2015

It's a Small World: Biogeography and Invasion in the Mouse Intestine

Ellen Lark

IT'S A SMALL WORLD: BIOGEOGRAPHY AND INVASION IN THE MOUSE INTESTINE

By

ELLEN LARK

MS, Mississippi State University, Starkville, Mississippi, 2004
DVM, Mississippi State University, Starkville, Mississippi, 2003
BS, Washington State University, Pullman, Washington, 1993

Dissertation

presented in partial fulfillment of the requirements
for the degree of

Doctorate of Philosophy

in Cellular, Molecular and Microbial Biology

The University of Montana
Missoula, MT

December 2015

Approved by:

Sandy Ross, Dean of The Graduate School
Graduate School

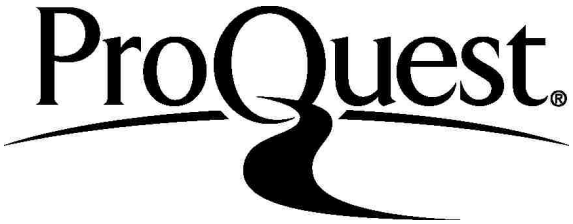
William E. Holben, Ph.D. Chair
Division of Biological Sciences

Douglas W. Raiford, Ph.D. Co-Chair
Department of Computer Science

Creagh W. Breuner, Ph.D.
Division of Biological Sciences

Scott R. Miller, Ph.D.
Division of Biological Sciences

L. Scott Mills, Ph.D.
Department of Forestry and Environmental Resources, North Carolina State University



Lark, Ellen, Ph.D., Fall 2015 Cellular, Molecular and Microbial Biology

It's a Small World: Biogeography and Invasion in the Mouse Intestine

Chairperson: William E. Holben, Ph.D.

Co-Chairperson: Douglas W. Raiford, Ph.D.

ABSTRACT: Host associated systems are of particular interest to many microbiologists because invasion of these systems can lead to disease. One important host-associated systems is the intestinal microbiome, but in many studies, including those on pathogenesis, this system is represented by samples from one location (generally the feces or cecum). This body of work was initiated in part because I wondered why a large and diverse ecosystem was being was being represented by samples from only one habitat.

The biogeography of living organisms has an impact on landscape ecology studies, including those in the field of invasion ecology. Despite several studies that specifically investigate the biogeography of the intestinal microbiome, there has been a general failure to describe the luminal biogeography of the lower intestinal tract, primarily due to "noise" introduced by inter-subject variation. Herein, the biogeography of the mouse lower intestinal tract was mapped using novel techniques to overcome problems caused by inter-subject variance. These techniques were then used to reveal nuances of invasion in the lower intestine by *Clostridium difficile*.

C. difficile is an invader of the intestinal microbiome that is well-known for its ability to cause disease following antibiotic treatment. I observed large changes with the introduction of antibiotics to this system, resulting in a series of "blooms" of various taxa, most likely an indication of successional changes due to the effects of antibiotics. I also found that without antibiotic treatment, *C. difficile,* is still associated with changes in the intestinal microbiome. This is an important development, as it suggests that small changes associated with normal colonization by introduced species may be compared with range expansion by the same species.

This body of work was primarily done in order to apply ecological theory to microbiome studies and in doing so gave rise to new techniques and new methods of looking at systems. It is my hope that these advances will result in contributions both to investigations of the intestinal microbiome as an ecological system as well as how as it relates to disease.

ACKNOWLEDGEMENTS

I wish to thank my committee for their support and encouragement. Without it this work would not have been possible. As a whole, I would like to thank them for their encouragement of a highly interdisciplinary project that took both patience and time on all our parts. The chair of my committee, Bill Holben has been an enthusiastic supporter of both my ideas and my work. His laboratory is a welcoming and a pleasant place to work (the view is incredible) and his door is always open. Doug Raiford, my co-chair is a wonderful, patient and encouraging person to work with. He truly believes that biologists serve as a resource for computer scientists and works hard to prove this. Creagh Breuner has been helpful, efficient and caring when needed. She is a true problem-solver and has helped both in committee sessions and office visits with advice that works. On many occasions, she has listened and then hit directly on a solution to a problem helping me to solve it more quickly. Scott Miller has been helpful both with questions about science and with some of the administrative problems that come with graduate school. He has asked some difficult questions that help me to think more clearly about my ideas, as well as my conclusions. Scott Mills is a poet. He has a way with words and an eidetic memory for all things ecological. He has made ecology come alive for me and thus encouraged my desire to do work that included ecology as well as medicine.

It has been a pleasure to work in the Holben Lab and I have been helped by many of the people I have met there, not the least of whom have been my confederates in crime, Sam Pannoni and Franny Gilman, both of whom have been supportive and caring. Their sense of humor has made grad school more bearable. Linda Hinze will tell you that she doesn't run the Holben Lab, but she does. Without her help several of these projects would not have gotten done. I owe her a great debt of gratitude. There have been many other people who have worked in the Holben lab and are gone doing other jobs and they were all helpful. One in particular who I'd like to mention is Marnie Rout. Her struggles with microbiology inspired me to do the same with ecology and she continues to be a friend and a valuable resource.

I have been lucky to interact with many undergraduates as a mentor. It was a pleasure to teach Phage Genomics, and HHMI course for 2 years. I met many great students and a couple of them worked on projects with me. Both Marlene Warner and Sam Pannoni worked with me on the C. difficile project and then went on to do their own work. Sam also worked with me on a project that involved the development of a novel technique. Inga Ortloff (also from phage genomics) and Elaina Weber both worked on a project to do with probiotics. Tampa Hutchins worked with me on the technique development project. Both Tampa and Elaina were funded by an NSF REU grant and did an amazing amount of work while they were here for a short while in the summer.

Finally, I'd like to acknowledge my friends and family. My father, Gordon Lark, in particular has been supportive and willing to talk science as well as read drafts of papers. I met Ylva Lekburg when taking her soil ecology course. She knew I wanted that course and provided it. She has remained a good friend. To all the others who helped along the way – Thank You!!

DEDICATION: To Cricket who was very, very patient and cared.

TABLE OF CONTENTS

Abstract...……..ii

Acknowledgements..……....iii

List of Tables...…….vi

List of Figures……..vii

CHAPTER 1: Introductory Overview of Biogeography and Invasion in the Mouse Intestine

Introduction……1

Glossary…….....1

Microbial Biogeography on Living Hosts……..……4

The Microbial Biogeography of Mammals……7

Invasion in the intestinal microbiome……...……20

Conclusions……..……28

Literature Cited……..……..32

CHAPTER 2: Development of a metagenomic DNA purification method for low-biomass samples and effective recovery of lactic acid bacterial DNA and the intestinal microbiome.*

Summary…….……..……..……...…42

Introduction……..……..……43

Experimental Procedures……...…..……48

Results……..……..……53

Discussion……...……...……59

Literature Cited……....……...……....64

CHAPTER 3: Location, location, location—Genus-level microbiome biogeography along the intestinal 'landscape'.

Abstract……...……..…….…….…….68

Background……...…….…….……..…..…….68

Materials and Methods……...……..…..…….71

Results……...…….…….…….79

Discussion……...……..…..…….87

Literature Cited……...…….……..…….90

Supplementary Figures……...…….……..93

CHAPTER 4: Carpe Diem – *C. difficile* invasion and the Intestinal Microbiome

Abstract…….……..……...…….103

Introduction……...……...……..103

Methods…….....……..106

Results…….……...……..…….....…….…...118

Discussion…….……....…….…..135

Literature Cited……..……...……..……...143

Supplementary Figures and Tables……..……..…...147

LIST OF TABLES

CHAPTER 2

Table 1: Left: Wilcoxin Rank Sum Results for LAB and genera found in all samples. Right: Genera found only in samples processed using one treatment

Table 2: Comparison of Bacteroidetes and LAB diversity between methods.

CHAPTER 3

Table 1: Selected genera visualized in Figs. 4 & 5*

Table 2: Core microbiome genera for each sampled lower intestinal tract location

CHAPTER 4

Table 1: List of experimental abbreviations used. Numbers following abbreviations denote experimental time points (refer to Fig. 1).

Supplementary Table 1: Table of PD_Whole_Tree, Chao1 and observed species diversity indices and averages.

LIST OF FIGURES

CHAPTER 2:

Figure 1: Average yields of DNA for each method by sample site.

Figure 2: Top: Abundance charts comparing evenness of the methods at each sampling site. Bottom: Side by side comparisons of mean number of reads for each extraction method for each sampling location.

Figure 3: Core microbiome indicated as a function of DNA recovery method. The smaller pie charts show the proportion of the core microbiome to non-core taxa (shown in white). The larger charts depict the core microbiome in more detail and indicate the number of reads for the more predominant genera. Top: MPS, Bottom: FTL

Figure 4: Core microbiome by sampling site and method. Small pie charts give the proportion of core to non-core (in white) for each sampling site/method.

CHAPTER 3:

Figure 1: Sampling sites along the lower intestinal tract.

Figure 2: Process flow diagram of the computational methods employed

Figure 3: Left: Two treatments in mice cannot be discriminated using a beta diversity measure (Unifrac) combined with PCoA. Right: Cages as the discriminating factor are shown, showing how cage effects can confound results.

Figure 4: Linear Discriminant Analysis (LDA) of the four main compartments sampled from C57Bl/6 strain mice (left panel) and CD-1 strain mice (right panel). Filled circles and open circles represent cohorts 1 and 2, respectively. ● Ileum, ● Caecum, ● Proximal Colon, ● Distal Colon. Black dots represent the centroid for each cluster and ellipses indicate 1 standard deviation. The arrows show the flow of digesta between chambers. The plots were made using vote-determined genera. The accuracies were 78.79% (62.12%) (left panel) and 63.93% (65.57%) (right panel). The first accuracies listed used vote-determined genera, while accuracies in parentheses were for genera identified using 'floating search within each fold'.

Figure 5: LDA of the Tip of the Cecum and Cecum (top panels) and Proximal, Mid and Distal Colon (bottom panels) for C57Bl/6 mice (left) and CD-1 mice (right). ● Cecum, ● Tip of the Cecum, ● Proximal Colon, ● Mid-Colon, ● Distal Colon. Filled circles and open circles represent cohorts 1 and 2, respectively. Black dots represent the centroid for each cluster and ellipses indicate 1 standard deviation. The plots were made using vote-determined general. The accuracies were 93.55% (77.42%) (top left), 71.88% (62.50%) (top right), 62.00% (52.00%) (bottom left), 58.70% (50.00%) (bottom right).

The first accuracies listed used vote-determined genera, while accuracies in parentheses were for genera identified using 'floating search within each fold'.

Supplementary Figure S1: Box plots showing cross-validation accuracies when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was Cohorts 1&2 with mouse strain B6 (top) and strain CD1 (bottom) filtered to 4 chambers: Ileum, Cecum, Proximal Colon, and Distal Colon. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature selection inside of cross-validation was compiled and used to select genera outside (before) cross-validation was performed.

Supplementary Figure S2: Box plots showing cross-validation accuracies when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was Cohorts 1&2 with mouse strain B6 (top) and strain CD1 (bottom) filtered to 2 chambers: Cecum and Tip of Cecum. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature selection inside of cross-validation was compiled and used to select genera outside (before) cross-validation was performed.

Supplementary Figure S3: Box plots showing cross-validation accuracies when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was Cohorts 1&2 with mouse strain B6 (top) and strain CD1 (bottom) filtered to 3 chambers: Proximal Colon, Mid Colon, and Distal Colon. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature selection inside of cross-validation was compiled and used to select genera outside (before) cross-validation was performed.

Supplementary Figure S4: Scatter plot of the 3D-Pareto frontiers when Supplementary Figure 1 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. **Green** points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.

Supplementary Figure S5: Scatter plot of the 3D-Pareto frontiers when Supplementary Figure 2 box plot data are optimized by median accuracy, lowest variance, and number of

dimensions. **Green** points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.

Supplementary Figure S6: Scatter plot of the 3D-Pareto frontiers when Supplementary Figure 3 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. **Green** points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.

Supplementary Figure S7: Example code listings for generation of LDA plots and cross validation, respectively.

CHAPTER 4:

Figure 1: Experimental groups and schematic diagram of the experimental set-up. The point of administration of *C. difficile* is day 0. The 6 sampling timepoints are indicated with asterisks (Chen et al., 2008).

Figure 2: Sampling sites along the mouse lower intestine.

Figure 3: Mean proportional differences in major phyla as a function of treatment. *Denotes antibiotic administration **Denotes *C. difficile* challenge. H_2O4 and ABX4 occur at the time of challenge, but the mice were not exposed (as indicated by bracketing asterisks), although they continued to be handled. H_2O: Negative controls (water or saline in place of treatments); ABX: Antibiotic controls; cH_2O: Challenged controls; cABX: Challenged with antibiotic treatment prior to *C. difficile* challenge; Van5: Vancomycin given for 4 days, starting 24 h after *C. difficile* challenge; Van6: Relapse after course of Vancomycin terminated.

Figure 4: Chao1 alpha diversity plots, showing alpha diversity for samples from different points in time. The plot for the Time1 (H2O1) controls is overlaid on each plot to enable comparisons. A: Timepoint 2, B: Timepoint 3, C: Timepoint 4, D: Timepoint 5, E: Timepoint 6. H2O: water treatment (negative controls); ABX: antibiotic treatment; Van5: vancomycin treatment; Van6: halting of vancomycin treatment (relapse).

Figure 5: Feature selection combined with LDA showing that within different sampling timepoints, treatments can be distinguished from one another. The plots were made using vote-determined genera. The first accuracies listed used vote-determined genera, while accuracies in parentheses were for genera identified using 'floating search within

each fold'. Black dots represent the centroid for each cluster and ellipses indicate 1 standard deviation. Cross-validation accuracies: A: Timepoint 2 accuracies: 83.33%(74.07%), 10 taxa; B. Timepoint 3 accuracies: 93.1% (74.14%), 12 taxa; C. Timepoint 4 accuracies: 81.51%(75.63%), 17 taxa; D. Timepoint 5 accuracies: 80.24%(68.86%),18 taxa; E. Timepoint 6 accuracies: 73.56%(81.03%), 18 taxa

Figure 6: Distribution of *Peptoclostridium difficile* reads across all sampling times/treatments

Figure 7: Distribution of *Clostridium* Group XI showing the more broad dispersal across treatments that implies that other members of Clostridium Group XI are present in addition to *C. difficile*.

Figure 8: Proportion of *Peptoclostridium* reads for individual mice from times 4, 5, and 6 showing variation of *Peptoclostridium* in individuals. Top: Mice at timepoint 4, previously treated with antibiotics, showing variable detection of *Peptoclostridium* between individual mice. Middle: *Peptoclostridium* was only detected in one mouse during vancomycin treatment (49.1). Bottom: Mice at timepoint 6 (during relapse). *Peptoclostridium* was detected in all mice, but variability between individuals can be seen. Mouse 49.1 (from timepoint5, during vancomycin treatment) is shown on the right to give an indication of the difference in the amount of *Peptoclostridium* found during vancomycin treatment and that found during relapse

Figure 9: Core plots of treatments. Genera represented in any core plot are present in all samples for that treatment at that timepoint. From Left to Right, columns are: timepoint1 (control), timepoint 2 (antibiotic treatment), timepoint 3 (clindamycin treatment), timepoint 4 (*C. difficile* challenge), timepoint 5 (vancomycin treatment), and timepoint 6 (Relapse (vancomycin removed). From top to bottom, rows represent treatments: 1. Negative controls (water only)), 2. antibiotic treatments only, 3. *C. difficile* challenge, 4. *C. difficile* challenge combined with antibiotic treatment, 5. Vancomycin treatment and relapse (far right bottom). The dark blue/purple in the relapse treatment is *Peptoclostridium*, the genus that contains *C. difficile* alongside *Escherichia*.

Figure 10: Distribution of *Lactobacillus* by treatment and location

Figure 11: Distribution of *Parabacteroides* by treatment and location

Figure 12: Distribution of *Bacteroides* by treatment and location

Figure 13: Distribution of *Alistipes* by treatment and location

Figure 14: Distribution of *Escherichia* by treatment and location

Supplementary Figure S1: Left: Unifrac for timepoint 1 and 4. Right Unifrac for total experiment.

Supplementary Figure S5: Scatter plot of the 3D-Pareto frontiers when Supplementary Figures 2 and 3 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. **Green** points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference. Box plots showing cross-validation accuracies when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was CDF time1 and 2 filtered to 2 treatments: Water and Antibiotics. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature selection inside of cross-validation was compiled and used to select genera outside (before) cross-validation was performed.

Supplementary Figure S3: Box plots showing cross-validation accuracies for 2 LDA runs when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was CDF filtered to 4 treatments: Water, Antibiotics Water, Cdiff and Antibiotics, Cdiff. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature selection inside of cross-validation was compiled and used to select genera outside (before) cross-validation was performed.

Supplementary Figure S4: Box plots showing cross-validation accuracies for 2 LDA runs when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was CDF filtered to 5 treatments: Water; Antibiotics; Water Cdiff, Antibiotics, Cdiff; Vancomycin, Cdiff. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature selection inside of cross-validation was compiled and used to select genera outside (before) cross-validation was performed.

Supplementary Figure S5: Scatter plot of the 3D-Pareto frontiers when Supplementary Figures 2 and 3 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. **Green** points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.

Supplementary Figure S6: Scatter plot of the 3D-Pareto frontiers when Supplementary Figures 4 and 5 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. **Green** points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.

Supplementary Figure S7: Scatter plot of the 3D-Pareto frontiers when Supplementary Figure 6 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. **Green** points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.

Supplementary Figure S8: Process flow diagram of the computational methods employed.

Supplementary Figure S9: The abundance and location of Lactobacillus spp. changes due to introduction of yogurt into the intestine. Crosshatching indicates *Lactobacillus delbrueckii*

Chapter 1: Introductory Overview of Biogeography and Invasion in the Mouse Intestine

Introduction:

Studying biogeography and invasion in a host-associated community system is important for several reasons: First of all, host-associated microbial communities (also called microbiomes) are colonized through natural processes and exhibit features of biogeography, invasion, and other ecological developments, and therefore are excellent model systems for examining basic ecological concepts at relatively small spatial and temporal scales. In addition, host–microbiome systems can be manipulated and replicated more readily than can most other ecosystems. Finally, these systems are relevant not only as ecosystem models, but medically as well, potentially leading to new therapeutics, prebiotics and probiotics to maintain host health and protect against invasion by pathogens. This manuscript is an attempt to provide a more detailed picture of the biogeography and invasion of the intestinal microbiome.

Glossary:

Microbial ecology integrates two disciplines: microbiology and ecology and microbial ecologists often try to employ key concepts and principles from general ecology to facilitate understanding of microbial systems. These two disciplines often have differing definitions for the terms used to describe biogeography and invasion. In part, this is because the history of microbiology in many ways is synonymous with that of medicine. In addition, microbes differ in many ways from macroorganisms. For instance,

they have shorter generation times, propagate differently from and are disseminated differently from macroorganisms. Because the body of work presented in this thesis seeks to explore how the mouse intestinal microbiome can be used as an ecosystem model, some ecological terms need to be contextually or operationally defined to understand how they apply to microbial systems:

Biogeography: Patterns of species abundance and distribution over space and time are commonly called biogeography. Thus, populations rather than communities are associated with biogeography. Here, because of the manner in which microbial communities are sampled and compared, using next generation sequencing techniques to identify community members together, the term, biogeography will be applied on the community level. There is some precedence for this with regard to systems in which the authors refer to community biogeography or to biogeographical provinces (Costello et al., 2009; Follows et al., 2007; Udvardy, 1975).

Microbiome: The term "microbiome" has been used to mean the total genetic component of the microorganisms associated with a host species (Hooper and Gordon, 2001). It has also been used to denote a characteristic microbial community associated with a specific habitat that has distinct physicochemical properties (Whipps, 1988). This manuscript examines only the bacterial communities associated with specific locations within the host (i.e. within specific compartments and sub-compartments of the mammalian lower gastrointestinal tract), thus the term "microbiome" will be used hereafter to denote bacterial communities that fit the latter definition.

Invasion: The fields of ecology and microbiology have differing definitions of what constitutes invasion. At this time, the most common ecological definition is that invasion is "the incursion of a novel organism in an ecosystem outside its host range and generally propagated by humans" (di Castri, 1990). Thus, an invasive organism is by definition, not a native (di Castri, 1990; Lockwood, 2013). In microbiology, the definition of invasion is that an organism must invade host tissues (Silva, 2012). Thus an invader could potentially be an organism native or endemic to an environment that invades a host also living in that environment (Ribet and Cossart, 2015; Todar, 2005). Additionally, as the term "colonization" merely implies the ability to live on and adhere to a host, both native and non-native microbes can colonize a given host, which can itself be thought of as an ecosystem. However, the first definition of invasion given above fails to integrate basic ecological theory regarding dispersal and colonization (succession) of a new environment by either natives or exotics (Davis et al., 2001). It also fails to account for natives (such as beavers or honey mesquite trees) that successfully expand their native ranges to become successful "invaders" in new areas adjacent to their former native range (Clements, 1991; Thompson et al., 1995; Wilson et al., 2001). Here, the ecological definition of 'invasion' will be integrated with basic community ecology to describe an organism, native or non-native that is able to successfully colonize the intestine and expand its range. Thus, in strictly medical terms *Clostridium difficile* is not invasive, but by ecological definition, with respect to the mouse intestinal microbiome, it is an invasive organism that is native to the intestine and may still successfully invade a new host or specific environment within the host intestinal tract.

Microbial Biogeography on Living Hosts:

General ecological theory predicts that different environments will select for different communities (Baas-Becking, 1934; Pocheville, 2015), producing patterns of species abundance and distribution that we call biogeography. Microorganisms have their own levels of community organization within environments that that may otherwise be abiotic, or may intersect with other living organisms. In common with other organisms, microbes disperse to environments in which they can live and multiply, The resulting patterns of distribution of microbial taxa (whether at species or higher levels, or even operationally defined) are the biogeography of microorganisms.

There are many microbial ecologists who might be surprised to hear that the concept of microbial biogeography is controversial (O'Malley, 2008; Whitfield, 2005). After all, species of microbes have differential distributions and some species cannot be found in some environments. Environmental factors affect microbes in a similar manner to the way they affect other organisms; however many biogeographers consider that biogeography is not just the study of how species are distributed at any given time, but must take into account the manner in which species dispersed and other geological, geographical and environmental factors affecting dispersion, endemicity and speciation (Fontaneto and Brodie., 2011; O'Malley, 2008).

Classical biogeography is the study of how organism dispersal is affected by geological features (Fontaneto and Brodie., 2011). Several studies have found that free-living microbes appear to be ubiquitous, with environmental factors not appearing to limit their dispersal (Finlay, 2002; Gibbons et al., 2013; MacDonell and Colwell, 1984).

This suggests that they are evenly dispersed and only environmental selection determines distribution. Thus, in terms of endemicity and speciation, there would be no biogeographical distribution of microbial organisms. These arguments are based in part on several studies done in aquatic systems (Finlay, 2002; Gibbons et al., 2013). Conversely, several studies also carried out in aquatic systems, have reached contradictory conclusions (Ghiglione et al., 2012; Hambright et al., 2015) finding that microbes are not necessarily ubiquitous and there are dispersal patterns indicating endemicity within specific areas.

Thus, microbes present a challenge to biogeographers. Their lineages are very old and they disperse easily, making it difficult to apply a dispersal pattern to modern species distributions. Because many microbes are considered 'cosmopolitan', it is difficult to determine whether geographical features are even factor in their distribution. Speciation is in microbial taxa is likewise difficult to determine due to horizontal gene transfer as well as the other attributes listed above (Rout, 2011).

Biogeographical analyses of microbes have also been hampered by problems of scale. A majority of the studies purporting to look at the biogeography of microorganisms do so at a human, or even super-human scale (Achtman, 2008; Finlay, 2002; Follows et al., 2007; Ghiglione et al., 2012; Hambright et al., 2015) rather than at the more intimate scale of the microorganisms themselves, where barriers for dispersal and distribution patterns might be more apparent. Analyses of the biogeography of microorganisms are further confounded by under-sampling, which stem from the inability to access the full diversity of any particular environment and the difficulty of identifying (i.e. classifying) many microorganisms found down to the species level.

Considering a living terrain as being composed of similar barriers to dispersal as a geological terrain is not a typically a consideration of biogeographers, but host associated microorganisms appear to demonstrate the existence of microbial biogeography. Most host organisms start by being either sterile or lightly colonized (Favier et al., 2002; Mitsuoka, 1996; Schaedler et al., 1965; Todar, 2005). This situation provides a perfect model for exploring the processes leading to biogeography, as any microbe that colonizes a new host must first disperse to that host, and then disperse on or within the host.

There are a number of suggested instances of how microbial biogeography applies to microbes living on other organisms (Achtman, 2008; Costello et al., 2009; Leff et al., 2015). It has been known for a long time that particular pathogens prefer certain environments. Tuberculosis affects the lungs. *Clostridium difficile* Associated Disease (CDAD) primarily affects the colon, while *Klebsiella* and *Shigella* prefer the environment of the small intestine. Examining how pathogens interact with a host on the scale of the individual host demonstrates that biogeography plays a part in host – microorganism interactions. Differential distributions of microbial pathogens associated with a host organism provide fundamental evidence that microbial biogeography exists and is important. Microorganisms are separated from the majority of the hosts' systems by a variety of physical and physiological barriers. If a barrier is breached, then the microorganism can, in theory, move throughout the host. Thus the host provides barriers in a similar manner to the barriers that exist to dispersal in the macro-environment. Beyond physical barriers to dispersal, it appears that there are preferred environments on or within the host, which microbial taxa find to be most optimal for their persistence.

Both physical barriers and preferred niches determine what communities exist both on different host organisms and also in different areas of a host organism (Todar, 2005).

While nutrient type and availability varies along the intestinal tract and has been considered to be a major factor driving microbial distribution (Gonzalez et al., 2011), there are other environmental factors that may limit or aid microbial dispersal in this system. Some of these include pH, the physical conformation of the gut, peristalsis, water and oxygen availability, as well as secretions from the host. These factors, as well as others comprise the barriers and niches that determine how far a microorganism will be able to disperse within the intestine and thereafter provide an environment that selects for those microorganisms best able to persist. Thus, microbiomes provide examples that free-living microorganisms can disperse and are prevented from doing so by environmental factors that are both biotic and abiotic. As such, microorganisms should be considered to exhibit biogeography.

