Comparison of the Fecal Microbiota of Horses Before and After Treatment for Parasitic Helminths: Massively Parallel Sequencing of the V4 Region of the 16S Ribosomal RNA Gene

Rachael Sirois

Submitted to the Department of Biological Sciences of Smith College
in partial fulfillment of the requirements for the degree of
Master of Science

Steven A. Williams, Master's Thesis Advisor

June 21, 2013

TABLE OF CONTENTS

ACKNOWLEDGMENTS....3

LIST OF FIGURES....4

LIST OF TABLES....5

ABSTRACT....6

INTRODUCTION....8
- The Microbiome of the Gastrointestinal Tract....8
- Equine Helminths and Current Antihelminthic Treatments....12
- The "Hygiene Hypothesis"....15
- Project Focus and Research Questions....17
- How I Will Address these Questions....18

EXPERIMENTAL METHODS....26
- Sample Collection....26
- Illumina MiSeq Genomic DNA Extraction....28
- 16S v4 PCR Amplification....30
- Illumina MiSeq Preparation and Sequencing....34
- Data Analysis....40

RESULTS....45
- 16S v4 Amplicon PCR....45
- 16S v4 Amplicon Purification....46
- Sequencing of Positive Controls....50
- MiSeq Primary Real-Time Analysis....52
- MiSeq Reporter Secondary Analysis....56
- QIIME Analysis....59

DISCUSSION....63
- 16S v4 Amplification....63
- MiSeq Run Quality....64
- Assessing Microbial Diversity....66

CONCLUSIONS....68

FUTURE DIRECTIONS....69

REFERENCES....71

APPENDICES....74

Acknowledgements:

I want to sincerely thank Steven Williams for giving me the chance to work on such a challenging project that has allowed me to learn so much. You were always there with advice and positive encouragement. You have afforded me the opportunity to really expand my knowledge base while working with such a terrific group of people.

Robert Merritt: For agreeing to be my third reader and being so patient as I continued to work on and finish my thesis. Your feedback and suggestions mean a lot.

Lou Ann Bierwert: Beginning to understand how the MiSeq even works was an overwhelming challenge and I am so thankful that you were there to bounce ideas off of from start to finish. Thank you for helping me and agreeing to be my third reader on such short notice.

Alissa Nolan: I am so incredibly thankful for your quick replies to my seemingly endless questions about library preparation and data quality while using the MiSeq.

Weam Zaky and Caroline Decker: You have truly been wonderful friends during my two years in this program and I know for much longer after I graduate. You both were so incredibly helpful and encouraging throughout the course of this project.

Lori Saunders: for always being so willing to help and teach with encouraging words. I learned so much during my first semester in your molecular biology lab and so many of the concepts you taught were incorporated into my own research.

Sue Haynes: For always being there to help me work through whatever problem may come my way. The lab would not be the same without you and everything you do.

Laura Katz and Jean-David Grattepanche: For introducing me to QIIME and helping me work through the program.

Ivy Mead: For being the best lab partner I could have asked for during my first year in this lab. I wish you the best of luck in all you do!

Jan Williams: For being so helpful throughout the project and a big thank you for the contributions by your own group of talented students.

SAW lab: For the encouragement and laughs. I know you will all go on to live exciting lives!

Breana Latty: For being the best roommate ever. My time here would not have been the same without you. You are amazing and will have a wonderful (and sparkling) future!

My family and friends: For your love and tireless encouragement. As I look forward to the next chapter in my life I know I may do so with your full support. **I love you.**

To all of you, I am forever grateful.

List of Figures:

Figure 1: Diagram illustrating the experimental workflow……...21
Figure 2: Diagram of the 16S rRNA gene in *E. coli*……..23
Figure 3: Schematic representation of the 16S rRNA gene……...24
Figure 4: Illumina Inc. sequencing-by-synthesis technology……25
Figure 5: Screenshot example of a "sample sheet" .csv file……..39
Figure 6. Flow chart illustrating primary and secondary analysis on the Illumina MiSeq……...41
Figure 7: Screenshot of read 1 mapping file for use in QIIME software……..42
Figure 8: Scan of unpurified 16S V4 amplicons for Fritz 8/25/11……....48
Figure 9: Scan of purified 16S V4 amplicons for Fritz 8/25/11……49
Figure 10: Screenshot of positive control BLAST result……...……51
Figure 11: Chart depicting % >Q30 by cycle……....54
Figure 12: Chart depicting base intensities by cycle…….55
Figure 13: Clusters graph for Metagenomics Run # 1……...56
Figure 14: Metagenomic pie chart for Fritz 8/25/11……..57
Figure 15: Metagenomics pie charts for Depp 8/25/11…….57
Figure 16: Metagenomics pie charts for Depp 12/3/11…….57
Figure 17: Histogram of sequence lengths and numbers.....60
Figure A1: An illustration depicting the life cycle of small strongyles…….74
Figure A2: 16S Amplicon PCR Forward Primer Sequence……...76
Figure A3: 16S Amplicon PCR Reverse Primer Sequences……..77
Figure A4: Photograph of amplified 16S product on 1.5 % agarose gel……...78
Figure A5: Agencourt AMPure XP Bead system overview……..79
Figure A6: Experimental photograph of Agencourt AMPure XP Bead system……80
Figure A7: 16S amplicon sequencing primers……...81
Figure A8: Taxonomy summary bar chart at the phylum level…….84
Figure A9: Taxonomy summary bar chart at the class level…….85
Figure A10: Taxonomy summary bar chart at the order level……...87
Figure A11 Taxonomy summary bar chart at the family level……..89
Figure A12: Taxonomy summary bar chart at the genus level……..93

List of Tables:

Table 1: General information on horses involved in study……26
Table 2: Horse medication dosage and fecal egg count before and after treatment……..27
Table 3: Reverse primer assignments for experimental samples……...30
Table 4: 25 µl master mix for the 16S V4 amplicon PCR reactions……...……..30
Table 5: PCR cycling conditions for the 16S V4 amplicon PCR reactions……...31
Table 6: 10 µl PCR mixture for sequencing the 16S V4 amplicon positive controls……....33
Table 7: PCR cycling conditions for sequencing of 16S V4 amplicon positive controls……..34
Table 8: Dilution of samples to 10nM concentration……35
Table 9: Dilution of samples to 2nM concentration……..35
Table 10: List of samples and corresponding 12 bp Golay barcodes……38
Table 11: Qubit results for purified 16S V4 amplicons……47
Table 12: Comparison of 260/280 and 260/230 ratios before and after purification…….47
Table 13: 16S metagenomics run summary and sequencing metrics per read……..…53
Table 14: Pyrosequencing metrics for demultiplexed samples read 1 and read 2…….59
Table 15: Classification of equine fecal bacteria before and after treatment……62
Table A1: Reagent volumes of 5 Prime HotMasterMix (2.5x)…….75

Abstract:

The advent of high-throughput genomic sequencing methods has become instrumental in the ability to study biological systems that are normally difficult to investigate using traditional culture-based techniques. Metagenomic analysis involves sequencing all DNA from a particular environment in order to reveal microbial diversity in a way that shotgun sequencing has not been able to do. The advantages of metagenomics are illustrated by this project that aimed to better characterize the equine gastrointestinal tract. Diseases affecting the gastrointestinal system are the main cause of mortality in horses and yet, our understanding of bacterial diversity and abundance is quite limited. To address these concerns, the project asks two questions: 1) which groups and species of bacteria are present in this environment before and after treatment with antihelminthics and in what abundance, and 2) does infection with parasitic helminths followed by antihelminthic treatment cause a shift in the composition of the equine GI tract microbiome?

A metagenomic study was conducted in order to address the above two questions. Genomic DNA was extracted from equine fecal samples (n=6) and assessed for purity and quantity. PCR was performed using a modified protocol from Caporaso *et al*. (2012) to amplify the v4 variable region of the 16S ribosomal RNA gene, which has been widely used to classify bacterial representatives inhabiting different environmental niches (Wang and Qian, 2009). PCR amplification of DNA samples proved to be difficult due to residual inhibitors and as a result, amplicons were generated for only three of the four study horses. Each horse had two representative samples: August 25, 2011 and December 3, 2011, which correspond to pre-treatment and post-treatment samples, respectively.

The 16S amplicons were sequenced by the Illumina MiSeq to assess bacterial diversity and abundance. Sequence data were analyzed using the software package, QIIME. The

GreenGenes database was used as reference for classification of sequences into Operational Taxonomic Units. A total of 3,361,963 sequences representing read 1 were classified into 3,528 OTUs. Bar charts were generated for each of the samples to visualize the OTU taxonomical data at the phylum, class, order, family, and genus levels. The two most abundant phyla were Bacteriodetes and Firmicutes. In the 'pre-treatment' samples, Bacteriodetes predominated (42.6%) followed by Firmicutes (27.1%) and Verrucomicrobia (12.7 %). Firmicutes was the most prevalent phylum among the 'post-treatment' samples accounting for 34.6 % of sequences, followed by Bacteriodetes (31.5%) and Verrucomicrobia (21.7%). However, statistical tests must be performed to determine if abundances are significantly different between treatment groups and individual horses.

INTRODUCTION

Horses are regularly exposed to treatments with antihelminthic medications in order to prevent infection by a variety of parasitic helminths. This project aims to investigate the effects of infection by these parasitic helminths and how the drugs used to treat them, namely antihelminthics, may shape the microbiome of the gastrointestinal (GI) tract. The GI tract is home to an intricate and microbe-rich environment that is constantly fluctuating in response to a variety of pressures, including host diet and disease (Fujimura *et al*., 2010). This introduction discusses the use of a metagenomic approach to study the equine GI microbiome in hopes of gaining a better understanding of this complex and relatively uncharacterized system. By studying this environment, we may also be provided with general insight into the relationship of host-microbe interactions that are not limited to the horse, but extend to humans as well.

The Microbiome of the Gastrointestinal Tract

The microbiome (i.e. the totality of microbes, their genetic elements (genomes), and environmental interactions) represents an essential component of the GI tract. Although the main function of the gastrointestinal tract is the conversion of food into usable nutrient components, the GI tract is also an ecosystem in which host cells and microbes interact extensively (Hooper *et al*., 2012). The complex microbial community of the GI tract consists of different groups of microbes, such as archaea, ciliate and flagellate protozoa, anaerobic phycomycete fungi, bacteriophage, and the most widely studied group, bacteria (Zoetendal *et al*., 2004).

The microbial population, particularly bacterial species, helps to maintain a balance within its host, contributing to the prevention of disorders, and forming a barrier against pathogens (Santos *et al*., 2011). Thus, it is not surprising that a significant amount of research has

demonstrated that disturbances in the balance of these microbial communities can disrupt intestinal homeostasis and are implicated in the pathogenesis of a number of gastrointestinal dysfunctions (Antonopoulos *et al*., 2009). For example, Milinovich *et al*. (2008) have reported that carbohydrate-induced laminitis in horses is correlated with drastic shifts in the composition of the hindgut microbiome, from a predominantly Gram-negative population to one dominated by Gram-positive bacteria. Similarly, gastric bacteria, particularly the *Heliobacter spp*., have been suggested as the cause for glandular stomach lesions in many animal species, including horses and humans (Husted *et al*., 2010).

The Microbiome of the Equine Gastrointestinal Tract

Although the composition and activities of the microbiome have been suggested to have a substantial effect on the health, growth, development, and performance of animals, the microbial community of the equine hindgut has received relatively little attention. The herbivorous diet of the horse requires extensive microbial fermentation for complete digestion. This demand is illustrated by the animal's physiological anatomy; the horse has a combination of a large cecum and an even larger colon where fermentation and absorption occur (Mackie and Wilkins, 1988). Understanding the effects of diet change, stress, disease, or drug treatment on the hindgut microbiome requires a basic knowledge of bacterial composition and abundance. Current information regarding the diversity and abundance of the microbes present in the equine gastrointestinal tract is still quite limited and this lack of a more comprehensive understanding could be the result of several factors. These factors include studies with narrow research interests, a focus on the human microbiome, and culture-based limitations.

Most studies on the gastrointestinal tract to date have been confined to specific research questions, as opposed to a comprehensive study of the gastrointestinal tract. There are few

reports that attempt to conduct a complete analysis of the equine microbiome, and those that do are often restricted to particular functional groups or health concerns. For example, Al Jassim *et al*. (2005) cultured 72 lactic acid producing bacterial isolates from the equine GI tract to study the progression of lactic acidosis to laminitis, a condition characterized by inflammation and degeneration of the lamellar membrane of the foot. The results from DNA analysis indicated that the majority of the isolates were closely related to species within the genera, *Lactobacillus* and *Mitsuokella*, including *L. salivarius, L. mucosae,* and *M. jalaudinii*.

Another shortcoming is the fact that most research aimed at studying digesta movement and microbial activity has long been conducted on humans and ruminant species, such as the pig (Clemens, *et al*., 1975). Current research projects examining the human microbiome, such as the National Institute of Health's (NIH) Human Microbiome Project, allow for the direct study of human disease, nutrition, and health. Additionally, recognition of numerous similarities between pigs and humans prompted extensive studies into the normal and abnormal nutritional and physiological states of the swine (Clemens, et al., 1975). In sharp contrast to these numerous studies, our understanding of the equine GI tract is quite narrow. Limited studies using new sequencing technologies in horses are available – a clear shortcoming since disease affecting the GI tract is the main cause of mortality in this species (Costa *et al*., 2012).

Another factor that has limited our understanding of this environment is that early microbiome studies traditionally utilized culture-based methods for microbe characterization. These methods rely on identification of organisms by phenotype and have several drawbacks. First, they can be used only for organisms that can be cultivated *in vitro*; many of the organisms that populate the gastrointestinal tract require strict anaerobic growth conditions and cannot be cultured under normal aerobic conditions. Second, culture-based methods are laborious, time

consuming, and may recover only a small fraction of the total microbial diversity present within the gut (Daly et al., 2001). The more recent application of molecular approaches to studying gastrointestinal microflora suggests that culture-based methods only allow for a superficial assessment of the components of the microbiome, which is a significant limitation, as a large component of the microbiome is thought to consist of unknown or unculturable microorganisms (Salzman et al., 2002).

The advent of high-throughput genomic methods has become instrumental in the ability to study biological systems that are normally difficult or impossible to investigate using traditional culture-based techniques. Metagenomic analysis involves sequencing all DNA from a particular environment in order to reveal microbial diversity in a way that shotgun sequencing (cutting DNA strands into manageable lengths and cloning them) has not been able to do (Handelsman, 2004). Metagenomics has revolutionized the way in which many scientific fields, including microbiology and molecular biology, operate (Kobayashi *et al*., 2004). Metagenomics has been used in several large-scale studies, such as the survey of deep-sea methane vents (Pernthaler, et al., 2008) and in the human distal gut microbiome study (Gill, et al., 2006). The accessibility and power of these molecular advances have put us in a position to better understand complex environments and the effects of disease and other pressures on those systems.

For example, a study by Costa *et al*. (2012) aimed to illustrate the changes in microbiome composition between two groups of horses, those in good health and those suffering from colitis (a condition characterized by inflammation of the colon). The species richness they reported indicates the complexity of the equine intestinal microbiome. Specifically, Costa *et al*. (2012) note the predominance of the bacterial phylum, Firmicutes, and demonstrate its importance in the

maintenance of a healthy gastrointestinal tract. The marked differences in the microbiome between healthy horses and horses with colitis indicate that colitis may be a disease of gut dysbiosis (or microbial imbalance) rather than one that occurs simply through the overgrowth of an individual pathogen. Additionally, the presence of the bacterial genus, *Fusobacteria*, in the colitis group but not in healthy horses, merits further investigation, as its role in equine colitis (See Appendix B for glossary of terms) is not currently known. By performing these types of studies, we can begin to understand how the microbial balance may shift as a result of disease. In this particular study, we propose to investigate the horse GI microbiome before and after treatment with an antihelminthic drug.

Equine Helminths and Current Antihelminthic Treatments

Horses worldwide are exposed to a complex assortment of intestinal helminths. Helminths (See Appendix B for glossary of terms) are parasitic worms that can cause a variety of infectious diseases (Matthews, 2008). The phylum, *Nematoda* (Order: Strongylidae), is the most important group of equine parasites because of its high prevalence and pathogenicity (Matthews, 2011). The nematodes (See Appendix B for glossary of terms) that primarily cause GI tract symptoms fall into two families: *Strongylinae* (large strongyles) and *Cyathostominae* (small strongyles). Small strongyles are currently the most common nematode species affecting equines worldwide and are comprised of over 50 species (Matthews, 2011). Despite a large range of species, only 12 species (*Cyathostomum catinatum*, *Cyathostomum pateratum*, *Coronocyclus coronatus*, *Coronocyclus labiatus*, *Coronocyclus labratus*, *Cylicocyclus nassatus*, *Cylicocyclus leptostomus*, *Cylicocyclus insigne*, *Cylicostephanus longibursatus*, *Cylicostephanus goldi*, *Cylicostephanus calicatus*, and *Cylicostephanus minutus*) are characterized as highly

prevalent and account for approximately 99% of the total cyathostome burden worldwide (Kornás *et al*., 2009).

The lifecycle of small strongyles is generally the same for all species. The lifecycle is direct and infection occurs via ingestion of the infective third stage larvae (See Appendix A, Figure 1). Ingested L3 stage larvae enter the mucosa of the large intestine and develop through a number of stages (L3 → L5), finally maturing to adult male and female worms. The eggs are excreted into the feces, which hatch and develop from L1 to L2 and then to the infective L3 stage (Matthews, 2011). Since a portion of the worm's lifecycle is spent in pastures, it is likely that essentially all grazing horses experience some level of GI tract nematode parasitism (Matthews, 2008). Additionally, co-infection with multiple species is not unusual and an individual horse may harbor upwards of 10 common species (Kornás *et al*., 2009).

Most small strongyle infections do not normally present as overt clinical disease but it is a general rule that the higher the level of infection, the more likely it is that an animal will develop clinical signs (Matthews, 2011). In those animals that do experience illness, it may manifest as colic (See Appendix B for glossary of terms), severe weight loss, decreased rates of growth, rough hair coat, and potentially fatal colitis with diarrhea (Kaplan, 2002). Additional complications can occur if the infective larvae become encysted within the lining of the large intestine causing a serious condition known as larval cyathostomiasis. The larvae remain dormant until they are signaled to reactivate and emerge in large numbers from the colonic mucosa, resulting in severe tissue damage, inflammation, sudden weight loss, and subcutaneous edema (Kaplan *et al*., 2004). This condition is more common in animals less than five-years-old and is fatal in up to 50% of cases despite treatment with antihelminthic medications (Matthews, 2011).

Currently, there are three classes of antihelminthic drugs licensed for use against cyathostome infection: macrocyclic lactones (MLs), tetrahydropyrimidines (THPs), and benzimidazoles (BZs). Before the regular use of antihelminthic drugs, the large strongyle parasite, *Strongylus vulgaris*, was responsible for the majority of helminth infections, with prevalence rates estimated at 80-100% (Bracken *et al*., 2012). Years of rigorous antihelminthic treatment have resulted in a substantial decrease in the prevalence of this large strongyle. However, recent studies have revealed that *S*. *vulgaris* remains a serious threat in areas where deworming is infrequent (Bracken *et al*., 2012).

