering Operational Taxonomic Units (OTUs) (https://github.com/thomasgurry/amplicon_sequencing_pipeline). With in the pipeline sequences of less than 250 bases were filtered out, 97% de novo OTUs were clustered using the UPARSE algorithm (2) in usearch8 (1), and OTUs with less than two reads were removed.

Shannon diversity was calculated using the command alpha_diversity.py with the metric option set to Shannon Diversity (-m shannon) (3). The Jensen-Shannon divergence was calculated using the dist_mat (metric='JS') command in pysurvey (https://bitbucket.org/yonatanf/pysurvey, https://github.com/swo/pysurvey).

S2 Text. Identification of predictive features in the microbiome and statistical analysis
All OTU abundance data were analyzed with a log transformation (--x_log). All metadata categories were analyzed without transformation. All parameters utilized can be found in supplemental methods. The analysis was run using three forests (-n 3), the data was shuffled 1000 times for p-value calculations (-s 1000), and downsampling was utilized for class balancing of classification models (-b downsample). All other parameters were default parameters.

S3 Text. Community state type assignment
In order to assign a community state type (CST) to each vaginal sample MCclassifier (https://github.com/pgajer/MCclassifier) was used to generate an environment specific species level table of read counts. A model directory of 16S rRNA V4 sequences corresponding to bacterial species commonly found in the vaginal microbiome was provided by the creators of this tool (see acknowledgements).

S4 Text. Description of tools and parameters utilized to generate phylogenetic trees.
Reference sequences for the genus Anaerococcus and the out group were downloaded from the SILVA database. Uncultured, unclassified, and candidate sequences were removed using custom python scripts. Aligned sequences used for tree building were trimmed of gaps using trimAL (version 1.2) with a gap threshold of 0.05 so gaps present in 95% or more of the sequences would be removed (-gt 0.05) (5). FastTree (6) was then used to build a phylogenetic tree from the trimmed sequences using nucleotide alignment (-nt) and a GTR+CAT model (-gtr).

Supplemental Figures and Tables

SFigure 1. Vaginal samples collected during the follicular and menses phases cluster by CST. We illustrate the hierarchical linkage clustering of samples based upon the community composition and OTU abundance for patient vaginal samples collected during the (A) follicular and (B) time points. The CST assigned to each patient is indicated as I-A, I-B, II, III-A, III-B, IV-A, IV-B, or V for each sample. Disease state of each patient is also indicated as endometriosis or control for each sample. OTU abundances are displayed as fractions of relative abundance.
A.
SFigure 2. Alpha diversity between endometriosis and control patient cohorts. Alpha diversity, calculated as Shannon Diversity, is shown for (A) rectal and (B) vaginal samples for the follicular and menses time points in the endometriosis (blue) and control (green) subject groups.
**Figure 3. Beta diversity between endometriosis and control patient cohorts.**

Pairwise beta diversity, calculated as Jensen-Shannon Divergence, is shown for (A) rectal and (B) vaginal samples collected during the follicular and menses time points in the endometriosis (blue) and control (green) subject groups. Follicular verses menses comparisons are calculated on samples obtained from the pairwise comparison of same patient across the two time points and not between patients. * Indicates a p-value less than 0.005, calculated using the Mann-Whitney U-Test.