The Microbial Biogeography of Mammals

In the last two decades, new tools have been developed that have revolutionized the field of microbial ecology. These techniques, which include new bench methods, computational tools and sequencing platforms, have expanded our knowledge of bacterial genetic and biochemical diversity and given us new insights into production of useful natural products, effects of bacterial communities on climate change, and biodegradation of harmful materials, as well as the potential health benefits from understanding roles and activities of the microbes that are host-associated. The health benefits to be derived from

understanding the human microbiome have been considered so potentially important that The National Institutes of Health (NIH) established the 5 year Human Microbiome Project (HMP) to investigate the impacts that the bacteria that live on and inside us have on our health (NIH HMP Working Group, 2009).

The HMP and other studies have sampled a variety of systems on and in humans and other animals. The basic considerations concerning biogeography from the microbiome of any physiological system can be applied to all of them. Most biogeographical studies focus on the skin, oral and intestinal microbiomes. Studies using animal models also focus on those microbiomes. For that reason, only a short synopsis of the skin and oral microbiomes are covered here.

Population and Small Group Oriented Studies of Biogeography

With the advent of culture independent studies to examine cross-sections of host associated bacterial communities, the idea of studying community biogeography of the host-associated landscape has become popular. Despite this, there are many biogeographical studies of the mammalian microbiome that use culture techniques to identify bacterial distribution patterns (Dubos and Schaedler, 1964; Keith et al., 1979; Lloyd et al., 1979; Montes and Wilborn, 1970). Of necessity, these examined the distribution of populations, or at most small groups, across the landscape provided by the host. In addition, there have been several studies done that use culture independent techniques such as Fluorescence in Situ Hybridization (FISH) to look at the biogeography of small groups of microbes or populations of one species in situ within the lower gut (Sarma-Rupavtarm et al., 2004; Swidsinski et al., 2005a; Swidsinski et al., 2005b). These

types of studies examined only small groups of bacterial taxa so they had an advantage that many later studies do not – They naturally eliminated the noise that is produced by analysis of a complete community. These population studies showed that there are differences in species patterns between different areas of the body, including distinct locations of the lower intestine in both humans and mice (Dubos et al., 1965; Sarma-Rupavtarm et al., 2004; Schaedler et al., 1965; Swidsinski et al., 2005a; Swidsinski et al., 2005b; Zilberstein et al., 2007).

Studies of the skin have distinguished between locations on the skin, as well as community differences due to the variety of niches that exist, such as hair/fur or sebaceous glands (Keith et al., 1979; Lloyd et al., 1979; Montes and Wilborn, 1970). In the oral cavity, distribution patterns have been shown to vary between the gingival plaque, the sub-gingival plaque, the teeth and other areas of the oral cavity (Gibbons and Van Houte, 1975; Minah et al., 1985; Van Houte et al., 1972). Although the intestinal tract is less accessible than either the skin or the mouth, its biogeography has also been explored using these methods (Sarma-Rupavtarm et al., 2004; Swidsinski et al., 2005a; Swidsinski et al., 2005b; Zilberstein et al., 2007). Mouse studies were among the first to confirm that the stomach has a microbiome and described differences in the microbiomes of discrete areas of the intestine (Dubos et al., 1965). The mouse studies are interesting in that they not only include culture-based methods but also include a variety of histochemical and microscopy protocols, including FISH. The FISH studies have been done more recently and sought to look at microbiome biogeography in situ within the gut. The first of these studies was done using mice (Swidsinski et al., 2005a), but the

technique has proven strong and has since then been used to distinguish between disease states in humans in conjunction with colonoscopy (Swidsinski et al., 2005b).

The main problem with the above techniques is that relatively few taxa can be explored. The culture-based methods suffer because only a minority of all environmental microbes can be cultured, even from within the gut. The FISH-based protocols suffer from the same problem, but with those methods, the problem is that a limited number of probes can be used without the probes interfering with one another. At this time, the limit is 6 – 12 organisms, but it is hoped that combinatoric techniques combined with better computational algorithms will enable more taxa to be investigated in situ (Valm et al., 2012). Despite this limitation, such techniques are useful for exploring the biogeography via bacterial populations of host-associated microorganisms.

Community Oriented Studies of Biogeography

The skin microbiome

As culture independent techniques have developed and become less expensive, they have become popular. They allow community analyses, on several levels such as population studies in situ using FISH (mentioned above), shot-gun analyses of potential functions, RNA analyses of gene expression and 16S rRNA (16S) analyses which give a "snapshot" of the community at a specific place and time. These methods have all been used to examine the mammalian microbiome.

One important lesson taken from these types of studies that have been done is that there is no definition of a healthy microbiome (Proctor, 2011). Instead, the definition of what is healthy for a host seems to vary between individuals, including identical twins

(Arumugam et al., 2011; Turnbaugh et al., 2009a). The HMP and other community studies have also reaffirmed what earlier studies had demonstrated – that different sites on the body have different bacterial communities (Costello et al., 2009; Ding and Schloss, 2014). Thus any one community on an individual host is specific to that host, as well as being specific for that location on the host. The definition of a healthy microbiome is highly individual and needs to be considered for the goals of personalized medicine. Indeed, medicine cannot only address how host genetics affect uptake of drugs, but must consider effects of the host's microbiome as well (Swanson, 2015). Despite this, the HMP and other studies have shown that it is possible to discern characteristics of disease states (Keku et al., 2014; Miyake and Yamamoto, 2013; Proctor, 2011; Rolig et al., 2013; Turnbaugh et al., 2009a). Some continuing concerns include whether a more diverse community protects the host from disease and if so what type(s) of disease, whether there are markers specific to the microbiome that delineate disease states/types and whether all changes to the host are reflected in its microbiome. These questions are dependent on the host landscape and therefore on the biogeography of the microbiome.

Our understanding of the biogeography of the skin microbiome and the oral microbiome benefited greatly from the HMP and other studies both associated with and independent of the HMP. Because both the skin and oral microbiomes can be accessed fairly easily, biogeographical studies have been fairly common. Studies of the skin have shown that there are a variety of environments on the skin surface that support their own communities and differ from one another due to differing environments. In addition specialized small niches of the skin have been explored and compared such as the skin

surface, sebaceous glands, sweat glands and hair follicles on both humans and mice and were found to have bacterial communities that differ from one another in many ways (Costello et al., 2009; Grice et al., 2009; Oh et al., 2014; Schommer and Gallo, 2013).

Oh et al., taking a metagenomic approach, describe differences in bacterial, viral and eukaryotic communities at different locations and niches provided by the skin (Oh et al., 2014). In particular, the sebaceous environment has a much greater viral component than do the other locations sampled, while the dry and moist environments seem to be more diverse than either the sebaceous or toenail environment. The latter could be seen as specialized environments that select for specific community elements. The populations of bacteria commonly found in most environments on the skin appear to have strain heterogeneity in the populations sampled across the landscape. In addition, differing communities also appear to have different functional potentials within the different environments – something that has not been suggested in other studies (Oh et al., 2014).

Recently, mass spectrometry in conjunction with 16S rRNA data has been used to build a landscape of microbial metabolites, human metabolites and the skin microbiome. This technique was used to look at approximately 400 sites on two subjects, making the sampling sites more continuous than in previous studies. One goal of the study was to correlate the microbiome with metabolites on the skin and build a 3D landscape giving the biogeography of the microbiome with regard to byproducts produced by bacteria on the skin. The group was not able to correlate community differences with metabolite differences, but they were able to associate some genera with metabolites on the skin surface. While it is still in its infancy, this technique shows great promise with regard to

being able to connect bacteria to their immediate environment; however a major problem with this type of study is the amount of computational effort needed (Bouslimani et al., 2015).

Oral microbiome

The oral cavity, or mouth, includes several distinct microbial habitats, such as teeth, gingival sulcus, attached gingiva, tongue, cheek, lip, hard palate, and soft palate. Similar results to studies on the skin have been obtained with differing communities found in different habitats (Dewhirst et al., 2010; Segata et al., 2012).

Groups studying the oral microbiome have developed several interesting techniques for both laboratory work and analysis. The first is that of Combinatorial Labeling and Spectral Imaging (CLASI) FISH in which two or more FISH probes are assigned to the same taxa and then combined during imaging analysis to discriminate between taxa in situ. This allows FISH of more than 6 phylotypes at one time (Valm et al., 2012). This technique, while promising, is still in early stages of development. Another technique, called oligotyping, allows discrimination of taxa at the strain level using the 16s rRNA gene. This technique is computational and was developed to get a resolution of taxa that goes beyond the genus level. Oligotyping uses the information that species and strain differences are generally associated with one area of the gene, while sequencing errors will be randomly distributed. This technique has been used to distinguish between hundreds of oral phylotypes, many of them at a species level or finer. Although, this still limits the number of taxa that can be identified at the species level, it allows the distribution of those species to be mapped (Eren et al., 2013).

In addition to the above techniques that are associated with sequence-based methods, a multivariate statistical approach to determining whether there are significant differences between communities has been developed. This technique is based on the consideration that Operational Taxonomic Units (OTUs) from microbial communities conform to a dirichlet distribution (Holmes et al., 2012). This technique can be used to test differences in treatments (such as location or disease). It also is used for power analyses that define the minimum number of subjects necessary to discriminate between treatments (La Rosa et al., 2012). In fact, power analyses done in this study showed that 20+ subjects would be needed to discriminate between locations in the oral cavity. As the community differences between locations appear to be stronger for the oral microbiome than the lower intestine, this implies that at least that many subjects should be used for intestinal studies as well. While this technique can be used for power analyses and hypothesis testing, it does not contain information about how communities differ from one another.

The skin and oral microbiomes have been sampled extensively. The communities have been identified down to the genus level and the biogeography of these areas explored in detail. In many cases, microbiome members have been identified down to the strain level, but this has been possible only for some community members. Other areas of the body that require invasive techniques, have seen less progress both on discriminating between locations and gaining a finer taxon resolution. One of these is the intestinal microbiome.

Biogeography of the Intestinal Microbiome

One of the mammalian microbiomes of primary interest is located within the host organism – the intestinal microbiome. The intestinal microbiome is considered so significant that the greatest number of returns from a search for "mammal and microbiome" on PubMed are articles on the intestinal microbiome (~7,500), and it has been suggested that it acts as a separate major organ within the host (Marchesi et al., 2015; Seksik and Landman, 2015; Swanson, 2015). In fact, recently some authors have used the term "human microbiome" (which refers to all environments on a human host) to mean the intestinal microbiome only (Beasley et al., 2015; Shen and Clemente, 2015). Yet we still don't have a good understanding of the composition and spatial organization of the various bacterial communities that make up the intestinal microbiome and how they are affected by exposure to allochthonous (introduced from outside) bacteria.

There are several reasons why our knowledge of the biogeography of intestinal microbiome has lagged behind that of other areas of the body. Some of these include that: 1. Fecal samples have been considered to be representative of the entire microbiome; 2. Samples other than fecal samples are generally much more difficult to obtain, requiring invasive techniques and extra expense in both humans and animal models; 3. Microbiome compositional differences (especially in the lower intestine) are swamped by 'noise' resulting from intersubject variation; and 4. Useful techniques for winnowing through the datasets to pick out relevant aspects of the information have lagged behind the techniques for generating large datasets. This has resulted in an incomplete picture of the biogeography of the human intestinal microbiome (Costello et al., 2009; Ding and Schloss, 2014; Segata et al., 2012), as well as for the intestinal microbiomes of several promising animal models (Isaacson and Kim, 2012; Turnbaugh et

al., 2009b; Yasuda et al., 2015). Further, this truncated picture of the biogeography of the intestinal microbiome likely affects our understanding of how bacteria that pass through the gut (i.e. transient populations) function to digest food, interact with more permanent members of the microbiome (i.e. resident populations) and potentially invade habitats within the gut.

It has been shown that the intestinal microbiome is a major barrier to colonization by bacterial pathogens (Stecher et al., 2013). In addition, it is established that pathogens have preferred locations within the intestine to invade (Aktan et al., 2007; Hoffmann et al., 2009), probably due to several factors, one of which may be variation in the intestinal microbiome by location. Thus, it has been suggested that the biogeography of host-associated bacteria needs to be delineated in order to examine the factors that contribute to colonization resistance, community changes due to both pathogens and non-pathogens, and better use of therapeutics (especially pre- and probiotics) (Costello et al., 2009).

That community differences in different locations of the intestine exist is suggested by the environmental differences that exist between different sections of the intestine. Changes in metabolites, nutrients and moisture content are visibly evident throughout the intestinal tract including the lower intestine. For instance, within the colon, digesta enter the colon as liquefied slurry but by the time the digesta exit the colon in most healthy mammals, it has solidified. There are color changes as well, indicating a different chemical compostion. pH also changes within the colon, with pH increasing from the proximal colon to the distal colon (Kawamata et al., 2006; Kohl et al., 2013). Thus, it should be possible to map community differences that occur with these changes. Nonetheless, this has yet to be achieved for the lower portion of the intestinal tract. Most

microbiome studies of the intestinal tract only examine one site as a proxy for the lower intestine because invasive techniques are needed to investigate the biogeography of the intestinal microbiome. The most popular sampling sites are the feces (humans, mice, wildlife), the tip of the cecum (mice) and the cecum (mice) (Arumugam et al., 2011; Cani et al., 2007; Ding and Schloss, 2014; Ley et al., 2005; Ley et al., 2008). Stool samples in particular are commonly used to represent the lower intestinal microbiome as a whole because they are easy to sample and inexpensive. While stool samples can and should be used to distinguish between broad health and disease states, there are other types of studies for which they are not so well suited. Studies that examine interactions within the microbiome, between the host and the microbiome or the microbiome and invasive organisms, should not use stool samples, as detailed analysis of either invasion or disease states are dependent on the biogeography of the intestinal microbiome (Costello et al., 2009; Schubert et al., 2015).

Eckburg (2005), based on characteristics of biopsies recovered during colonoscopy of healthy people indicated that the bacterial biogeography of the human colon appeared to be patchy and heterogenous (Eckburg et al., 2005), but human studies have suffered from several problems. One is that samples can only represent a small portion of the contents of the human intestine, or for that matter the mucosal layer. Another is that biogeographical studies use biopsies recovered during colonoscopy as samples. The preparation for colonoscopy requires purging of the intestinal contents using substances that are somewhat inflammatory. There is some evidence that this type of inflammation may change the composition of the intestinal microbiome (Jalanka et al., 2015). Finally, the process of getting samples from humans is highly invasive and

expensive, so most human studies are small (for instance the study by Eckberg et al above, has only 3 subjects), thereby automatically not allowing for discrimination between contiguous sites due to noise caused by intersubject variation (La Rosa et al., 2012).

Several recent studies of the human intestinal microbiome have been done that examine the biogeography of the lower intestinal tract. One of the studies sampled the cecum and rectum of 9 healthy subjects and also compared the lumen of the intestine to the mucous layer and epithelial layers using terminal restriction fragment length polymorphism (T-RFLP). This study found no differences in communities longitudinally but did find differences between luminal and mucosal communities (Lavelle et al., 2013). Two other studies with 10 healthy subjects apiece purported to find differences in communities located along the intestine. One of these used a variety of statistical techniques but the results were not convincing (Zhang et al., 2014). The second used correspondence analysis to examine differences based on location in the intestine as well as differences due to sex. While they found differences longitudinally within the intestine, the results merely state that certain genera and/or families are different between communities without showing clear differences (Aguirre de Carcer et al., 2011).

Despite advances in mouse and animal models as well as sampling techniques, very few animal studies have been done that examine the lower intestine in detail, taking samples both within compartments as well as between compartments. One mouse study took 1cm samples from the stomach to the cecum, but neglected to continue the fine resolution sampling for the cecum and colon (Turnbaugh et al., 2009b). Another sampled 3 sites in the lower intestine (cecum, proximal and distal colon) as a part of a study on the

effects of an infectious agent. The esophagus and stomach were found to have distinct differences, differences within the two compartments were not found. The small intestine was found to have community differences longitudinally which corresponded with nutrient digestion and the influx of bile elements and other digestive enzymes (Hoffmann et al., 2009). Both studies noted significant differences between luminal and mucosal samples. Despite strong visual evidence that the contents of the colon vary greatly between the proximal and distal portions of the colon, no longitudinal differences were found within the lower intestine at the genus level. A rhesus macaque study had similar results, at the genus level (Yasuda et al., 2015). Both human and animal studies have noted that differences between subjects were greater than longitudinal differences in the lower intestine (Eckburg et al., 2005; Hoffmann et al., 2009; Lavelle et al., 2013; Turnbaugh et al., 2009b; Yasuda et al., 2015).

Conclusion

It is interesting to note that despite the early advances made in understanding the biogeography of host-associated microbial communities, the field has not advanced much between 2008 and the present. In 2009, Costello et al. produced a study using 7 – 9 subjects and up to 27 body sites, most of them the skin. The Costello paper, clearly gives the following as recommendations based on their conclusions: 1. The body habitat should be specified when conducting in-patient microbial surveillance studies designed to examine the flow of normal and pathogenic organisms into and out of different body sites in patients and their health care providers; 2. The local biotic and abiotic conditions of subsites of a given body habitat should be determined to understand why some subsites

are more or less resistant to invasion; and 3. Those sites that are amenable to transplantation of microbial communities with natural or engineered metabolic capacities that would be beneficial to a host should be designated (Costello et al., 2009). In 2014, Ding and Schloss examined HMP data from 300 healthy adults and the 18 body sites sampled in the HMP. They came to the same conclusions regarding the biogeography of the human microbiome, as the earlier study (Ding and Schloss, 2014). Both studies examined biogeography over time as well as at different locations. The main difference is that the 2014 study included a larger number of subjects. Both studies therefore link knowledge of microbial biogeography to invasion and pathogenesis of disease.

Invasion in the intestinal microbiome

Many studies have been done on invasion of the intestinal tract by microorganisms. Most studies of invasion focus on pathogens because of the medical relevance. Generally, these studies examine direct interactions between the host and the invader. Here, the emphasis will be on interactions between invaders and the intestinal communities. One advantage to studying host-associated invasion is that the system is a naturally occuring intact ecosystem. As the invasion occurs over a shorter period of time than in larger systems, there can be many replicates of the experiments. By careful selection of an animal model, they occur over both time and spatial scales that are relevant to the microorganisms rather than the host as well as being observed in context of a natural ecosystem. The main advantage gained by studying invasion on a microscopic scale is that it is possible to examine how invaders gain a foothold in order

to successfully colonize, something that has been quite difficult to study at larger scales. Pathogenic invasion is well known both in domestic animals and humans, and the mechanisms of invasion in many cases have been explained.

Colonization

The first step of invasion, colonization, depends on many of the same mechanisms within the intestine as invasion of any landscape. In general, two basic factors of the invaded community are thought to impact colonization and invasion. The first, whether the diversity of the invaded community plays a part in invasion has been tested with ambiguous results (Kennedy et al., 2002; Levine and D'Antonio, 1999). The second, whether disturbance of the invaded community plays a part in invasion, is also to some degree controversial. This is due to differing types and amounts of disturbance having different effects on invasion (Lockwood, 2013). Several studies have indicated that perhaps the major factor in invasion is that the invaded habitat be similar to the native habitat (Fitzpatrick et al., 2007; Montemayor et al., 2015; Thompson et al., 1995).

Some invaders create environmental changes that help them to invade more easily. These types of invaders are called ecosystem engineers (Lockwood, 2013). Microbes are good ecosystem engineers. They are capable of using the host immune response as well as toxins that they release to effectively clear the ground so that they can increase their range. The mechanism by which *Salmonella typhimurium* achieves a foothold has been known for a while and in common with many other enteric pathogens depends on individual cells causing an inflammatory response through Type III secretion systems (TSS) and flagella (Ackermann et al., 2008; Ribet and Cossart, 2015).

Essentially *S. typhimurium,* causes an inflammatory response from the host that clears the indigenous community from that area. This sacrifices the *Salmonella* that cause the response as well commensal bacteria. *Salmonella* uses phenotypic noise to overcome this problem. Although the population may be clonal, and most of the population expresses TSS, a small proportion of the population do not and therefore are not attacked by the immune system. Thus Salmonella takes advantage of the variation in phenotypic traits by sacrificing a large portion of the infecting propagule and killing the commensal community while allowing a small portion of the population to persist and infect the host without competition from the indigenous community. The sacrifice of a portion of the population contributes to the good of the population and therefore is able to persist even though a proportion of the population dies. The mechanism of sacrificing a portion of the population is fairly common among bacterial pathogens. Similar enteric pathogens, such as *Escherichia coli* also use secretion factors to clear an area of competing indigenous community members. A number of other pathogens must release their inflammatory signals by cell lysis, but the mechanism is essentially the same. One of these is *Clostridium difficile*, which must lyse itself to release TcdA, a toxin that causes inflammation in the host cells. While C. difficile may use its ability to do this in order to compete with other commensals, it apparently relies on other forms of disturbance to spread throughout the intestine (Ackermann et al., 2008; Koenigsknecht et al., 2015; Lessa et al., 2012).

Pathogens often utilize types of disturbance that occur naturally or from sources foreign to the intestine. These disturbances may be quite noticeable or somewhat subtle. *C. difficile* utilizes the community disturbance caused by antibiotics(Koenigsknecht et al.,

2015; Lessa et al., 2012). Without antibiotics to cause a disturbance, it either passes through the intestine or may become a permanent member without causing disease. Studies of how antibiotics affect the intestinal microbiome show that there are large community shifts when an antibiotic is given. The community changes are in part dependent on the antibiotic given and generally last between 2 – 6 weeks before gradually shifting back towards the original community composition, without ever quite returning to the composition that was present before being given antibiotics (Hill et al., 2010; Jernberg et al., 2010). While some forms of antibiotics appear to be fairly innocuous, others may create large enough shifts in community composition that a pathogen such as *C. difficile* may invade. A recent study suggests that these types of community shifts combined with the original community composition may play a large part in whether *C. difficile* can successfully invade the large intestine (Schubert et al., 2015).

Nutritional changes also cause shifts in community composition that permit invasion of the intestine. *Campylobacter jejuni* is a pathogen that lives as a commensal in all species other than humans. In humans it is a pathogen. In mice, which have a colonization resistance to *C. jejuni*, alteration of the intestinal microbiome allows colonization by *C. jejuni*. By increasing the percentage of *E. coli* in mice, colonization resistance to *C. jejuni* is eliminated, resulting in successful colonization and symptoms of disease (Haag et al., 2012). While humanized mice normally have colonization resistance to Campylobacter, that can be overcome by feeding them human food. The change in diet causes changes in the intestinal microbiome, including an increase in proteobacteria, leading to successful invasion by *C. jejuni* (Bereswill et al., 2011).

There are other forms of disturbance that may also favor colonization by pathogens. For instance young, old and immune compromised animals are more prone to colonization by pathogens. In all these cases, the immune system does not function optimally. This might be seen as being comparable either to the enemy release hypothesis used to explain successful colonization and spread of macroorganisms in larger ecosystems or to a change in climate at the microscopic scale. Essentially, selective pressure on an organism is removed or mediated to favor the invasive organism and thus the invader can outcompete indigenous organisms that still must contend with that stress. A variety of pathogens that normally would not cause disease are considered pathogens to the elderly or people with autoimmune diseases including indigenous microorganisms that normally are not considered pathogens (Brode et al., 2015; Fernandez-Natal et al., 2015; Ihde and Armstrong, 1973).

One unique aspect of microscopic studies that is seen rarely or not at all at larger scales is invasion of one community by another community. Most people have heard about fecal transplants being used to treat *C. difficile*. Treating a person with recurrent C. difficile with another person's feces can reverse the community changes caused by the disease (Brandt, 2015; Konturek et al., 2015). Another example that is just as interesting is that of mixtures of "probiotic" organisms in yogurt, kefir and other fermented foods. These communities of bacteria have been marketed to the general public as healthful and people consume them every day. The two examples are interesting in that they are very different from one another. The bacterial communities that are eaten as probiotics, tend not to colonize the intestine and therefore to get their benefits, one has to eat them every

day. On the other hand, the community changes caused by fecal transplants have been shown to cause permanent changes in the host intestinal community.

Probiotic foods, particularly yogurt (known to be probiotic), introduce a small artificially created community into the intestinal tract. There have been several studies that examine how the probiotic bacteria function on their own when introduced into the intestine (Lee et al., 2013; Majamaa et al., 1995; Nyangale et al., 2015). In addition, there have been numerous studies indicating that the probiotic species in yogurt do confer benefits to the host (Nabavi et al., 2014; Pei et al., 2015). There is no indication that the probiotics used in yogurt are able to colonize the intestine or repopulate the gut after taking antibiotics (Derrien and van Hylckama Vlieg, 2015; Shahani and Ayebo, 1980). Despite this, probiotics may have an effect on antibiotic associated diarrhea, but the studies done have been small and are somewhat controversial (Fox et al., 2015; Patro-Golab et al., 2015; Shahani and Ayebo, 1980). Not many studies have been done on the bacteria from yogurt and how they interact with the intestinal microbiome. Until recently there was no real evidence that these bacteria modulated the intestinal microbiome directly (Eloe-Fadrosh et al., 2015).

Another use for community inoculations has been the treatment of disease. Recently, people with recurrent *C. difficile* have been treated with feces from people who were healthy. In most cases, the person treated has recovered. Due to the relatively small number of cases complications from this treatment are undetermined. Still, in one case, a patient underwent a fecal transplant to treat recurrent *C. difficile* and then gained a large amount of weight (Alang and Kelly, 2015). In the case of fecal transplant, it was found that the complex community of a healthy person would essentially change the

intestinal community created by *C. difficile* to one where *C. difficile* is unable to act as a pathogen. A more controlled study where mice were inoculated with 6 bacterial species found normally in the intestine, had the same result (Lawley et al., 2012). It is clear that a small number of bacteria are capable of restoring health under these circumstances. This may be because the intestines of people suffering from recurrent *C. difficile* are less diverse than those of healthy people (Gu et al., 2015).

Spread of invasion

For a successful invasion, an invader must not only gain a foothold within the landscape, but must either become integrated into the indigenous community or increase its range at the expense of the native community. Most studies of invasion in the intestinal microbiome have to do with colonization and establishment of the invader within the intestine and possible spread from intestine to other parts of the body of the host. Very few studies concern themselves with the spread of the invader to other portions of the intestine. As most invasions involve pathogens, this focus on colonization makes sense, as the purpose of understanding invasions by pathogens is either to interfere with colonization or to treat disease once it is established. Most enteric disease can be prevented by not allowing exposure of the host to the pathogen (for instance by providing a clean water source) so the focus is on prevention. Once a pathogen has established itself, stopping it or modulating its behavior is difficult. Typically for bacterial disease, treatment has consisted of antibiotic therapy but antibiotic resistance has become a problem with all pathogens treated this way. In addition, antibiotic resistance of the intestinal microbiome has also become a concern, as the resistance can potentially be transferred to pathogens within the gut (Jernberg et al., 2010).