Although *S*. *vulgaris* infections declined in the equine population as a result of the frequent antihelminthic dosing, prevalence of small strongyle infections has increased. When first introduced, all classes of drugs exhibited good to excellent efficacy against cyathostomes. However, reports of drug-resistant cyathostomes are becoming increasingly more common (Kaplan *et al*., 2004). It has recently been reported that single cyathostome populations have been identified that exhibit resistance to all three classes of drugs (Matthews, 2011). Drug resistance may be due to a variety of factors. Firstly, frequent acquisition of resistance could result from excessive and unnecessary dosing of horses. The concept of strategic parasite control for horses was introduced over 40 years ago in a program known as the interval dose system, whereby horses were treated every six to eight weeks to prevent parasite maturation (Kaplan, 2002). Although this strategy was successful in terms of a marked decrease in *S*. *vulgaris* infections, it has inadvertently resulted in the selection of drug-resistant cyathostomes. All too often, horses are treated with antihelminthics as a prophylactic measure without considering the actual level of parasite burden.

In addition to general overuse, imprecise dosing due to a lack of accurate tests for measurement of parasite load is also a concern. The Fecal Egg Count (FEC) is the principal diagnostic tool for equine intestinal parasite load and is often the sole determinant of the host's parasite load (Donecker *et al*., 2007). The FEC is a quantitative fecal analysis that determines the specific number of parasite eggs per gram (EPG) of feces (Cornell University, 2012). Although it is sometimes referred to as the "gold standard", the test does have limitations. There is no universal FEC threshold for initiating antihelminthic treatment and counts were proposed when large strongyles were prevalent and have not been modified since the rise in small strongyle prevalence (Donecker *et al*., 2007). Promoting further ineffectual measurement is the fact that the FEC does not take into consideration all parasite stages, for example, the encysted larval stages. As such, horses may be receiving treatment at a much lower dose than is required to efficiently control the infection.

The "Hygiene Hypothesis"

Despite causing a variety of clinical disorders, bacteria and other microbes are an essential and significant component of the equine gastrointestinal tract - this is true for all mammals (Hooper *et al*., 2012). In the lower intestine, particularly, these microorganisms reach extraordinary densities and function to degrade plant polysaccharides and other metabolites (Eckburg *et al*., 2005). However, over millions of years of coevolution, invaluable interconnections between the physiologies of microbial communities and their hosts have formed that extend beyond metabolic functions (Hooper *et al*., 2012). The mammalian immune system plays an essential role in maintaining homeostasis with resident microbial communities. At the same time, resident bacteria profoundly shape mammalian immunity (Hooper *et al*., 2012). The components of this system – the host cells, resident bacteria, and parasitic worms – can each be

thought of as individual pieces of a puzzle; if one of the pieces is missing, the final product is incomplete. Such an imbalance could have serious health implications by means of immune system dysfunction (Costa *et al*., 2012).

In developed countries, such as the United States and those in Western Europe, the escalating incidence of human autoimmune and inflammatory disease is a major public health concern. The 'Hygiene Hypothesis' infers that many of our so-called modern illnesses, such as asthma and Crohn's disease, have increased exponentially since industrialization because improved sanitation practices have made the body's ecology too sterile (Berglund, 2012). The hypothesis speculates that those individuals who have not been exposed to parasites and other microorganisms in sufficient quantity or early enough to properly prime the immune system, may be more likely to develop autoimmune diseases (Yazdanbakhsh *et al*., 2002).

The existence of this relationship has translated to clinical investigations aimed at the safe and controlled reintroduction of helminthic exposure to patients suffering from autoimmune diseases (so-called 'helminthic therapy') in an effort to mitigate the inflammatory response (Wolff *et al*., 2012). The company, Autoimmune Therapies, is one such corporation that actively employs the putative healing capabilities of helminthic therapy. The company, founded in 2007 by Jasper Lawrence, infects its clients with a minor helminth, for example hookworm (*Necator americanus*), in order to treat a variety of autoimmune disorders, including multiple sclerosis. This "treatment" costs a client about $3,000 USD (Autoimmune Therapies, 2009). Since the founding of Autoimmune Therapies, other companies dedicated to the study of helminthic therapy have been established around the globe, including Ovamed GmbH in Barsbüttel, Germany and Worm Therapy in Tijuana, Mexico.

Project Focus and Research Questions

As previously discussed, microorganisms, including parasitic worms and bacteria, interact with the immune system to ensure its proper development and priming. Microorganisms have co-evolved with host animals over an extensive period of time and this suggests a symbiotic, rather than commensal relationship between microbe and host (Berglund, 2012). It is already known that microorganisms perform a range of useful functions, including the fermentation of unused energy substrates, a role in mucosal defense, and the production of vitamins for the host (such as biotin and vitamin K) and hormones to direct fat storage (Salzman et al., 2002). The removal of gut parasites could change the environment and makeup of the gastrointestinal tract, perhaps altering microbiome composition and thus, immune system function.

Diseases affecting the gastrointestinal system are the main cause of mortality in horses (Costa *et al*., 2012). Despite the clear importance of the microbiome, our understanding of bacterial diversity, abundance, and fluctuations as a result of different pressures is to date, quite limited. To address these concerns, the aims of this project are to better understand the composition of the equine intestinal microbiome as a whole, as well as study the possible effects of helminths and anti-helminthic medications on this environment. By tracking bacterial populations in the GI tract, we can learn not only about its general composition but also monitor fluctuations that may occur as a consequence of treatment with antihelminthic medications. Additionally, this project aims to compare the microbiome of different horses to see if there are significant differences in microbial populations between horses. This is likely, as it has been shown that microbiome structures differ significantly between humans (Caporaso *et al*., 2011).

To accomplish this, the project asks two main questions:

1) Which groups and species of bacteria are present in this environment before and after treatment with antihelminthics and in what abundance?
2) Does infection with parasitic helminths followed by antihelminthic treatment cause a shift in the composition of the equine GI tract microbiome?

Although some of the bacterial species of the equine GI tract are known, a comprehensive metagenomic study of this size will identify many previously uncharacterized species that are unculturable under normal aerobic conditions.

How I plan to address these questions

A metagenomic study will be conducted in order to address the above two questions concerning the equine gastrointestinal tract. Metagenomics is a rapidly growing field of research that is particularly useful in the study of microbe diversity, function, cooperation, and evolution in various environments such as soil, water, or the digestive tract of animals (Huson *et al*., 2009). The key approach in metagenomics is large-scale sequencing of environmental samples, involving the direct isolation of genomic DNA from the environment, thus allowing it to circumvent traditional organism isolation and culturing methods (Handelsman, 2004).

For years, shotgun Sanger sequencing was the main technology used in metagenomics, but new sequencing technologies and the substantial reduction in the cost of sequencing have boosted the development of this field (Gori *et al*., 2011), allowing for the study of increasingly complex environmental sample data sets. This, in turn, has increased development of various bioinformatics tools for the analysis and comparison of these complicated data sets (Simon and Daniel, 2011). Bioinformatics is an important factor to consider when conducting a

metagenomic study, because unless the data is analyzed properly, it may be very difficult to extract useful and relevant conclusions (Tucker *et al*., 2009).

The environmental samples from which the 'metagenome' will be extracted are equine fecal samples, representing gastrointestinal tract content. The horse will serve as the research model for this project and is a suitable choice for several reasons. Horses are a nonruminant species, and so digestive function in the stomach and small intestine occurs in a similar fashion as other monogastric animals, such as humans. Additionally, like humans, or any other host organism for that matter, horses have evolved alongside their parasites for a long time and, relative to that relationship, treatment with antihelminthic medications is fairly recent. As far as actual treatments, the horses in this study have been treated using medications that are also common in human medicine, including Fenbendazole and Ivermectin. By using similar drugs, there is the possibility of future studies to look at other aspects of this system, such as resistance to these common drugs – a concern not limited to equine health.

Equine metagenomic DNA was previously analyzed in our laboratory using a different molecular technique to assess bacterial diversity and their fluctuation in the equine GI tract. This first approach was a series of bacterial fingerprinting tests known as Automated Ribosomal Intergenic Spacer analysis (ARISA). The ARISA is commonly used to provide a general overview on the structure of the bacterial communities within the environment to be studied (Cardinale *et al*., 2004). ARISA is a PCR fingerprinting technique that uses a universal primer set designed to amplify the 16S-23S rRNA intergenic transcribed spacers (Popa *et al*., 2009). One of the primers is affixed with a fluorescent tag, which allows for fragment analysis, producing spectral data representing the bacterial community. Spectral data was analyzed using the computer-based program, GeneMapper®. According to Cardinale *et al*. (2004), each peak

matches PCR fragments in the samples that likely correspond to an individual species or strain of bacteria in that environment. ARISA is particularly useful for the comparison of a large number of samples, as it is very cost-efficient. Furthermore, the repeatability of the ARISA technique has been previously demonstrated, so comparisons across different times and equine samples are valid (Cardinale *et al*., 2004). However, one considerable limitation to the ARISA is that the spectra alone are not sufficient for positive identification of the bacterial species in each peak.

Due to this limitation, massively parallel sequencing was performed, the results of which are the focus of this thesis. Massively parallel sequencing, commonly referred to as Next-Gen Sequencing (NGS), has advanced significantly over the past ten years and has helped make sequencing and analysis of large genomes faster and more cost-effective (Tucker *et al*., 2009). Massively parallel sequencing allows for the simultaneous sequencing of a large number of DNA samples, which lends well to a metagenomic study (Tucker *et al*., 2009). In its basic concept, the bases of small fragments of DNA are sequentially identified from signals that are emitted, as each fragment is re-synthesized from a DNA template strand (Illumina, Inc®, 2012).

The sequencing instrument we will be using for this part of the project is an Illumina MiSeq Personal Sequencer. The MiSeq system is a fully integrated sequencer that simplifies sample preparation through sequencing, automated data analysis, and storage of data in the BaseSpace cloud (Illumina, Inc®, 2012) (See Figure 1 for experimental workflow). 16S ribosomal DNA (rDNA) amplicons will be generated through PCR amplification and subsequently sequenced using the MiSeq technology to assess bacterial diversity and abundance.

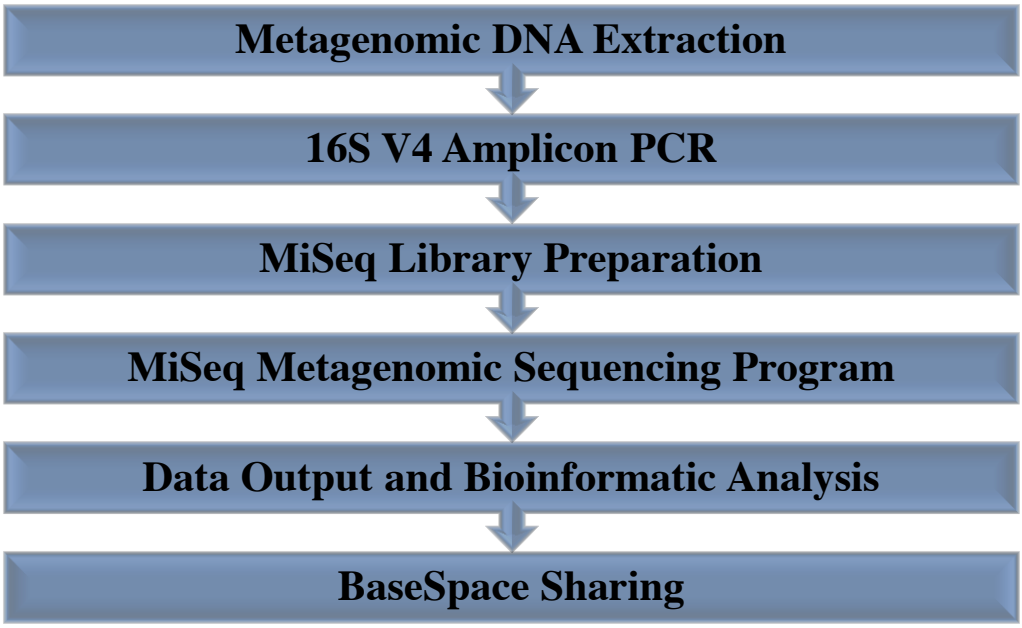

Figure 1. Diagram illustrating the experimental workflow.

Bacterial 16S rDNA amplicons have been widely used to classify bacterial representatives inhabiting different environmental niches (Wang and Qian, 2009) (Figure 2). The 23S/16S rRNA represents more than 80% of the total bacterial transcriptome and consists of interspersed conserved and variable regions, making it suitable for PCR amplification and sequencing (Figure 3). The primers, based on the protocol by Caporaso *et al*. (2012) will be designed to hybridize to the conserved regions of the 16S rRNA gene, allowing for amplification and sequencing of a particular variable (V) region. Previous studies suggest that the V4 variable region will yield optimal community clustering (See Appendix B for glossary of terms) (Caporaso *et al*., 2012, Supplementary Information).

The V4 region will be amplified with region-specific primers that include adapters to hybridize the sequences to the Illumina flow cell and a 12 base pair (bp) barcode (reverse primer) sequence to allow samples to be pooled prior to sequencing (Figure 4a). During sequencing, a

third indexing primer will sequence the 12 bp barcode to allow for sample identification (Figure 4b). Illumina sequencing uses a "reverse terminator-based method" by which the templates are sequenced using a four-color DNA sequencing-by-synthesis technology that employs reversible terminator dNTPs with removable fluorescent dyes. High-sensitivity fluorescence detection is subsequently achieved using laser excitation and high speed scanning optics (Figure 4c) (Illumina, Inc®, 2012).

Using the MiSeq system, we hope to gain a much more in-depth understanding of the equine GI tract and be able to track possible fluctuations in response to antihelminthic treatment. As we begin to sift through the massive amount of data that will be provided, the aims for this part of the project will be to identify bacterial diversity by group and species and obtain an estimate of their relative abundance based on the 16S V4 sequences. Of course, with any research approach, there will be disadvantages and limitations. Massively parallel sequencing is much more expensive than traditional approaches and will produce a vast output of data that needs to be properly organized and analyzed to be useful. However, with the proper bioinformatic tools, this massively parallel metagenomics approach should prove to be an efficient method to analyze the equine GI microbiome metagenome.

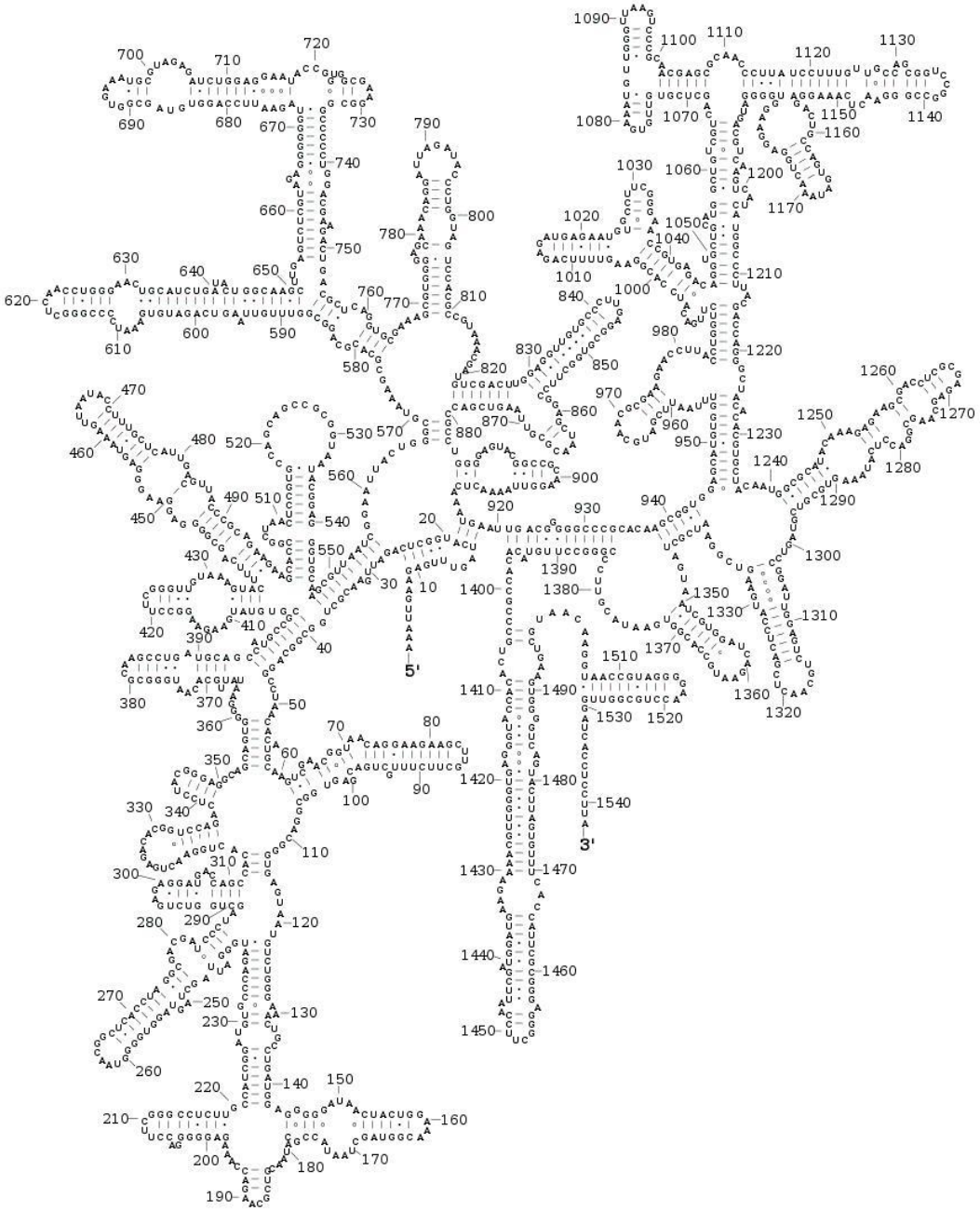

Figure 2. Diagram of *Escherichia coli* 16S (SSU) 5' rRNA. The SSU of the ribosome (the site of protein synthesis in all living cells) contains the 16S rRNA, which is transcribed from a ribosomal operon in the bacterial genome. The rRNA folds into an intricate 3D structure and is incorporated into a protein-RNA complex that is critical for ribosomal function. The 16S gene is a multi-copy gene in most bacteria (SSU = small subunit) (http://rna.ucsc.edu/).

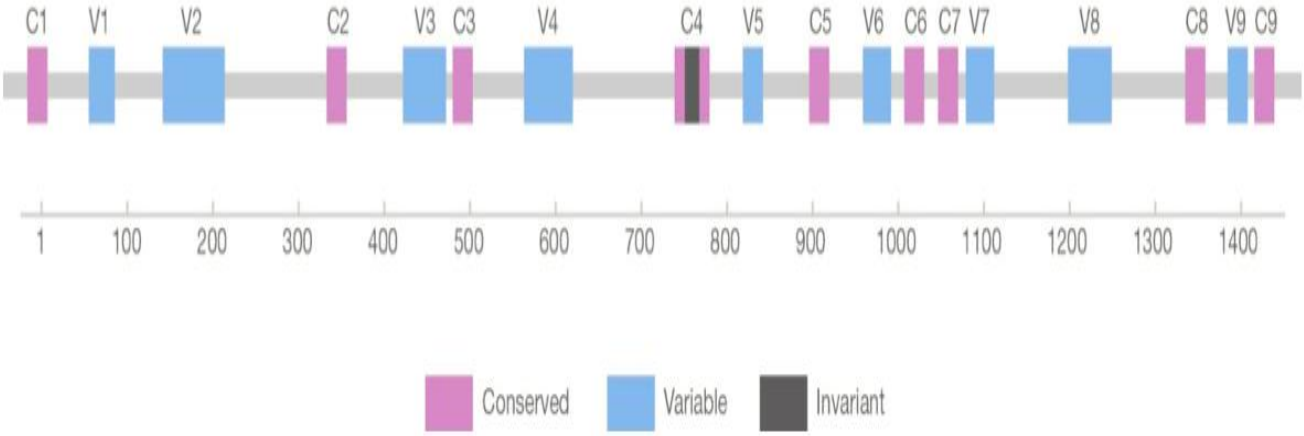

Figure 3. Schematic representation of the 16S rRNA gene. Location of variable (blue) and conserved (purple) regions in a canonical bacterial 16S rRNA. The grey region is invariant in all bacteria (Illumina, Inc.®, 2012).