**Stable 1. De-identified clinical data for (A) endometriosis case and (B) control subject cohorts.** De-identified metadata for each patient used in this study is displayed including age, self reported race, body mass index (BMI), disease state, and infertility are reported for both case and control subjects. Additional health issues are reported for control subjects, and stage of disease is reported for case subjects. Blank fields in the table represent a category where there is no data for that subject.
<table>
<thead>
<tr>
<th>Patient</th>
<th>Age</th>
<th>Race</th>
<th>BMI</th>
<th>Disease State</th>
<th>Infertility</th>
<th>Additional Health Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>37</td>
<td>white</td>
<td>26.9</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>16</td>
<td>25</td>
<td>mixed race</td>
<td>26.9</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>19</td>
<td>47</td>
<td>white</td>
<td>20.6</td>
<td>Control</td>
<td>no</td>
<td>diagnostic laparoscopy</td>
</tr>
<tr>
<td>20</td>
<td>47</td>
<td>mixed race</td>
<td>25.5</td>
<td>Control</td>
<td>no</td>
<td>hysterectomy</td>
</tr>
<tr>
<td>22</td>
<td>33</td>
<td>white</td>
<td>23.7</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>30</td>
<td>35</td>
<td>white</td>
<td>20.8</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>32</td>
<td>41</td>
<td>mixed race</td>
<td>25</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>33</td>
<td>32</td>
<td>white</td>
<td>22.7</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>35</td>
<td>36</td>
<td>mixed race</td>
<td>22.4</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>40</td>
<td>41</td>
<td>black</td>
<td>20.9</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>42</td>
<td>42</td>
<td>white</td>
<td>22.2</td>
<td>Control</td>
<td>no</td>
<td>ovarian serous cystadenoma</td>
</tr>
<tr>
<td>43</td>
<td>26</td>
<td>black</td>
<td>23.4</td>
<td>Control</td>
<td>no</td>
<td>ovarian teratoma</td>
</tr>
<tr>
<td>45</td>
<td>21</td>
<td>mixed race</td>
<td>24.6</td>
<td>Control</td>
<td>no</td>
<td>ovarian teratoma</td>
</tr>
<tr>
<td>46</td>
<td>46</td>
<td>mixed race</td>
<td>23.4</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>51</td>
<td>40</td>
<td>mixed race</td>
<td>25.3</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>57</td>
<td>35</td>
<td>white</td>
<td>24.3</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>58</td>
<td>31</td>
<td>black</td>
<td>22.6</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>60</td>
<td>34</td>
<td>white</td>
<td>23.7</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>61</td>
<td>29</td>
<td>mixed race</td>
<td>22.4</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>63</td>
<td>26</td>
<td>mixed race</td>
<td>30.9</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>64</td>
<td>34</td>
<td>white</td>
<td>24.6</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>67</td>
<td>38</td>
<td>mixed race</td>
<td>27.8</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>71</td>
<td>36</td>
<td>black</td>
<td>30.5</td>
<td>Control</td>
<td>no</td>
<td>tubal ligation</td>
</tr>
<tr>
<td>Patient</td>
<td>Age</td>
<td>Race</td>
<td>BMI</td>
<td>Disease State</td>
<td>Disease Stage</td>
<td>Infertility</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>--------</td>
<td>------</td>
<td>---------------</td>
<td>---------------</td>
<td>-------------</td>
</tr>
<tr>
<td>5</td>
<td>46</td>
<td>white</td>
<td>33.3</td>
<td>Endo</td>
<td>4</td>
<td>no</td>
</tr>
<tr>
<td>6</td>
<td>38</td>
<td>white</td>
<td>23.2</td>
<td>Endo</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>35</td>
<td>white</td>
<td>20.5</td>
<td>Endo</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>32</td>
<td>white</td>
<td>19.8</td>
<td>Endo</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>11</td>
<td>37</td>
<td>white</td>
<td>27.3</td>
<td>Endo</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>13</td>
<td>38</td>
<td>white</td>
<td>20.4</td>
<td>Endo</td>
<td>3</td>
<td>yes</td>
</tr>
<tr>
<td>15</td>
<td>36</td>
<td>white</td>
<td>21.6</td>
<td>Endo</td>
<td>3</td>
<td>yes</td>
</tr>
<tr>
<td>17</td>
<td>40</td>
<td>white</td>
<td>23</td>
<td>Endo</td>
<td>4</td>
<td>no</td>
</tr>
<tr>
<td>26</td>
<td>28</td>
<td>white</td>
<td>25.1</td>
<td>Endo</td>
<td>4</td>
<td>no</td>
</tr>
<tr>
<td>27</td>
<td>39</td>
<td>white</td>
<td>25.1</td>
<td>Endo</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>29</td>
<td>33</td>
<td>black</td>
<td>22.2</td>
<td>Endo</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>31</td>
<td>37</td>
<td>white</td>
<td>25.3</td>
<td>Endo</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>36</td>
<td>28</td>
<td>white</td>
<td>25.2</td>
<td>Endo</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>37</td>
<td>34</td>
<td>white</td>
<td>22.8</td>
<td>Endo</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>38</td>
<td>34</td>
<td>white</td>
<td>19.1</td>
<td>Endo</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>39</td>
<td>27</td>
<td>white</td>
<td>35.6</td>
<td>Endo</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>48</td>
<td>36</td>
<td>white</td>
<td>26.4</td>
<td>Endo</td>
<td>4</td>
<td>no</td>
</tr>
<tr>
<td>49</td>
<td>32</td>
<td>black</td>
<td>23.3</td>
<td>Endo</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>50</td>
<td>36</td>
<td>white</td>
<td>22.04</td>
<td>Endo</td>
<td>4</td>
<td>yes</td>
</tr>
<tr>
<td>52</td>
<td>35</td>
<td>black</td>
<td>20</td>
<td>Endo</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>53</td>
<td>23</td>
<td>white</td>
<td>19.3</td>
<td>Endo</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>54</td>
<td>23</td>
<td>white</td>
<td>23.7</td>
<td>Endo</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>55</td>
<td>34</td>
<td>white</td>
<td>23.3</td>
<td>Endo</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>59</td>
<td>35</td>
<td>mixed race</td>
<td>30.9</td>
<td>Endo</td>
<td>4</td>
<td>no</td>
</tr>
<tr>
<td>62</td>
<td>47</td>
<td>white</td>
<td>28.2</td>
<td>Endo</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>65</td>
<td>38</td>
<td>white</td>
<td>27.1</td>
<td>Endo</td>
<td>4</td>
<td>no</td>
</tr>
<tr>
<td>66</td>
<td>22</td>
<td>white</td>
<td>23.2</td>
<td>Endo</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>68</td>
<td>32</td>
<td>black</td>
<td>28.4</td>
<td>Endo</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>70</td>
<td>40</td>
<td>white</td>
<td>21.8</td>
<td>Endo</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>72</td>
<td>47</td>
<td>black</td>
<td>34.1</td>
<td>Endo</td>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>73</td>
<td>39</td>
<td>black</td>
<td>23.7</td>
<td>Endo</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>74</td>
<td>37</td>
<td>white</td>
<td>32.3</td>
<td>Endo</td>
<td>4</td>
<td>no</td>
</tr>
<tr>
<td>75</td>
<td>33</td>
<td>white</td>
<td>20.2</td>
<td>Endo</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>76</td>
<td>36</td>
<td>mixed race</td>
<td>22</td>
<td>Endo</td>
<td>4</td>
<td>yes</td>
</tr>
<tr>
<td>77</td>
<td>21</td>
<td>white</td>
<td>20.9</td>
<td>Endo</td>
<td>4</td>
<td>yes</td>
</tr>
<tr>
<td>78</td>
<td>49</td>
<td>mixed race</td>
<td>30.1</td>
<td>Endo</td>
<td>2</td>
<td>no</td>
</tr>
</tbody>
</table>