Despite this, there are reasons to study the spread of invasion within the intestine. One of these is that antibiotic resistance has become common and is no longer a treatment option in many cases. Study of how a pathogen behaves after it gains a foothold may inform new therapies. Another is that not all colonization may end in disease and understanding why this happens may lead to understanding why colonization resistance varies between hosts. There is evidence that the host microbiome may be integral to this, with variation of the microbiome leading to differences in how invasion may progress (Schubert et al., 2015). Although there aren't many studies that explore how an invader spreads, a couple of them have shown that even highly virulent diseases which need only a few propagules to colonize the intestine have distinct patterns of invasion (Hoffmann et al., 2009; Koenigsknecht et al., 2015) and tend to reshape the intestinal microbiome as a whole. In addition, in some cases, recovery from an invasion depends on retaining certain members of the intestinal microbiome. In cases where essential members of the community are lacking, it appears that recovery is uncertain (David et al., 2015; Schubert et al., 2015).

By necessity, most studies of invasion focus on the invader. When commensals are considered, they may be considered peripherally and not in detail. Determination of which taxa are important at the genus and species level and what they contribute has not been successful. While some taxa are considered to be important, they are not found in all subjects, implying that either a specific function that they fulfill is absent or that the function is fulfilled by another taxon. Where absent, the host may or may not be susceptible to disease. Consortia of commensals are thought to be important, but in most cases taxon resolution is only taken to the family level. The protective effects of a

specific taxon may not be able to be resolved at that level. As many families, (indeed genera) contain both pathogens and probiotic organisms, family resolution may not be enough to fully understand deficits in the indigenous community that allow invasion. Comparison of functionality to try and understand whether certain taxa have functions that are interchangeable with other taxa in another host is difficult because of the amount of information available, while the lack of resolution gives a blurred picture at best of a functioning community.

Conclusions

Both studies of biogeography and of invasion in the gut have been hindered by several problems: 1. Cost restraints and invasive techniques have kept subject numbers low. 2. Laboratory methods are not universally effective in extracting DNA from all community members. 3. Intersubject variation in many cases overwhelms subtle distinctions between locations.

One problem is that due to cost restraints and invasive techniques, subject numbers in most studies are still very low, making discrimination between treatments difficult. This has been further obstructed by not having a method for doing power analyses for microbiome studies. With the recent development of a method for doing that, new studies should be able to determine how many subject they need before starting a study and plan accordingly (La Rosa et al., 2012). As the minimum number of subjects needed is often between 20 and 40, this may not change the problem of expense. One solution is for more laboratories to cooperate on laboratory methods so that analysis can be done over multiple data sets.

It is well known that differing laboratory methods contribute to perceived differences in community composition (Delmont et al., 2012; Salonen et al., 2010). In particular, reconciling the need to recover DNA from particular taxa while maintaining an accurate picture of the entire community has not been adequately addressed. This is an important problem in studies where the bacterial taxa of interest might be either recalcitrant to lysis by normal means or may consist of sporulating forms where discrimination between spores and vegetative forms may cause problems. This problem is being addressed by the development of laboratory methods that address these problems directly ((Kuske et al., 1998; Zhang et al., 2012) See chapter 2).

Noise created both by intersubject variation and large amounts of data is increased under conditions where there are few subjects. One common comment made on biogeographical studies of the lower intestine is that longitudinal variation in the lower intestine is less than intersubject variation (Lavelle et al., 2013; Turnbaugh et al., 2009b; Yasuda et al., 2015). This does not mean that longitudinal variation is negligible and can be ignored. Rather, it means that the noise created by differences between subjects swamps out the information about biogeography that might be important. This may result in small, but important community differences being lost. Computational analysis reduces the problem to some extent, but there are still problems with determining which data is relevant. While new computational tools exist, they have been primarily used to increase the numbers of sequences that can be analyzed and to provide analyses that may show differences between communities exposed to different medical treatments. This may be why biogeographical studies of the lower intestinal lumen fail to distinguish community differences between locations that have distinct visible environmental

differences. As discussed above, advances in distinguishing between subtle treatments have not occurred. For example, although differences between locations in the lower intestine may be observed, they cannot be discerned by typical methods of sequence analysis.

One method for getting around some of these problems is to reduce the resolution of the analysis. This is one reason for many analyses being done at the family, or even phylum level. This may not be helpful when different species or strains of bacteria perform differently from one another creating yet another form of noise. For instance different strains of *E. coli* and different species of *Lactobacillus* have differing impacts on the lower intestine. *E. coli* as a species consists of strains that cause acute and life threatening disease, as well as strains that are considered probiotic (Stecher et al., 2013). The genus Lactobacillus also contains species that cause disease as well as ones that are considered probiotic (Cannon et al., 2005). Using family level resolution means that discrimination between these forms is lost and that only broad differences between treatments may be discerned. Another method for dealing with the problems created by too much data is to reduce the amount of data gained. Thus, many studies use microarrays to determine what sequences are present. But this may result in missing important community components due to individual nature of the data.

As communities that have fewer members (such as the small intestine) seem to be amenable to analysis, the above problems may result from the complexity of the communities being analyzed. Perhaps typical ecological methods of community analysis cannot work with communities that are as complex as the ones in the lower intestine. Thus, some method of reducing the complexity without reducing resolution must be

found. In the next few chapters we propose new workflows that use tested methods. The first is used in the laboratory to find possible under-represented taxa within a community. The second is a new computational workflow that filters noisy data to find taxa that discriminate between treatments. These two workflows have allowed us to re-examine the luminal biogeography of the lower intestine, and to examine invasion of this community with regard to the biogeography of the lower intestine.

Literature Cited:

Achtman, M. (2008). Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annual review of microbiology* 62, 53-70.

Ackermann, M., Stecher, B., Freed, N.E., Songhet, P., Hardt, W.D., and Doebeli, M. (2008). Self-destructive cooperation mediated by phenotypic noise. *Nature* 454, 987-990.

Aguirre de Carcer, D., Cuiv, P.O., Wang, T., Kang, S., Worthley, D., Whitehall, V., Gordon, I., McSweeney, C., Leggett, B., and Morrison, M. (2011). Numerical ecology validates a biogeographical distribution and gender-based effect on mucosa-associated bacteria along the human colon. *The ISME Journal* 5, 801-809.

Aktan, I., La Ragione, R.M., and Woodward, M.J. (2007). Colonization, persistence, and tissue tropism of Escherichia coli O26 in conventionally reared weaned lambs. *Applied and Environmental Microbiology* 73, 691-698.

Alang, N., and Kelly, C.R. (2015). Weight gain after fecal microbiota transplantation. *Open Forum Infectious Diseases 2*, ofv004.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.M., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174-180.

Baas-Becking, L.G.M. (1934). *Geobiologie of Inleiding Tot de Milieukunde [Geobiology or Introduction to the Science of the Environment]* (The Hague, The Netherlands: W. P. Van Stockum and Zoon).

Beasley, D.E., Koltz, A.M., Lambert, J.E., Fierer, N., and Dunn, R.R. (2015). The Evolution of Stomach Acidity and Its Relevance to the Human Microbiome. *PloS One* 10, e0134116.

Bereswill, S., Plickert, R., Fischer, A., Kuhl, A.A., Loddenkemper, C., Batra, A., Siegmund, B., Gobel, U.B., and Heimesaat, M.M. (2011). What you eat is what you get: Novel Campylobacter models in the quadrangle relationship between nutrition, obesity, microbiota and susceptibility to infection. *European Journal of Microbiology & Immunology* 1, 237-248.

Bouslimani, A., Porto, C., Rath, C.M., Wang, M., Guo, Y., Gonzalez, A., Berg-Lyon, D., Ackermann, G., Moeller Christensen, G.J., Nakatsuji, T., et al. (2015). Molecular cartography of the human skin surface in 3D. *Proceedings of the National Academy of Sciences of the United States of America* 112, E2120-2129.

Brandt, L.J. (2015). Fecal Microbiota Transplant: Respice, Adspice, Prospice. *Journal of Clinical Gastroenterology* 49 Suppl 1, S65-68.

Brode, S.K., Jamieson, F.B., Ng, R., Campitelli, M.A., Kwong, J.C., Paterson, J.M., Li, P., Marchand-Austin, A., Bombardier, C., and Marras, T.K. (2015). Increased risk of mycobacterial infections associated with anti-rheumatic medications. *Thorax* 70, 677-682.

Cani, P.D., Neyrinck, A.M., Fava, F., Knauf, C., Burcelin, R.G., Tuohy, K.M., Gibson, G.R., and Delzenne, N.M. (2007). Selective increases of bifidobacteria in gut microflora improve high-fat-diet-induced diabetes in mice through a mechanism associated with endotoxaemia. *Diabetologia* 50, 2374-2383.

Cannon, J.P., Lee, T.A., Bolanos, J.T., and Danziger, L.H. (2005). Pathogenic relevance of Lactobacillus: a retrospective review of over 200 cases. European *Journal of Clinical Microbiology & Infectious Diseases* : official publication of the European Society of Clinical Microbiology 24, 31-40.

Clements, C. (1991). Beavers and riparian ecosystems. *Rangelands* 13, 277 - 279.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science* 326, 1694-1697.

David, L.A., Weil, A., Ryan, E.T., Calderwood, S.B., Harris, J.B., Chowdhury, F., Begum, Y., Qadri, F., LaRocque, R.C., and Turnbaugh, P.J. (2015). Gut microbial succession follows acute secretory diarrhea in humans. *mBio* 6, e00381-00315.

Davis, M.A., Thompson, K., and Grime, J.P. (2001). Charles S. Elton and the Dissociation of Invasion Ecology from the Rest of Ecology. *Diversity and Distributions* 7, 97-102.

Delmont, T.O., Prestat, E., Keegan, K.P., Faubladier, M., Robe, P., Clark, I.M., Pelletier, E., Hirsch, P.R., Meyer, F., Gilbert, J.A., et al. (2012). Structure, fluctuation and magnitude of a natural grassland soil metagenome. *The ISME Journal* 6, 1677-1687.

Derrien, M., and van Hylckama Vlieg, J.E. (2015). Fate, activity, and impact of ingested bacteria within the human gut microbiota. *Trends in Microbiology* 23, 354-366.

Dewhirst, F.E., Chen, T., Izard, J., Paster, B.J., Tanner, A.C., Yu, W.H., Lakshmanan, A., and Wade, W.G. (2010). The human oral microbiome. *Journal of Bacteriology* 192, 5002-5017.

di Castri, F. (1990). On invading species and invaded ecosystems: the interplay of historical chance and biological necessity. In *Biological Invasions in Europe and the Mediterranean Basin*, F. di Castri, A.J. Hansen, and M. Debussche, eds. (Springer Netherlands), pp. 3-16.

Ding, T., and Schloss, P.D. (2014). Dynamics and associations of microbial community types across the human body. *Nature* 509, 357-360.

Dubos, R., and Schaedler, R.W. (1964). The Digestive Tract as an Ecosystem. *The American Journal of the Medical Sciences* 248, 267-272.

Dubos, R., Schaedler, R.W., Costello, R., and Hoet, P. (1965). Indigenous, Normal, and Autochthonous Flora of the Gastrointestinal Tract. *The Journal of Experimental Medicine* 122, 67-76.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A. (2005). Diversity of the human intestinal microbial flora. *Science* 308, 1635-1638.

Eloe-Fadrosh, E.A., Brady, A., Crabtree, J., Drabek, E.F., Ma, B., Mahurkar, A., Ravel, J., Haverkamp, M., Fiorino, A.M., Botelho, C., et al. (2015). Functional dynamics of the gut microbiome in elderly people during probiotic consumption. *mBio* 6.

Eren, A.M., Maignien, L., Sul, W.J., Murphy, L.G., Grim, S.L., Morrison, H.G., and Sogin, M.L. (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution / British Ecological Society* 4.

Favier, C.F., Vaughan, E.E., De Vos, W.M., and Akkermans, A.D. (2002). Molecular monitoring of succession of bacterial communities in human neonates. *Applied and Environmental Microbiology* 68, 219-226.

Fernandez-Natal, M.I., Saez-Nieto, J.A., Medina-Pascual, M.J., Valdezate-Ramos, S., Guerra-Laso, J.M., Rodriguez-Pollan, R.H., and Soriano, F. (2015). First report of bacteremia by Janibacter terrae in humans. *Infection* 43, 103-106.

Finlay, B.J. (2002). Global dispersal of free-living microbial eukaryote species. *Science* 296, 1061-1063.

Fitzpatrick, M.C., Weltzin, J.F., Sanders, N.J., and Dunn, R.R. (2007). The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? Global Ecology and Biogeography 16, 24-33.
Follows, M.J., Dutkiewicz, S., Grant, S., and Chisholm, S.W. (2007). Emergent biogeography of microbial communities in a model ocean. *Science* 315, 1843-1846.

Fontaneto, D., and Brodie., J. (2011). Why biogeography of microorganisms? *Biogeography of Microscopic Organisms* (Cambridge University Press).

Fox, M.J., Ahuja, K.D., Robertson, I.K., Ball, M.J., and Eri, R.D. (2015). Can probiotic yogurt prevent diarrhoea in children on antibiotics? A double-blind, randomised, placebo-controlled study. *BMJ Open* 5, e006474.

Ghiglione, J.F., Galand, P.E., Pommier, T., Pedros-Alio, C., Maas, E.W., Bakker, K., Bertilson, S., Kirchmanj, D.L., Lovejoy, C., Yager, P.L., et al. (2012). Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America* 109, 17633-17638.

Gibbons, R.J., and Van Houte, J. (1975). Bacterial adherence in oral microbial ecology. *Annual Review of Microbiology* 29, 19-44.

Gibbons, S.M., Caporaso, J.G., Pirrung, M., Field, D., Knight, R., and Gilbert, J.A. (2013). Evidence for a persistent microbial seed bank throughout the global ocean. *Proceedings of the National Academy of Sciences of the United States of America* 110, 4651-4655.

Gonzalez, A., Clemente, J.C., Shade, A., Metcalf, J.L., Song, S., Prithiviraj, B., Palmer, B.E., and Knight, R. (2011). Our microbial selves: what ecology can teach us. *EMBO Reports* 12, 775-784.

Grice, E.A., Kong, H.H., Conlan, S., Deming, C.B., Davis, J., Young, A.C., Program, N.C.S., Bouffard, G.G., Blakesley, R.W., Murray, P.R., et al. (2009). Topographical and temporal diversity of the human skin microbiome. *Science* 324, 1190-1192.

Gu, S., Chen, Y., Zhang, X., Lu, H., Lv, T., Shen, P., Lv, L., Zheng, B., Jiang, X., and Li, L. (2015). Identification of key taxa that favor intestinal colonization of Clostridium difficile in an adult Chinese population. *Microbes and Infection* / Institut Pasteur.

Haag, L.M., Fischer, A., Otto, B., Plickert, R., Kuhl, A.A., Gobel, U.B., Bereswill, S., and Heimesaat, M.M. (2012). Intestinal microbiota shifts towards elevated commensal Escherichia coli loads abrogate colonization resistance against Campylobacter jejuni in mice. *PloS One* 7, e35988.

Hambright, K.D., Beyer, J.E., Easton, J.D., Zamor, R.M., Easton, A.C., and Hallidayschult, T.C. (2015). The niche of an invasive marine microbe in a subtropical freshwater impoundment. The ISME journal 9, 256-264.
Hill, D.A., Hoffmann, C., Abt, M.C., Du, Y., Kobuley, D., Kirn, T.J., Bushman, F.D., and Artis, D. (2010). Metagenomic analyses reveal antibiotic-induced temporal and spatial changes in intestinal microbiota with associated alterations in immune cell homeostasis. *Mucosal Immunology* 3, 148-158.

Hoffmann, C., Hill, D.A., Minkah, N., Kirn, T., Troy, A., Artis, D., and Bushman, F. (2009). Community-wide response of the gut microbiota to enteropathogenic Citrobacter rodentium infection revealed by deep sequencing. *Infection and Immunity* 77, 4668-4678.

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. PloS one 7, e30126.

Hooper, L.V., and Gordon, J.I. (2001). Commensal host-bacterial relationships in the gut. *Science* 292, 1115-1118.

Ihde, D.C., and Armstrong, D. (1973). Clinical spectrum of infection due to Bacillus species. *The American Journal of Medicine* 55, 839-845.

Isaacson, R., and Kim, H.B. (2012). The intestinal microbiome of the pig. *Animal Health Research Reviews* / Conference of Research Workers in Animal Diseases 13, 100-109.

Jalanka, J., Salonen, A., Salojarvi, J., Ritari, J., Immonen, O., Marciani, L., Gowland, P., Hoad, C., Garsed, K., Lam, C., et al. (2015). Effects of bowel cleansing on the intestinal microbiota. *Gut* 64, 1562-1568.

Jernberg, C., Lofmark, S., Edlund, C., and Jansson, J.K. (2010). Long-term impacts of antibiotic exposure on the human intestinal microbiota. *Microbiology* 156, 3216-3223.

Kawamata, K., Hayashi, H., and Suzuki, Y. (2006). Chloride-dependent bicarbonate secretion in the mouse large intestine. *Biomedical Research* 27, 15-21.

Keith, W.A., Jr., Smiljanic, R.J., Akers, W.A., and Keith, L.W. (1979). Uneven distribution of aerobic mesophilic bacteria on human skin. *Applied and Environmental Microbiology* 37, 345-347.

Keku, T.O., Dulal, S., Deveaux, A., Jovov, B., and Han, X. (2014). The Gastrointestinal Microbiota and Colorectal Cancer. *American Journal of Physiology Gastrointestinal and Liver Physiology*, ajpgi 00360 02012.

Kennedy, T.A., Naeem, S., Howe, K.M., Knops, J.M., Tilman, D., and Reich, P. (2002). Biodiversity as a barrier to ecological invasion. *Nature* 417, 636-638.

Koenigsknecht, M.J., Theriot, C.M., Bergin, I.L., Schumacher, C.A., Schloss, P.D., and Young, V.B. (2015). Dynamics and establishment of Clostridium difficile infection in the murine gastrointestinal tract. *Infection and Immunity* 83, 934-941.

Kohl, K.D., Stengel, A., Samuni-Blank, M., and Dearing, M.D. (2013). Effects of anatomy and diet on gastrointestinal pH in rodents. Journal of experimental zoology Part A, *Ecological Genetics and Physiology* 319, 225-229.

Konturek, P.C., Haziri, D., Brzozowski, T., Hess, T., Heyman, S., Kwiecien, S., Konturek, S.J., and Koziel, J. (2015). Emerging role of fecal microbiota therapy in the treatment of gastrointestinal and extra-gastrointestinal diseases. *Journal of Physiology and Pharmacology* : an official journal of the Polish Physiological Society 66, 483-491.

Kuske, C.R., Banton, K.L., Adorada, D.L., Stark, P.C., Hill, K.K., and Jackson, P.J. (1998). Small-Scale DNA Sample Preparation Method for Field PCR Detection of

Microbial Cells and Spores in Soil. *Applied and Environmental Microbiology* 64, 2463-2472.

La Rosa, P.S., Brooks, J.P., Deych, E., Boone, E.L., Edwards, D.J., Wang, Q., Sodergren, E., Weinstock, G., and Shannon, W.D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS One* 7, e52078.

Lavelle, A., Lennon, G., Docherty, N., Balfe, A., Mulcahy, H.E., Doherty, G., D, O.D., Hyland, J.M., Shanahan, F., Sheahan, K., et al. (2013). Depth-dependent differences in community structure of the human colonic microbiota in health. *PloS One* 8, e78835.

Lawley, T.D., Clare, S., Walker, A.W., Stares, M.D., Connor, T.R., Raisen, C., Goulding, D., Rad, R., Schreiber, F., Brandt, C., et al. (2012). Targeted restoration of the intestinal microbiota with a simple, defined bacteriotherapy resolves relapsing Clostridium difficile disease in mice. *PLoS Pathogens* 8, e1002995.

Lee, S.M., Donaldson, G.P., Mikulski, Z., Boyajian, S., Ley, K., and Mazmanian, S.K. (2013). Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* 501, 426-429.

Leff, J.W., Del Tredici, P., Friedman, W.E., and Fierer, N. (2015). Spatial structuring of bacterial communities within individual Ginkgo biloba trees. *Environmental Microbiology* 17, 2352-2361.

Lessa, F.C., Gould, C.V., and McDonald, L.C. (2012). Current status of Clostridium difficile infection epidemiology. *Clinical Infectious Diseases* : an official publication of the Infectious Diseases Society of America 55 Suppl 2, S65-70.

Levine, J.M., and D'Antonio, C.M. (1999). Elton revisited: a review of evidence linking diversity and invasibility. *Oikos*, 15-26.

Ley, R.E., Backhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., and Gordon, J.I. (2005). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America* 102, 11070-11075.

Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, J.S., Schlegel, M.L., Tucker, T.A., Schrenzel, M.D., Knight, R., et al. (2008). Evolution of mammals and their gut microbes. *Science* 320, 1647-1651.

Lloyd, D.H., Dick, W.D., and Jenkinson, D.M. (1979). Location of the microflora in the skin of cattle. *The British Veterinary Journal* 135, 519-526.

Lockwood, J.L., Hoopes, Martha F., Marchetti, Michael P. (2013). An introduction to invasion ecology. In *Invasion Ecology* (USA, UK: John Wiley and Sons, Ltd.), pp. 1 -- 23.

MacDonell, M.T., and Colwell, R.R. (1984). Identical 5S rRNA nucleotide sequence of Vibrio cholerae strains representing temporal, geographical, and ecological diversity. *Applied and Environmental Microbiology* 48, 119-121.

Majamaa, H., Isolauri, E., Saxelin, M., and Vesikari, T. (1995). Lactic acid bacteria in the treatment of acute rotavirus gastroenteritis. *Journal of Pediatric Gastroenterology and Nutrition* 20, 333-338.

Marchesi, J.R., Adams, D.H., Fava, F., Hermes, G.D., Hirschfield, G.M., Hold, G., Quraishi, M.N., Kinross, J., Smidt, H., Tuohy, K.M., et al. (2015). The gut microbiota and host health: a new clinical frontier. *Gut*.

Minah, G.E., Solomon, E.S., and Chu, K. (1985). The association between dietary sucrose consumption and microbial population shifts at six oral sites in man. *Archives of Oral Biology* 30, 397-401.

Mitsuoka, T. (1996). Intestinal flora and human health. *Asia Pacific Journal of Clinical Nutrition* 5, 2-9.

Miyake, Y., and Yamamoto, K. (2013). Role of gut microbiota in liver diseases. *Hepatology Research* : the official journal of the Japan Society of Hepatology 43, 139-146.

Montemayor, S.I., Dellape, P.M., and Melo, M.C. (2015). Predicting the potential invasion suitability of regions to cassava lacebug pests (Heteroptera: Tingidae: Vatiga spp.). *Bulletin of Entomological Research* 105, 173-181.

Montes, L.F., and Wilborn, W.H. (1970). Anatomical location of normal skin flora. *Archives of Dermatology* 101, 145-159.

Nabavi, S., Rafraf, M., Somi, M.H., Homayouni-Rad, A., and Asghari-Jafarabadi, M. (2014). Effects of probiotic yogurt consumption on metabolic factors in individuals with nonalcoholic fatty liver disease. *Journal of Dairy Science* 97, 7386-7393.

NIH HMP Working Group (2009). The NIH Human Microbiome Project. *Genome Research* 19, 2317-2323.

Nyangale, E.P., Farmer, S., Cash, H.A., Keller, D., Chernoff, D., and Gibson, G.R. (2015). Bacillus coagulans GBI-30, 6086 Modulates Faecalibacterium prausnitzii in Older Men and Women. *The Journal of Nutrition* 145, 1446-1452.

O'Malley, M.A. (2008). 'Everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography. *Studies in History and Philosophy of Biological and Biomedical sciences* 39, 314-325.

Oh, J., Byrd, A.L., Deming, C., Conlan, S., Program, N.C.S., Kong, H.H., and Segre, J.A. (2014). Biogeography and individuality shape function in the human skin metagenome. *Nature* 514, 59-64.

Patro-Golab, B., Shamir, R., and Szajewska, H. (2015). Yogurt for treating antibiotic-associated diarrhea: Systematic review and meta-analysis. *Nutrition* 31, 796-800.

Pei, R., Martin, D.A., DiMarco, D.M., and Bolling, B.W. (2015). Evidence for the Effects of Yogurt on Gut Health and Obesity. *Critical Reviews in Food Science and Nutrition,* 0.

Pocheville, A. (2015). The ecological niche: history and recent controversies. In Handbook of Evolutionary Thinking in the Sciences, T.H. Heams, Philippe; Lecointre, Guillaume et al., ed. (Dordrecht: Springer), pp. 547–586.
Proctor, L.M. (2011). The Human Microbiome Project in 2011 and beyond. *Cell Host & Microbe* 10, 287-291.

Ribet, D., and Cossart, P. (2015). How bacterial pathogens colonize their hosts and invade deeper tissues. *Microbes and Infection* 17, 173-183.

Rolig, A.S., Cech, C., Ahler, E., Carter, J.E., and Ottemann, K.M. (2013). The degree of Helicobacter pylori-triggered inflammation is manipulated by preinfection host microbiota. *Infection and Immunity* 81, 1382-1389.

Rout, M.E. (2011). The role of microbial endosymbionts in sorghum halepense invasions: evidence of a new invasion strategy, microbially enhanced competitive ability (MECA) (Department of Environmental Studies, Bangor University).

Salonen, A., Nikkila, J., Jalanka-Tuovinen, J., Immonen, O., Rajilic-Stojanovic, M., Kekkonen, R.A., Palva, A., and de Vos, W.M. (2010). Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *Journal of Microbiological Methods* 81, 127-134.

Sarma-Rupavtarm, R.B., Ge, Z., Schauer, D.B., Fox, J.G., and Polz, M.F. (2004). Spatial distribution and stability of the eight microbial species of the altered schaedler flora in the mouse gastrointestinal tract. *Applied and Environmental Microbiology* 70, 2791-2800.

Schaedler, R.W., Dubos, R., and Costello, R. (1965). The Development of the Bacterial Flora in the Gastrointestinal Tract of Mice. *The Journal of Experimental Medicine* 122, 59-66.

Schommer, N.N., and Gallo, R.L. (2013). Structure and function of the human skin microbiome. *Trends in Microbiology* 21, 660-668.

Schubert, A.M., Sinani, H., and Schloss, P.D. (2015). Antibiotic-Induced Alterations of the Murine Gut Microbiota and Subsequent Effects on Colonization Resistance against Clostridium difficile. *mBio* 6.

Segata, N., Haake, S.K., Mannon, P., Lemon, K.P., Waldron, L., Gevers, D., Huttenhower, C., and Izard, J. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biology* 13, R42.

Seksik, P., and Landman, C. (2015). Understanding Microbiome Data: A Primer for Clinicians. *Digestive Diseases* 33 Suppl 1, 11-16.