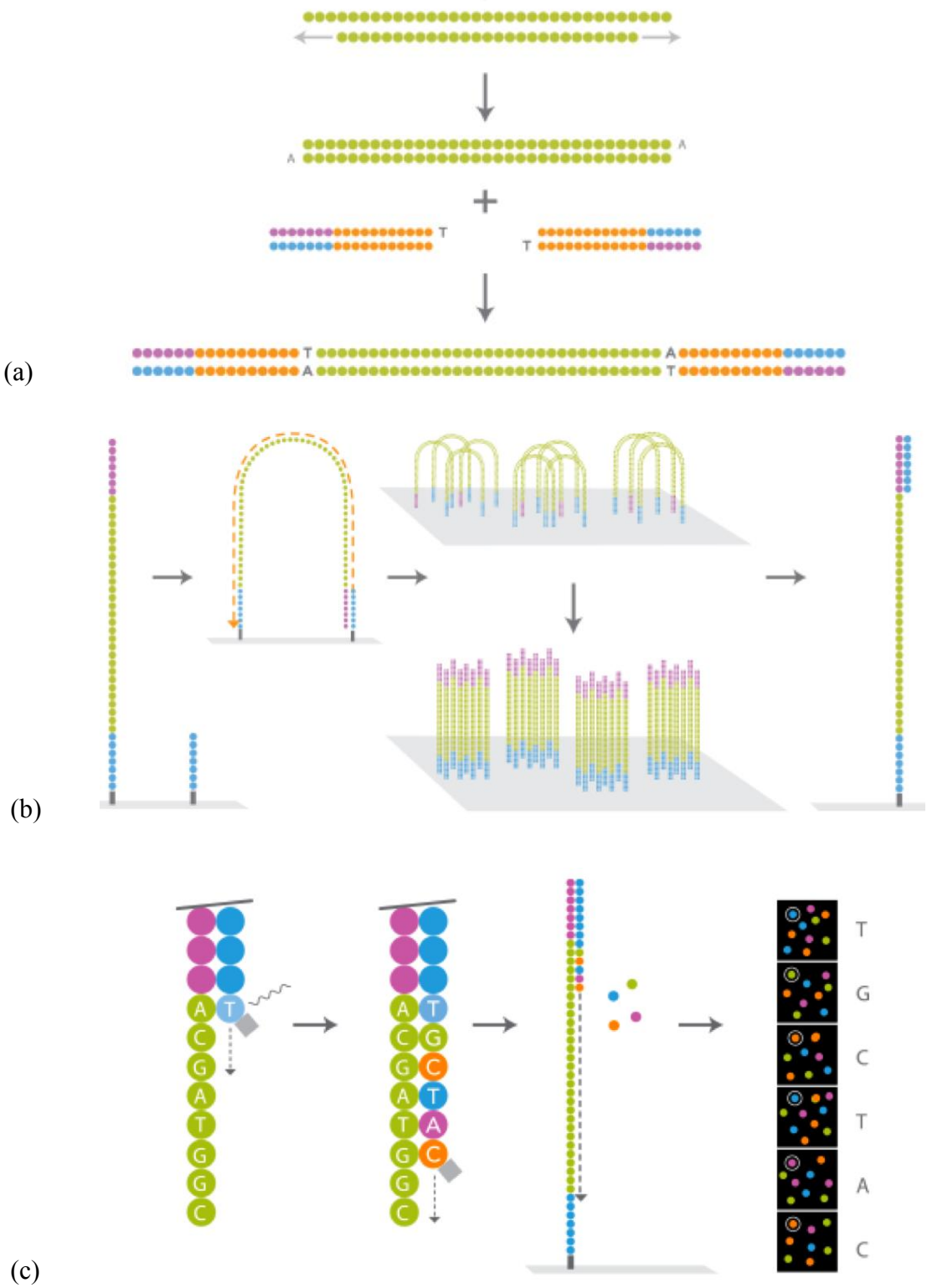

Figure 4. Illumina sequencing technology (a) Sample library preparation (b) Cluster generation (c) Sequencing-by-synthesis technology (Illumina Sequencing Technology) (Illumina, Inc®, 2012).

EXPERIMENTAL METHODS

Sample Collection

Environmental Data

Samples to be used in this study were collected from four horses at Coyote Run Farm in North Hatfield, Massachusetts (Table 1). These animals are readily available to us and what makes this set of animals even more appropriate is that they come from a relatively well-controlled environment. The horses are cared for by the same person, they are exposed to the same living conditions, they are fed on a similar diet and schedule, they receive the same veterinary care, and are regularly dosed with the same antihelminthic treatments.

Table 1. General information on each of the four horses used in the research project, including breed, sex, and age.

	Name	**Breed**	**Sex**	**Year of Birth**
Horse # 1	Desiderata (Desi)	Warmblood	Female	2000
Horse # 2	Syrtaki (Taki)	Warmblood	Gelding	2007
Horse # 3	Fritz William * (Fritz)	Hanoverian	Male	2010
Horse # 4	Despriano * (Depp)	Hanoverian	Male	2010

- Note that Despriano and Fritz William were born in Germany and were moved to Coyote Run Farm six months following birth.

Fecal samples were collected from the ground as soon as possible after dropping. Samples were sealed in a clean plastic resealable bag and kept in a cooler until transported to the laboratory at Smith College. Once at the laboratory, samples were stored at -20°C. Samples were collected on August 25, 2011 and December 3, 2011 and represent pre-treatment samples and post-treatment samples, respectively.

Antihelminthic Treatment

Each of the horses was treated with the standard dosing regimen of the antihelminthic, Fenbendazole, over the course of five days beginning on the 23 November 2011. The horses' treatment was further supplemented with a dose of Ivermectin on the fifth and final day. Following treatment, samples were collected daily at approximately the same time from December 3, 2011 until December 6, 2011. These samples mark the Post-treatment samples in the study and were selected for use based on fecal egg count results. Janet Williams, of Elms College in Chicopee, Massachusetts, performed the fecal egg counts (Table 2).

Table 2. Individual horse medication dosage based on weight of the horse and results from the fecal egg counts before and after treatment with Fenbendazole.

Horse Treated	Date Treated	Medication Weight Dosage (Fenbendazole)	Medication Weight Dosage (Ivermectin)	Fecal Egg Count for Pre-treatment Sample (eggs/gram)	Fecal Egg Count for Post-treatment Sample (eggs/gram)
Desi	11/23/11	*	1200 lb	0	0
Taki	11/23/11	1100 lb	1000 lb	> 500	0
Fritz	11/23/11	1000 lb	1000 lb	> 500	0
Depp	11/23/11	900 lb	900 lb	> 500	0

- Note that Horse #1 was not treated with Fenbendazole due to a dosage protocol regarding non-infected horses.

Horses #2, #3, and #4 were infected with parasites ("small strongyle-like eggs" were observed). A value greater than 500 eggs/gram is indicative of a high load of parasite infection in the horse. Following treatment, no evidence of parasite infection was reported. Indication of parasite

infection was not observed in Horse #1 either before or after treatment and thus served as a control horse for the study.

Illumina MiSeq Genomic DNA Extraction

Metagenomic DNA was extracted from frozen horse fecal samples using the QIAamp DNA Stool Mini Kit for DNA Purification from Stool Samples (Qiagen, Catalogue #51504), following the kit protocol for "Isolation of DNA from Stool for Pathogen Detection". Approximately 200 mg of each frozen fecal sample was suspended in Buffer ASL and vortexed continuously for one minute (note that the exact weight of the prepared sample was difficult to accurately assess due to water weight as a result of the frozen storage method). The samples were heated at 90°C for five minutes until a homogenous solution had been obtained (note that this first heating step was increased from 70°C, as suggested by the manufacturer, in order to better lyse cell types that are difficult to lyse, including some bacteria and parasites). The samples were centrifuged for one minute to pellet stool particles [note that all centrifugation steps were carried out at room temperature (15 - 25°C) at 20,000 X g (approximately 14,000 rpm)].

An InhibitEX matrix tablet was added to 1.2 mL of the supernatant and immediately vortexed until the tablet was completely suspended. The samples were allowed to sit at room temperature for one minute to allow adsorption of DNA-damaging inhibitors to the InhibitEX matrix. The samples were centrifuged for three minutes to pellet inhibitors bound to the InhibitEX matrix and the supernatant incubated with 15 µl Proteinase K for 10 minutes at 70°C to allow for digestion of contaminating protein. The samples were purified using equal amounts of ethanol-based wash buffers AW1 and AW2 (note that wash buffer AW2 contains 0.04%

sodium azide and should be appropriately handled as a toxic substance). The genomic DNA samples were eluted in 200 µl Buffer AE (AE buffer is a mixture of 10 mM Tris-Cl and 0.5 mM EDTA; pH 9.0) and stored at -20°C.

To prevent interference by PCR inhibitors, an additional purification step was found to be necessary (Radstrom *et al*., 2004). The inhibitors were extracted using Chelex 100 Resin (Biorad, Catalogue #142-1253). A mixture of 0.5 g Chelex and 10 ml elution buffer AE (Qiagen, Catalogue #51504) was prepared in a 15 ml conical tube. Then, 30 µl of the Chelex mixture was added for every 50 µl of genomic DNA. Samples were gently mixed, briefly centrifuged, and then incubated for 20 minutes at room temperature. Samples were centrifuged at 3,000 RPM for three minutes to pellet the Chelex resin. The supernatant was transferred into a new 1.7 ml microfuge tube, taking care to avoid the Chelex beads. The centrifugation and transfer steps were repeated, and then the samples were stored at -20°C. At this step, all samples were reassessed for quality and quantity using the NanoDrop™ 1000 Spectrophotometer (Thermo-Fisher).

Genomic DNA samples were assessed for purity and quantity prior to their use in downstream applications. DNA concentration and purity was measured using the NanoDrop™ 1000 Spectrophotometer (Thermo-Fisher) and the Qubit® 2.0 Fluorometer (Life Technologies). If Genomic DNA samples showed a 260/280 value below 1.60, the extraction process for that sample was repeated. Samples were visualized on a 1.5% agarose gel in 1X TAE gel running buffer to further assess DNA quality and size. After purification, most samples showed improved 260/280 values and performed better in downstream PCR applications.

16S v4 rRNA PCR Amplification

16S Amplicon PCR Mixture and Cycling Conditions

Amplicon PCR and sequencing reactions were performed using a modified version of the protocol used in Caporaso *et al*. (2012) that was then adapted for use with the Illumina MiSeq platform. In brief, the V4 variable region of the 16S rRNA gene was amplified with region-specific primers that included the Illumina flow cell adapter sequences. The reverse amplification primer also contained a twelve base barcode sequence that supports pooling of up to 2,167 different samples in each lane (See Table 3 for reverse primer assignments) (See Appendix A, Figures 2 and 3 for 16S amplicon PCR primer sequences). The 25 µl PCR reaction mix is based on a modified protocol from Caporaso *et al*. (2012), as shown in Table 4.

Table 3. Reverse primer assignments for experimental samples

Sample	Bar-coded Reverse Primer Assignment
Taki 8/25/11	806rcbc0
Desi 8/25/11	806rcbc1
Fritz 8/25/11	806rcbc2
Depp 8/25/11	806rcbc3
Taki 12/3/11	806rcbc4
Desi 12/3/11	806rcbc5
Fritz 12/3/11	806rcbc6
Depp 12/3/11	806rcbc7

Table 4. The 25 µl master mix for the 16S V4 Amplicon PCR reactions.

PCR Reagent	Volume (µL)
ddH_20 (RT-PCR grade water)	12.0
5 Prime Hot Master Mix (with self-adjusting $MgCl_2$)*	10.0
Primer (Forward) (10µM) *	1.0
Primer (Reverse) (10µM) *	1.0
Template Genomic DNA (concentration)	1.0

- See Appendix A, Table 1 for vendor 5 Prime HotMasterMix components and volumes

All PCR reactions were run on a Veriti® 96 Well Thermal Cycler (Applied Biosystems) using a PCR cycling program based on Caporaso *et al*. (2012) (Table 5). Amplification of predicted PCR products was confirmed using gel electrophoresis. A volume of 7.0 µl of each PCR product was run on a 1.5% agarose gel in 1X TAE running buffer at approximately 75 volts for two hours. A 100 bp DNA ladder was run with the samples to estimate amplicon size (Cat. # N3231L, New England Biolabs). Those samples in which a band of ~400 bp was visualized were then subjected to purification in preparation for pooling and sequencing on the MiSeq (See Appendix, Figure 4).

Table 5. PCR cycling conditions for the 16S Amplicon PCR reactions.

Cycle	Temperature (°C)	Time	Number of Cycles
Initial Denaturation	94	3 minutes	1
Denaturation	94	45 seconds	35
Annealing	50	1 minute	
Extension	72	1.5 minutes	
Final Extension	72	10 minutes	1
Hold	8	∞	1

16S V4 Amplicon Purification and Quantification

Each sample (n=6) was amplified in replicates of ten and replicates were pooled in a 1.5 ml microcentrifuge tube. In order to remove residual PCR artifacts and contaminants, amplicons were purified using the Agencourt® AMPure® XP Bead System (Beckman Coulter, Catalogue # A63880) following New England Biolabs (Ipswich, MA) protocol for the "Purification of Double-Stranded cDNA using 1.8X Agencourt® AMPure® XP Beads (See Appendix A, Figure 5 for process overview). The AMPure® XP solution was briefly vortexed to resuspend beads before 1.8X (441 µl) of the beads were added to the amplicon reactions. Samples were mixed

well by pipetting up and down at least ten times and incubated for five minutes at room temperature. The tubes were placed on a magnetic stand and rotated 360° (90° at a time) to separate the beads from the supernatant until the solution was clear (about five minutes) (See Appendix A, Figure 6). The supernatant was carefully removed so as to not disturb the beads that contain DNA targets. 200 µl of freshly prepared 80% ethanol was added to the tube while in the magnetic stand and mixed by pipetting up and down. The samples were incubated for 30 seconds at room temperature before the supernatant was removed and discarded. The ethanol step was repeated and the beads were allowed to air dry for ten minutes while the tube was on the magnetic stand with the lid open. The DNA target was eluted from the beads into 52 µl of nuclease free water. The solution was mixed well and the supernatant (~50 µl) was transferred to a clean 1.5 ml microcentrifuge tube.

The concentrations and 260/280 ratios of the purified samples were measured using the Qubit® 2.0 Fluorometer (Life Technologies) and the NanoDrop™ 1000 Spectrophotometer (Thermo-Fisher), respectively (See Figures 6 and 7 for comparison of NanoDrop scans). In order to be used on the MiSeq platform, samples had to have a 260/280 ratio between 1.8 and 2.0. A composite sample for sequencing was pooled by combining equimolar ratios of amplicons from the individual samples.

Positive Control Amplicon PCR and Sequencing

To ensure that the amplicon reactions were working properly, positive control DNA template from *Escherichia coli* strain CFT073 (uropathogenic) was used. Carolyn Dehner (Ph.D., Smith College) had previously isolated the genomic DNA using a standard TE buffer extraction with an ethanol precipitation and eluted in nuclease-free water. The concentration was

determined by Qubit® 2.0 Fluorometer (Life Technologies) to be 40.0 ng/μL. The DNA was diluted in a 1:10 fashion to a final concentration of 4.0 ng/μL. Positive sample amplicon reactions were carried out in the same manner as the fecal samples (See Tables 4 and 5).

Positive controls that showed successful amplification after gel electrophoresis were subjected to Sanger dideoxy sequencing to ensure that the correct product was amplified. The mixture (Table 6) and cycling conditions (Table 7) for the sequencing reactions were determined after communication with Weam Zaky (M.S., Smith College). All sequencing reactions were run using a Veriti® 96 Well Thermal Cycler (Applied Biosystems).

Following cycling, the sequencing reactions were purified using the Performa® DTR Gel Filtration Cartridges (EdgeBio, Catalogue # 98780) following the manufacturer's instructions. The Performa gel filtration cartridges were centrifuged for three minutes at 850 x g. The cartridges were transferred to 1.5 mL microcentrifuge tubes and the sequence samples were added to the center of the packed columns. The cartridge was centrifuged for three minutes at 850 X g and the eluate was retained for sequencing. Sequencing of purified samples was completed on an ABI 3130xl Genetic Analyzer (Applied Biosystems).

Table 6. The 10 μl PCR mixture for sequencing of the 16S V4 amplicon positive controls

PCR Reagent	**Volume (μL) (per PCR reaction)**
Big Dye	4.0
Reverse Primer (0.8 pm/μl)	2.0
PCR Product *	2.0
ddH_20 (RT-PCR grade water)	2.0

- Note: If the PCR band on the 1.5% gel was weak, 4.0 μl of PCR product was used instead of 2.0 μl.

Table 7. Cycling conditions for sequencing of the 16S amplicon positive controls

Cycle	Temperature (°C)	Time	Number of Cycles
Initial Denaturation	96	1 minute	1
Denaturation	96	10 seconds	25
Annealing	50	5 seconds	
Extension	60	4 minutes	
Hold	8	∞	1

The sequencing chromatogram files acquired from sequencing were exported into the Finch TV program. The command <export FASTA sequence> was used to generate and export FASTA files of the chromatograms to the desktop. These FASTA files were used to run a BLAST nucleotide search and compare the positive controls to sequences in the NCBI GenBank database.

Illumina MiSeq Library Preparation and Sequencing

Denature and Dilute DNA Samples

The DNA to be sequenced was quantified using the Qubit® 2.0 Fluorometer (Life Technologies). The concentration estimate was used to first dilute the sample to 10 nM (Table 8), and then to 2 nM (Table 9) using serial dilutions. To denature the DNA into single strands, a 1 ml volume of 0.2 N NaOH was prepared and 10 µl NaOH was added to 10 µl of each of the 2 nM DNA samples. The samples were vortexed briefly, centrifuged at 280 x g for one minute, and incubated for five minutes at room temperature. The samples were pooled and 20 µl was used for sequencing (the remaining DNA sample was stored at -20° C for future use). Pre-chilled HT1 Hybridization Buffer (980 µl) was added to the denatured DNA to give a concentration of 20 pM. To dilute the denatured DNA to a final concentration of 15 pM (1.5

mM), 750 µl of the 20 pM DNA were added to 250 µl pre-chilled HT1. The solution was inverted several times, briefly centrifuged, and kept on ice for later use.

Table 8. Dilution of DNA samples to a 10nM concentration

Sample	Total nM	Dilution	Volume of sample added (µl)	Volume of Nuclease-Free H_20 Added (µl)
Desi 8/25	614.72	1: 61.472	5	302.36
Fritz 8/25	655.16	1: 65.5	5	322.58
Depp 8/25	683.47	1:68.3	5	336.5
Desi 12/3	744.13	1: 74.4	5	367
Fritz 12/3	36.96	1:3.696	5	13.48
Depp 12/3	279.65	1:27.65	5	134.83

Table 9. Dilution of DNA samples to 2nM final concentration

Sample	Concentration (nM)	Dilution	Volume of sample added (µl)	Volume of Nuclease-free H20 added (µl)
Desi 8/25	10	1: 5	5	20
Fritz 8/25	10	1: 5	5	20
Depp 8/25	10	1:5	5	20
Desi 12/3	10	1: 5	5	20
Fritz 12/3	10	1:5	5	20
Depp 12/3	10	1:5	5	20

Denature and Dilute PhiX Control

ØX174 DNA was used as an internal control to help balance the extreme base bias present in low-diversity 16S amplicon samples. To dilute the PhiX DNA to 2 nM, 2 µl ØX174 (10 µM) was combined with 8 µL 10 mM Tris-Cl, pH 8.5 with 0.1% Tween 20. The ØX174 DNA was further diluted to 1 nM by combining 10 µl 2 nM ØX174 DNA with 10 µl 0.2 N NaOH. The DNA was vortexed briefly, centrifuged at 280 x g for one minute, and incubated for five minutes at room temperature to denature the PhiX template DNA into single strands. Pre-chilled HT1 Hybridization Buffer (980 µl) was added to the denatured DNA (20 µl) to give a final concentration of 20 pM. The denatured DNA was kept on ice until use.