SFigure 4. Statistics for machine-learning classifications performed for metadata categories. Area under the curve (AUC), Fisher p-values, and Out of Bag (OOB) estimated error rates calculations for various metadata categories are shown.
### Early vs late disease

<table>
<thead>
<tr>
<th>Sample type</th>
<th>Time point</th>
<th>AUC</th>
<th>Fisher p-value</th>
<th>OOB (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rectal</td>
<td>follicular</td>
<td>0.49</td>
<td>0.66</td>
<td>44.87</td>
</tr>
<tr>
<td></td>
<td>menses</td>
<td>0.43</td>
<td>1</td>
<td>53</td>
</tr>
<tr>
<td>vaginal</td>
<td>follicular</td>
<td>0.36</td>
<td>1</td>
<td>48.15</td>
</tr>
<tr>
<td></td>
<td>menses</td>
<td>0.89</td>
<td>0.019</td>
<td>16.67</td>
</tr>
</tbody>
</table>

### Disease state

<table>
<thead>
<tr>
<th>Sample type</th>
<th>Time point</th>
<th>AUC</th>
<th>Fisher p-value</th>
<th>OOB (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rectal</td>
<td>follicular</td>
<td>0.44</td>
<td>1</td>
<td>45.24</td>
</tr>
<tr>
<td></td>
<td>menses</td>
<td>0.59</td>
<td>0.075</td>
<td>33.33</td>
</tr>
<tr>
<td>vaginal</td>
<td>follicular</td>
<td>0.34</td>
<td>0.08</td>
<td>57.78</td>
</tr>
<tr>
<td></td>
<td>menses</td>
<td>0.42</td>
<td>0.49</td>
<td>44.44</td>
</tr>
</tbody>
</table>

### Early disease vs control

<table>
<thead>
<tr>
<th>Sample type</th>
<th>Time point</th>
<th>AUC</th>
<th>Fisher p-value</th>
<th>OOB (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rectal</td>
<td>follicular</td>
<td>0.44</td>
<td>0.73</td>
<td>54.55</td>
</tr>
<tr>
<td></td>
<td>menses</td>
<td>0.66</td>
<td>0.02</td>
<td>31.03</td>
</tr>
<tr>
<td>vaginal</td>
<td>follicular</td>
<td>0.29</td>
<td>0.09</td>
<td>68.52</td>
</tr>
<tr>
<td></td>
<td>menses</td>
<td>0.68</td>
<td>0.12</td>
<td>34.52</td>
</tr>
</tbody>
</table>

### Late disease vs control

<table>
<thead>
<tr>
<th>Sample type</th>
<th>Time point</th>
<th>AUC</th>
<th>Fisher p-value</th>
<th>OOB (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rectal</td>
<td>follicular</td>
<td>0.46</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>menses</td>
<td>0.36</td>
<td>0.66</td>
<td>53.33</td>
</tr>
<tr>
<td>vaginal</td>
<td>follicular</td>
<td>0.41</td>
<td>0.53</td>
<td>40.74</td>
</tr>
<tr>
<td></td>
<td>menses</td>
<td>0.47</td>
<td>0.63</td>
<td>45.83</td>
</tr>
</tbody>
</table>