Shahani, K.M., and Ayebo, A.D. (1980). Role of dietary lactobacilli in gastrointestinal microecology. *The American Journal of Clinical Nutrition* 33, 2448-2457.

Shen, N., and Clemente, J.C. (2015). Engineering the Microbiome: a Novel Approach to Immunotherapy for Allergic and Immune Diseases. *Current Allergy and Asthma Reports* 15, 39.

Silva, M.T. (2012). Classical labeling of bacterial pathogens according to their lifestyle in the host: inconsistencies and alternatives. *Frontiers in Microbiology* 3, 71.

Stecher, B., Berry, D., and Loy, A. (2013). Colonization resistance and microbial ecophysiology: using gnotobiotic mouse models and single-cell technology to explore the intestinal jungle. *FEMS Microbiology Reviews* 37, 793-829.

Swanson, H.I. (2015). Drug Metabolism by the Host and Gut Microbiota: A Partnership or Rivalry? *Drug Metabolism and Disposition*: the biological fate of chemicals 43, 1499-1504.

Swidsinski, A., Loening-Baucke, V., Lochs, H., and Hale, L.P. (2005a). Spatial organization of bacterial flora in normal and inflamed intestine: a fluorescence in situ hybridization study in mice. *World Journal of Gastroenterology* : WJG 11, 1131-1140.

Swidsinski, A., Weber, J., Loening-Baucke, V., Hale, L.P., and Lochs, H. (2005b). Spatial organization and composition of the mucosal flora in patients with inflammatory bowel disease. *Journal of Clinical Microbiology* 43, 3380-3389.

Thompson, K., Hodgson, J.G., and Tim, C.G.R. (1995). Native and Alien Invasive Plants: More of the Same? *Ecography* 18, 390-402.

Todar, K. (2005). Todar's Online Textbook of Bacteriology. http://textbookofbacteriology.net/colonization.html. Accessed 24 Aug 2015

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009a). A core gut microbiome in obese and lean twins. *Nature* 457, 480-484.

Turnbaugh, P.J., Ridaura, V.K., Faith, J.J., Rey, F.E., Knight, R., and Gordon, J.I. (2009b). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine* 1, 6ra14.

Udvardy, M.D.F. (1975). A classification of the hiogeographical provinces of the world. In lUCN Occasional Paper (Morges, Switzerland: International Union for Conservation of Nature and Natural Resources).

Valm, A.M., Mark Welch, J.L., and Borisy, G.G. (2012). CLASI-FISH: principles of combinatorial labeling and spectral imaging. *Systematic and Applied Microbiology* 35, 496-502.

Van Houte, J., Gibbons, R.J., and Pulkkinen, A.J. (1972). Ecology of human oral lactobacilli. *Infection and Immunity* 6, 723-729.

Whipps, J.M., Lewis, K., Cooke, RC (1988). Mycoparasitism and plant disease. In *Fungi in Biological Control Systems* M.N. Burge, ed. (NY, USA: Manchester University Press), pp. 161 - 187.

Whitfield, J. (2005). Biogeography. Is everything everywhere? *Science* 310, 960-961.

Wilson, T.B., Webb, R.H., and Thompson, T.L. (2001). Mechanisms of range expansion and removal of mesquite in desert grasslands of the southwestern United States.

Yasuda, K., Oh, K., Ren, B., Tickle, T.L., Franzosa, E.A., Wachtman, L.M., Miller, A.D., Westmoreland, S.V., Mansfield, K.G., Vallender, E.J., et al. (2015). Biogeography of the intestinal mucosal and lumenal microbiome in the rhesus macaque. *Cell Host & Microbe* 17, 385-391.

Zhang, S.S., Chen, D., and Lu, Q. (2012). An improved protocol and a new grinding device for extraction of genomic DNA from microorganisms by a two-step extraction procedure. *Genetics and Molecular Research* : GMR 11, 1532-1543.

Zhang, Z., Geng, J., Tang, X., Fan, H., Xu, J., Wen, X., Ma, Z.S., and Shi, P. (2014). Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *The ISME Journal* 8, 881-893.

Zilberstein, B., Quintanilha, A.G., Santos, M.A., Pajecki, D., Moura, E.G., Alves, P.R., Maluf Filho, F., de Souza, J.A., and Gama-Rodrigues, J. (2007). Digestive tract microbiota in healthy volunteers. *Clinics* 62, 47-54.

Chapter 2: Development of a metagenomic DNA purification method for low-biomass samples and effective recovery of lactic acid bacterial DNA and the intestinal microbiome.*

Ellen Lark[1], Eric M. Spaulding[2], Sam Pannoni[3], Tampa O. Hutchins[5], Douglas W. Raiford[2,4], William E. Holben[1,4]

[1]Cellular, Molecular and Microbial Biology Program; [2]Department of Computer Science; [3]Department of Wildlife Biology; [4]Systems Ecology Program, University of Montana, Missoula, MT, USA 59812; [5]University of South Florida, Tampa, FL, 33620

*Manuscript is formatted for submission to the Journal of Microbiological Methods

Summary

Studies that are concerned with interactions between a microbiome and selected taxa of bacteria have a unique problem in that they need to provide both a broad 'view' of the total bacterial community sampled, and at the same time accurately represent the taxa of interest. Thus, the method of DNA purification is critical as it must adequately capture DNA from the total microbiome and at the same time very effectively capture genomic DNA from the taxa of interest. It has been shown previously that different DNA purification methods result in differing 'views' of bacterial communities based on effectiveness of lysis and recovery of DNA from all of the various taxa that comprise that microbiome. Here, we describe a novel method developed to efficiently obtain lactic acid bacterial DNA along with general microbiome DNA and compare it to a commonly

used commercial 'kit-based' method developed for purifying microbiome DNA. Our method provides far greater yields of total microbiome DNA as well as recovering a greater proportion of Lactic Acid Bacteria DNA, perhaps at the expense of DNA from easily lysed non-target bacterial taxa such certain Gram-negative bacteria. The kit-based method appears to primarily recover DNA from bacteria that are easily lysed, while underrepresenting more recalcitrant bacteria such as the Lactic Acid Bacteria.

Introduction

Microbiome studies are dependent on community ecology analyses to discriminate between treatments. As in most community studies, it is important to sample accurately and to understand where sampling bias occurs, so that it can be corrected for, or at least taken into consideration. If a given DNA recovery method affects sampling (i.e. is biased) then the results gained will not be strictly quantitative, but can still be quite useful in comparative analyses. This appears to be a truism, but sampling methods can be elaborate, especially in the field of microbial ecology, where the methods used downstream of the original sampling are, in effect, a form of sampling themselves. Thus, DNA purification, amplification and even sequencing are all biased forms of sampling that may affect how the original community is perceived and analyzed.

Most studies of bacterial communities rely on analysis of community composition as determined from some sort of phylogenetic classification of DNA sequence reads. It is well known that different DNA purification protocols give rise to different results regarding community assemblages – even when the original samples are considered

identical (Delmont et al., 2012; Kennedy et al., 2014; Salonen et al., 2010), although these differences may appear to be minor or negligible (Delmont et al., 2012).

Methods that purport to examine the effects of one or more taxa on a community (such as those examining effects of pathogens or probiotics) are especially vulnerable to differences in microbiome DNA recovery methods as even slight differences in efficiency of recovery may result in misrepresentation of the taxa of interest, or of total community composition. It is often difficult to determine in exactly what ways different methods result in different community profiles other than on a very broad level. For instance, previous studies have shown that DNA yields, alpha diversity measures and community compositions can differ when different methods are used on identical samples (Delmont et al., 2012; Maukonen et al., 2012; Salonen et al., 2010). Exactly how these differences should be interpreted and what they mean for any particular study is unknown; however, the protocol chosen should fit the study needs, providing a reasonably accurate representation of the community in question as well as an accurate representation of the taxa of interest. As more studies attempt to address questions regarding community and population interactions, growth, functionality and effects due to treatments, it is necessary to know that the methods used will be relevant to the study questions.

One problem with community and population sampling is that particular species of interest may be underrepresented. That is, they are either not detected, or are detected in reduced numbers not accurately reflecting their true population size. This is a common problem in ecological studies—when populations are small, sampling may be limited due to time or financial constraints, and/or species that are difficult to accurately sample (Gu

and Swihart, 2004; MacKenzie et al., 2004). This can be a problem in microbiome studies, where some areas of the body, such as the skin, present challenges to detecting particular species (e.g. those low in relative abundance or difficult to lyse) and therefore present the possibility of underrepresentation in the data obtained (Garcia-Garcera et al., 2013).

Another example highly relevant to the research presented herein is when working with intestinal samples of small laboratory animals, where, even though it is possible to sample the entire contents of the intestine, the samples do not yield equivalent amounts of DNA when different DNA recovery methods are used ((Ferrand et al., 2014); also see Figure 1). Some problems that may result in underrepresentation of particular microbes include failure to lyse sporulated bacteria as well as bacteria having cell walls that are resistant to ordinary lysis methods (Ferrand et al., 2014; Filippidou et al., 2015; Kuske et al., 1998; Zhang et al., 2012). Methods that produce higher yields of DNA, as well as DNA from greater numbers of recalcitrant and sporulating bacteria, would seem to be optimal for analysis of bacterial communities rather than methods that result in low yields of DNA. This is particularly true where individual samples have low or variable biomass between individuals. Nevertheless, there are caveats. Examples of such include where such methods may result in misrepresentation of the remainder of the community, or where kit-based protocols are much faster and less expensive yet are suitable for use in comparative analyses. Thus, the method used for recovery of microbiome DNA should be selected to fit the purpose of the study.

The mammalian intestinal microbiome consists of a wide variety of organisms that vary depending on time of day, nutritional status, and other environmental factors

(Arumugam et al., 2011; Zarrinpar et al., 2014). Changes in the intestinal microbiome have been associated with changes in the health status of the host (Earley et al., 2015; Henao-Mejia et al., 2013; Turnbaugh et al., 2006; West et al., 2014). The intestinal microbiome has been shown to affect not only the intestinal tract, but other organ systems within the body, including the liver, circulatory system and nervous system (El Aidy et al., 2016; Ma et al., 2015; Vinje et al., 2014).

Profound changes to the intestinal microbiome may be induced via food, water or therapeutic regimens (Derrien and van Hylckama Vlieg, 2015; Hill et al., 2010; Jernberg et al., 2010; Wolf et al., 2014). One method by which these changes may be effected is via the introduction of new microbial taxa that interact with the microbiome in any of a variety of ways. For example, many pathogens affect both the microbiome and the host (Ackermann et al., 2008; Ribet and Cossart, 2015). Another example is the introduction of probiotic bacteria, commonly introduced via food, as well as on their own (Eloe-Fadrosh et al., 2015).

One of the largest groups of probiotic bacteria is the lactic acid bacteria (LAB) that are found in many cultured or fermented foods such as yogurt. In addition, the LAB typically form a large proportion of the autochthanous (indigenous) flora of the gut in many mammals (Hooda et al., 2012; Isaacson and Kim, 2012; Minamoto et al., 2012; Reuter, 2001; Tomas et al., 2012). Having an accurate representation of the composition and abundance of LAB in the gut is essential both to studies of the effects of probiotic species on the intestinal microbiome and also to studies where LAB make up a significant proportion of the microbiome or perform essential community functions.

LAB are very difficult to lyse or even permeate for analysis with oligonucleotide probes due to the thickness and density of their cell walls (Quevedo et al., 2011; Scornec et al., 2014). Thus evaluation of the identity and numbers of LAB in a bacterial community, or an evaluation of how introduced LAB interact with rest of the microbiome is dependent on the use of a metagenomic DNA recovery method that is able to effectively lyse these bacteria. A search of the primary literature found that many DNA recovery methods rely on the enzyme lysozyme to weaken or lyse LAB cell walls in order to permeate or lyse them (Bianchi et al., 2004; Brown et al., 1962; Chassy and Giuffrida, 1980; Ferrand et al., 2014; Quevedo et al., 2011; Scornec et al., 2014). Thus, we hypothesized that a metagenomic DNA recovery method that included lysozyme to lyse the cells comprising the microbial community should result in more complete lysis (and subsequent detection) of LAB than other physical and/or chemical lysis methods. In addition, initial work with luminal contents from various locations along the mouse lower intestinal tract showed that the most widely used kit for microbiome DNA recovery did not reliably obtain microbiome DNA from many of these samples, suggesting low efficiency of recovery with that approach. Since mouse models are central to essentially all of the studies presented herein, and LAB were key foci in some of the work, we deemed it necessary to develop a superior method for recovery of metagenomic DNA from mouse luminal microbiome samples including LAB.

To accomplish this, we modified an earlier protocol devised by this group for recovery of microbiome DNA from intestinal microbiome samples from various animals (Apajalahti et al., 1998). The earlier protocol combines freezing and thawing with chemical lysis by lysozyme (hereafter, FTL) to break bacterial cell walls. The

modificiation was to subsequently purify metagenomic DNA from this process using anion exchange columns from Qiagen (Valencia, CA) rather than the cesium chloride-ethidium bromide equilibrium density gradients employed in the original protocol.

This modified protocol was directly compared to use of the MoBio Power Soil kit (hereafter MPS), which is very widely used in the microbiome research community for extraction of metagenomic DNA from fecal samples. MPS uses bead beating combined with a proprietary surfactant as well as SDS for lysis of bacterial cell walls, followed by purification using proprietary silicon membrane columns. The FTL extraction method resulted in greater total yields of metagenomic DNA, as well as enhanced detection of LAB due to more effective recovery of DNA from these cells.

Experimental Procedures:

Animal Models and Sampling:

We employed laboratory rats for this study in order to reliably have larger individual intestinal samples that could be split in two to allow direct comparison of results from the two protocols. All animal care and treatment protocols were approved by the Institutional Animal Care and Use Committee (IACUC) at the University of Montana under AUP# LAR 009-11. All animals were housed, fed and otherwise treated identically in order to control for variability that might result from differing environmental conditions. Samples of opportunity were taken from eight 6 – 8 week old Dolly-Sprague rats from 3 litters that were euthanized by The University of Montana Laboratory Animal Facility. Luminal intestinal samples were taken from three locations in the lower intestine: the cecum (hereafter, Ce), the proximal colon (hereafter, PC) and the distal colon (hereafter, DC), resulting in 24 samples.

The luminal contents of surgically removed samples were recovered by cutting out the intestinal section with the contents intact and placing each into an individual sterile microcentrifuge tube. All samples were immediately placed on ice during collection and then stored at -70 ˚C prior to downstream processing. The three samples from within each rat were considered to be linked and therefore were processed for microbiome DNA recovery, PCR amplification and amplicon sequencing together. All rats were given a number and sets of samples were identified by that number and processed randomly using a random number table (Rand Corporation, 2001).

Microbiome DNA Recovery, Amplification and Sequencing:

Microbiome DNA was recovered from each sample using each of the two protocols: 1. The FTL protocol uses differential centrifugation, five freeze-thaw cycles and lysozyme treatment combined with the Qiagen genomic tip protocol. This is a modification of the Apajalahti *et al*. protocol (Apajalahti et al., 1998) that was shown to provide highly effective recovery of bacterial DNA from chicken GI tract. 2. The MPS protocol was performed exactly according to the manufacturer's protocol including the additional steps for recalcitrant samples.

In order to provide identical samples for both methods, the digesta were first cut from the intestinal tissue aseptically and cut in half lengthwise in sterile dishes on ice to keep them frozen. The half-samples to be processed that day were kept on ice, while the other halves of the samples were immediately placed back at -70˚ C until further processing. All samples were weighed before further processing. Sample weights and

nanophotometer (Implen P 300, Implen, Inc., Westlake Village, CA) readings were used to calculate DNA yields.

FTL digesta samples were placed into sterile Oak Ridge tubes containing 10 mL of sterile wash buffer (0.5 M sodium phosphate [pH 8.0]; 0.1% Tween-80) and washed 4 times as follows. Samples were vortexed briefly before being shaken at high speed on a reciprocating shaker for 10 min. Next, samples were centrifuged at 30,000 x *g* for 15 min at room temperature, after which the supernatant was removed by aspiration and the samples resuspended in 10 ml of wash buffer. Following the final wash step and centrifugation, the samples were resuspended in 3 ml of Qiagen Buffer B1 (50 mM sodium EDTA; 50 mM Tris base [pH 8.0]; 0.5% Tween-20; 0.5% Triton X-100; Qiagen, Valencia, CA) to which RNase A was added to a final concentration of 200 μg/ml, then stored at -70˚C to initiate the 5 freeze-thaw cycles that facilitate bacterial cell lysis. The samples were thawed and refrozen a total of 5 times by being placed in a water bath at 40˚C for 15 min, then placed back at -70 ˚C for at least 1 h before being thawed again. Following the final thaw, 50 μL of lysozyme (200 mg/ml) and 90 μL of proteinase K (20 mg/ml) were added. Samples were then incubated in a water bath at 37˚C for 45 min, after which 1 mL of Qiagen B2 buffer (3 M guanidine HCl, 20% Tween-20) was added. The samples were incubated in a water bath at 50˚C for 45 min, then centrifuged for 10 min at 5,000 x *g* at 4˚C. The supernatant was transferred to a sterile microcentrifuge tube and vortexed for 10 sec. At this point, the Qiagen Genomic Tip 20G protocol was followed precisely to elute microbiome DNA, except that 1 extra 70% ethanol wash was performed. Finally, the dried samples were resuspended in 50 ml of TE (10 mM Tris [pH 8.0], 1 mM EDTA).

MPS samples were processed exactly according to directions contained in the MoBio PowerSoil kit with the following modifications. During the bead-beating step, the samples were beaten for a total of 15 minutes (5 extra minutes) as per the manufacturer's instructions for lysis of potentially recalcitrant material. Sterile TE was used to elute the DNA at the final step. Both sets of samples were quantified using the Implen nanophotometer.

Partial 16S/18S rRNA gene sequences encompassing regions V4 & V5 were PCR amplified from the microbiome DNA (25 ng) using the highly conserved primers 536f and 907r (*25*), which were barcoded for pyrosequencing. Where samples were not of sufficient concentration to provide 25 ng for PCR, 3 µl of purified microbiome DNA was used as template. The resulting 16S amplicons were gel purified from 18S and other spurious products using the Qiagen Gel Purification kit per manufacturer's instructions, then further purified through two successive rounds using Agencourt AMPure XP magnetic beads per manufacturer's instructions (Beckman Coulter Inc., Brea CA). Purified DNA was quantified, multiplexed, and sequenced by the Utah State University Center for Integrated Biosystems using the 454 Roche GS FLX system (454 Life Sciences, Roche Diagnostics, Indianapolis, IN).

Data Analysis and Statistics:

Determination of DNA yields

DNA yields for each sample were calculated using the weight of the sample in grams and the DNA concentration (mg/ml) as determined by spectrophotometry.

Identification of OTUs and taxa summary tables

Sequences were 'denoised' and identified to the genus level using the Quantitative Insights Into Microbial Ecology (QIIME) pipeline (Caporaso et al., 2010) and the number of reads per sample determined. OTUs were discriminated at 97% similarity.

A taxa summary table was built containing a row for each sample (one for each locational sampling point for each mouse) and a column for each genus identified. The entries in the matrix were the number of reads in the sample represented by the associated genus. Any reads that were unclassified at the genus level were removed from further consideration. Frequency matrices were created for both cohorts. A further matrix was made using the proportion of reads following the same method. This matrix was used to determine the proportions of the pie charts for the core microbiome (see Fig. 4).

Test for Differences in Population Means

The two-tailed Wilcoxin Rank Sum Test was used to test whether there were differences in the population means for 5 genera that were found in all samples. *Streptococcus* was added to the analysis even though it does not occur in all samples, because it is one of the LAB. A Bonferroni correction was used and the alpha-level was set at 0.005 (Table 1).

Abundance Plots and Bar Charts

Abundance plots of treatments at all locations were made using the mean number of reads for all genera for each sampling site (location). These were then separated by method. A histogram was made using the mean number of reads for each genus in each

location and method. For each location, those genera that have zero reads for a method were marked with an asterisk (see Figure 2).

For side-by-side bar charts of mean abundance, the number of reads for all genera at any sampling location were found and then separated by method. The arithmetic mean was calculated for each genus at each location and method. The plots were made using a $\log_{10}$ transform for the x-axis.

Core Microbiome Determination

The core microbiome for each of the treatments as well as sampling locations within the lower GI tract was determined to the genus level. A genus was considered a part of the core microbiome if all samples for that treatment (whether by method or location) from all animals contained that genus.

Diversity of Bacteroidetes and LAB

The diversity of the phylum Bacteroidetes and of LAB for both methods was determined by using the taxa summary tables produced from the QIIME pipeline. LAB were selected for the reasons described above, which include resistance to lysis, presence in the intestinal microbiome and biological significance. Bacteroidetes were selected for this analysis because they are easily lysed, present in large numbers within intestine and of biological significance (Maukonen et. al., 2012, Turnbaugh et. al., 2006). Taxa of interest were transferred to Excel where the sums of the proportions in all samples for each taxon were calculated. Taxa were considered to be present in a method if the sum was greater than zero.

Results:

DNA yields were significantly greater in the cecum and distal colon for FTL than for MPS (Fig. 1) ranging from 1 – 2 orders of magnitude greater than those of MPS. The yields for FTL were also more variable, especially for the proximal colon, where the range of the yields is almost 1 order of magnitude (with a commensurate loss of statistical significance between the two methods for this chamber).

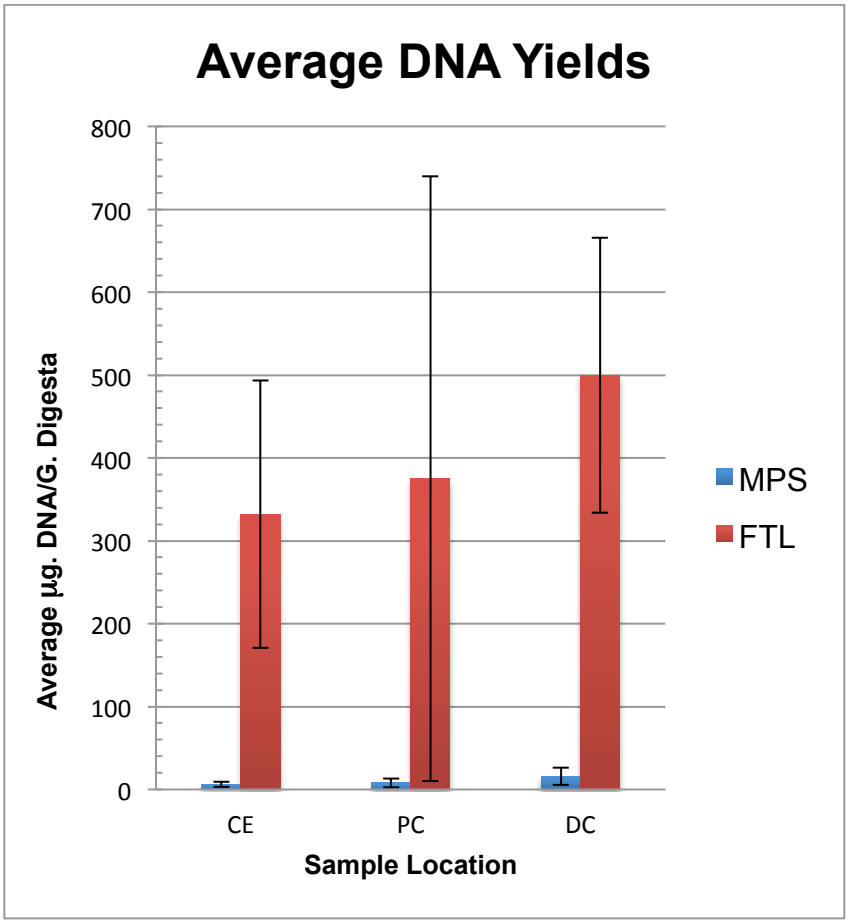


Figure 1: Average yields of DNA for each method by sample site.

The mean average number of reads of *Lactobacillus* in each of the FTL samples at all locations was greater than all other genera in those samples and much greater than that of *Lactobacillus* in the MPS samples. Evenness of the communities can be seen in the

first set of abundance plots (Fig. 2 top), which are truncated to mitigate the much greater abundance of *Lactobacillus* in the FTL processed samples for visualization purposes. With the amount of *Lactobacillus* plotted to full scale with the rest of the taxa, the evenness in the abundance plots for each treatment at each location appear similar (Fig. 2 top).

In order to more closely examine meaningful differences in abundance resulting from the different extraction protocols, we created 'side-by-side bar charts' which show the number of reads of the genera for each method at each of the sampling sites (Fig. 2 bottom). Of the 36 identified genera, only *Bacteroides*, *Lactobacillus*, *Ruminococcus*, *Parabacteriodes* and *Oscillospira* were found in all samples. Of these, there are significant differences in the amount of *Lactobacillus, Bacteroides* and *Parabacteroides* between the two DNA recovery methods. In addition, *Streptococcus* abundance was also monitored, as a member of LAB (Table 1). Some of the more scarce genera are absent entirely from samples processed by one method or another (Table 1), but whether these differences are biologically significant was difficult to determine.

Interestingly, *Facklamia* another member of the LAB is only present in samples processed using the FTL method (Table 1). As expected, the numbers of LAB were higher in the FTL processed samples than in the MPS processed samples and these differences are collectively the most noticeable aspect of all of the abundance graphs. Of the 5 families identified belonging to the order *Lactobacillales*, only *Lactobacillus* and *Streptococcus* were present in the samples processed with the MPS method.