Combine Sample DNA and PhiX Control

For most samples, Illumina recommends a low-concentration PhiX control spike-in at 1%. However, for metagenomics or low diversity libraries, Illumina recommends increasing the PhiX control spike-in to 20-25%. For a 20% PhiX spike-in, 200 µl of the denatured ØX174 control DNA was added to 800 µl of the denatured sample DNA and kept on ice until it was ready to load onto the MiSeq reagent cartridge.

Loading the Sample Library onto Cartridge

For sequencing 16S rRNA amplicons, the 300-cycle MiSeq Reagent Cartridge v2 kit was used (Illumina, Inc., Catalogue # MS-102-2001). The reagent cartridge was thawed in a bath of room temperature ultrapure water, no higher than the water line, for 1-1.5 hrs before use. Once the cartridge was thawed, it was kept at 4°C until use. The Hybridization Buffer HT1 was thawed on the bench top and also kept at 4°C until use. When preparing the sequencing cartridge, the foil was pierced with a pipette tip and 600 µl of the denatured sample DNA plus

ØX174 DNA was added to the "Load Sample" well. Next, 3.4 µl of the Index Sequencing Primer (100µM), 3.4 µl of the Read 1 Sequencing Primer (100 µM) and 3.4 µl of the Read 2 Sequencing Primer (100 µm) were added to reservoirs 13, 12, and 14, respectively [note that primers were diluted in Buffer EB (Qiagen, Catalogue # 19086)]. The contents of each of the reservoirs were mixed to ensure that the custom primers were mixed with the standard Illumina primers already in the reservoirs. The sequencing primers were designed to be complementary to the V4 amplification primers to avoid sequencing of the primers, and the barcode is read using the third sequencing primer in an additional cycle (See Appendix A, Figure 7). The amplification primers were adapted from the Caporaso *et al*. (2010) protocol to include nine extra bases in the adapter region of the forward amplification primer that support paired-end sequencing on the MiSeq. The amplification and sequencing primers additionally contain a new pad region to avoid primer-dimer formation with the modified adapter. It should also be noted that library preparation and sequencing was performed at four separate times before quality data were obtained.

Cleaning and Loading the Flow Cell

The flow cell is provided in a separate container and stored in a storage buffer. Prior to loading the flow cell on the MiSeq it was rinsed with laboratory-grade water. This ensured that both the glass and plastic cartridge were rinsed of excess salts, which can affect flow cell seating on the instruments and affect imaging if allowed to dry on the imaging area. The flow cell was dried and the glass was then cleaned using an alcohol wipe, making sure the glass was free of streaks and fingerprints (Note: it is important to avoid using the alcohol wipe on the flow cell port gasket). Excess alcohol was dried with a lint-free lens cleaning tissue. The flow cell, with the Illumina label facing upward, was set on the MiSeq stage and the flow cell latch closed.

Sample Sheet Input

The MiSeq uses a "Sample Sheet" .csv file (set-up through the Illumina Experiment Manager to define the analysis parameters) for each run (Figure 5). When creating the sample sheet for this run, under "Select Workflow", "Targeted Resequencing" was selected, followed by "16S Metagenomics". Under the field "Select Compatible Assay", "TruSeq LT" was selected and the number listed on the cartridge to be used was entered in the "Sample Sheet Name*" (e.g. MS2017966-300V2). The option, "Paired End," 1 Index Read, Index Cycles 6, and 151x151bp was selected (note that despite the barcodes being 12 bases, I set Index Cycles to 6 in this step – this was corrected manually in a subsequent step). On the next screen, the Sample ID was entered and one of the standard barcodes (e.g. A001) is selected. Once this .csv file was created, it was edited manually to instruct the MiSeq to conduct a 12 bp index read. This was achieved by opening the appropriate sample sheet for the run in the text editor Notepad and under [Data], the 6 bp barcode was replaced with the appropriate 12 bp barcode to indicate a 12 bp index read (Table 10).

Table 10. List of samples and corresponding 12 bp Golay barcodes.

Sample	12 Base Pair Golay Barcode
Taki 8/25	TCC CTT GTC TCC
Desi 8/25	ACG AGA CTG ATT
Fritz 8/25	GCT GTA CGG ATT
Depp 8/25	ATC ACC AGG TGT
Taki 12/3	TGG TCA ACG ATA
Desi 12/3	ATC GCA CAG TAA
Fritz 12/3	GTC GTG TAG CCT
Depp 12/3	AGC GGA GGT TAG

Sample Sheet

Header

IEMFileVersion	4
Investigator Name	Rachael Sirois
Project Name	Equine Metagenomics
Experiment Name	Equine Metagenomics 16S Run 1
Date	3/22/2013
Workflow	Metagenomics
Application	Metagenomics 16S rRNA
Assay	TruSeq LT
Description	
Chemistry	Default

Reads

151
151

Settings

Adapter	AGATCGGAAGAGCACACGTC

Data

Sample ID	Sample Name	Sample Plate	Sample Well	I7_Index_ID	Index
1	Desi 8/25	ms2023250-300	A01	1	ACGAGACTGATT
2	Fritz 8/25	ms2023250-300	A02	2	GCTGTACGGATT
3	Depp 8/25	ms2023250-300	A03	3	ATCACCAGGTGT
4	Desi 12/3	ms2023250-300	A04	4	ATCGCACAGTAA
5	Fritz 12/3	ms2023250-300	A05	5	GTCGTGTAGCCT
6	Depp 12/3	ms2023250-300	A06	6	AGCGGAGGTTAG

Figure 5. Screenshot example of a “sample sheet” .csv file after editing run parameters.

MiSeq Data Analysis

Sequence data were analyzed using two different methods. First, sequence data underwent primary and secondary analysis on the MiSeq instrument (Figure 6). MiSeq Real-Time Analysis (RTA) is a software application that helps perform primary analysis for Illumina's sequencing instruments. RTA runs locally on the instrument control personal computer and performs imaging, measures intensities, base calling, and quality scoring. The analysis is performed during the chemistry and imaging cycles of a sequencing run and data output is in the form of .bcl files ("base calls").

MiSeq Reporter, the secondary analysis software on the MiSeq, was launched automatically after RTA completed primary analysis. Under the DNA workflow, the 'Targeted Resequencing' application was selected, and 'Metagenomics 16S rRNA Analysis' was applied to the run. The Metagenomics workflow enables analysis of 16S rRNA to determine which organisms are present. The Illumina 16S rRNA data store is populated by sequences in the May 2011 release of the 'GreenGenes 16S rRNA' database that provides users with a curated taxonomy based on *de novo* tree inference (McDonald *et al*., 2012). The main output of this workflow is a classification of reads at several taxonomic levels and provides a clusters graph, samples table, and a metagenomics pie chart. The clusters graph provides information about the number of clusters that are detected during sequencing characterized by the following descriptions: total, passing filter, unaligned, unindexed, and duplicates. The Metagenomics pie chart provides a visualization of the number of clusters from each sample that were assigned to a category at each taxonomic level.

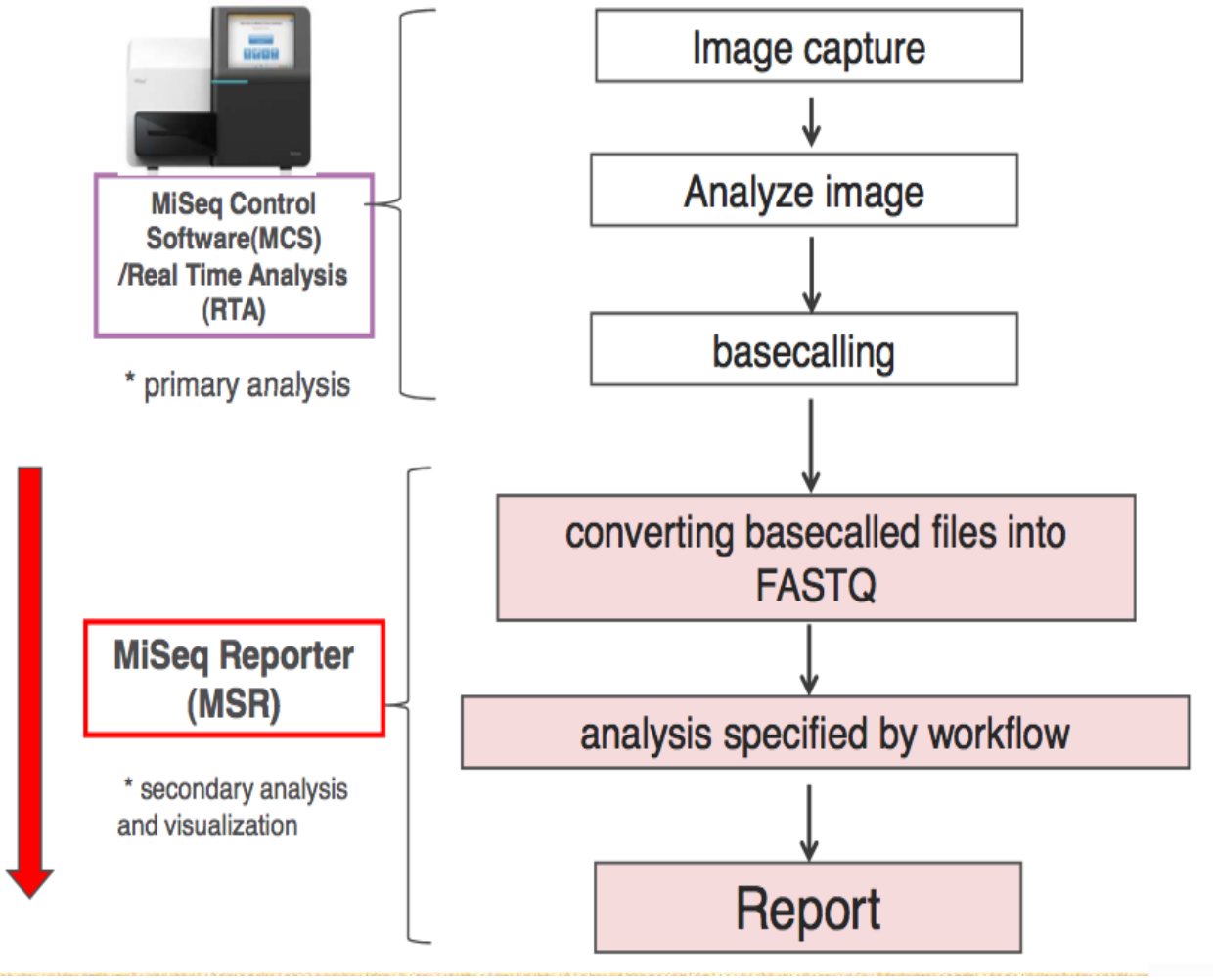

Figure 6. Flow chart illustrating primary and secondary analysis on the Illumina MiSeq (Illumina, Inc., 2012).

I also used the software package, QIIME version 1.7.0, to analyze the sequencing data (Caporaso *et al*., 2011). QIIME (Quantitative Insights Into Microbial Ecology) is an open source software package for comparison and analysis of microbial communities, primarily based on high-throughput amplicon sequencing data (such as SSU rRNA) generated on a variety of platforms (QIIME Team, 2011). QIIME can be used with raw sequencing output to perform initial analyses such as Operational Taxonomic Unit (OTU) picking (See Appendix B for glossary of terms), taxonomic assignment, and construction of phylogenetic trees from representative sequences of OTUs. Sample FASTQ files and index FASTQ files were first generated before applying QIIME to demultiplex and process the data.

The QIIME program uses a command line interface via the Terminal (a program on an Apple computer) to interact with the computer. Commands are then typed into the computer in order to analyze the data as desired. The data were uploaded into the program as FASTQ files, which provide the sequence FASTA files as well as Phred quality scores. Using TextEdit, or any other simple word processor, a mapping file was constructed for each read that designated a particular barcode and primer for each sample (Figure 7). The mapping file locations were verified by the QIIME program using the command, < $ check_id_map.py –m map.txt >, where map.txt is the name of the mapping file.

#SampleID	BarcodeSequence	LinkerPrimerSequence	Treatment
Desi. 8.25F	ACGAGACTGATT	GTGTGCCAGMGCCGCGGTAA	before
Desi.12.3F	ATCGCACAGTAA	GTGTGCCAGMGCCGCGGTAA	after
Fritz.8.25F	GCTGTACGGATT	GTGTGCCAGMGCCGCGGTAA	before
Fritz.12.3F	GTCGTGTAGCCT	GTGTGCCAGMGCCGCGGTAA	after
Depp.8.25F	ATCACCAGGTGT	GTGTGCCAGMGCCGCGGTAA	before
Depp.12.3F	AGCGGAGGTTAG	GTGTGCCAGMGCCGCGGTAA	after

Figure 7. Screenshot of read 1 (forward) mapping file for use in the QIIME split_libraries_.fasta.py command.

The next command split the libraries during which each of the sequences were evaluated based on the barcode and were binned as a specific sample. For example, all sequences that were found with the barcode <ACGAGACTGATT> would be labeled as Desi 8/25. The sequences were analyzed by read (either read 1 or read2) and the final command for these parameters is listed below, where -i is the input file, -b is the barcode (index) file, -m is the mapping file, and -o is the output file:

Read 1(Forward):

```
$ split_libraries_fastq.py -i
     Undetermined_S0_L001_R1_001.fastq.gz -b
     Undetermined_S0_L001_I1_001.fastq.gz -m
     Mapping_forward.txt -o test2/ --
     rev_comp_mapping_barcodes
```

Read 2 (Reverse):

```
$ split_libraries_fastq.py -i
     Undetermined_S0_L001_R2_001.fastq.gz -b
     Undetermined_S0_L001_I1_001.fastq.gz  -m
     Mapping_reverse.txt -o test2/ --
     rev_comp_mapping_barcodes
```

Reads were assigned to OTUs using a closed-reference OTU (See Appendix B for glossary of terms) picking protocol using the QIIME toolkit (Caporaso *et al*., 2010) where Uclust (Edgar, 2010) was applied to search sequences against a subset of the GreenGenes database (DeSantis *et al*., 2006) filtered at 97% identity. Reads were assigned to OTUs based on their best match to this database at greater than or equal to 97% sequence identity. Reads that did not match a reference sequence were discarded. Taxonomy was assigned to each read by accepting the GreenGenes taxonomy string of the best matching GreenGenes sequence. The final command for these parameters is listed below:

Read 1 (Forward):

```
$ pick_closed_reference_otus.py -i read1_demulti/seqs.fna -
     r gg_12_10_otus/rep_set/97_otus.fasta -t
     gg_12_10_otus/taxonomy/97_otu_taxonomy.txt -o
     read1crotus
```

Read 2 (Reverse):

```
$ pick_closed_reference_otus.py -i read2_demulti/seqs.fna -
     r gg_12_10_otus/rep_set/97_otus.fasta -t
     gg_12_10_otus/taxonomy/97_otu_taxonomy.txt -o
     read2crotus
```

This command provided .txt files, which reported those sequences successfully classified into an OTU, as well as sequences that failed to match a reference in the GreenGenes database. The OTU picking command also presented an OTU table in Biological Observation Format (BIOM), which is designed to be a general-use format for representing biological sample by observation contingency tables. The workflow script, summarize_taxa_through_plots, was used to visualize the taxonomy data in a chart form. The results of this script are folders containing taxonomy summary files (at different taxonomic levels) and a folder containing taxonomy summary plots. The final command for these parameters is listed below:

Read 1 (Forward):

```
$ summarize_taxa_through_plots.py -o taxa_summaryr1 -i
     out_tabler1.biom
```

Read 2 (Reverse):

```
$ summarize_taxa_through_plots.py -o taxa_summaryr2 -i
     out_tabler2.biom
```

RESULTS

16S V4 Amplicon PCR

Despite prior purification steps, PCR amplification of DNA samples proved to be difficult due to residual inhibitors present in the fecal samples. As a result, 16S V4 amplicons were generated for only three of the four study horses (Desi, Fritz, and Depp). The DNA samples for Taki were unable to be amplified with the same PCR protocol successfully used for the other samples and will hopefully be included in a future study. Each horse had two representative samples: August 25, 2011 and December 3, 2011, which correspond to pre-treatment and post-treatment samples, respectively. For each sample (n=6), 16S V4 amplicon PCR reactions were completed in replicates of ten to ensure a sufficient volume of PCR product following purification protocols.

All 16S V4 amplicon PCR reactions were run on 1.5 % agarose gels, and differences in band size between PCR products of different samples were visually identified (See Appendix A, Figure 4 for an example). An amplicon product size of approximately 382 base pairs was expected, as the primer pair amplifies a region of 253 base pairs, and the forward and reverse primers themselves are 60 and 68 nucleotides, respectively. Each of the successful PCRs for Fritz 8/25 (lanes 2, 3, and 4) show a band running just below the 400 base pair marker of the 100 bp DNA Ladder. Lane 6, which represents the positive control, presents the brightest band also just smaller than 400 bp. The negative control in lane 8 shows no band of any size except for primer dimer between 100 and 200 base pairs in size. Additionally, the positive control shows a secondary product at approximately 800 bp (note that the control template was added at a higher concentration). The results for the other successful PCRs (gel photographs not provided) are consistent with the findings for this particular sample with the exception that some of the sample lanes also show a secondary product at the same 800 bp size.

16S V4 Amplicon Purification

The purification of the 16S v4 amplicons was very successful using the AMPure® XP Bead System. As with any purification protocol, some of the DNA was lost but for the purposes of the MiSeq library preparation, there was still an excess of sample. Additionally, the concentration for Fritz 12/3 (9.14 ng/µl) was much lower in comparison to the other samples, which ranged from 69.2 ng/µl (Depp 12/3) to 184 ng/µl (Desi 12/3) (Table 11). This range of concentrations was eventually standardized when samples were diluted to a 2nM final concentration. The NanoDrop scans for the unpurified samples (See figure 8 for a representative scan) showed very low 260/230 ratios indicating protein contamination; this was most likely due to the Taq polymerase used in the 16S v4 PCR. After purification with the AMPure® beads, the NanoDrop scans (See figure 9 for a representative scan) show that the 260/230 ratios have gone up drastically, a sign that contaminating protein was successfully removed. Additionally, the 260/280 ratios also went up for the majority of samples, ensuring that all of the samples met the 1.7-1.9 ratio standard required by the MiSeq (See Table 12 for comparison of 260/280 and 260/230 ratios).

Table 11. Qubit results for the AMPure® purified 16S V4 amplicons.

Sample	Concentration (ng/μl)	Total Quantity (μg)
Desi 8/25	152.0	7.14
Fritz 8/25	162.0	7.61
Depp 8/25	169.0	7.94
Desi 12/3	184.0	8.65
Fritz 12/3	9.14	0.43
Depp 12/3	69.2	3.25

Table 12. Comparison of 260/280 and 260/230 ratios for samples before and after AMPure® XP Bead purification.