### Stage 4 vs other disease stages

<table>
<thead>
<tr>
<th>Sample type</th>
<th>Time point</th>
<th>AUC</th>
<th>Fisher p-value</th>
<th>OOB (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rectal</td>
<td>follicular</td>
<td>0.47</td>
<td>0.35</td>
<td>24.36</td>
</tr>
<tr>
<td></td>
<td>menses</td>
<td>0.51</td>
<td>1</td>
<td>28.79</td>
</tr>
<tr>
<td>vaginal</td>
<td>follicular</td>
<td>0.42</td>
<td>1</td>
<td>38.27</td>
</tr>
<tr>
<td></td>
<td>menses</td>
<td>0.87</td>
<td>0.0088</td>
<td>10</td>
</tr>
</tbody>
</table>
References


Appendix C
Supplemental Information for Chapter 4

Supplementary Methods

**SText 1. Sequencing data processing and OTU analysis**
Forward and reverse read FASTQ files were merged and the estimated error per read was determined using the usearch (version 8) command -fastq_mergepairs with options -eetabbeout, -fastq_ascii 33 (1). Merged reads were then filtered for an estimated error rate of 1 using the usearch8 command -fastq_filter (-fastq_maxee 1).

Further processing and clustering Operational Taxonomic Units (OTUs) was performed with an inhouse pipeline (https://github.com/thomasgurry/amplicon_sequencing_pipeline). With in the pipeline sequences of less than 250 bases were filtered out, 97% denovo OTUs were clustered using the UPARSE algorithm (2) in usearch8 (1), and OTUs with less than two reads were removed.

After initial processing a total of 11,036,040 sequence reads were generated with an average of 91,967 reads per sample (±112,735, n=120). Shannon diversity was calculated using the command alpha_diversity.py with the metric option set to Shannon Diversity (-m shannon) (3). The Jensen-Shannon divergence was calculated using the dist_mat (metric='JS') command in pysurvey (https://bitbucket.org/yonatanf/pysurvey, https://github.com/swo/pysurvey).

**SText 2. Identification of predictive features in the microbiome**
All OTU abundance data were analyzed with a log transformation on the OUT abundances by including the parameter -x_log. All metadata categories were analyzed without transformation. The analysis was run using three forests by including the parameter -n 3. The data was shuffled 100 times for p-value calculations by including the parameter -s 100. Class balancing was performed via downsampling using the parameter -b downsample. All other parameters were default parameters.

**SText 3. Description of tools and parameters utilized to generate phylogenetic trees.**
Aligned sequences used for tree building were trimmed of gaps using trimAL (version 1.2) with a gap threshold of 0.05 so gaps present in 95% or more of the sequences would be removed (-gt 0.05)(4). FastTree (5) was then used to build a phylogenetic tree from the trimmed sequences using nucleotide alignment (-nt), a GTR+CAT model (-gtr), and constraint file created from RDP classifications described above (-constraint). The constraint file was generated using custom a Python script (Python 3.5.2).
Supplementary Figures and Tables

SFigure 1. Chloroplast associated reads differ significantly between high and low graded Line A chickens. A darker tinted background indicates a significant difference within that genetic line between high and low chickens. Significance is defined as a p-value less than or equal to 0.05 using a Mann-Whitney U-test.

<table>
<thead>
<tr>
<th>Genetic Line</th>
<th>Line A</th>
<th>Line B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight Gain</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
</tr>
<tr>
<td>Feed Conversion</td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
</tr>
</tbody>
</table>

Relative Abundance (log)

SFigure 2. Shannon Diversity does not significantly differ between high and low graded chickens in most cases. In this figure we present the alpha diversity (calculated as Shannon diversity) for high and low graded Line A (A) and Line B (B) chickens; as well as Line A (C) and Line B (D) chickens graded using the feed conversion metric. Significance is defined as a p-value less than or equal to 0.05 using a Mann-Whitney U-test.

A. p-value: 0.62
B. p-value: 0.01
C. p-value: 0.07
D. p-value: 0.53
Table 1. Classification model statistics

<table>
<thead>
<tr>
<th>Grading:</th>
<th>Feed conversion</th>
<th>Weight gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic Line:</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>AUC</td>
<td>0.85</td>
<td>0.76</td>
</tr>
<tr>
<td>Fisher p-value</td>
<td>0.041</td>
<td>0.003</td>
</tr>
<tr>
<td>OOB error rate (%)</td>
<td>20</td>
<td>11.76</td>
</tr>
</tbody>
</table>

References