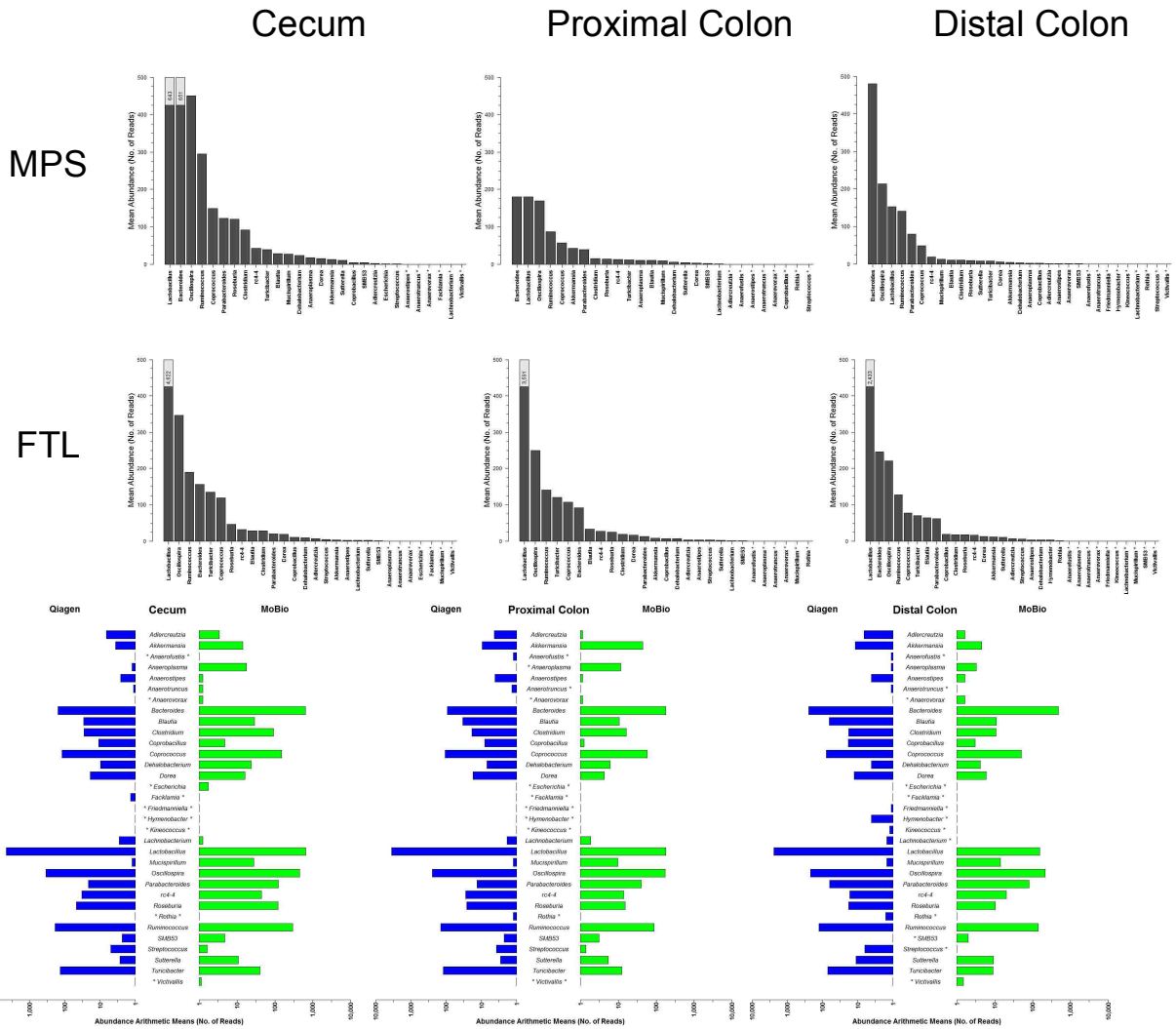


Figure 2: Top: Abundance charts comparing evenness of the methods at each sampling site. Bottom: Side by side comparisons of mean number of reads for each extraction method for each sampling location.

Table 1: Left: Wilcoxin Rank Sum Results for LAB and genera found in all samples. Right: Genera found only in samples processed using one treatment

A. Wilcoxin Rank Sum Results		**B. Genera Present for One Method Only**	
Genera Tested	**∞Significant a-level = 0.005**	**FTL**	**MPS**
*Lactobacillus**	Yes (FTL>MPS)	*Anaerofustis*	*Anaerovorax*
Bacteroides	Yes (MPS>FTL)	*Facklamia**	*Escherichia*
*Streptococcus**	Yes (FTL>MPS)	*Friedmanniella*	*Victivallis*
Oscillospira	No (P = 0.0797)	*Hymenobacter*	
Ruminococcus	No (P = 0.4833)	*Kineococcus*	
Parabacteroides	Yes (MPS>FTL)	*Rothia*	

∞A. Bonferroni Correction = 0.05/6 = 0.008. P-values are below the level of detection unless otherwise indicated.

B: List of genera detected from one method only.

*Denotes members of the LAB group classified to the genus level.

Core microbiomes were generated for each method as a whole by combining all sequence reads from all three compartments for each method (Fig. 3). The core microbiome for the FTL method was a higher proportion of the whole than for the MPS approach, and is more diverse. *Lactobacillus* itself dominates the FTL core microbiome, while *Bacteriodes* and *Lactobacillus* together make up the largest proportion of the core for the MPS method. All genera that comprise the core microbiome for MPS are also included in the core microbiome of FTL, which, in addition, contains *Blautia, Dorea, Coprococcus, and rc4-4.*

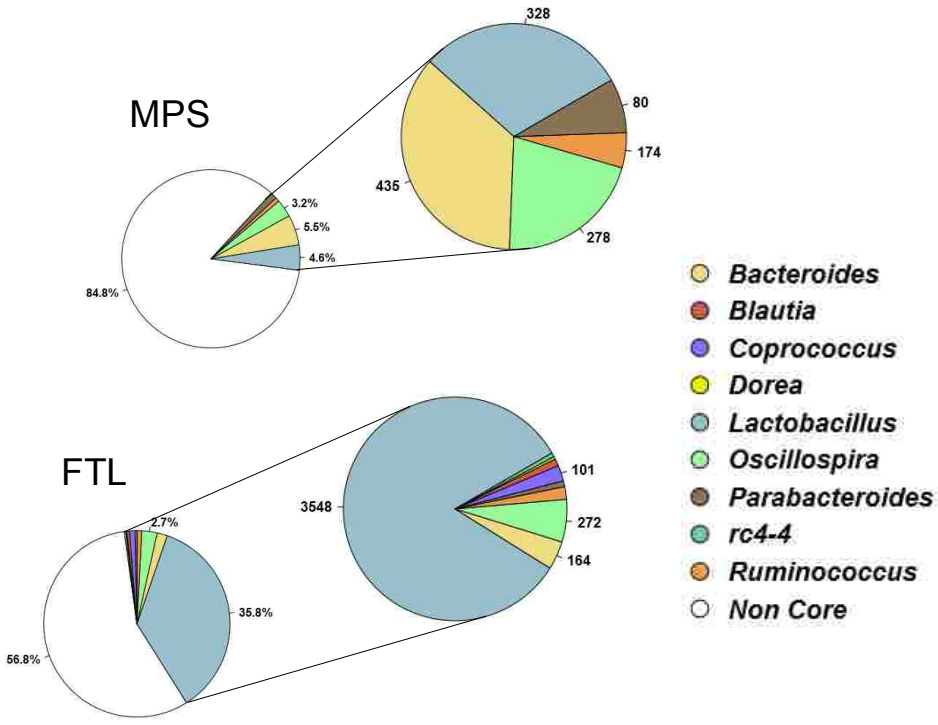


Figure 3: Core microbiome indicated as a function of DNA recovery method. The smaller pie charts show the proportion of the core microbiome to non-core taxa (shown in white). The larger charts depict the core microbiome in more detail and indicate the number of reads for the more predominant genera. Top: MPS, Bottom: FTL

More detailed analysis of the core microbiomes by both sampling location and DNA recovery method clearly reflects the predominance of *Lactobacillus* in the metagenomic DNA recovered by the FTL method of DNA extraction (Fig. 4). The core taxa derived from metagenomic DNA recovered by the MPS method are dominated by *Lactobacillus* and *Bacteroides,* along with a slightly smaller proportion of *Oscillospira*. The core to non-core ratios across all sites for each method were relatively consistent.

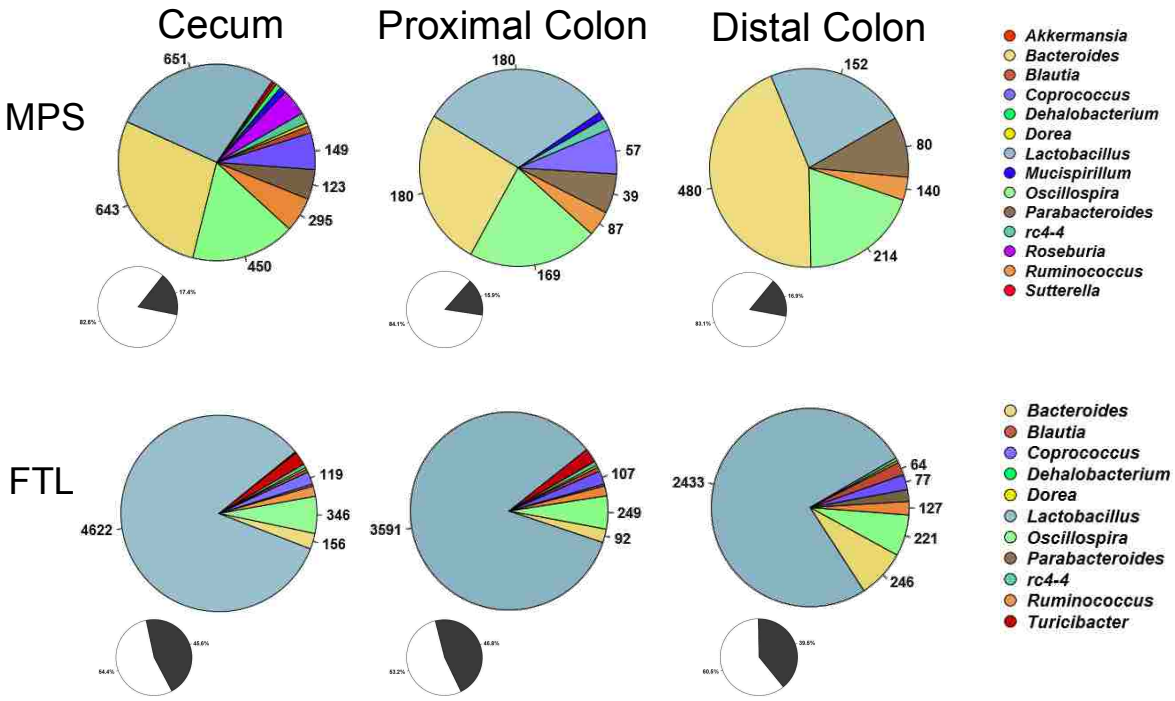


Figure 4: Core microbiome by sampling site and method. Small pie charts give the proportion of core to non-core (in white) for each sampling site/method.

We compared the indicated diversity of two important groups of bacteria (Bacteroidetes and LAB) to assess whether there were differences in the data obtained using the two methods. Five groups of Bacteroidetes were present, with three identified to the genus level, in samples extracted by either of the two methods (Table 2). Thus, the indicated diversity of this group was consistent for both methods. By contrast, five families of LAB were indicated from FTL-recovered DNA, with three of those classified

to the genus level (Table 2), while only 3 were detected from DNA recovered using the MPS approach.

Table 2: Comparison of Bacteroidetes and LAB diversity between methods.

	FTL	MPS
Bacteroidetes	Bacteroides	Bacteroides
	Parabacteroides	Parabacteroides
	Hymenobacter	Hymenobacter
	Rikenellaceae (F)*	Rikenellaceae (F)
	S24-7 (F)	S24-7 (F)
LAB	Lactobacillus	Lactobacillus
	Streptococcus	Streptococcus
	Facklamia	Not present
	Enterococcaceae (F)	Not present
	Carnobacteriaceae (F)	Not present

*Taxa classified only to the family level are denoted by (F); otherwise they are classified to the genus level.
Taxa not detected in samples extracted by one of the methods are marked "Not present".

Discussion:

Numerous studies have demonstrated how metagenomic DNA extraction methods affect the indicated composition of the sampled community following sequence analysis, and some examined whether some particular feature of the extraction process had more effect than another (Carrigg et al., 2007; Ferrand et al., 2014; Kennedy et al., 2014; Maukonen et al., 2012; Yuan et al., 2012). Yet very few of those studies have described the problems associated with studying a particular taxon or group of bacteria of interest, while maintaining an accurate representation of the remainder of the bacterial community (Ferrand et al., 2014; Kuske et al., 1998; Maukonen et al., 2012). This topic is potentially of great importance in forensics, (where often it is necessary to recover taxa of interest while not destroying community DNA which may provide contextual information), in

studies intended to examine introduced taxa and how they affect total community dynamics, or in studies that examine the role(s) that particular taxa play in overall community dynamics (Filippidou et al., 2015; Kuske et al., 1998; Maukonen et al., 2012).

Lysis of the bacterial cell wall is an important aspect of DNA recovery methods and many of the studies done on community DNA extraction have focused on cell wall lysis being the determining factor in how extraction techniques result in differing community compositions (Carrigg et al., 2007; Maukonen et al., 2012). The relative contribution of lysis techniques to variations in the perceived community composition caused by differing extraction methods is somewhat controversial. At least one study found that lysis methods don't contribute to community differences (Kennedy et al., 2014). By contrast, other studies have found that the method of lysis is a strong driver in determining how various metagenomic DNA recovery methods result in differing DNA yields as well as different indications of bacterial community composition (Carrigg et al., 2007; Maukonen et al., 2012; Salonen et al., 2010).

The lysis approach that has been determined to be the most generally successful by many groups is mechanical lysis, such as bead beating, although it may result in more DNA shearing than other methods (Olson and Morrow, 2012; Salonen et al., 2010). Although the FTL approach has lower throughput and is more time consuming, its gentler method of combined freeze-thaw and enzymatic cell wall degradation represents an alternative method of bacterial cell lysis, that produces greater yields of DNA. presumably due to less physical disruption of DNA molecules following release from cells.

The two methods compared in this paper resulted in greatly different DNA yields with the FTL method giving metagenomic DNA yields one to two orders of magnitude greater than is obtained from identical samples by the MPS method. These results are congruent with other studies showing that the MPS protocol produced lower yields of DNA than many other DNA extraction methods (Delmont et al., 2012; Kennedy et al., 2014). On the other hand, a limitation of the FTL method is that it is more expensive in time and materials than MPS and other kit-based DNA recovery approaches (this work).

At least one study showed that various kit methods appeared to give community profiles that are similar to one another (Kennedy et al., 2014). Further, as discussed above, reliable but less quantitative recovery of metagenomic DNA can suffice for some comparative analyses of samples. Thus, unless a particular study asks direct questions about the relative abundance of LAB or other difficult to lyse microbes, or the samples have limited extractable biomass or otherwise produce very low yields of metagenomic DNA, the MPS technique (or a comparable kit-based approach) may suffice for some analyses of bacterial community composition with less expense and time. However, the additional time and expense necessary for the FTL extraction process can be offset if one is able to subsequently forgo using qPCR or other techniques in order to make sure that taxa of interest are represented accurately.

A major difference between the two methods was the enhanced detection of LAB taxa in the samples processed using the FTL method (Table 1), including 3 families not found in any of the samples processed with the MPS approach. LAB are significant in their own right as autochthonous members of the intestinal microbiome (Hooda et al., 2012; Isaacson and Kim, 2012; Minamoto et al., 2012; Reuter, 2001). LAB also are an

important constituent of the oral microbiome (Gibbons and Van Houte, 1975; Quevedo et al., 2011). They are also highly valuable to the food industry and as potential probiotics (Lefeber et al., 2011; Messaoudi et al., 2013; Pei et al., 2015; Wouters et al., 2013). Thus, being able to accurately detect LAB taxa, both autochthonous and introduced, in both control and experimental subjects can be paramount for some studies, particularly those that examine how introduced LAB affect the host's indigenous microbiome. In these cases, employing the FTL extraction method is highly advisable and would reward the extra time, effort and expense invested into metagenomic DNA recovery. Such is the case with much of the work described in the remaining chapters of this thesis dissertation.

Interestingly, the enhanced detection of LAB with the FTL approach is concurrent with decreases in the proportion of *Bacteroides* and *Parabacteroides* detected, as compared to the proportions found in the communities when using the MPS method of extraction. One explanation for the proportional decreases in these genera is that they are easily lysed and thus are lost due to excessive lysis and DNA degradation in the FTL extraction process. A more likely explanation is that the larger total yields of DNA along with larger proportional amounts of DNA from LAB (Lactobacillus, in particular) effectively "dilutes" the abundance of target DNA representing the Bacteroidetes phylum. This conclusion is supported by the maintenance of diversity in that phylum despite the decrease in the proportion of those taxa (in terms of numbers of sequence reads) obtained from samples processed using the FTL approach.

Another way to consider this point is that once metagenomic DNA is recovered from a set of samples, the laboratory and analytical workflow downstream can be thought of as a 'closed system'. Thus, a significant increase in the proportional amounts of DNA

from one group of taxa (e.g. LAB) in a metagenomic DNA sample would result in a concurrent decrease in the proportional amounts of DNA from other taxa.

Despite the more effective detection of LAB taxa afforded by our approach, when the large amounts of *Lactobacillus* (in this case) are taken into account, the overall composition of the communities determined by the FTL method are not especially different from the composition indicated by the MPS method. While the proportional amount of DNA from Bacteroidetes is decreased in the FTL samples, the diversity of this phylum is maintained. With respect to rare genera, samples from both methods contain genera not present in samples processed by the other method. Indeed, the MPS method resulted in greater loss of rare taxa than the FTL method.

The FTL method of metagenomic DNA recovery proved very useful in the additional studies described herein, where the use of the MPS approach with samples from mouse models produced inconsistent metagenomic DNA yields and many samples that couldn't be PCR amplified in a preliminary study (data not shown). The FTL approach is also recommended for studies where the numbers of LAB within a community is important, such as studies of the oral and intestinal microbiome, as well as studies examining effects of introducing probiotic bacteria on the indigenous microbiome as described in the next Chapter.

As with other comparisons of DNA extraction methods, this investigation underscores the importance of selecting an appropriate DNA recovery method and of using it in a normalized and consistent way where results from multiple experiments, treatments and/or timepoints are to be pooled and compared with one another. Caution should be

taken when comparing seemingly related studies within the literature that did not use the same method of DNA recovery.

Literature Cited:

Ackermann, M., Stecher, B., Freed, N.E., Songhet, P., Hardt, W.D., and Doebeli, M. (2008). Self-destructive cooperation mediated by phenotypic noise. *Nature* 454, 987-990.

Apajalahti, J.H., Sarkilahti, L.K., Maki, B.R., Heikkinen, J.P., Nurminen, P.H., and Holben, W.E. (1998). Effective recovery of bacterial DNA and percent-guanine-plus-cytosine-based analysis of community structure in the gastrointestinal tract of broiler chickens. *Applied and Environmental Microbiology* 64, 4084-4088.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.M., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174-180.

Bianchi, M.A., Del Rio, D., Pellegrini, N., Sansebastiano, G., Neviani, E., and Brighenti, F. (2004). A fluorescence-based method for the detection of adhesive properties of lactic acid bacteria to Caco-2 cells. *Letters in Applied Microbiology* 39, 301-305.

Brown, W.C., Sandine, W.E., and Elliker, P.R. (1962). Lysis of lactic acid bacteria by lysozyme and ethylenediaminetetraacetic acid. *Journal of Bacteriology* 83, 697-698.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7, 335-336.

Carrigg, C., Rice, O., Kavanagh, S., Collins, G., and O'Flaherty, V. (2007). DNA extraction method affects microbial community profiles from soils and sediment. *Applied Microbiology and Biotechnology* 77, 955-964.

Chassy, B.M., and Giuffrida, A. (1980). Method for the lysis of Gram-positive, asporogenous bacteria with lysozyme. *Applied and Environmental Microbiology 39*, 153-158.

Delmont, T.O., Prestat, E., Keegan, K.P., Faubladier, M., Robe, P., Clark, I.M., Pelletier, E., Hirsch, P.R., Meyer, F., Gilbert, J.A., et al. (2012). Structure, fluctuation and magnitude of a natural grassland soil metagenome. *The ISME Journal* 6, 1677-1687.

Derrien, M., and van Hylckama Vlieg, J.E. (2015). Fate, activity, and impact of ingested bacteria within the human gut microbiota. *Trends in Microbiology* 23, 354-366.

Earley, Z.M., Akhtar, S., Green, S.J., Naqib, A., Khan, O., Cannon, A.R., Hammer, A.M., Morris, N.L., Li, X., Eberhardt, J.M., et al. (2015). Burn Injury Alters the Intestinal Microbiome and Increases Gut Permeability and Bacterial Translocation. *PloS One* 10, e0129996.

El Aidy, S., Stilling, R., Dinan, T.G., and Cryan, J.F. (2016). Microbiome to Brain: Unravelling the Multidirectional Axes of Communication. *Advances in Experimental Medicine and Biology* 874, 301-336.

Eloe-Fadrosh, E.A., Brady, A., Crabtree, J., Drabek, E.F., Ma, B., Mahurkar, A., Ravel, J., Haverkamp, M., Fiorino, A.M., Botelho, C., et al. (2015). Functional dynamics of the gut microbiome in elderly people during probiotic consumption. *mBio* 6.

Ferrand, J., Patron, K., Legrand-Frossi, C., Frippiat, J.P., Merlin, C., Alauzet, C., and Lozniewski, A. (2014). Comparison of seven methods for extraction of bacterial DNA from fecal and cecal samples of mice. *Journal of Microbiological Methods* 105, 180-185.

Filippidou, S., Junier, T., Wunderlin, T., Lo, C.C., Li, P.E., Chain, P.S., and Junier, P. (2015). Under-detection of endospore-forming Firmicutes in metagenomic data. *Computational and Structural Biotechnology Journal* 13, 299-306.

Garcia-Garcera, M., Garcia-Etxebarria, K., Coscolla, M., Latorre, A., and Calafell, F. (2013). A new method for extracting skin microbes allows metagenomic analysis of whole-deep skin. *PloS One* 8, e74914.

Gibbons, R.J., and Van Houte, J. (1975). Bacterial adherence in oral microbial ecology. *Annual Review of Microbiology* 29, 19-44.

Gu, W., and Swihart, R.K. (2004). Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. *Biological Conservation* 116, 195-203.

Henao-Mejia, J., Elinav, E., Thaiss, C.A., Licona-Limon, P., and Flavell, R.A. (2013). Role of the intestinal microbiome in liver disease. *Journal of Autoimmunity* 46, 66-73.

Hill, D.A., Hoffmann, C., Abt, M.C., Du, Y., Kobuley, D., Kirn, T.J., Bushman, F.D., and Artis, D. (2010). Metagenomic analyses reveal antibiotic-induced temporal and spatial changes in intestinal microbiota with associated alterations in immune cell homeostasis. *Mucosal Immunology* 3, 148-158.

Hooda, S., Minamoto, Y., Suchodolski, J.S., and Swanson, K.S. (2012). Current state of knowledge: the canine gastrointestinal microbiome. *Animal Health Research Reviews / Conference of Research Workers in Animal Diseases* 13, 78-88.

Isaacson, R., and Kim, H.B. (2012). The intestinal microbiome of the pig. *Animal Health Research Reviews / Conference of Research Workers in Animal Diseases* 13, 100-109.

Jernberg, C., Lofmark, S., Edlund, C., and Jansson, J.K. (2010). Long-term impacts of antibiotic exposure on the human intestinal microbiota. *Microbiology* 156, 3216-3223.

Kennedy, N.A., Walker, A.W., Berry, S.H., Duncan, S.H., Farquarson, F.M., Louis, P., Thomson, J.M., Consortium, U.I.G., Satsangi, J., Flint, H.J., et al. (2014). The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PloS One* 9, e88982.

Kuske, C.R., Banton, K.L., Adorada, D.L., Stark, P.C., Hill, K.K., and Jackson, P.J. (1998). Small-Scale DNA Sample Preparation Method for Field PCR Detection of Microbial Cells and Spores in Soil. *Applied and Environmental Microbiology* 64, 2463-2472.

Lefeber, T., Gobert, W., Vrancken, G., Camu, N., and De Vuyst, L. (2011). Dynamics and species diversity of communities of lactic acid bacteria and acetic acid bacteria during spontaneous cocoa bean fermentation in vessels. *Food Microbiology* 28, 457-464.

Ma, H.D., Wang, Y.H., Chang, C., Gershwin, M.E., and Lian, Z.X. (2015). The intestinal microbiota and microenvironment in liver. *Autoimmunity Reviews* 14, 183-191.

MacKenzie, D.I., Bailey, L.L., and Nichols, J. (2004). Investigating species co-occurrence patterns when species are detected imperfectly. *Journal of Animal Ecology* 73, 546-555.

Maukonen, J., Simoes, C., and Saarela, M. (2012). The currently used commercial DNA-extraction methods give different results of clostridial and actinobacterial populations derived from human fecal samples. *FEMS Microbiology Ecology* 79, 697-708.

Messaoudi, S., Manai, M., Kergourlay, G., Prevost, H., Connil, N., Chobert, J.M., and Dousset, X. (2013). Lactobacillus salivarius: bacteriocin and probiotic activity. *Food Microbiology* 36, 296-304.

Minamoto, Y., Hooda, S., Swanson, K.S., and Suchodolski, J.S. (2012). Feline gastrointestinal microbiota. *Animal Health Research Reviews / Conference of Research Workers in Animal Diseases* 13, 64-77.

Olson, N.D., and Morrow, J.B. (2012). DNA extract characterization process for microbial detection methods development and validation. *BMC Research Notes* 5, 668.

Pei, R., Martin, D.A., DiMarco, D.M., and Bolling, B.W. (2015). Evidence for the Effects of Yogurt on Gut Health and Obesity. *Critical Reviews in Food Science and Nutrition,* 0.

Quevedo, B., Giertsen, E., Zijnge, V., Luthi-Schaller, H., Guggenheim, B., Thurnheer, T., and Gmur, R. (2011). Phylogenetic group- and species-specific oligonucleotide probes for single-cell detection of lactic acid bacteria in oral biofilms. *BMC Microbiology* 11, 14.

Reuter, G. (2001). The Lactobacillus and Bifidobacterium microflora of the human intestine: composition and succession. *Current Issues in Intestinal Microbiology* 2, 43-53.

Ribet, D., and Cossart, P. (2015). How bacterial pathogens colonize their hosts and invade deeper tissues. Microbes and Infection 17, 173-183.

Salonen, A., Nikkila, J., Jalanka-Tuovinen, J., Immonen, O., Rajilic-Stojanovic, M., Kekkonen, R.A., Palva, A., and de Vos, W.M. (2010). Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *Journal of Microbiological Methods* 81, 127-134.

Scornec, H., Tichit, M., Bouchier, C., Pédron, T., Cavin, J.-F., Sansonetti, P.J., and Licandro-Seraut, H. (2014). Rapid 96-well plates DNA extraction and sequencing procedures to identify genome-wide transposon insertion sites in a difficult to lyse bacterium: Lactobacillus casei. *Journal of Microbiological Methods* 106, 78-82.

Tomas, J., Langella, P., and Cherbuy, C. (2012). The intestinal microbiota in the rat model: major breakthroughs from new technologies. *Animal Health Research Reviews / Conference of Research Workers in Animal Diseases* 13, 54-63.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027-1031.

Vinje, S., Stroes, E., Nieuwdorp, M., and Hazen, S.L. (2014). The gut microbiome as novel cardio-metabolic target: the time has come! *European Heart Journal* 35, 883-887.

West, C.E., Jenmalm, M.C., and Prescott, S.L. (2014). The gut microbiota and its role in the development of allergic disease: a wider perspective. *Clinical and Experimental Allergy* : Journal of the British Society for Allergy and Clinical Immunology.