	Before Purification		After Purification	
Sample	**260/280**	**260/230**	**260/280**	**260/230**
Desi 8/25	1.88	0.93	1.90	2.21
Fritz 8/25	1.87	0.95	1.90	2.23
Depp 8/25	1.91	0.92	1.91	2.21
Desi 12/3	1.90	0.94	1.90	2.22
Fritz 12/3	1.83	1.01	1.80	2.11
Depp 12/3	1.83	1.00	1.85	2.19

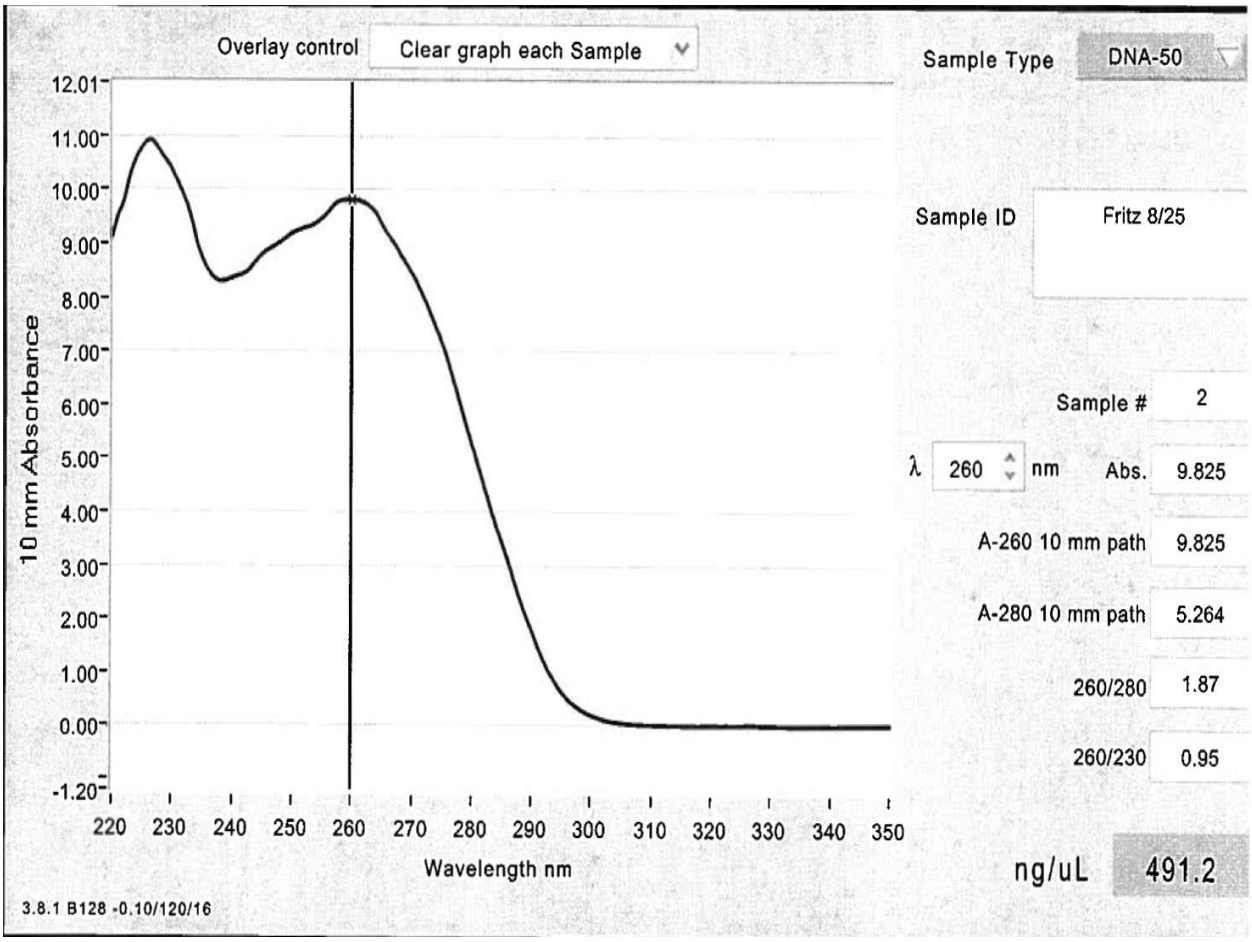

Figure 8. NanoDrop™ 1000 Spectrophotometer scan of unpurified 16S V4 amplicons for Fritz 8/25/11 (February 27, 2013).

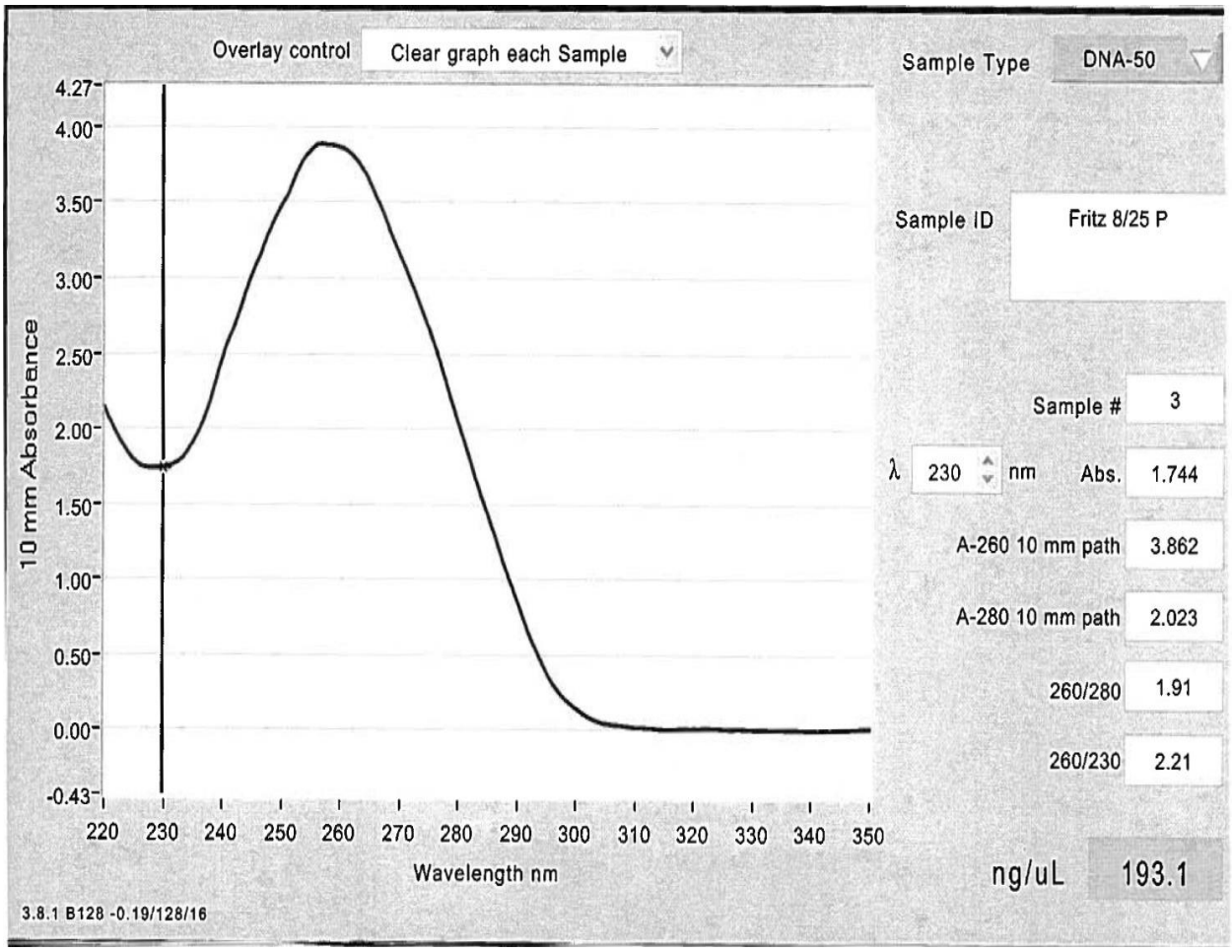

Figure 9. NanoDrop™ 1000 Spectrophotometer scan of AMPure® XP purified 16S V4 amplicons for Fritz 8/25/11 (February 27, 2013).

Sequencing of Positive Controls

A total of ten 16S V4 positive controls were sequenced to ensure that the primers were amplifying the correct region at the proper amplicon length [for reference, the primer pair 515F/806R amplifies the region 533 – 786 in the *E. coli* strain 83972 sequence (Greengenes accession no. prokM-SA_id:470367)] (Caporaso *et al*., 2011). Therefore, the correct amplicon would have a length of 253 bp. The results from one of the ten BLAST nucleotide searches are provided in Figure 10. The sequences from the ten positive controls correctly aligned to partial sequences of the *E. coli* strain KUBWB218 16S rRNA sequence, verifying that the primer pair had correctly amplified the 16S amplicon. Additionally, since all of the sequences gave positive matches to *E. coli* sequences, it can be concluded that all of the products were 16S amplicons, including those bands at 800 bp. These 800 bp bands are likely PCR dimers of the ~400 bp product.

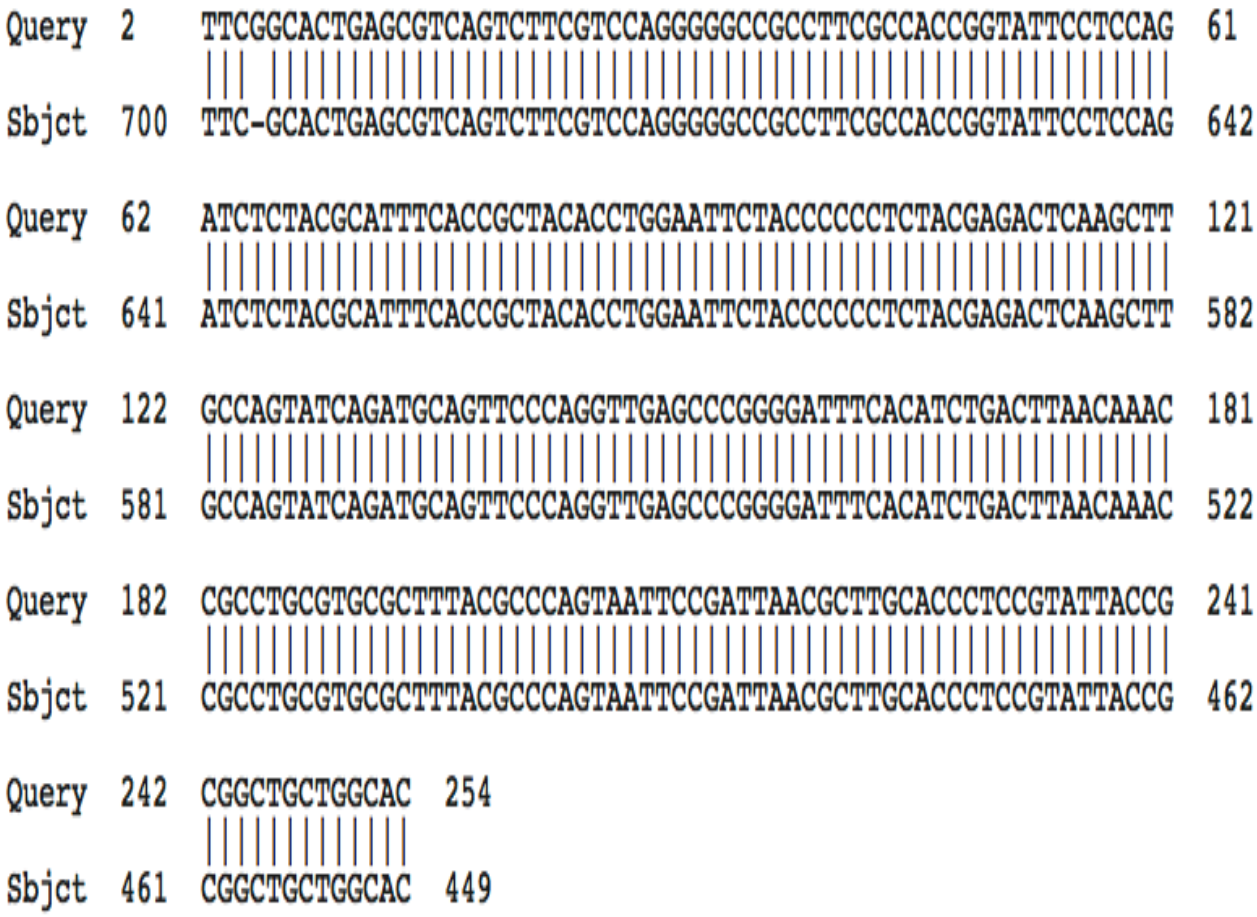

Figure 10. Screenshot of a positive control BLAST result, where the query is the positive control (*E. coli* strain CFT073) and the subject is the *E. coli* strain KUBWB218 16S ribosomal RNA gene, partial sequence, accession no. JQ266004.1.

MiSeq Primary Real-Time Analysis

The run summary and sequencing metrics per read following application of quality control filters are presented in Table 13. The yield total represents the number of bases that were sequenced during the length of the run. The Q score is an integer mapping of *P*, the probability that the corresponding base call is incorrect, with higher Q scores indicating lower error rates. The %>= Q30 is the percentage of bases from clusters that passed filter with a quality score of 30 or greater (probability of incorrect base call is 1/1000). A chart depicting the % Q30 at each cycle is presented in Figure 11. At the start of the first 150 cycles, the percentage of reads that have a quality score above 30 is well over 90%. As the instrument moves through the 150 cycles, the percentage begins to exhibit a negative trend. However, it picks back up again at the start of the second 150 cycles.

The percent aligned is the percentage of the samples that aligned to the PhiX genome and a calculated error rate of reads that aligned to PhiX is provided. The density values indicate the density of clusters (in thousands per mm2) that are detected by image analysis, +/- one standard deviation. The percentage of those clusters that passed the Chastity filter is also presented. The number of reads (in millions) and how many of those reads passed filter are provided. A chart illustrating the base intensities called during each cycle is present in Figure 12. Each of the bases move in a gradual upward trend. However, the intensity of an individual base at any given cycle is quite inconsistent. In this case, these results are expected and are a product of the low diversity samples.

Table 13. 16S metagenomics run summary and sequencing metrics per read

	Read 1	**Read 2 (I)**	**Read 3**	**Total**
Cycles	151	12	151	314
Yield Total	1.4 G	104.6 M	1.4 G	3.0 G
% >= Q30	91.4	72.5	94.2	92.0
Aligned (%)	23.46	0.00	22.98	23.22
Error Rate (%)	0.76	0.00	0.49	0.62
Density (K/mm^2)	524 ± 12	524 ± 12	524 ± 12	
Cluster Passing Filter (%)	90.10 ± 0.82	90.10 ± 0.82	90.10 ± 0.82	
Reads (Millions)	10.56	10.56	10.56	
Read Passing Filter (M)	9.51	9.51	9.51	

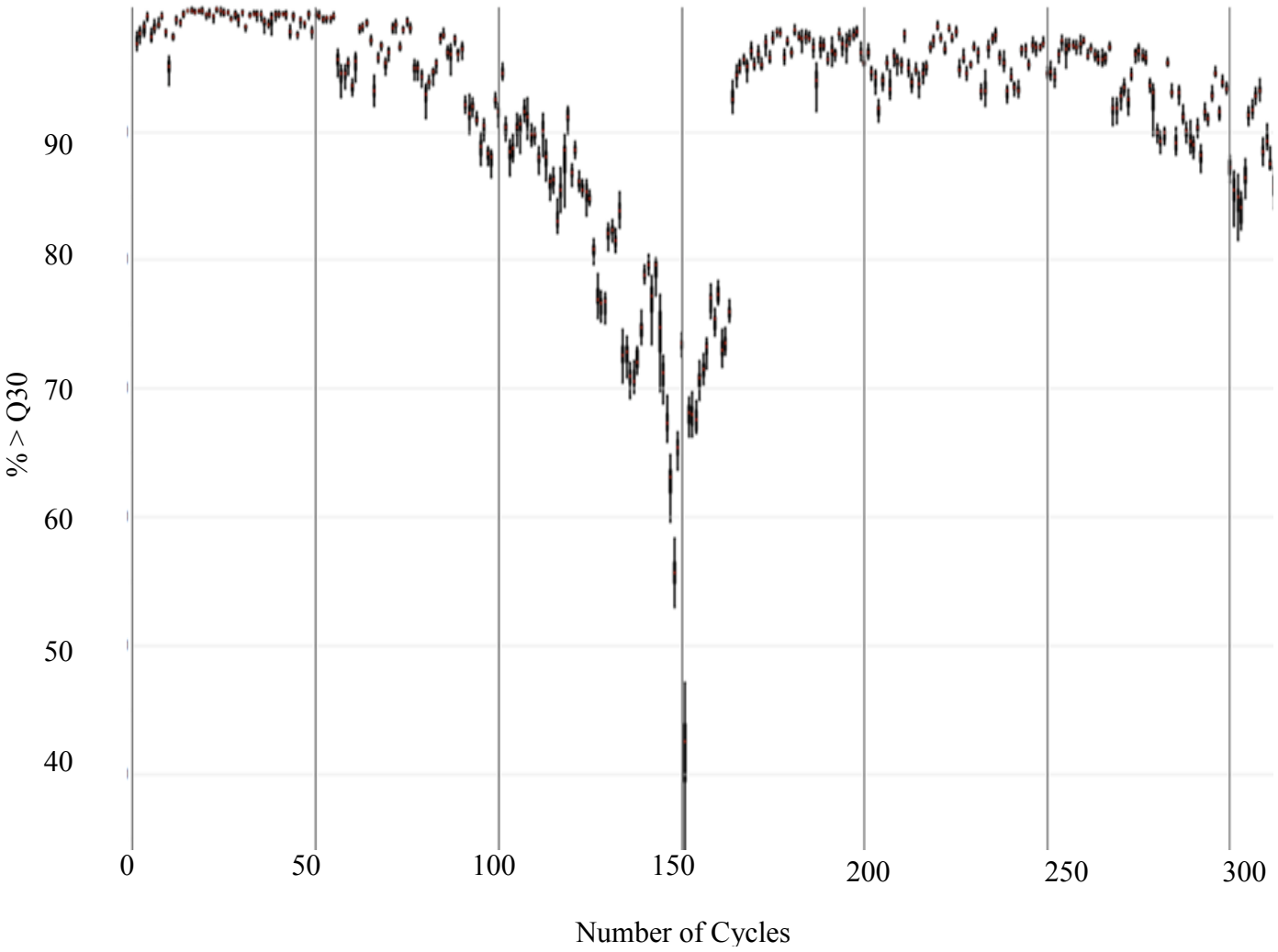

Figure 11. Chart depicting the % > Q30 by cycle.