Wolf, K.J., Daft, J.G., Tanner, S.M., Hartmann, R., Khafipour, E., and Lorenz, R.G. (2014). Consumption of acidic water alters the gut microbiome and decreases the risk of diabetes in NOD mice. *The Journal of Histochemistry and Cytochemistry* : Official journal of the Histochemistry Society 62, 237-250.

Wouters, D., Grosu-Tudor, S., Zamfir, M., and De Vuyst, L. (2013). Bacterial community dynamics, lactic acid bacteria species diversity and metabolite kinetics of traditional Romanian vegetable fermentations. *Journal of the Science of Food and Agriculture* 93, 749-760.

Yuan, S., Cohen, D.B., Ravel, J., Abdo, Z., and Forney, L.J. (2012). Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PloS One* 7, e33865.

Zarrinpar, A., Chaix, A., Yooseph, S., and Panda, S. (2014). Diet and feeding pattern affect the diurnal dynamics of the gut microbiome. *Cell Metabolism* 20, 1006-1017.

Zhang, S.S., Chen, D., and Lu, Q. (2012). An improved protocol and a new grinding device for extraction of genomic DNA from microorganisms by a two-step extraction procedure. *Genetics and Molecular Research* : GMR 11, 1532-1543.

Chapter 3: Location, location, location—Genus-level microbiome biogeography along the intestinal 'landscape'.

Authors: Ellen Lark[1], Eric M Spaulding[2], Douglas W. Raiford[2,3], Alden H. Wright[2] William E. Holben[1,3]*

Affiliations: [1] Cellular, Molecular and Microbial Biology Program; [2]Department of Computer Science, and [3]Systems Ecology Program, University of Montana, Missoula, MT, USA 59812

*Correspondence to: bill.holben@mso.umt.edu

** This chapter has been formatted for submission to Cell, Host and Microbe

Abstract: The mammalian gastrointestinal (GI) microbiome is involved in host health, nutrition and immunological status via intimate, sometimes intricate, interactions with the host and the intestinal environment. Despite increasing awareness and extensive research on the roles and importance of the GI microbiome, a clear vision of how individual bacterial taxa are distributed along the 'landscape' of the lower GI tract is almost entirely lacking. Using a mouse model, key ecological concepts and a novel combination of computational, statistical and bioinformatic tools, we show that there is a clear biogeographical distribution of bacterial taxa down to the genus level along the lower intestinal tract. We expect that intestinal microbiome biogeography is widespread in other mammalian and non-mammalian hosts and, given the immunological, nutritional and health-related roles ascribed to host-associated microbes, will prove to be a key element governing host physiology and disease etiology.

Background:

Within the GI tract microbes mediate many key metabolic reactions of which the host is incapable, thereby facilitating digestion to the nutritional benefit of both microbes

and host. GI bacteria are also increasingly shown to be critical in priming and tuning the host immune response to commensals, pathogens and self (Chang et al., 2014; Rolig et al., 2013; West et al., 2014). Further, there is progressively more evidence that GI microbiome composition and stability is an important determinant of the outcome of infection by pathogens and parasites (Rolig et al., 2013; Schubert et al., 2015; Yurist-Doutsch et al., 2014), as well as being involved in a number of multifactorial diseases such as irritable bowel disease (IBD), diabetes, rheumatoid arthritis, obesity and cancer (Cani, 2014; Cani et al., 2012; Li et al., 2015; McLean et al., 2015; Zackular et al., 2013).

General ecological theory predicts that different environments will select for different communities (Baas-Becking, 1934; Pocheville, 2015), producing patterns of species abundance and distribution (i.e. biogeography). The lumen of the lower intestine in mammals exhibits pronounced regional differences in pH, oxygen availability, moisture content and nutrient composition (Heinzmann and Schmitt-Kopplin, 2015; Kawamata et al., 2006; Louis et al., 2014), yet, to date, no reports have clearly demonstrated spatial patterns in the distribution of its resident microbes. One obstacle to intestinal biogeography analyses is that inter-subject variation obscures patterns of regional differentiation (Eckburg et al., 2005; Lavelle et al., 2013; Rogers et al., 2014). It has also been suggested that mixing of luminal contents by peristalsis would obscure biogeographical distributions resulting from local environmental differences (Zoetendal et al., 2002). In addition, mouse studies have found that differences due to housing mice in separate facilities, cages (cage effects), as well as those due to different litters may be just as significant as intersubject variation and also confound the ability to discriminate between treatments (Ericsson et al., 2015; Rogers et al., 2014).

Some human studies have indicated differences in microbiome composition between luminal and mucosal communities, but failed to detect longitudinal differences (Lavelle et al., 2013). Others based on colonoscopy samples suggest longitudinal differences only for bacteria closely associated with the intestinal mucosa (Aguirre de Carcer et al., 2011; Eckburg et al., 2005; Zhang et al., 2014; Zoetendal et al., 2002), which is not surprising given the intense purging of luminal contents prior to the procedure. Mouse studies have not fared much better, tending to be either broadly comparative, focused on particular taxa, or both (Hu et al., 2010; Nava et al., 2011; Sarma-Rupavtarm et al., 2004; Swidsinski et al., 2005a; Turnbaugh et al., 2009b). As a result, microbiome studies to date have lacked taxonomic and locational resolution, generally describing 'dysbioses' between healthy and diseased individuals wherein the proportions of phylum- or family-level groups differ between healthy and diseased individuals (Li et al., 2015; Turnbaugh et al., 2006).

In the current study, a mouse model, next generation sequencing (NGS) and a novel computational and bioinformatic workflow were used to demonstrate that distinct luminal microbial communities can be differentiated at the genus level among the various compartments and regions comprising the lower GI tract. We hypothesized that spatial variation in physicochemical properties such as pH, oxygen availability, water content and nutrient composition along the intestinal tract would influence which bacterial genera predominate at different locations. The rationale is that bacterial genera have fairly coherent physiological, biochemical and metabolic properties and thus would have preferred locations in this system. Herein, we demonstrate differential distribution of bacterial genera (i.e. biogeography) along the lower intestinal tract, providing a high-

resolution demonstration of biogeography for the luminal microbiome of the mammalian lower GI tract.

Materials and Methods:

Animal Models and Sampling:

All animal treatments were approved by the Institutional Animal Care and Use Committee (IACUC) at the University of Montana under AUP# Holben 007-12. Ten week old female C57Bl/6 and ICR (here called CD-1) mice were provided by Taconic Laboratories (Hudson, NY) and held for 3 weeks to establish that their health was stable. In independent experiments, two cohorts a year apart (consisting of 10 and 24 mice, respectively), with each cohort comprised of both inbred C57Bl/6 mice (6 in cohort 1, 12 in cohort 2) and outbred CD-1 mice (4 in cohort 1, 12 in cohort 2) were analyzed, producing a total of 204 samples from six sampling locations. To maintain their native microbiome, the mice were isolated from environmental bacteria using HEPA-filtered air in positive pressure cages. The mice were given sterile food, bedding and water, with all cage changes performed aseptically. The food (NIH-31) was the same diet they received at Taconic. Following the 3-week holding period, the mice were euthanized humanely and luminal intestinal samples taken from the indicated locations (see Fig. 1). To control for unintended environmental variables, a random number table (Rand Corporation, 2001) was used to select cages at the same time on 3 separate days (for each cohort) and all samples for that day were collected within 1 h of each other.

Total microbiome DNA was purified from intestinal contents (digesta) at six sampling sites along the lower intestinal tract of two cohorts of mice, with each cohort

comprised of both inbred C57Bl/6 mice and outbred CD-1 mice to increase host genetic variation. All mice were housed, fed and otherwise treated identically to minimize variability that might result from differing environmental conditions. Sampling points were the distal ileum (defined as the last 3 cm of the ileum), the cecum, the tip of the cecum, the proximal colon (the region nearest the cecum containing liquid digesta), the mid-colon (the mid-portion containing the first-formed, soft, 'pre-fecal' pellets), and the distal colon (defined as the last 2 cm including fully formed fecal pellets and the rectum) (Fig. 1).

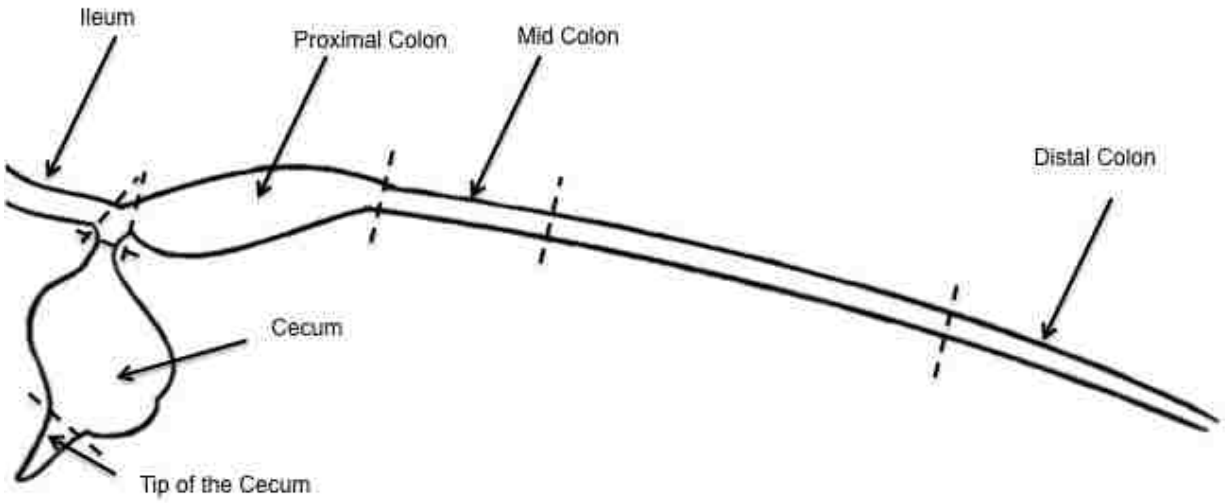


Figure 1: Sampling sites along the lower intestinal tract.

The luminal contents of surgically removed samples were recovered by gentle squeezing into sterile microcentrifuge tubes, which were kept on ice during collection then stored at -70 ˚C prior to downstream processing. All samples from within each mouse were considered to be linked and therefore were processed for microbiome DNA recovery, PCR amplification and sequencing together. Sample sets were identified by

cage number and mouse number and processed in an order determined by a random number table (Rand Corporation, 2001).

Microbiome DNA Recovery, Amplification and Sequencing:

Microbiome DNA was recovered using a protocol adapted from Apajalahti *et al*. (Apajalahti et al., 1998) that was shown to provide highly effective recovery of bacterial DNA from GI tract samples. Digesta samples were placed into sterile Oak Ridge tubes containing 10 mL of sterile wash buffer (0.5 M sodium phosphate [pH 8.0], 0.1% Tween-80) and washed 4 times as follows. Samples were vortexed briefly before being shaken at high speed on a reciprocating shaker for 10 min. Next, samples were centrifuged at 30,000 x *g* for 15 min at room temperature, after which the supernatant was removed by aspiration and the samples resuspended in 10 ml of wash buffer. Following the final wash step and centrifugation, the samples were resuspended in 3 ml of Qiagen Buffer B1 (50 mM sodium EDTA, 50 mM Tris base [pH 8.0], 0.5% Tween-20, 0.5% Triton X-100; Qiagen, Valencia, CA) to which RNase A (200 mg/L) was added to a final concentration of 200 µg/ml, then stored at -70˚C to initiate the 5 freeze-thaw cycles that facilitate bacterial cell lysis. The samples were thawed and refrozen a total of 5 times by being placed in a water bath at 40˚C for 15 min, then placed back at -70 ˚C for at least 1 h before being thawed again. Following the final thaw, 50 µL of lysozyme (200 mg/ml) and 90 µL of proteinase K (20 mg/ml) were added. Samples were then incubated in a water bath at 37˚C for 45 min, after which 1 mL of Qiagen B2 buffer (3 M guanidine HCl, 20% Tween-20) was added. The samples were incubated in a water bath at 50˚C for 45 min, then centrifuged for 10 min at 5,000 x *g* at 4˚C. The supernatant was transferred

to a sterile microcentrifuge tube and vortexed for 10 sec. At this point, the Qiagen Genomic Tip 20G protocol was followed precisely to elute microbiome DNA, except that 1 extra 70% ethanol wash was performed. Finally, the dried samples were resuspended in 50 ml of TE (10 mM Tris [pH 8.0], 1 mM EDTA) and quantified using a nanophotometer (Implen P 300, Implen, Inc., Westlake Village, CA).

Partial 16/18S rRNA gene sequences encompassing regions V4 & V5 were PCR amplified from the microbiome DNA (25 ng) using the highly conserved primers 536f & 907r (*25*), which were barcoded for pyrosequencing. Where samples were not of sufficient concentration to provide 25 ng for PCR, 3 µl of purified microbiome DNA was used as template. The resulting 16S-sized amplicons were gel purified using the Qiagen Gel Purification kit per manufacturer's instructions, then further purified through two successive rounds using Agencourt AMPure XP magnetic beads per manufacturer's instructions (Beckman Coulter Inc., Brea CA). Purified DNA was quantified, multiplexed, and sequenced by the Utah State University Center for Integrated Biosystems using the 454 Roche GS FLX system (454 Life Sciences, Roche Diagnostics, Indianapolis, IN).

Data Analysis and Statistics:

Sequences were 'denoised' and identified to the genus level using the Quantitative Insights Into Microbial Ecology (QIIME) pipeline (Caporaso et al., 2010) and the number of sequences per sample determined. Sample datasets containing fewer than 750 or 1000 sequences were eliminated from cohort 1 and 2, respectively. A taxa summary table was built containing a row for each sample (one for each locational sampling point for each mouse) and a column for each genus identified. The entries in the

matrix were the proportion of reads in the sample represented by the associated genus. Any reads that were unclassified at the genus level were removed from further consideration. Frequency matrices were created for both cohorts.

A combined cohort 1 & 2 frequency matrix was created for each of the comparisons that were to be plotted. Each plot was for either the C57Bl/6 or CD-1 mouse strain, and was for a specific subset of longitudinal sample locations (ileum, caecum, proximal colon and distal colon; the caecum and tip of caecum; and the proximal, mid, and distal colon). These matrices provided the data for the three pairs of plots depicted in Figs. 2 and 3.

An important part of this analysis was identifying which genera were associated with each longitudinal location along the GI tract. We accomplished this with a feature selection algorithm known as floating search (described below; in this case the important features being identified are the genera). This approach presented a challenge for the analysis. Because we used a classifier (Linear Discriminant Analysis, LDA) to determine whether microbial community composition could be used to discriminate between sample locations, cross-validation was required to provide confidence in the results. In a two-step process like this (feature selection followed by classification), the feature selection should be performed for each fold in a cross-fold validation process, just as the training phase is performed during each fold. This meant that, potentially, a different subset of genera might have been identified during each fold of the cross-fold-validation process, leading to the problem of which genera to use for visualization. We employed a computational voting process (described below) to identify which genera to use during visualization in the LDA scatter plots. The more often a given genus was selected as being important for

discrimination during the cross-validation process, the more likely it was ultimately used for visualizing the results.

Another challenge was identification of the proper number of features (genera) to utilize in the analyses. In machine learning, this is often accomplished by examining classifier performance across different numbers of features, and choosing the number of features that provide the best classifier performance. The observation of classifiers operating best at a specific number of features is known as the 'peaking phenomenon' (Trunk, 1979). This also helps prevent over-fitting (the classifier becoming overly sensitive to nuances in the data). The approach we chose for identifying the number of genera, or dimensions, to use was to run the classifier on various datasets (cohort 1, 2, and 1 & 2) at various numbers of dimensions (from 7 to 23 genera) with various levels of 'pre-pruning.' Pre-pruning involved the removal of genera if they were not present in at least 1 sample (equivalent to no pruning), and in 3%, 5%, 8%, and 16% of all samples. This provided 15 accuracies at each number-of-dimensions tested for each classifier. A different classifier was used for each LDA scatter plot. These accuracies were visualized in a boxplot format (Supplementary Figs. S1 - S3).

To remove human bias from the process, we determined the appropriate number of features by choosing the number with good performance (classifier median accuracy), low variation (accuracies at that number of features tend to have low variance), and a larger number of dimensions. This was accomplished with a Pareto-front-analysis (Hwang and Masud, 1979) (Supplementary Figs S4 - S6). All data points on the Pareto front are of equivalent multi-objective quality. A representative number of dimensions

was chosen for visualization from the set on the Pareto front and presented in Figs. 4 and 5.

This strategy also facilitated the voting process. The genera chosen in each of the folds of the leave-one mouse-out cross-validation runs (because samples from the same mouse were considered linked) for each of the three datasets (cohorts 1, 2, and 1 & 2) and for each of the five pre-pruning levels contributed to the voting tallies. Those with the highest normalized tallies were used for visualization purposes. By ‘normalized’ it is meant that each vote was weighted according to its achieved accuracy. That is, a genus chosen during a fold that achieved a greater accuracy was given more weight than one that did not perform as well. A schematic diagram of the bioinformatic and computational workflow is presented in Fig. 2, while example code listings for generation of LDA plots and cross validation, respectively, are presented in Supplementary Fig. S7.

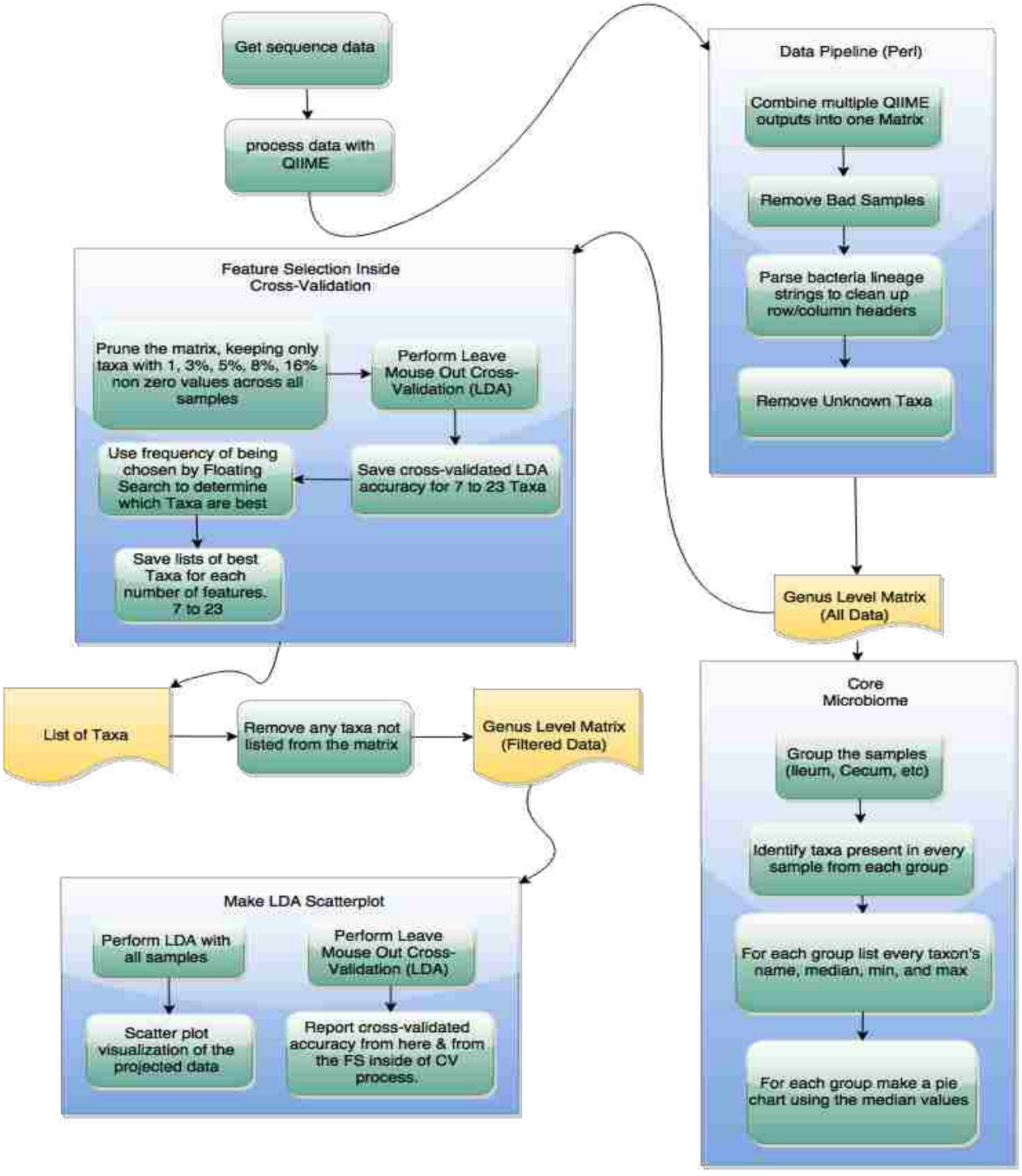


Figure 2: Process flow diagram of the computational methods employed.

Core Microbiome Determination:

The core microbiome for each of the six sampling locations within the lower GI tract was determined to the genus level for Cohorts 1 & 2. A genus was considered a part of the core microbiome if all samples for that location from all animals for both mouse strains contained that genus. The effect of sampling depth on core microbiome composition was determined by successively eliminating all sample data sets with less than 3,000, 2,500, 2,000, or 1,500 sequences, respectively.

Results:

Conceiving of the ~10 cm-long mouse lower intestinal tract as a bacterial 'landscape' is facilitated by considering that bacteria are roughly 1/2,000,000 the size of humans (~1 µm versus ~2 m). Indeed, 1 cm of distance between two bacterial cells is equivalent to 20 km of physical separation between two humans, with all of the potential environmental variations that come into play at that scale (e.g. rivers, highways, mountains, altitude, local weather). In mice, intersubject differences may be due to genetic, nutritional, housing or other environmental effects. Differences in housing have been shown to be significant, and possibly as great as those caused by intersubject variation. Cage effects in particular have been shown to be a significant source of noise when doing community comparisons even when treatments don't involve the subtle differences found along the lower intestine (Ericsson et al., 2015; Rogers et al., 2014).

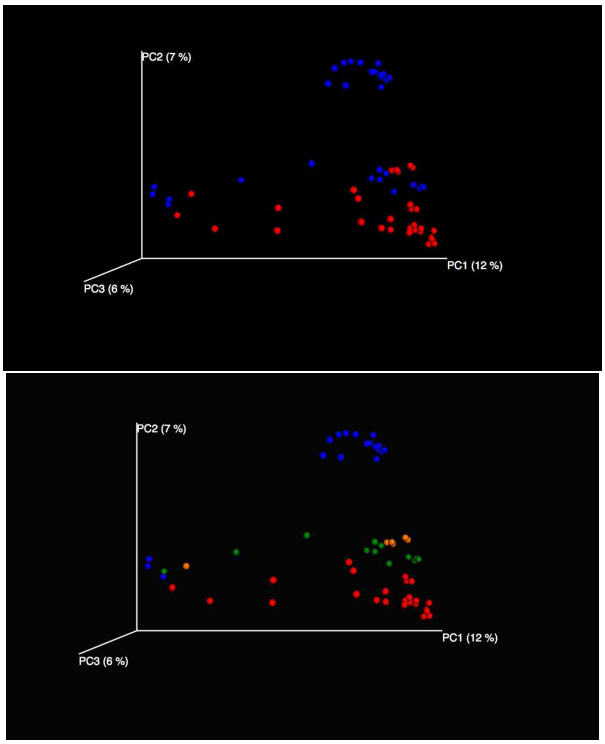


Figure 3: Left: Two treatments in mice cannot be discriminated using a beta diversity measure (Unifrac) combined with PCoA. Right: Cages as the discriminating factor are shown, showing how cage effects can confound results.

For this intestinal landscape study, discerning differences between locations along this landscape required overcoming the challenges posed by inter-subject compositional variation, cage effects and other causes of noise. This was accomplished using a novel bioinformatic workflow that employs a computational 'feature selection' technique to identify genera that differentiate between intestinal locations, in combination with a 'locational classifier' that predicts location based on microbiome composition.

The feature selection step employed the 'floating search' algorithm (Pudil et al., 1994), while the classification step involved the use of Linear Discriminant Analysis (LDA) applied to the selected genera to predict the location from which the sample was taken. To ensure the validity of this approach, each of the feature selection and locational

classification steps were performed within a cross-fold validation process. Each fold involved withholding data from the sample sites within a single mouse, then testing the classifier's ability to predict the location from which the withheld samples were drawn. For visualization purposes, a computational voting process identified the relevant genera across all instances of feature selection (Table 1). Two separate overall accuracies were reported for each plot (Figs. 4 and 5) based on whether the genera were selected using floating search within each fold of the cross-validation process (compiled to give the overall accuracy), or using the voting approach. The 'floating search within each fold' accuracy is the more conservative technique, while the vote-generated set of genera was used for visualization because a fixed set of genera is required when plotting results.

Table 1: Selected genera visualized in Figs. 4 & 5*

Fig. 2 (left panel)	Fig. 2 (right panel)	Fig. 3 (upper left panel)	Fig. 3 (upper right panel)	Fig. 3 (lower left panel)	Fig. 3 (lower right panel)
Oscillibacter	*Lactobacillus*	*Lactobacillus*	*Sporacetigenium*	*Oscillibacter*	*Bifidobacterium*
Lactobacillus	*Dorea*	*Parabacteroides*	*Lactobacillus*	*Robinsoniella*	*Dorea*
Robinsoniella	*Turicibacter*	*Bacteroides*	*Butyricicoccus*	*Dorea*	*Parasutterella*
Ruminococcus	*Oscillibacter*	*Turicibacter*	*Ruminococcus*	*Butyricimonas*	*Turicibacter*
Barnesiella	*Sporacetigenium*	*Robinsoniella*	*Coprobacillus*	*Alistipes*	*Anaerotruncus*
Dorea	*Robinsoniella*	*Butyricimonas*	*Oscillibacter*	*Ruminococcus*	*Akkermansia*
Coprobacillus	*Akkermansia*	*Mucispirillum*	*Parabacteroides*	*Barnesiella*	*Lactobacillus*
Coprococcus	*Marvinbryantia*	*Barnesiella*	*Asaccharobacter*	*Bacteroides*	*Sporobacter*
Butyricimonas	*Asaccharobacter*	*Holdemania*	*Johnsonella*	*Coprobacillus*	*Coprobacillus*
Blautia	*Anaerotruncus*	*Lactonifactor*	*Turicibacter*	*Enterorhabdus*	*Bacteroides*
Turicibacter	*Bacteroides*	*Anaerovorax*	*Roseburia*	*Blautia*	*Robinsoniella*
Mucispirillum	*Butyricicoccus*	*Marvinbryantia*	*Dorea*	*Holdemania*	*Parabacteroides*
Anaerotruncus	*Coprobacillus*	*Sporobacter*		*Parabacteroides*	*Johnsonella*
Parabacteroides	*Papillibacter*				*Holdemania*
	Sporobacter				*Allobaculum*
					Marvinbryantia

*Genera within each column are listed in their rank order, which represents the frequency that the floating search chose them, affirmed by cross-fold validation, and weighted by LDA performance in the voting process.

When LDA was applied to feature selected data from four sites that are well separated and anatomically distinct—the ileum, caecum, proximal colon and distal

colon—we found that the selected genera separated readily into clusters, thereby indicating biogeography with respect to microbiome composition (Fig. 4). The process is robust given that the inclusion of samples from two very different strains of mice (one inbred, the other outbred), and from two independent experiments a year apart, produced highly similar results (Fig. 4).

An attractive and important outcome of the combined feature selection and LDA analysis is that the visualization reflects the biogeography of the sampled communities. That is, data-point clusters from sample sites closer together in the intestinal tract tend to be situated near one another, while clusters from locations farther apart tend to be more separated (Fig. 4). Thus, the biogeographical relationships between microbiome communities located longitudinally along the mouse lower intestine appear to be a function of the degree of observable environmental differentiation between locations. For example, the ileum, a compartment physically proximal to, but environmentally distinct from, the caecum is widely separated from that compartment based on this analysis (Fig. 4).

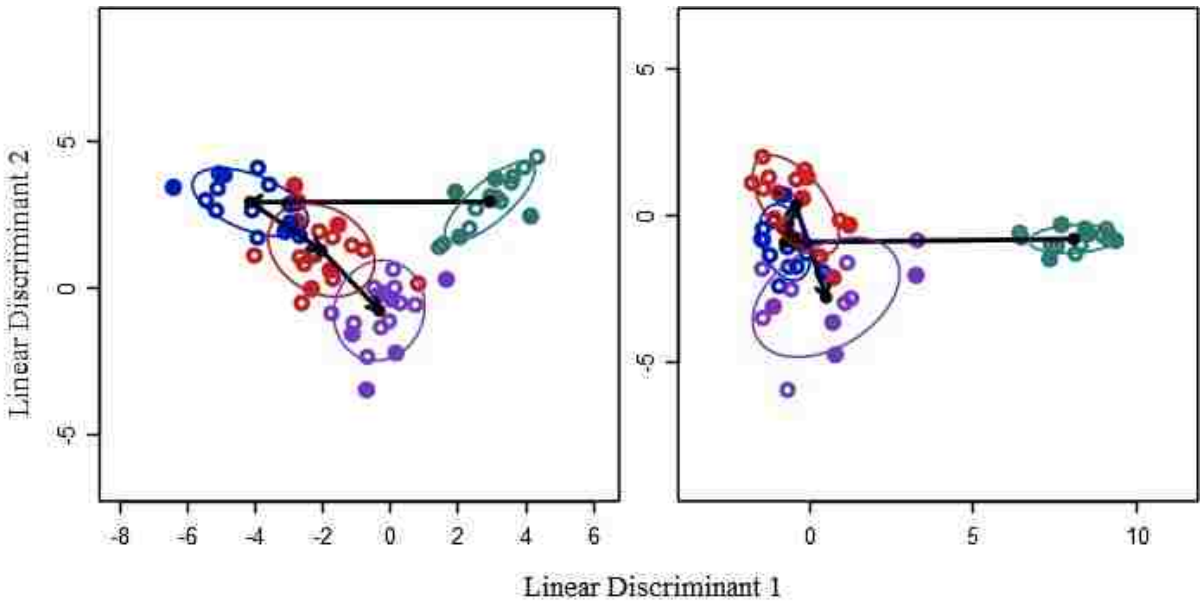


Figure 4: Linear Discriminant Analysis (LDA) of the four main compartments sampled from C57Bl/6 strain mice (left panel) and CD-1 strain mice (right panel). Filled circles and open circles represent cohorts 1 and 2, respectively. ● Ileum, ● Caecum, ● Proximal Colon, ● Distal Colon. Black dots represent the centroid for each cluster and ellipses indicate 1 standard deviation. The arrows show the flow of digesta between chambers. The plots were made using vote-determined genera. The accuracies were 78.79% (62.12%) (left panel) and 63.93% (65.57%) (right panel). The first accuracies listed used vote-determined genera, while accuracies in parentheses were for genera identified using 'floating search within each fold'.

In terms of biogeographical resolution, this study further shows that samples from different regions within the same chamber (e.g. within the caecum or within the colon) can also be differentiated, suggesting that localized physicochemical variations within compartments are sufficiently strong to produce differences in microbiome composition. For example, microbiome composition of the main body of the caecum is distinct from that of the tip of the caecum (Fig. 5, top panels), presumably reflecting differences such as moisture content, solidity and composition of digesta, and greater bacterial density in the tip (as indicated by microbiome DNA yield per mg of digesta; not shown).

Microbiome composition also differed among the proximal, middle and distal colon, which exhibited visibly distinct features (morphology, digesta/fecal pellet appearance) that were used to select sampling sites (Fig. 5).

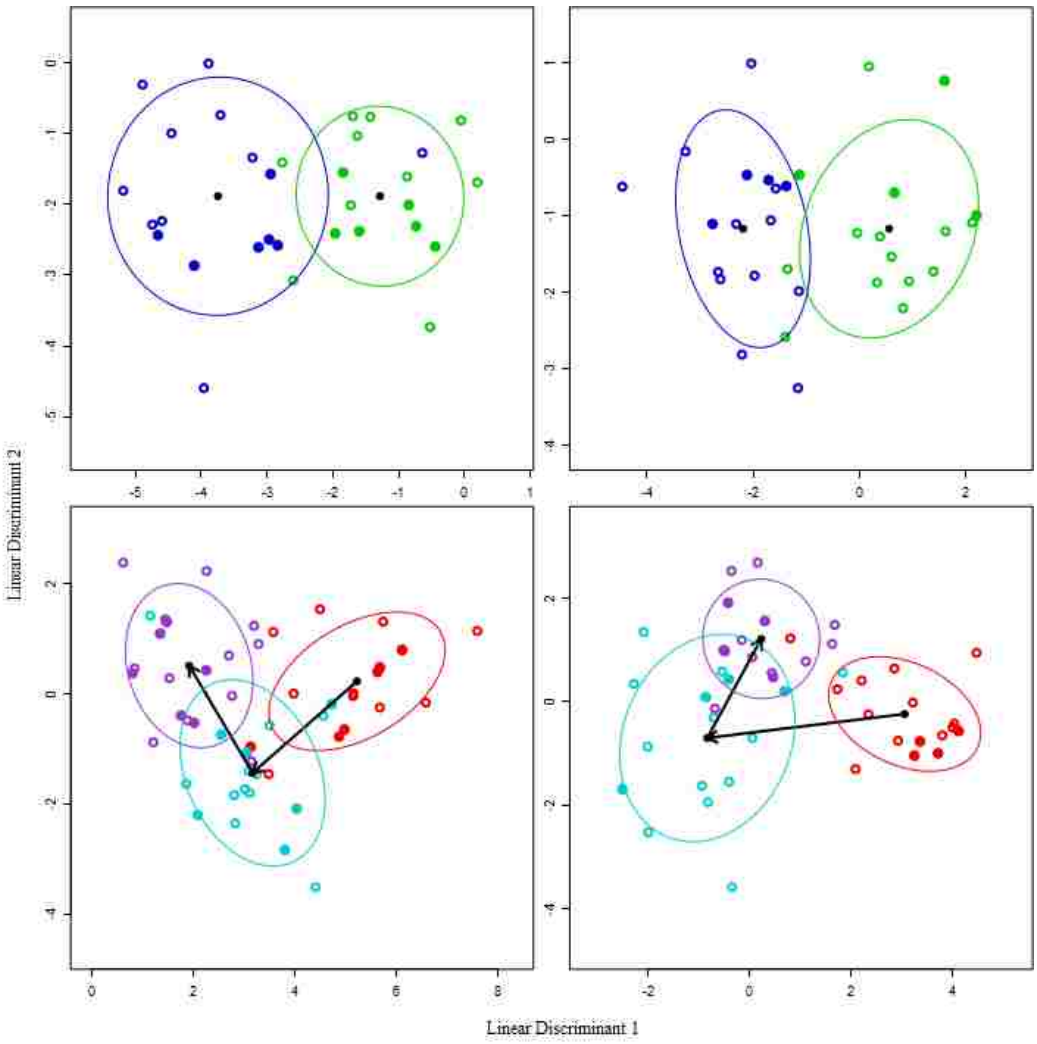


Figure 5: LDA of the Tip of the Cecum and Cecum (top panels) and Proximal, Mid and Distal Colon (bottom panels) for C57Bl/6 mice (left) and CD-1 mice (right). ● Cecum, ● Tip of the Cecum, ● Proximal Colon, ● Mid-Colon, ● Distal Colon. Filled circles and open circles represent cohorts 1 and 2, respectively. Black dots represent the centroid for each cluster and ellipses indicate 1 standard deviation. The plots were made using vote-determined general. The accuracies were 93.55% (77.42%) (top left), 71.88% (62.50%) (top right), 62.00% (52.00%) (bottom left), 58.70% (50.00%) (bottom right). The first accuracies listed used vote-determined genera, while accuracies in parentheses were for genera identified using 'floating search within each fold'.

Given the genus-level resolution afforded by this approach, we also assessed whether there was a distinguishable 'core microbiome' for the different sampling locations. To detect core genera, we examined all samples from each location in cohorts 1 and 2 to determine whether there were any genera present in all individuals from both mouse strains at that location. Even given these strict criteria, some genera qualified as core constituents of the microbiome for each intestinal location (Table 2). Differences in the core microbiome for each location support the differences seen in the LDA plots, providing further evidence for longitudinal biogeography along the lower intestinal tract. The core microbiome for each location did not change as a function of sequence depth as determined for samples having 3,000, 2,500, 2,000, or 1,500 sequences (not shown), suggesting that the identified core genera and the criteria used to detect them are insensitive to sampling depth, both in terms of identity and proportional composition. Interestingly, core genera that reside in more than one intestinal location (e.g. *Lactobacillus*) represent a different proportion of the microbiome depending on location, which we suggest is a manifestation of broader tolerance of varying environmental conditions (e.g. pH and oxygen), while still exhibiting a preferred or optimal site (niche).

To date, it has not been possible to determine a 'universal' laboratory mouse core microbiome due to differences between strains and environments at different animal facilities (Ericsson et al., 2015). Similarly, it seems likely that efforts to establish wild mouse, human, or other core microbiomes might be challenged by differences in local foods and environmental parameters. Nonetheless, the data herein support the concept of a core microbiome at each sampling location for both cohorts and both strains of mice in

our study, likely because the experimental design tightly controlled for potentially confounding environmental factors.

Table 2: Core microbiome genera for each sampled lower intestinal tract location

Ileum	Caecum	Tip of Caecum	Proximal Colon	Mid Colon	Distal Colon
Lactobacillus **60.76%*** **(12.54-90.62%)§**	***Lactobacillus*** **11.80%** **(2.55-21.12%)**	***Lactobacillus*** **8.62%** **(0.53-18.11%)**	***Lactobacillus*** **14.83%** **(2.61-40.73%)**	***Lactobacillus*** **28.94%** **(6.36-57.15%)**	***Lactobacillus*** **22.90%** **(1.52-63.46%)**
Turicibacter **7.19%** **(0.17- 57.63%)**	***Dorea*** **0.58%** **(0.18-1.45%)**	***Dorea*** **0.74%** **(0.19-1.90%)**	***Dorea*** **0.39%** **(0.08-1.54%)**	***Dorea*** **0.24%** **(0.08-1.08%)**	
	Oscillibacter **1.61%** **(0.86-4.16%)**	***Oscillibacter*** **1.60%** **(0.52-2.76%)**	***Oscillibacter*** **1.07%** **(0.18-3.57%)**	***Oscillibacter*** **0.34%** **(0.05-2.04%)**	
	Robinsoniella **0.75%** **(0.07-3.05%)**	***Robinsoniella*** **0.85%** **(0.05-3.81%)**	***Bacteroides*** **0.51%** **(0.05-2.46%)**	***Bacteroides*** **1.16%** **(0.10-7.49%)**	***Bacteroides*** **0.88%** **(0.09-4.27%)**
		Parabacteroides **0.54%** **(0.04-1.68%)**	***Coprobacillus*** **0.24%** **(0.01-1.53%)**		***Holdemania*** **0.31%** **(0.03-1.87%)**

*Percentage given is the median value for cohorts 1 and 2 and both strains.

§Percentages in parentheses are the range of values for cohorts 1 and 2 and both mouse strains.

Discussion:

Previously, a variety of methods have been used to distinguish between locations in the lower intestinal tract that either involve beta diversity measures, basic statistical analyses with PCoA or PCA used to visualize biogeography in the lower intestinal tract. These methods have been unsuccessful at discriminating differences between locations. One particular issue is the lack of whole community assessments that give a sense of which taxa contribute to the differences between treatments. At most, this type of study

has specified that selected genera are present or absent based on evidence that the taxa in question may be physiologically important to the host (Ding and Schloss, 2014; Zhang et al., 2014). The most successful study to date that addresses these problems is a human-based on microarray analysis from colonoscopy biopsies to examine biogeography; however, in addition to using methods from numerical ecology, it used pairwise comparisons between sites rather than multivariate methods (Aguirre de Carcer et al., 2011). One could also argue that the use of a micro-array is, in itself, somewhat biased as a form of subsampling taxa of interest.

The current study shows that we can discriminate genus-level differences in microbiome composition among geographically distinct intestinal environments, i.e. biogeography. The notion of a biogeographically-focused microbiome controlled by the local environment is consistent with the current vision of the ecophysiology of the GI tract, where food becomes progressively digested, degraded and transformed through host and microbial activities. Importantly, this approach identifies the genera that are exhibiting differential localization (Table 1). The observed localization supports the hypothesis that these genera are responsive to their local surroundings and potentially key in microbiome functionality rather than just transitory taxa introduced with food or from the environment (i.e. just passing through).

This demonstration of genus-level biogeography was possible because of the novel combination of computational, statistical, and bioinformatic approaches, namely feature selection, LDA, and voting, that comprise the data analysis workflow described herein. Through broader application to existing datasets, this new workflow may add value to prior studies where the effects of various diseases, treatments, or environmental

parameters on microbiome or microbial community composition were obscure or of lower resolution.

Biogeography in the intestinal microbiome very likely influences host health, immune system function, and the ability to digest food and absorb microbial and host metabolites. For example, it has been shown that nutrients that affect bacteria in the GI tract may have a profound impact on multifactorial diseases such as cancer or obesity simply by changing how bacterial metabolites are distributed longitudinally along the colon (Cani, 2014; Louis et al., 2014). In addition, microbiomes that are able to block pathogens from colonizing their preferred environments through competitive exclusion or predation may provide host resistance to pathogens, thereby explaining why some individuals are less susceptible to diseases than others. Enhanced knowledge of the distribution and activities of intestinal microbiome taxa afforded by this and other approaches should lead to greater understanding of host health, better insights into the etiology of many pathogen-based and chronic illnesses, and new or better-targeted therapeutics, prebiotics or probiotics.

Literature Cited:

Aguirre de Carcer, D., Cuiv, P.O., Wang, T., Kang, S., Worthley, D., Whitehall, V., Gordon, I., McSweeney, C., Leggett, B., and Morrison, M. (2011). Numerical ecology validates a biogeographical distribution and gender-based effect on mucosa-associated bacteria along the human colon. *The ISME journal 5*, 801-809.

Apajalahti, J.H., Sarkilahti, L.K., Maki, B.R., Heikkinen, J.P., Nurminen, P.H., and Holben, W.E. (1998). Effective recovery of bacterial DNA and percent-guanine-plus-cytosine-based analysis of community structure in the gastrointestinal tract of broiler chickens. *Applied and environmental microbiology 64*, 4084-4088.

Baas-Becking, L.G.M. (1934). *Geobiologie of Inleiding Tot de Milieukunde [Geobiology or Introduction to the Science of the Environment]* (The Hague, The Netherlands: W. P. Van Stockum and Zoon).

Cani, P.D. (2014). Metabolism in 2013: The gut microbiota manages host metabolism. *Nature reviews Endocrinology 10*, 74-76.

Cani, P.D., Osto, M., Geurts, L., and Everard, A. (2012). Involvement of gut microbiota in the development of low-grade inflammation and type 2 diabetes associated with obesity. *Gut microbes 3*, 279-288.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I*., et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods 7*, 335-336.

Chang, P.V., Hao, L., Offermanns, S., and Medzhitov, R. (2014). The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. *Proceedings of the National Academy of Sciences of the United States of America 111*, 2247-2252.

Ding, T., and Schloss, P.D. (2014). Dynamics and associations of microbial community types across the human body. *Nature 509*, 357-360.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A. (2005). Diversity of the human intestinal microbial flora. *Science 308*, 1635-1638.

Ericsson, A.C., Davis, J.W., Spollen, W., Bivens, N., Givan, S., Hagan, C.E., McIntosh, M., and Franklin, C.L. (2015). Effects of vendor and genetic background on the composition of the fecal microbiota of inbred mice. *PloS one 10*, e0116704.

Heinzmann, S.S., and Schmitt-Kopplin, P. (2015). Deep metabotyping of the murine gastrointestinal tract for the visualization of digestion and microbial metabolism. *Journal of proteome research 14*, 2267-2277.

Hu, S., Wang, Y., Lichtenstein, L., Tao, Y., Musch, M.W., Jabri, B., Antonopoulos, D., Claud, E.C., and Chang, E.B. (2010). Regional differences in colonic mucosa-associated microbiota determine the physiological expression of host heat shock proteins. *American journal of physiology Gastrointestinal and liver physiology 299*, G1266-1275.

Hwang, C.L., and Masud, A.S.M. (1979). *Multiple Objective Decision Making, Methods and Applications: A State-of-the-Art Survey* (Springer-Verlag).

Kawamata, K., Hayashi, H., and Suzuki, Y. (2006). Chloride-dependent bicarbonate secretion in the mouse large intestine. *Biomedical research 27*, 15-21.

Lavelle, A., Lennon, G., Docherty, N., Balfe, A., Mulcahy, H.E., Doherty, G., D, O.D., Hyland, J.M., Shanahan, F., Sheahan, K.*, et al.* (2013). Depth-dependent differences in community structure of the human colonic microbiota in health. *PloS one 8*, e78835.

Li, J., Butcher, J., Mack, D., and Stintzi, A. (2015). Functional impacts of the intestinal microbiome in the pathogenesis of inflammatory bowel disease. Inflammatory bowel diseases *21*, 139-153.

Louis, P., Hold, G.L., and Flint, H.J. (2014). The gut microbiota, bacterial metabolites and colorectal cancer. *Nature reviews Microbiology 12*, 661-672.

McLean, M.H., Dieguez, D., Jr., Miller, L.M., and Young, H.A. (2015). Does the microbiota play a role in the pathogenesis of autoimmune diseases? *Gut 64*, 332-341.

Nava, G.M., Friedrichsen, H.J., and Stappenbeck, T.S. (2011). Spatial organization of intestinal microbiota in the mouse ascending colon. *The ISME journal 5*, 627-638.

Pocheville, A. (2015). The ecological niche: history and recent controversies. In *Handbook of Evolutionary Thinking in the Sciences*, T.H. Heams, Philippe; Lecointre, Guillaume et al., ed. (Dordrecht: Springer), pp. 547–586.

Pudil, P., Novovicova, J., and Kittler, J. (1994). Floating Search Methods in Feature-Selection. *Pattern Recognition Letters 15*, 1119-1125.

Rand Corporation (2001). A Million Random Digits with 100,000 Normal Deviates (Glencoe, IL USA: Rand Corporation), pp. B1 - 8.

Rogers, G.B., Kozlowska, J., Keeble, J., Metcalfe, K., Fao, M., Dowd, S.E., Mason, A.J., McGuckin, M.A., and Bruce, K.D. (2014). Functional divergence in gastrointestinal microbiota in physically-separated genetically identical mice. *Scientific reports 4*, 5437.

Rolig, A.S., Cech, C., Ahler, E., Carter, J.E., and Ottemann, K.M. (2013). The degree of *Helicobacter pylori*-triggered inflammation is manipulated by preinfection host microbiota. *Infection and immunity 81*, 1382-1389.

Sarma-Rupavtarm, R.B., Ge, Z., Schauer, D.B., Fox, J.G., and Polz, M.F. (2004). Spatial distribution and stability of the eight microbial species of the altered schaedler flora in the mouse gastrointestinal tract. *Applied and environmental microbiology 70*, 2791-2800.

Schubert, A.M., Sinani, H., and Schloss, P.D. (2015). Antibiotic-induced alterations of the murine gut microbiota and subsequent effects on colonization resistance against *Clostridium difficile*. *mBio 6*.

Swidsinski, A., Loening-Baucke, V., Lochs, H., and Hale, L.P. (2005). Spatial organization of bacterial flora in normal and inflamed intestine: a fluorescence in situ hybridization study in mice. *World Journal of Gastroenterology* : WJG *11*, 1131-1140.

Trunk, G.V. (1979). A problem of dimensionality: a simple example. *Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-1*, 306 - 307.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature 444*, 1027-1031.

Turnbaugh, P.J., Ridaura, V.K., Faith, J.J., Rey, F.E., Knight, R., and Gordon, J.I. (2009). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine 1*, 6ra14.

West, C.E., Jenmalm, M.C., and Prescott, S.L. (2014). The gut microbiota and its role in the development of allergic disease: a wider perspective. *Clinical and experimental allergy : Journal of the British Society for Allergy and Clinical Immunology*.

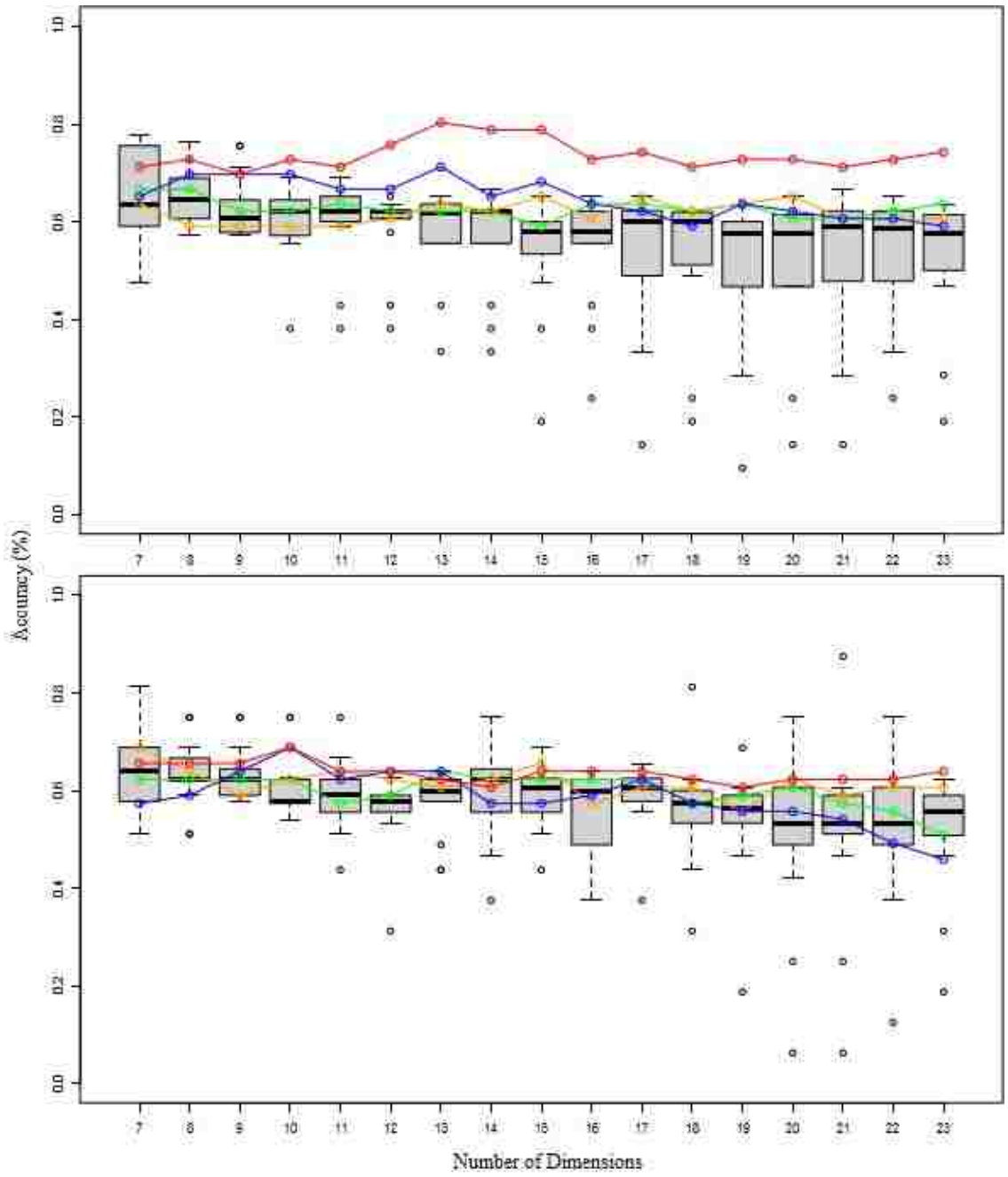
Yurist-Doutsch, S., Arrieta, M.C., Vogt, S.L., and Finlay, B.B. (2014). Gastrointestinal Microbiota-Mediated Control of Enteric Pathogens. Annual review of genetics.

Zackular, J.P., Baxter, N.T., Iverson, K.D., Sadler, W.D., Petrosino, J.F., Chen, G.Y., and Schloss, P.D. (2013). The gut microbiome modulates colon tumorigenesis. *mBio 4*, e00692-00613.

Zhang, Z., Geng, J., Tang, X., Fan, H., Xu, J., Wen, X., Ma, Z.S., and Shi, P. (2014). Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *The ISME journal 8*, 881-893.