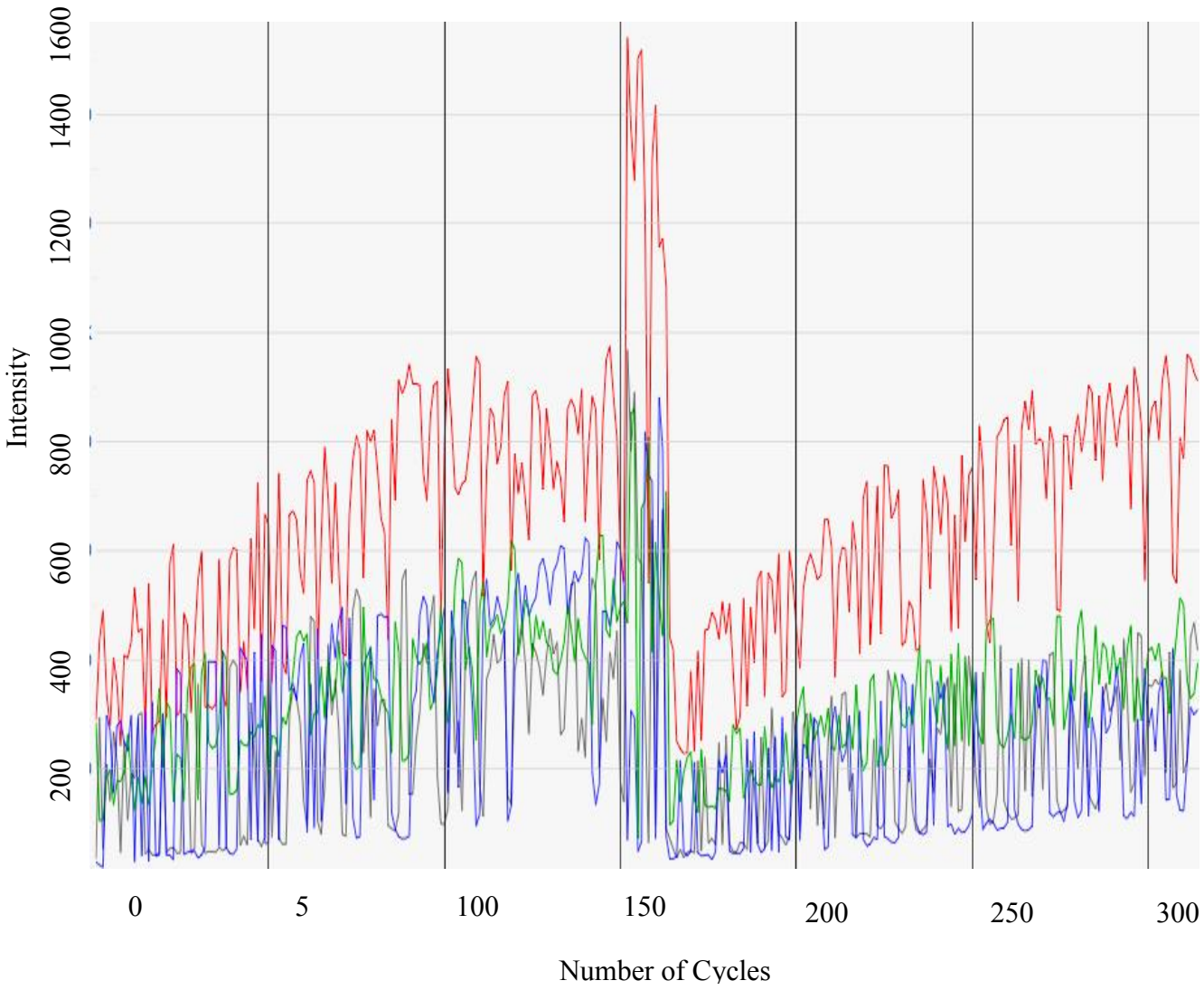

Figure 12. Chart depicting base intensities during each cycle (**A** / **C** / **G** / **T**).

MiSeq Reporter Secondary Analysis

The results for the amount and groups of clusters that were detected during sequencing are presented in Figure 13. Ten million clusters were detected during sequencing and nine million of those clusters passed the quality filter. Unfortunately, all of the clusters (except for five clusters) that passed filter were left unindexed with a 12 bp barcode (See Table 10 for sample barcode assignments). After consultation with Illumina Technical Support, it was ascertained that MiSeq Reporter was unable to effectively process the 12 bp barcode, as the instrument is programmed to recognize a 6 bp barcode. Therefore, the clusters that passed filter could not be assigned to a sample (Desi 8/25, Fritz 8/25, Depp 8/25, Desi 12/3, Fritz 12/3, Depp 12/3).

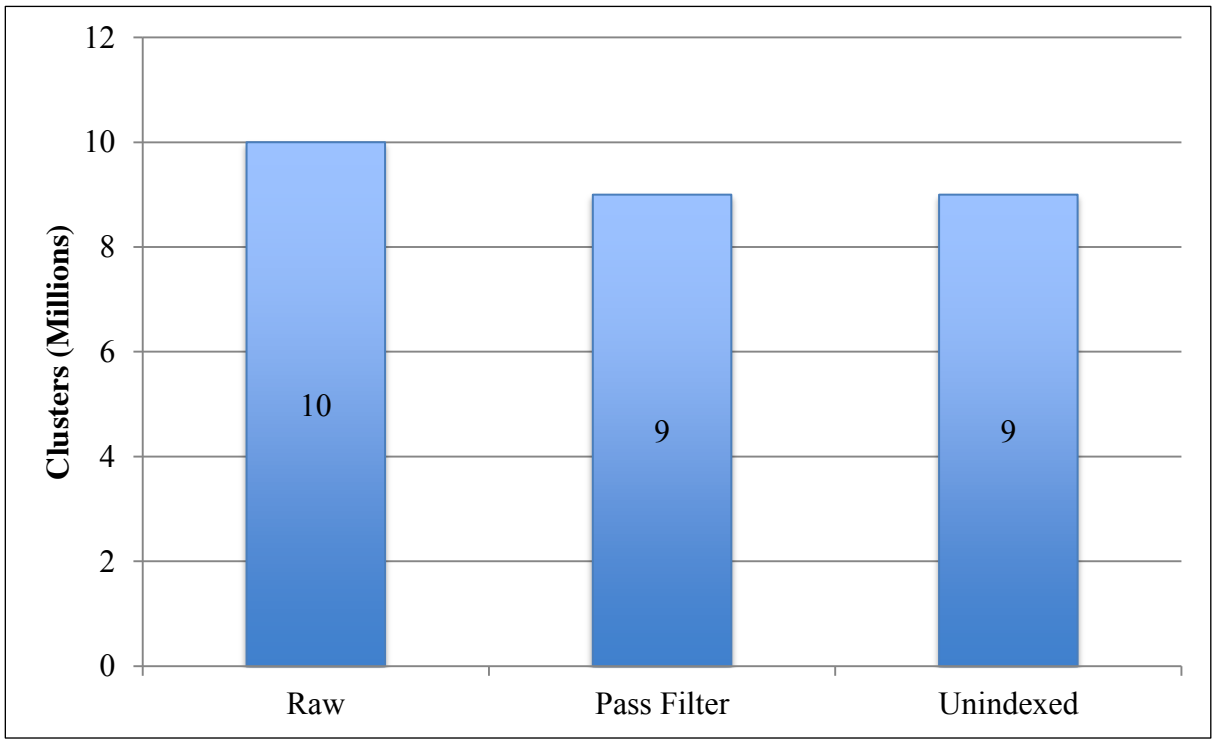

Figure 13. Clusters graph generated by MiSeq Report Secondary Analysis

Due to the fact that the great majority of clusters were not assigned to a sample, graphical data could not be generated. However, the barcode for five of the pass filter clusters were recognized and therefore were provided with minimal taxonomical data. The metagenomics pie chart for Fritz 8/25/11 (Figure 14) shows only that the cluster was bacterial. The pie charts for Depp 8/25 (Figure 15) shows that the cluster was bacterial and was also determined to represent the phylum, Proteobacteria. Similarly, Depp 12/3 (Figure 16) is characterized as being Bacteria in the phylum Proteobacteria.

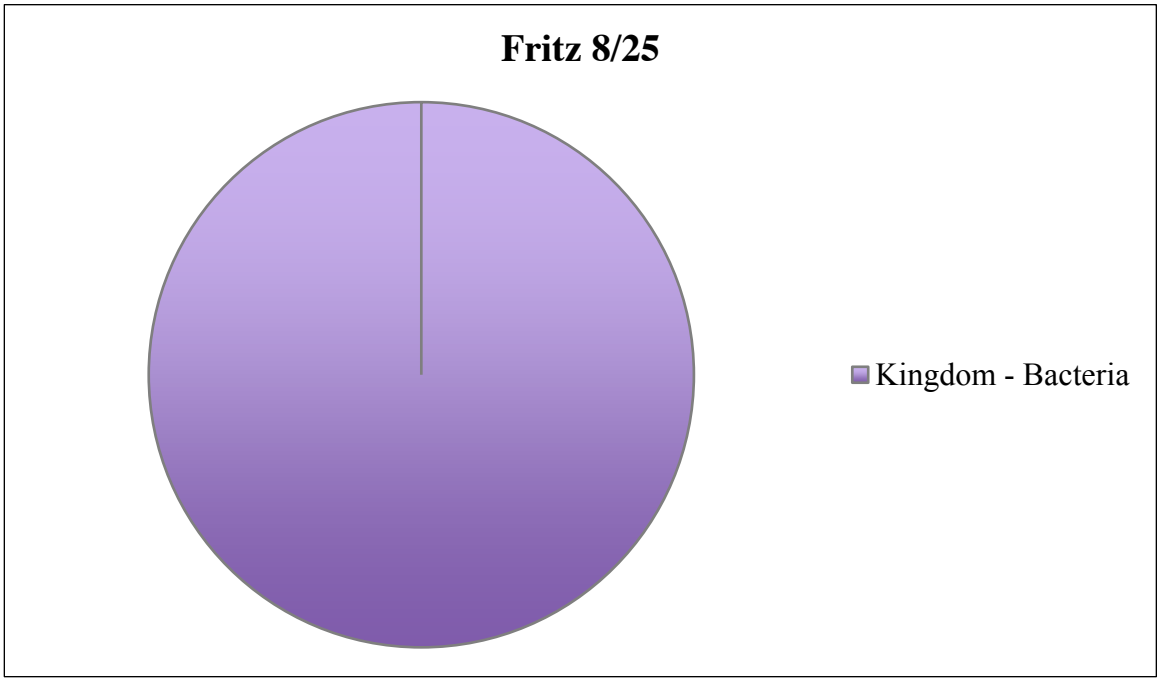

Figure 14. Metagenomics pie chart generated for Fritz 8/25/11 by MiSeq Reporter Secondary Analysis.

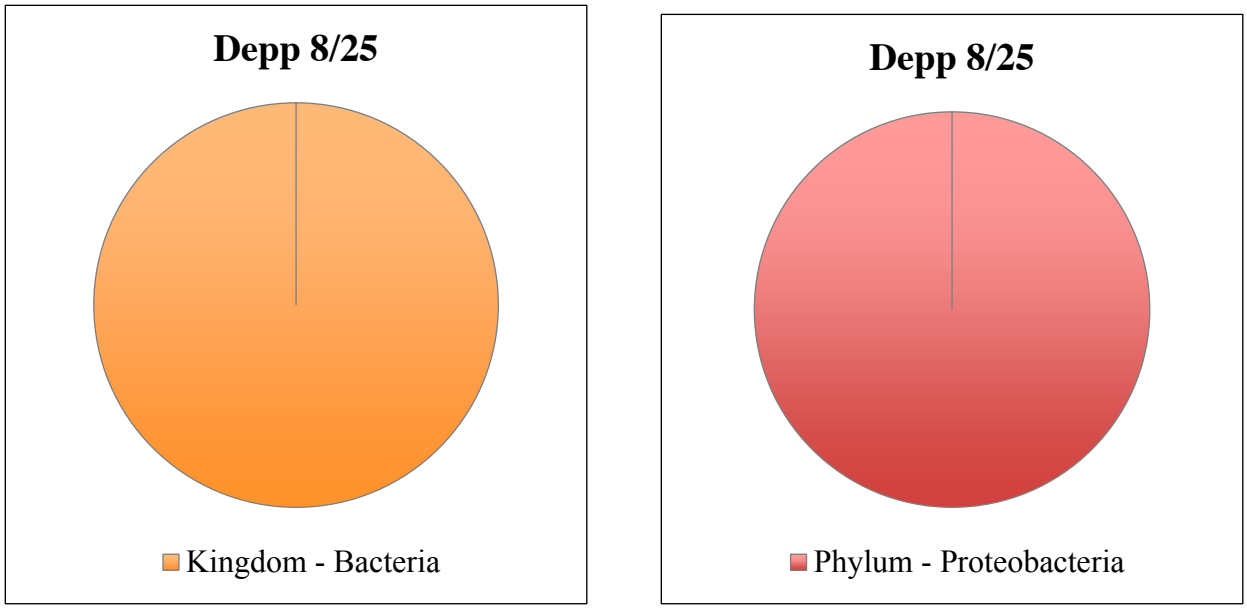

Figure 15. Metagenomics pie charts for Depp 8/25/11 generated by MiSeq Reporter Secondary Analysis.

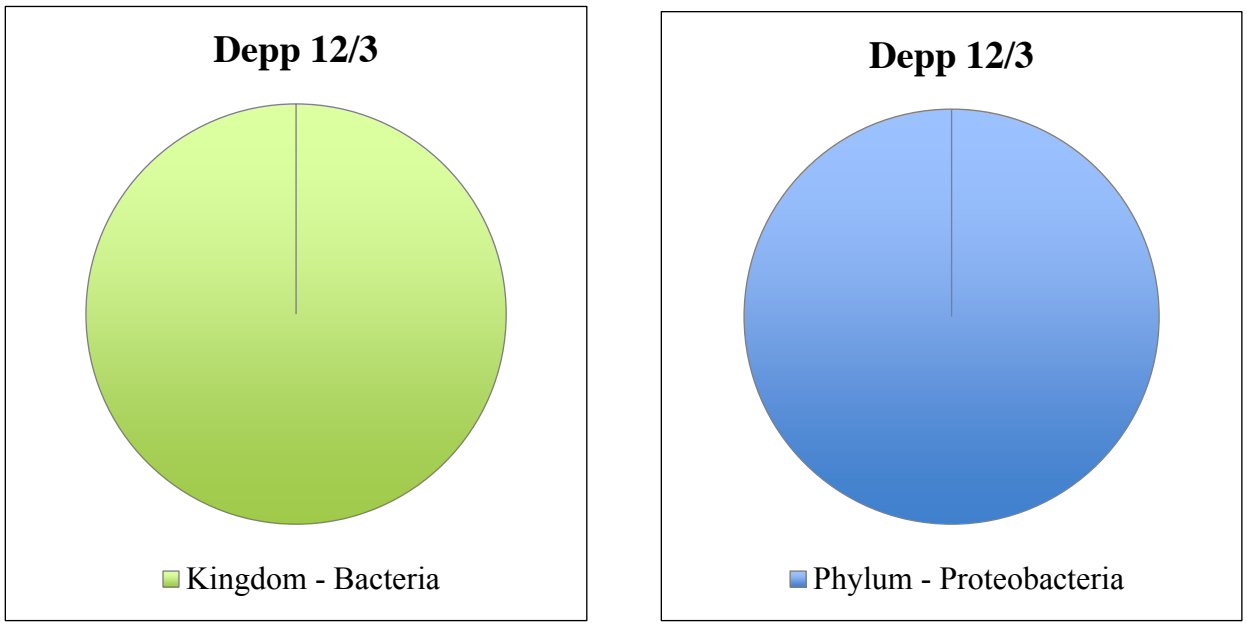

Figure 16. Metagenomics pie charts for Depp 12/3/11 generated by MiSeq Reporter Seconday Analysis.

QIIME Analysis

Metrics

The total number of input sequences, reads with a barcode not in the mapping file, reads that were too short, the median sequence length, and the number of sequences for each fecal sample are presented in Table 14. A total of 5,978,952 and 6,408,791sequences were written for reads 1 and 2, respectively, and therefore used for OTU picking and taxonomical classification. A histogram illustrating sequence length and the number of sequences of that length is presented in Figure 17.

Table 14. Pyrosequencing metrics for demultiplexed samples for read 1 (forward) and read 2 (reverse).

	Read 1	**Read 2**
Total Number of Input Sequences	9,510,926	
Barcode not in Mapping File	173,061	
Reads too short after Quality Truncation	856,102	419,789
Median Sequence Length	151.0	150.0
Desi 8/25 Sequences	1,144,511	1,203,492
Desi 12/3 Sequences	1,120,203	966,571
Fritz 8/25 Sequences	940,739	1,013,687
Fritz 12/3 Sequences	742,930	790,539
Depp 8/25 Sequences	1,144,511	1,230,840
Depp 12/3 Sequences	1,130,879	1,203,662
Total Number of Sequences Written	5,978,952	6,408,791

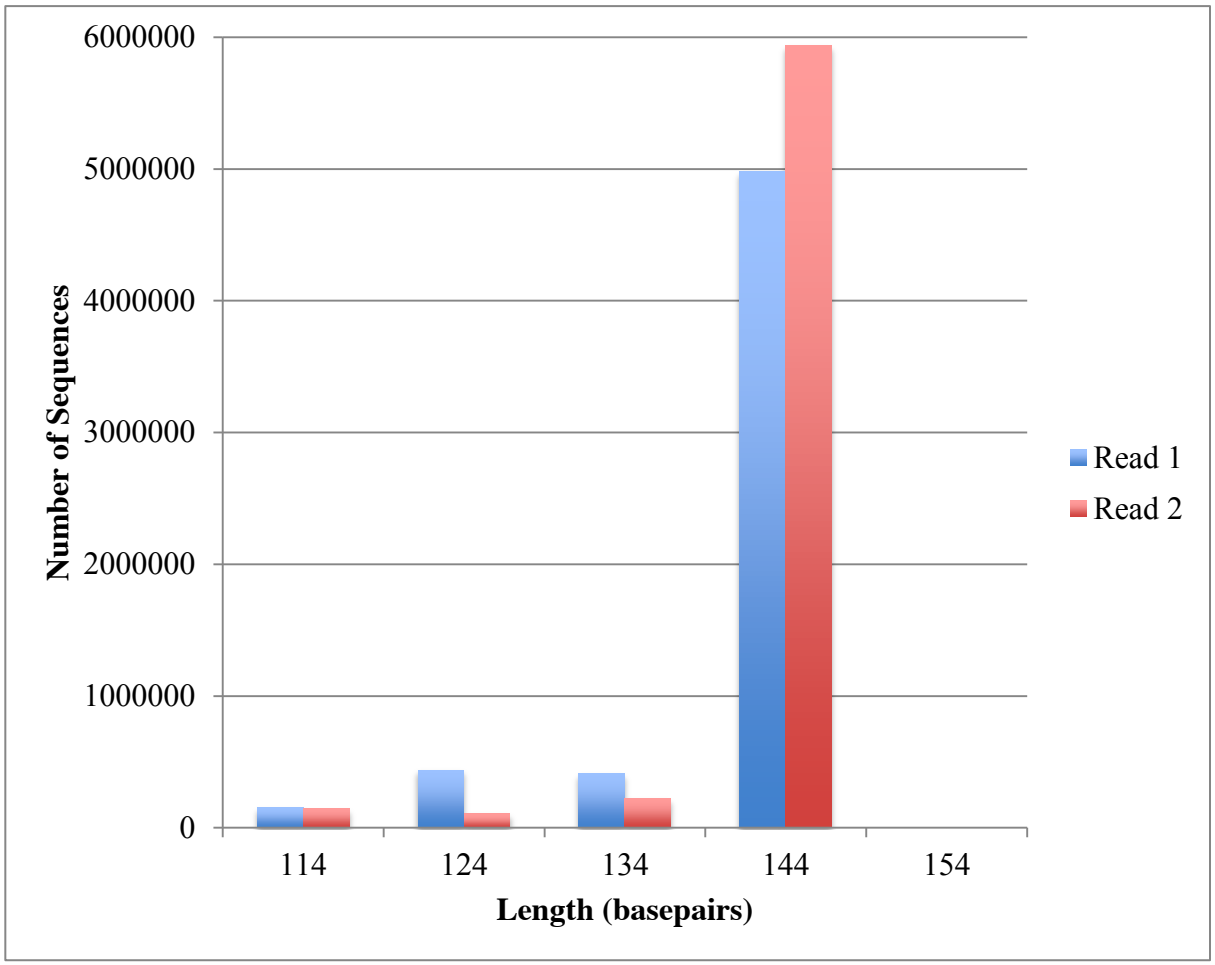

Figure 17. Histogram of sequence lengths and numbers generated during demultiplexing of read 1 and read 2 in QIIME.

Operational Taxonomic Unit Picking

A total of 3,361,963 sequences representing read 1 were classified into 3,528 OTUs while 2,616,989 sequences failed to be grouped into an OTU. Additionally, all of the sequences from read 2, with the exception of one, failed to be classified into an OTU. As a result, the taxonomy summary command was performed for read 1 OTUs but was not run for the read 2 sequences.

Relative Abundances

The complete QIIME taxonomy summary, including bar charts at the phylum, class, order, family, and genus level and the respective legends, is presented in Appendix C. Additionally, a web browser version is available at the following address < file:///Volumes/science.DATAVOL/DROPBOX/BIO/sawlab/taxa_summaryr1/taxa_summary_plots/bar_charts.html#k__Bacteria;%20p__>. The html address allows users to interact with the results in order to understand which colors correspond to which groups of bacteria. The sequences that were categorized into an OTU were classified into a total of 11 phyla. The phyla included ten bacterial phyla and a single archaea phylum. Bacteria phyla representing more than 1 % of total reads are presented in Table 15. For each sample, the most abundant phylum is one of two groups: Bacteriodetes and Firmicutes. In the 'pre-treatment' samples, Bacteriodetes predominated (42.6%) followed by Firmicutes (27.1%) and Verrucomicrobia (12.7 %); these results are conclusive at both the mean and individual level. Firmicutes was the most prevalent phylum among the 'post-treatment' samples accounting for 34.6 % of sequences, followed by Bacteriodetes (31.5%) and Verrucomicrobia (21.7%); these results reflect that of the mean and differ in some samples at the individual level.

Table 15. Classification of equine fecal bacteria before and after treatment with an antihelminthic (in percentages).

Sample	Bacteriodetes	Fibrobacteres	Firmicutes	Proteobacteria	Spirohcaetes	Verruco-microbia
Desi 8/25	46.0	0.3	28.0	0.2	5.0	19.1
Fritz 8/25	39.0	7.3	30.0	0.4	4.5	17.8
Depp 8/25	42.8	7.2	23.2	0.5	7.2	18.1
Mean	42.6	4.9	27.1	0.4	5.6	12.7
Desi 12/3	31.2	4.5	32.2	7.6	5.6	17.6
Fritz 12/3	19.4	1.1	42.2	0.3	1.3	35.2
Depp 12/3	43.9	9.3	29.4	0.1	4.5	12.3
Mean	31.5	5.0	34.6	2.7	3.8	21.7

DISCUSSION

16S v4 Amplification

A total of six samples had 16S v4 amplicons generated that were subsequently sequenced using the Illumina MiSeq. Once a suitable PCR protocol had been optimized, consistent amplification of the desired 16S region was achieved. The problem with optimizing the PCR protocol was not so much volumes and concentrations, but rather efficiency of the enzyme in the presence of inhibitors from fecal samples. The fecal samples were incredibly sensitive and so I found that the most successful amplifications occurred when the enzyme had not undergone several freeze-thaw cycles. In order to avoid these freeze-thaw cycles, I made several aliquots of the enzyme for one time use – this guaranteed that the enzyme would be working at its peak efficiency.

Despite optimizing a PCR protocol that was successful for the majority of the samples (six out of the eight samples), I was unable to generate 16S v4 amplicons for either the pre-treatment (8/25) or post-treatment (12/3) samples for Taki. This is most likely the result of residual inhibitors that were not removed despite several steps of purification. There are several approaches for successful amplification of the v4 region in these samples. First, the PCR protocol may have to be optimized specifically for this set of difficult samples. This could include changing the primer concentrations, increasing the volume of the *Taq* enzyme, or adding an additional reagent to stabilize the reaction such as Bovine Serum Albumin (BSA) or Dimethyl Sulfoxide (DMSO). Additionally, less template could be used in an effort to reduce inhibitor levels while still providing enough template for amplification.

The samples for which amplicons were successfully generated were visualized on a 1.5% agarose gel to verify amplification of the correct product. In the gel photographs (See Figure A4 for example), a band at approximately 400 bp is clearly visible in each of the samples and is

most prominent in the positive control lane. The 800 bp secondary product is most likely a larger 16S v4 amplicon that was generated by dimerization of the ~400 bp product during PCR. This is supported by the fact that the product is just under 800 bp in size, which is twice as large as the desired product. Additionally, the genes coding for the three rRNAs in prokaryotes (both *Bacteria* and *Archaea*) are normally present in multiple copies in the genome (Antón *et al*., 1998). Genes encoding the 16S rRNA (*rrs*), 23S rRNA (*rrl*), and 5S rRNA (*rrf*) are typically arranged into an operon (*rrn* operon) with an intergenic spacer region (ISR) located between the 16S and 23S rRNA genes (Luz *et al*., 1998). In contrast to the general pattern of single-copy genes in bacterial and archaeal genomes, each of the ribosomal RNA encoding genes may be present in 1-15 copies, for example, *Escherichia coli* K12 and *Salmonella enterica* LT2, each possess seven copies of the operon (Lee *at al*., 2008).

MiSeq Run Quality

There were several factors to consider when assessing the quality of the MiSeq run. Different quality values are expected depending on the specific workflow the instrument was instructed to perform. For low diversity, metagenomics workflows, such as in this project, there are certain values expected in order to accept the data for use in further analysis with QIIME. As this was a 2 X 150 run (300 cycle), Illumina technical support confirmed that I should get a yield that exceeds 1 Gb of data and indeed, my run generated a total of 3.0 Gb of data.

Even though the Q30 score decreased across the 150 cycles, the percentage of bases greater than Q30 is averaged across the entire run and not on a per-read or per-cycle basis and so the %>Q30 for the total run was 92.0 %. It should also be noted that a decrease in the %Q30 is perfectly normal and is just inherent to the sequencing technology. The %Q30 decreases over time as the clusters degrade due to the heating/cooling cycles, exposure to different enzymes, and

phasing/ pre-phasing (the percentage of molecules in a cluster for which sequencing falls behind or jumps ahead of the current cycle within a read, respectively). The %Q30 jumps back up at the turnaround (after cycle 150 for a 2x150 run) because during this time, re-growth of some of the clusters occurs so the quality is almost as good as when they were first generated. These newly grown clusters then begin to degrade as the cycles progress for the reasons listed above.

Another value to be considered is the percent of sequences aligned. When library preparation was done, 20% of the PhiX control DNA was added in order to get better results with these low diversity samples. The percent aligned should be equal to the percentage of PhiX that was spiked into the sample. The alignment for this run was 23.2%, which means that a greater portion of the library aligned to the PhiX genome than was added, although not by much. This is most likely due to degradation of the sample DNA resulting in the addition of a lower concentration than expected.

This theory is also supported by the cluster density value. During sample preparation, the denatured sample DNA was diluted to a final concentration of 15 pM in hopes of obtaining a cluster density at the ideal 800 K/mm^2. The acceptable range is anywhere from 500-1300 K/mm2, however problems can occur at both extremes of this range. A cluster density that is too low may not provide the amount of information needed to make accurate base calls or to perform specified workflows. For example, the metagenomics workflow requires a cluster density of at least 500 K/mm^2 to perform the analysis. On the other hand, a cluster density that is so high has the possibility of overloading the instrument – that is the lawn of clusters are too dense that the instruments imager can not accurately capture the color that is being emitted and therefore fails to call bases properly. The cluster density for the fecal samples was 524 +/- 12 K/mm^2. It must be noted that this value also includes the clusters generated for the PhiX control and so in reality,

my cluster density fell below the minimum threshold. However, the clusters that were generated were of high quality since 90% of the clusters passed the quality filter; Illumina suggests that a data quality score of 75% or greater is acceptable.

The last data quality parameter that needs to be discussed is the base intensity chart. Figure 11 illustrates the intensity of each base at each cycle during the run. Although there is a general increasing trend, the base intensities at any given cycle seem erratic and inconsistent. This result is inherent to the type of samples used for this run. These samples were prepared using only 16S amplicons for sequencing and thus represent very low diversity samples. At each cycle, the majority of the clusters will show the same base (since the sequence of the 16S region is very similar in most bacterial species). Thus the intensity of the base at any given position will be very high in relation to the other three bases. This is why the results show such dips and spikes along the intensity plot lines. In this case, these results are expected and are a product of the low diversity samples.

Assessing Microbial Diversity

These results characterize the fecal microbiome of three horses before and after treatment with a common antihelminthic medication. Three phyla of bacteria: Bacteriodetes, Firmicutes, and Verrucomicrobia, represented the three most abundant groups across all six samples. This finding may suggest that members of these groups play a major role in shaping the core structure of the equine gastrointestinal tract. Bacteriodetes was found to be the major bacterium phylum populating the gastrointestinal environment of horses prior to treatment with an antihelminthic medication, followed by Firmicutes. Conversely, Firmicutes was the predominant phylum in 'post-treatment' samples and then Bacteriodetes. However, statistical tests must be performed in

order to determine whether phylum abundance was significantly different between the treatment groups.

The number of horses used in this study was small and as only one samples per animal (for each treatment) was analyzed, some of the differences between groups may be due to interhorse variation. However, the similarities among the values of 'pre-treatment' samples and 'post-treatment' samples suggest that interhorse variation may not be great, at least at the phylum level. Therefore, this study serves as the basis for further studies using larger sample sizes to look at the effect of antihelminthic treatment on the gut microbiome.

CONCLUSION

The greater part of this thesis stems from research that began when The Center for Molecular Biology (CMB) at Smith College confirmed that it would be getting an Illumina MiSeq Sequencer. Since that time, this equine metagenomic project has focused on using massively parallel sequencing (MSP) to better understand this complex microbiome. First, successful protocols for working with difficult environmental samples were devised, including DNA extraction, purification, and quantification. Second, downstream applications, such as PCR, were demonstrated to be successful with the fecal DNA samples. A PCR program for the amplification of the 16S V4 rRNA region was optimized for fecal samples from three of the four research subjects, resulting in a total of six samples for MSP.

The six samples were run a total of four times on the MiSeq sequencer in order to develop a sample library preparation protocol that best served the goals of this research project. Quality scores and measurements were obtained for the run, indicating that a good quality run was performed using the MiSeq instrument. Primary and secondary analyses were performed for the sequencing data using the MiSeq Real-Time Analysis and MiSeq Reporter software, respectively. Although taxonomical results were inconclusive, the data did indicate that the quality of the run was excellent and led to the use of QIIME software for further analysis.

Lastly, initial information regarding the structure of the equine GI tract microbiome was obtained and analyzed using the QIIME software. The GreenGenes database was used as the reference tool to classify sequences into particular OTUs. Bar charts were generated for each of the samples to visualize the OTU taxonomical data at the phylum, class, order, family, and genus levels.

FUTURE DIRECTIONS

Although much time and effort was expended over the course of this study, more needs to be done in order to truly understand the environment of the equine gastrointestinal tract and how helminths and anti-helminthic drugs affect the structure of this environment. First, the results obtained from the QIIME analysis must be more closely analyzed. I believe a good way to begin to understand how the drug treatment may be affecting the microbiome would be to establish which species of bacteria are present in the major groups of genera. The GreenGenes database does not have reference data down to the species level, however, the idea that one could develop their own database sounds reasonable. In order to accomplish this, 16S rRNA data for each of the major genera would have to be compiled using sequences from the GenBank database. This file could then be used as the reference file for the OTU picking command.

Another suggestion for a more comprehensive study would be to have a larger sample size. A larger sample size would allow us to draw general inferences about the equine gastrointestinal environment as a whole, as opposed to the very specific conclusions that can be drawn with a sample size of n=6. A larger sample size with a greater number of sampling days is necessary in order to properly identify trends in bacterial populations. The horse was originally used for this research project because samples from which DNA could be extracted were readily available, but the numbers we would need to increase the sample size significantly may not be possible with such a large and expensive research model. Therefore, it may be useful to apply this research model to another organism that is readily available but also easily controlled and inexpensive to maintain, such as the gerbil or mouse.

Additional information is also necessary regarding other microorganisms that inhabit the gastrointestinal tract. The bacterial populations that are present, although vast in number, only

represent a fraction of the organisms that contribute to the gastrointestinal environment. Similar techniques could be used to study the 28S rRNA gene in order to gain insight into the parasites, protozoa, and other eukaryotic microbes that are present in the gastrointestinal tract microbiome.

I believe researching new primers for the amplification of the 16S amplicon is merited for two reasons. First, the current forward and reverse primers are very long at 60 and 68 nucleotides, respectively, so they are expensive. Moreover, a different reverse primer must be ordered for each sample due to the fact that the identifying barcode is put on the amplicon with this primer. For this reason, I think researching shorter primers should be a priority. A second reason to look into new 16S primers would be to amplify a different region of the 16S rRNA region. According to Chakravorty *et al*. (2007), no single region can differentiate all bacteria. The 16S rRNA hypervariable regions (9) exhibit different degrees of sequence diversity and so combining these regions may provide sufficient sequence diversity to identify a greater number of species. For example, Costa *et al*. (2012) compared the fecal microbiota of healthy horses and horses with colitis by sequencing the v3-v5 region of the 16S rRNA gene. By way of literature research, we can evaluate the benefits of using different combinations of hypervariable regions to decide which ones will best resolve bacterial diversity of a particular microbiome.

REFERENCES

Al Jassim, R., Scott, P., Trebbin, A., Trott, D., and C. Pollitt. (2005). The genetic diversity of lactic acid producing bacteria in the equine gastrointestinal tract. *FEMS Microbiology Letters*, 248(1), 75-81.

Antón, A., Martínez-Murcia, A., and F. Rodrígues-Valera. (1998). Sequence diversity in the 16S-23S intergenic spacer region (ISR) of the rRNA operons in representatives of the *Escherichia coli* ECOR collection. *Journal of Molecular Evolution*, 47, 62-72.

Antonopoulos, D.A., Huse, S.M., Morrison, H.G., Schmidt, T.M., Sogin, M.L., and V.B. Young. (2009). Reproducible community dynamics of the gastrointestinal microbiota following antibiotic perturbation. *Infection and Immunity*, 77(6), 2367-2375.

Berglund, J. (2012). Alternative therapies: desperate measures. *Nature*, 484(7393), S11.

Bracken, M., Wohlk, C., Petersen, S., and M. Nielsen. (2012). Evaluation of conventional PCR for detection of Strongylus vulgaris on horse farms. *Veterinary Parasitology*, 184(2-4), 387-391.

Cardinale, M., Brusetti, L., Quatrini, P., Borin, S., Puglia, A., Rizzi, A., Zanardini, E., Solini, C., Corselli, C., and D., Daffoncho. (2004). Comparison of Different Primer Sets for Use in Automated Ribosomal Intergenic Spacer Analysis of Complex Bacterial Communities. *Applied and Environmental Microbiology*, 73(2), 659-662.

Caporaso, J., Lauber, C., Walters, W., Berg-Lyons, D., Lozupone, C., Turnbaugh, P., Fierer, N., and R. Knight. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *PNAS*, 108, 4516-4522.

Caporaso, J., Lauber, C., Walters, W., Berg-Lyons, D., Huntley J., Fierer, N., Owens, S., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J., Smith, G., and R. Knight. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The International Society for Microbial Ecology Journal*, 6, 1621-1624.

Chakravorty, S., Helb, D., Burday, M., Connell, N., and D. Alland. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiology Methods*

Clemens, E.T., Steven, C.E., and M. Southworth. (1975). Sites of organic acid production and pattern of digesta movement in the gastrointestinal tract of swine. *Journal of Nutrition*, 105, 759-68.

Costa, M.C., Arroyo, L.G., Allen-Vercoe, E., Stämpfil, H.R., Kim, P.T., Sturgeon, A., and J.S. Weese. (2012). Camparison of the fecal microbiota of healthy horses and horses with colitis by high throughput sequencing of the V3-V5 region of the 16S gene. *PLoS ONE*, 7(7), 1-12.

Daly, K., Stewart, C.S., Flint, H.J., and S.P. Shirazi-Beechey. (2001) Bacterial diversity within the equine large intestine as revealed by molecular analysis of cloned 16S rRNA genes. *FEMS Microbiology Ecology*, 38 (2-3), 141-151.

DeSantis, T., Hugenholtz, P., Keller, K., Brodie, E.L., Larsen, N., Piceno, Y.M., Phan, R., and G.L. Anderson. (2006) Greengenes, a chimera-checked 16S rRNA database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72(7), 5069-5072.

Donecker, J.M. and R.E. Holland. (2007). Significance of fecal egg counts in managing equine intestinal parasites. *Pfizer Animal Health*. http://www.pfizer.com.

Eckburg, P., Bik, E., Bernstein, C., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S., Nelson, K., and D. Relman. (2005). Diversity of the Human Intestinal Microbial Flora. *Science*, 308(5728), 1635-1638.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), 2460-2461.

Fujimura, K., Slusher, N., Cabana, M., and S. Lynch. (2010). Role of the gut microbiota in defining human health. *Expert Review of Anti-Infective Therapy*, 8(4), 435-454.

Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., an K.E., Nelson. (2006). Metagenomic analysis of the human distal gut microbiome. *Science*, 312, 1355-1359.

Gori, F., Folino, G., Jetten, M., and E. Marchiori. (2011). MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. *Bioinformatics*, 27(2), 196-203.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, 68(4), 669-685.

Hooper, L. V., Littman, D.R., and A.J. Macpherson. (2012). Interactions between the microbiota and the immune system. *Science*, 336, 1268-1273.

Husted, L., Jenson, T.K., Olsen, S.N., and L. Molbek. (2010). Examination of equine glandular stomach lesions for bacteria, including *Helicobacter spp* by fluorescence *in situ* hybridization. *BMC Microbiology*, 10, 84-91.

Kaplan, R. (2002). Anthelmintic resistance in nematodes of horses. *Veterinary Research*, 33(5), 491-507.

Kobayashi, Y., Koike, S., Taguchi, H., Itabashi H., Kam, D.K., and J.K. Ha. (2004). Recent advances in gut microbiology and their possible contribution to animal health and production-a review. *Asian-Australasian Journal of Animal Sciences*, 17(6), 877-884.

Kornás, S., Gawor, J., Cabaret, J., Molenda, K., Skalska, M., and B. Nowosad. (2009). Morphometric identification of equid cyathostome (nematoda: Cyathostominae) infective larvae. *Veterinary Parasitology*, 162, 290-294.

Lee, Z., Bussema, C., and T. Schmidt. *rrn*DB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Research*, 37, 489-493.

Luz, S., Rodrígues-Valera, F., Lan, R., and P. Reeves. Variation of the ribosomal operon 16S-23S gene spacer region in representative of *Salmonella enterica* subspecies. *Journal of Bacteriology*, 180(8), 2144-2151.

Mackie, R. and C. Wilkins. (1988). Enumeration of anaerobic bacterial microflora of the equine gastrointestinal tract. *Applied and Environmental Microbiology*, 54(9), 2155-2160.

Matthews, J. (2008). An update on cyathostomins: antihelmintic resistance and worm control. *Equine Veterinary Education*, 20(10), 552-560.

Matthews, J. (2011). HBLB's advances in equine veterinary science and practice: facing the threat of equine parasitic disease. *Equine Veterinary Journal*, 43, 126-132.

McDonald, D., Price, M., Goodrich, J., Nawrocki, E., DeSantis, J., Probst, A., Anderson, G. Knight, R., and P. Hugenholtz. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, 6, 610-618.

Men, A., Forrest, S., and K. Siemering. (2011). Metagenomics and beyond: new toolboxes for microbial systematics. *Microbiology Australia*, 32(2), 86-89.

Milinovich, G., Burrell, P., Pollitt, C., Klieve, A., Blackall, L., Ouwerkerk, D., Woodland, E., and D. Trott. (2008). Microbial ecology of the equine hindgut during oligofructose-induced laminitis. *ISME Journal*, 2, 1089-1100.

Pernthaler, A., Dekas, A., Brown, C., Goffredi, S., Embaye, T., and V. Orphan. (2008) Diverse syntrophic partnerships from deep-sea methane vents revealed by direct cell capture and metagenomics. *PNAS*, 105(19), 7052-7057.

Popa, R., Popa, R., Marshall, M.J., Nguyen, H., Tebo, B.M., and S. Brauer. (2009) Limitations and benefits of ARISA intra-genomic diversity fingerprinting. *Journal of Microbiological Methods* ,78(2), 111-118.

Radstrom, P., Knutsson, R., Wolffs, P., Lovenklev, M., and C. Lofstrom. (2004). Pre-PCR processing: strategies to generate PCR-compatible samples. *Molecular Biotechnology*, 26(2), 133-146.

Salzman, N. H., De Jong, H., Paterson, Y., Harmsen, H., Welling, G., and N. Bos. (2002). Analysis of 16S libraries of mouse gastrointestinal microflora reveals a large new group of mouse intestinal bacteria. *Microbiology*, 148(11), 3651-3660.

Santos, A. S., Rodrigues, M., Bessa, R., Ferreira, L., and W. Martin-Rosset. (2010) Understanding the equine cecum-colon ecosystem: current knowledge and future perspectives. *Animal*, 5(1), 48-56.

Simon, C., and R. Daniel. (2011). Metagenomic analyses: past and future trends. *Applied and Environmental Microbiology*, 77(4), 1153-1161.

Tucker, T., Marra, M., and J.M. Friedman. (2009). Massively parallel sequencing: the next big thing in genetic medicine. *The American Journal of Human Genetics*, 85, 142-154.

Wang, Y. and P. Qian. (2009). Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS ONE*, 4(10), 1-9.

Wolff, M., Broadhurst, M., and P. Loke. (2012). Helminthic therapy: improving mucosal barrier function. *Trends in Parasitology*, 28(5), 187-192.

Yazdanbakhsh, M., Kremsner, P., and R. Ree. (2002). Allergy, parasites, and the hygiene hypothesis. *Science*, 296, 490-494.

Zoetendal, E. G., et al. (2004). Molecular microbial ecology of the gastrointestinal tract: from phylogeny to function. *Current Issues in Intestinal Microbiology*, 5(2), 31-48.

APPENDIX A: Supplemental Figures and Tables

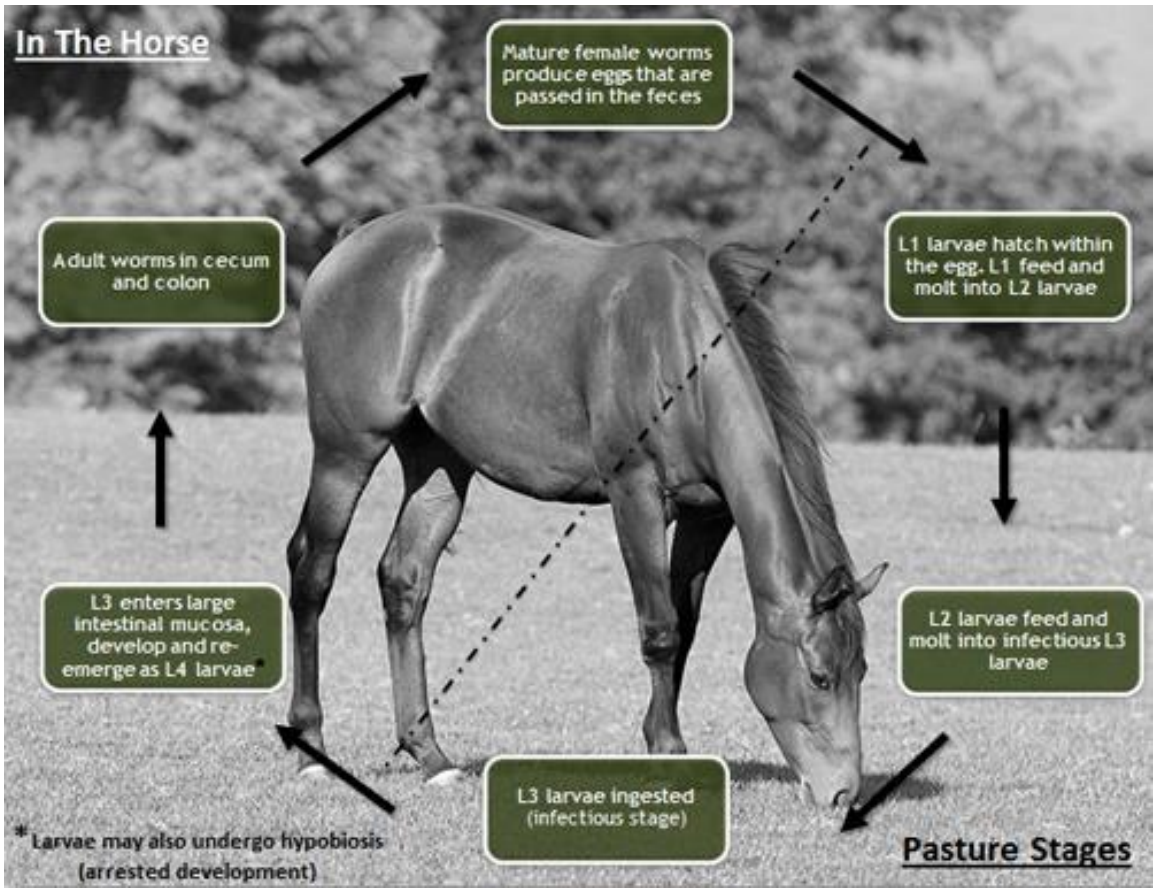

Figure A1. An illustration depicting the life cycle of small strongyles in both the horse host and the pasture stage (Eggzamin™, 2010 - 2012).

Table 1. List of reagents and respective volumes in the 5 PRIME HotMasterMix.

PCR Component	Volume
HotMaster *Taq* DN Polymerase	50 U/ml
HotMaster *Taq* Buffer (pH 8.5)	2.5x
$Mg(OAc)_2$	6.25 mM
dNTPs	500 μM (each)

5’ – AATGATACGGCGACCACCGAGATCTACAC
TATGGTAATT GT GTGCCAGCMGCCGCGGTAA – 3’

Figure A2. 16S Amplicon 515F (forward) PCR primer sequence [Field number (space-delimited) and description]. 1. Reverse complement of 3' Illumina adapter, 2. Reverse primer pad, 3. Reverse primer linker, 4. Reverse primer (Caporaso *et al*., 2011).

806rcbc0

5’ – CAAGCAGAAGACGGCATACGAGAT TCCCTTGTCTCC
AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT – 3’

806rcbc1

5’ – CAAGCAGAAGACGGCATACGAGAT ACGAGACTGATT
AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT– 3’

806rcbc2

5’ – CAAGCAGAAGACGGCATACGAGAT GCTGTACGGATT
AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT– 3’

806rcbc3

5’ – CAAGCAGAAGACGGCATACGAGAT ATCACCAGGTGT
AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT– 3’

806rcbc4

5’ – CAAGCAGAAGACGGCATACGAGAT TGGTCAACGATA
AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT– 3’

806rcbc5

5’ – CAAGCAGAAGACGGCATACGAGAT ATCGCACAGTAA
AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT– 3’

806rcbc6

5’ – CAAGCAGAAGACGGCATACGAGAT GTCGTGTAGCCT
AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT– 3’

806rcbc7

5’ – CAAGCAGAAGACGGCATACGAGAT AGCGGAGGTTAG
AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT– 3’

Figure A3. 16S Amplicon 806 (reverse) GoLay barcoded PCR primer sequences. Each primer is followed by a barcode identifier. [Field number (space-delimited) and description]. 1. Reverse complement of 3' Illumina adapter, 2. Golay barcode, 3. Reverse primer pad, 4. Reverse primer linker, 5. Reverse primer (Caporaso *et al*., 2012).

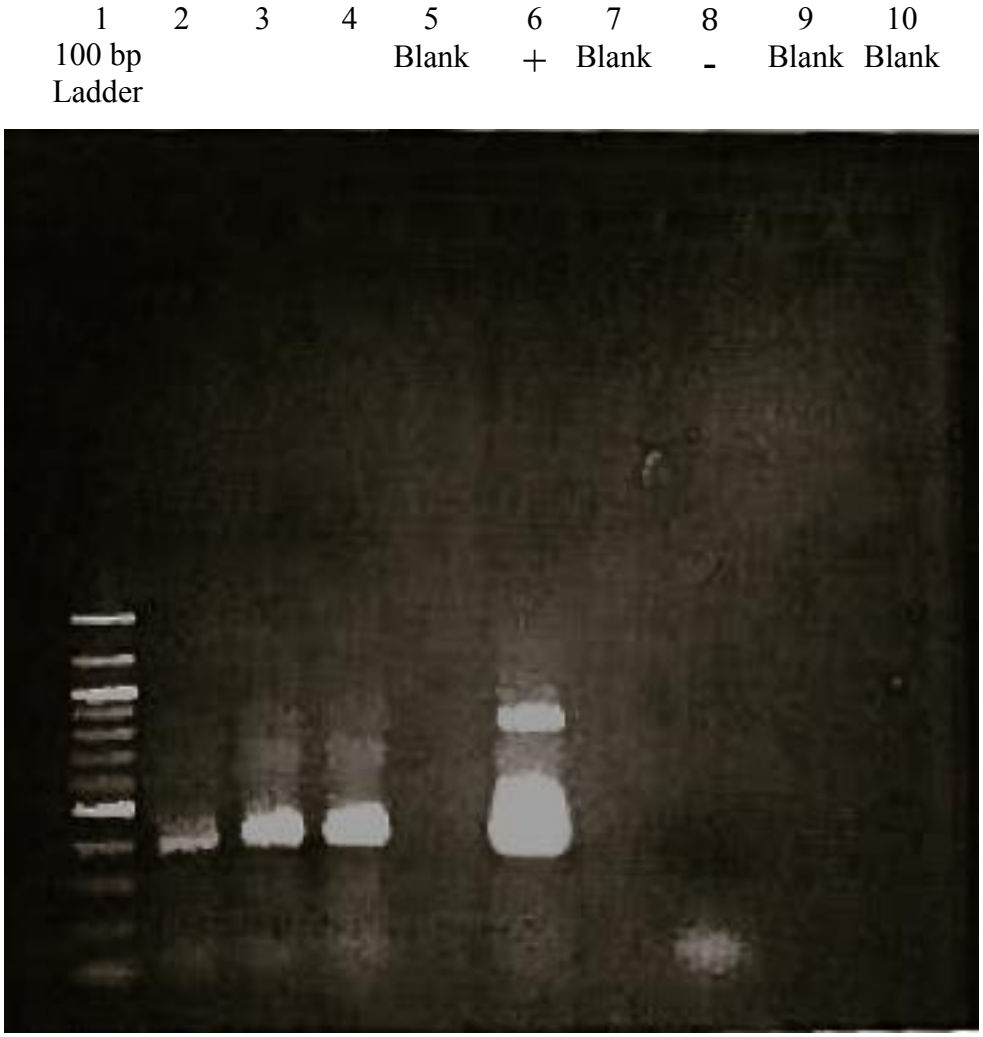

Figure A4. Photograph of amplified 16S product on a 1.5 % agarose gel run at 75 V for two hours (lane 1: 100 bp DNA Ladder; lane 2: Fritz 8/25; lane 3: Fritz 8/25; lane 4: Fritz 8/25; lane 5: blank; lane 6: positive control; lane 7: blank; lane 8: negative control; lane 9: blank; lane 10: blank). Confirmation of ~400 bp product in sample lanes and positive control. Note secondary amplicon product in positive control lane resulting from dimerization of 400 bp product during PCR (February 7, 2012).

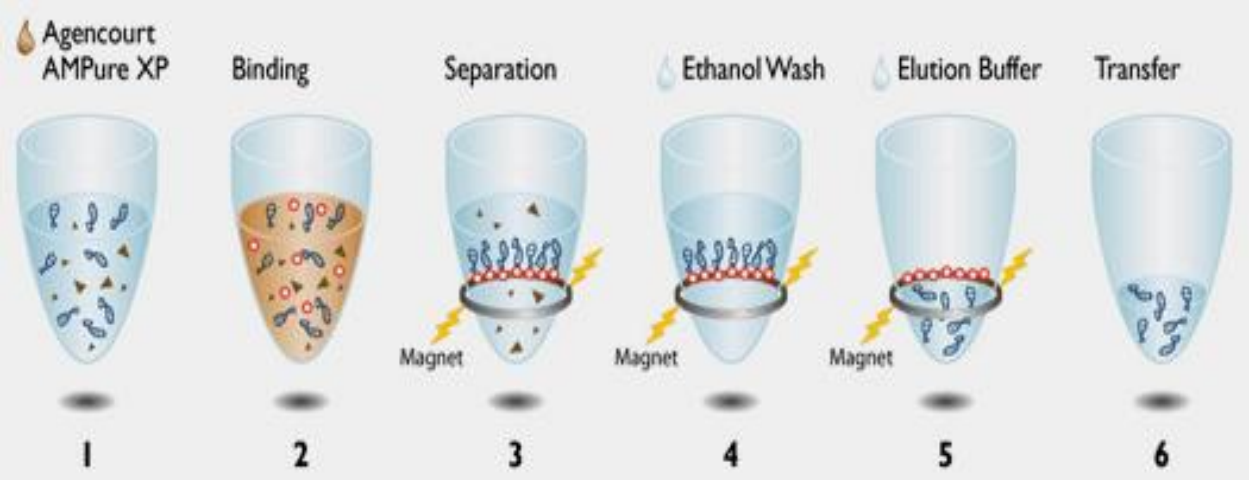

Figure A5. Agencourt AMPure XP Bead system overview (Beckman-Coulter, 2012).

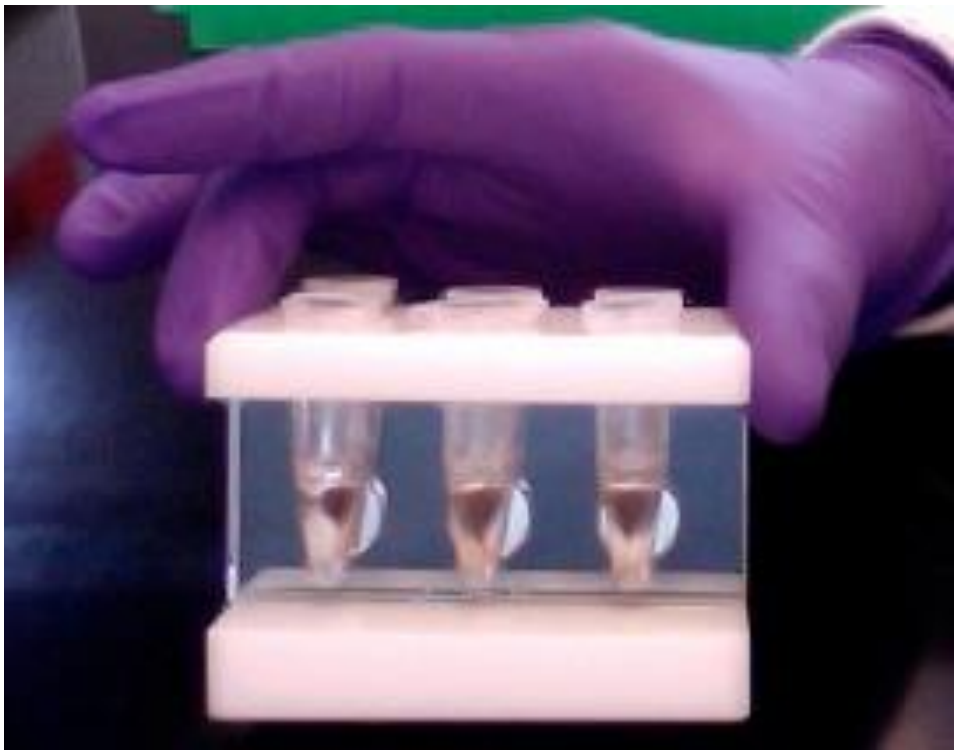

Figure A6. Photograph of Agencourt AMPure XP Bead system (Photo Credit: Weam Zaky, March 1, 2013).

(a) TATGGTAATT GT GTGCCAGCMGCCGCGGTAA

(b) AGTCAGTCAG CC GGACTACHVGGGTWTCTAAT

(c) ATTAGAWACCCBDGTAGTCC GG CTGACTGACT

Figure A7. 16S amplicon sequencing primers (a) Read 1 sequencing primer (b) Read 2 sequencing primer (c) Index sequencing primer. [Field number (space-delimited) and description]. 1. Primer pad, 2. Primer linker, 3. Primer (Caporaso *et al*., 2011).

APPENDIX B: Glossary of Terms

Closed-reference OTU picking	In a closed-reference OUT picking process, reads are clustered against a reference sequence collection and any reads which do no hit a sequence in the reference sequence collection are excluded from downstream analyses. If the user provides taxonomic assignments for sequences in the reference database, those are assigned to OTUs (versus open-reference OUT picking and De novo picking). (http://qiime.org/tutorials/otu_picking.html)
Colic	A term that indicates clinical signs of pain in the abdominal cavity; it is not a specific disease but rather a combination of signs that signal the presence of abdominal pain in horses; these signs can range from mild to severe and can rapidly become a life-threatening situation. (http://www.vetmed.ucdavis.edu/ceh/docs/)
Colitis	Inflammation of the large intestine (colon); can have many different causes, including: infections, including those caused by a virus, parasite, and food poisoning due to bacteria, inflammatory disorders (ulcerative colitis and Crohn's disease), and lack of blood flow (ischemic colitis). (www.ncbi.nlm.nih.gov)
Community clustering	Sequencing templates are immobilized on a proprietary flow cell surface designed to present the DNA in a manner that facilitates access to enzymes while ensuring high stability of surface-bound template and low non-specific binding of fluorescently labeled nucleotides. Solid-phase amplification creates up to 1,000 identical copies of each single template molecule in close proximity. (http://www.ucl.ac.uk/cancer/supportservices/SupportDocs/SS_DNAsequencing.pdf)
De novo OTU picking	In a de novo OTU picking process, reads are clustered against one another without any external reference sequence collection. It includes taxonomy assignment, sequence alignment, and tree-building steps. (http://qiime.org/tutorials/otu_picking.html)

Helminth	A general term meaning worm and may cause parasitic disease; helminths are invertebrates characterized by elongated, flat or round bodies. In medically oriented schemes the flatworms or platyhelminths (platy from the Greek root meaning "flat") include flukes and tapeworms. Roundworms are nematodes (nemato from the Greek root meaning "thread"). These groups are subdivided for convenience according to the host organ in which they reside, e.g., lung flukes, extraintestinal tapeworms, and intestinal roundworms. (www.ncbi.nlm.nih.gov/books/NBK8282/)
Nematode	Simple roundworms; Colorless, un-segmented, and lacking appendages, nematodes may be free-living, predaceous, or parasitic. Many of the parasitic species cause important diseases of plants, animals, and humans. Other species are beneficial in attacking insect pests, mostly sterilizing or otherwise debilitating their hosts. (www.biocontrol.entomology.cornell.edu/)
Open-reference OTU picking	In an open-reference OTU picking process, reads are clustered against a reference sequence collection and any reads which do no hit the reference sequence collection are subsequently clustered de novo. It includes taxonomy assignment, sequence alignment, and tree-building steps. (http://qiime.org/tutorials/otu_picking.html)
Operational Taxonomic Unit	A cluster of sequences based on a user-defined similarity threshold. Sequences that are similar at or above the threshold level are taken to represent the presence of a taxonomic unit (e.g., a genus) in the sequence collection. (http://qiime.org/scripts/pick_otus.html)