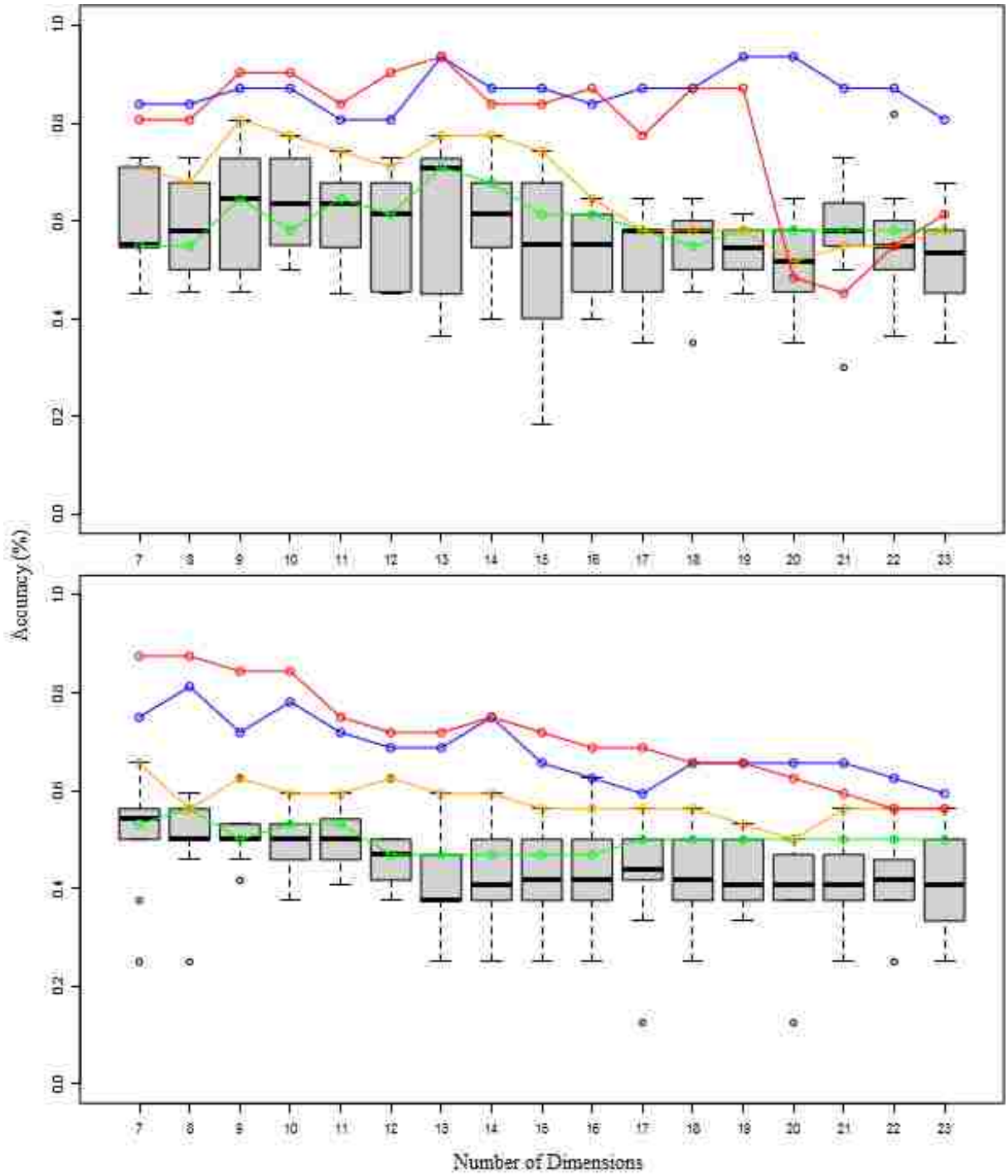
Zoetendal, E.G., von Wright, A., Vilpponen-Salmela, T., Ben-Amor, K., Akkermans, A.D., and de Vos, W.M. (2002). Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces. *Applied and environmental microbiology 68*, 3401-3407.

Supplementary Figures:



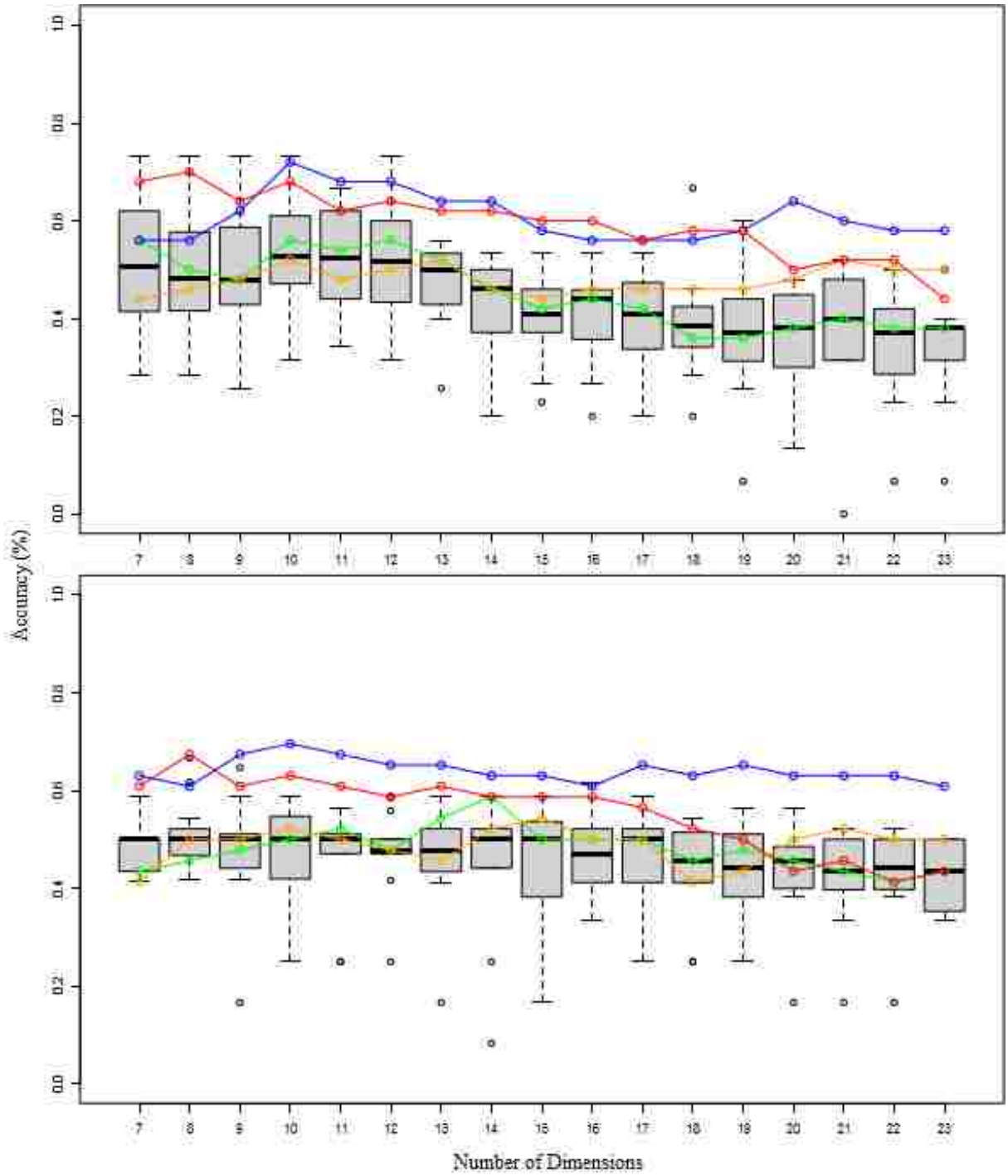
Supplementary Figure S1: Box plots showing cross-validation accuracies when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was Cohorts 1&2 with mouse strain B6 (top) and strain CD1 (bottom) filtered to 4 chambers: Ileum, Cecum, Proximal Colon, and Distal Colon. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature

selection inside of cross-validation was compiled and used to select genera outside (before) cross-validation was performed.



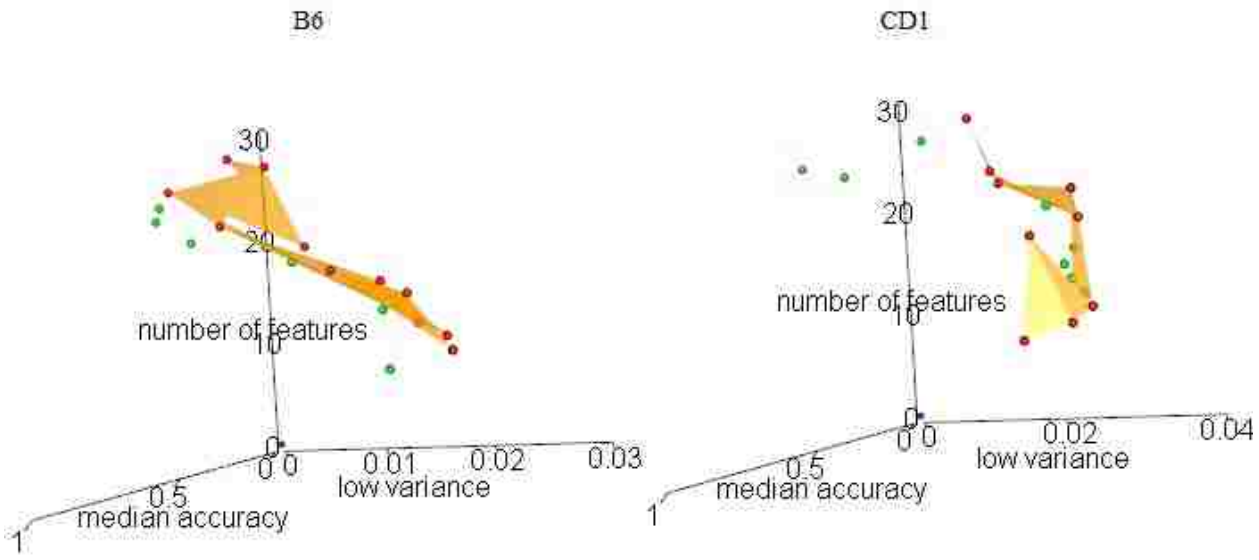
Supplementary Figure S2: Box plots showing cross-validation accuracies when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was Cohorts 1&2 with mouse strain B6 (top) and strain CD1 (bottom) filtered to 2 chambers: Cecum and Tip of Cecum. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature selection inside of

cross-validation was compiled and used to select genera outside (before) cross-validation was performed.

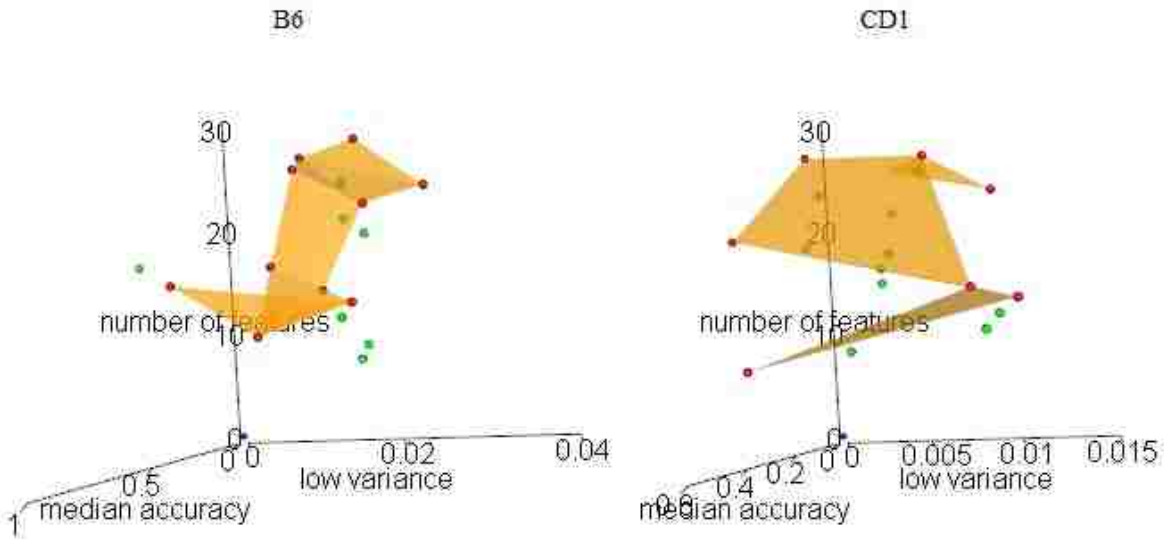


Supplementary Figure S3: Box plots showing cross-validation accuracies when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was Cohorts 1&2 with mouse strain B6 (top) and strain CD1 (bottom) filtered to 3 chambers: Proximal Colon, Mid Colon, and Distal Colon. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature

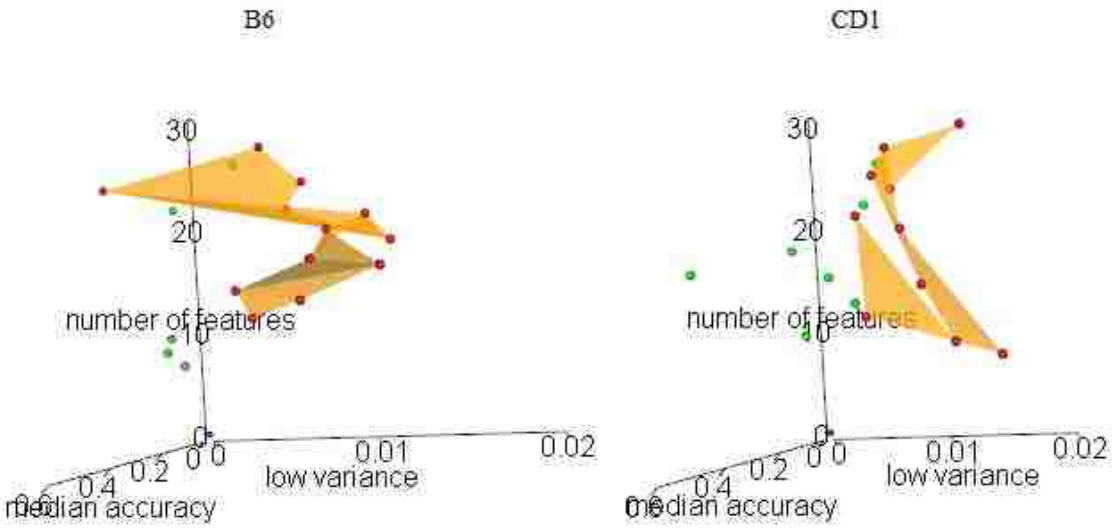
selection inside of cross-validation was compiled and used to select genera outside (before) cross-validation was performed.



Supplementary Figure S4: Scatter plot of the 3D-Pareto frontiers when Supplementary Figure 4 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. Green points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.



Supplementary Figure S5: Scatter plot of the 3D-Pareto frontiers when Supplementary Figure 5 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. Green points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.



Supplementary Figure S6: Scatter plot of the 3D-Pareto frontiers when 6 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. Green points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.

Code Listing 1

```
Function LMO-CV( matrix, n_genera ):
    n_samples = number of rows of matrix
    n_correct = 0
    for each mouse m (LMO fold):
        The test samples are those from mouse m
        The training samples are all remaining samples
        Do floating search on training samples using n_genera genera
        Build LDA model using training data filtered to selected genera
        Predict test data using LDA model
        n_correct += number samples correctly classified
    return n_correct/n_samples
```

Code Listing 2

```
Function choose_genera( matrix, n_genera, information generated by LMO-CV )
    for each LMO-CV:
        cv_accuracy = accuracy returned by LMO-CV
        for each genus g:
            count_folds[g] = 0
            for each fold of CV:
                if fold chooses g then count_folds[g] += 1
            accuracy[g] += count_folds[g]*cv_accuracy
    sort genera into descending order by accuracy[g]
    Return set of the best n_genera genera
```

Supplementary Figure S7: Example code listings for generation of LDA plots and cross validation, respectively.

Chapter 4: Carpe Diem – *C. difficile* invasion and the Intestinal Microbiome

Abstract:

Clostridium difficile associated disease has a high (5-9% in 2011) and growing prevalence rate in hospitalized patients and appears to be moving to increased prevalence in long-term care and community reservoirs. It is an excellent model for the study of interactions between opportunistic pathogens and the intestinal microbiome, in part because the antibiotic treatment that enables *C. difficile* invasion in the intestine can be experimentally manipulated. Here, a detailed examination of the distribution of and associations between CDAD and commensal bacterial taxa (the microbiome) combined with the biogeography of the intestinal microbiome provides insight into how such associations facilitate or mitigate the progression of *Clostridium difficile* associated disease. Major changes were found in both the core micriobiome and distribution maps of major taxa during both antibiotic treatment and *C. difficile* infection. In addition we found that C. difficile appears to affect the intestinal microbiome when introduced without prior antibiotic treatment. Finally, we found that *C. difficile* distribution along the intestinal tract changed during vancomycin treatment, suggesting that in some individuals *C. difficile* may be sequestered in the cecum and appendix during vancomycin treatment, thus resulting in recurrence of the disease following the discontinuation of treatment.

Introduction:

Opportunistic bacterial pathogens such as *Clostridium difficile (C. difficile)* exhibit many hallmarks of invasive macroorganisms such as plant and animal pests that

operate at landscape scales of meters, kilometers and greater. The GI tract represents a landscape of environmental parameters that vary at centimeter to meter scales yet produce quite distinct environments ('compartments') distinguished by factors such as pH, enzymatic activities, oxygen status, nutrients, and water availability. Like larger invasive organisms, invading opportunistic bacteria must become established in the face of the indigenous community (here, the host microbiome) to cause disease. However, little is known regarding the biogeographical distribution of indigenous microbiome taxa across the GI landscape and, until recently, studies have failed to examine the positive and negative associations between *C. difficile* and the component taxa of the microbiome across this landscape.

C. difficile is a good model for the study of mechanisms that predispose for disease caused by opportunistic pathogens. It is one of the most emergent nosocomial (i.e. hospital-acquired) diseases in the world (Clements et al., 2010; Shin et al., 2008; Tae et al., 2009; Weiss et al., 2009). Of the individuals that contract the disease, 20 – 30% will develop recurrent disease, leading to life-threatening complications. Basic conditions for precipitating *C. difficile* associated disease (CDAD) have been described and involve gross changes to the intestinal microbiome due to prior antibiotic treatment for other infections (Bartlett, 1979; Viswanathan et al., 2011). Several good mouse models for CDAD have been developed that exhibit a similar etiology to the human disease, making unraveling the invasion and establishment processes more tractable (Chen et al., 2008; Lawley et al., 2009; Reeves et al., 2011).

While the clinical conditions that precipitate changes in the intestinal microbiome leading to CDAD are reasonably well understood, how these changes permit *C. difficile*

to initially invade or allow its recurrence is not well understood, although recent studies done on *C. difficile* colonization and the intestinal microbiome suggest that the composition of the microbiome has a large effect on whether the initial colonization by *C. difficile* will be successful (Schubert et al., 2015). It is also unknown whether or where *C. difficile* becomes sequestered during therapeutic treatment in those who suffer from recurrent infections, though the appendix has been suggested as a potential site for sequestration (Bollinger et al., 2007; Laurin et al., 2011).

Interestingly, in many invasive systems, including plants, animal pests, and disease systems, invasive organisms have been shown to be capable of altering the invaded environment (so-called 'ecological engineering') to make it more hospitable to the invader and more hostile to its indigenous competitors (Cuddington and Hastings, 2004; Karatayev et al., 2007; Knodler et al., 2010; Nolte, 2011). *C. difficile* has also been shown to be an ecosystem engineer—using a form of 'phenotypic noise' to clear an area for colonization (Ackermann et al., 2008). In this scenario, a proportion of the *C. difficile* population lyses itself to release TcdA, a toxin which causes inflammation in host cells, allowing the remaining viable *C. difficile* population to invade the newly cleared area. How important the ability to engineer colonization is when compared to the greater generalized impact of community perturbation due to antibiotic use, is difficult to determine, but changes in the microbiome due to *C. difficile* alone might play a part in both initial colonization and recurrent disease. Thus, although we know that the intestinal microbiome as a whole changes when antibiotics are administered, and that *C. difficile* appears to take advantage of these changes, we have no knowledge of what interactions occur between *C. difficile* and other members of the intestinal microbiome,

either during disease, when *C. difficile* is present without disease, or during the phenomenon of recurrence.

Although excellent work has been done recently on the effect of microbiome interactions on *C. difficile* invasion, those models relied on fecal samples alone to explore the interactions (Schubert et al., 2015). In addition, the effects and outcomes of *C. difficile* introduction alone when not preceded by antibiotic treatment have not been studied in detail. The primary goal of the current study were to examine a recurrence model of *C. difficile* to determine whether recurrence could be due, at least in part, to sequestration of *C. difficile* in a particular gut compartment. Another goal was to determine whether the composition and biogeography of the lower gut microbiome plays a role in successful colonization. Finally, the effects of *C. difficile* when not preceded by antibiotic treatment were examined to assess whether *C. difficile* itself has an effect on the intestinal microbiome.

Methods:

Animal Model and Sampling:

Animal Model

All animal treatments were approved by the Institutional Animal Care and Use Committee (IACUC) at the University of Montana under AUP# 045-13. Three week old-female C57Bl/6 mice were purchased from Envigo (http://www.envigo.com/about-envigo/ (formerly Harlan Laboratories)) and housed locally until 21 weeks of age to allow their intestinal microbiome to acclimate to environmental conditions in the animal care facility at the University of Montana. During this time, they were handled on a daily

basis, and in addition to their regular food (NIH 31), were fed small amounts of fresh apple each day.

Two weeks prior to the start of the experiment, the mice were weighed and categorized into three weight classes: low (15 – 21 g), medium (> 21 – 30 g) and high (> 30 g) to facilitate the administration of antibiotics and other treatments. The housing was rearranged so that mice from one weight range were caged together in groups of 3 per cage. The day prior to the start of the experiment, the mice were weighed again. The three weight classes were maintained to allow for more accurate dosage of the mice with antibiotics, but to accommodate the experimental protocol of 6 mice per group, the middle class was subdivided either into the low weight class (≤ 27 G. ea.) or high weight class (> 27 G. ea.). These two groups consisted of equal numbers of cages, which were randomly assigned into experimental groups using a random number table (Rand Corporation, 2001), but allowing for one high- and one low-weight cage (2 cages, 6 mice total) per sampling time for a total of 36 mice per group (refer to Fig. 1).

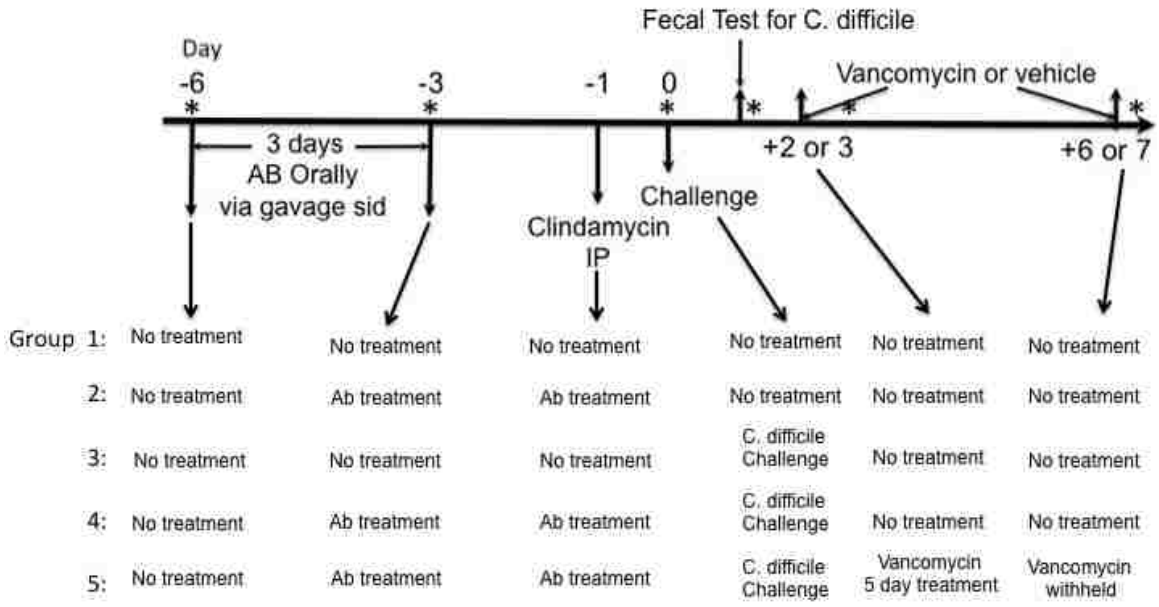


Figure 1: Experimental groups and schematic diagram of the experimental set-up. The point of administration of *C. difficile* is day 0. The 6 sampling timepoints are indicated with asterisks (Chen et al., 2008).

All animal treatments followed the experimental protocol described by Chen et al. (2008) model of *C. difficile* infection in mice (Fig. 1) with the following modifications: All mice received small amounts of pasteurized applesauce in addition to their food, and all food and bedding was pre-sterilized from this point on in the study. According to weight class, mice were given a cocktail of antibiotics (kanamycin (40 mg/kg), gentamicin (3.5 mg/kg), colistin (4.2mg/kg), metronidazole (21.5 mg/kg), and vancomycin (4.5 mg/kg), or a sterile water placebo via gavage each day over a 3-day period. The relevant groups of mice (groups 2, 4 and 5) were given clindamycin IP (10 mg/kg) 30 h prior to challenge with *C. difficile*. Cages of mice to be challenged with *C. difficile* (groups 3, 4 and 5) were moved into the BSL II area of the animal care facility. These mice were inoculated with 10^5 CFU of *C. difficile* strain BI17 spores in sterile water via gavage. Strain BI17 was used because it had been demonstrated in this model to cause CDAD without concurrent mortality (Chen et al., 2008). One day (24 h) post-

inoculation, vancomycin (50 mg/kg) was administered to the relapse group (group 5) once a day for 7 days. Following this, the mice were observed for clinical signs of CDAD twice daily to monitor for indications of relapse (e.g. febrility, weakness, diarrhea).

Table 1: List of experimental abbreviations used. Numbers following abbreviations denote experimental time points (refer to Fig. 1).

H_2O1	Negative control, time point 1
ABX2	Antibiotic control, time point 2 (received antibiotic but not challenged)
cH_2O4	*C. difficile* control (challenged, but no antibiotic treatment), time point 4
cABX 4	received antibiotics prior to challenge with *C. difficile*, time point 4
Van5	Antibiotic treatment, challenged with *C. difficile* and treated with vancomycin, time point 5
Van6	Antibiotic treatment, challenged with *C. difficile*, treated with vancomycin, then sampled after vancomycin therapy was stopped (relapse), time point 6
time1, 2, etc.	Shorthand for time point 1, 2, etc.

C. difficile *Spore Preparation and Plating*

BI17 cells were received frozen from Dale N Gerding, MD and were used to produce spores for gavage. Spores were produced in one large batch. Brain-Heart Infusion (BHI) plates were reduced overnight using the anaerobic GasPak system and chambers (Becton, Dickinson and Company, Franklin Lakes, NJ). One hundred ml of glycerol stock cells were used to inoculate 4 Blood Agar Plates (BAP) (company ,state), which were placed in anaerobe chambers with 2 GasPaks and an anaerobic indicator strip, then incubated at 37˚ C for 6 d.

The plates were scraped of cells and spores using a sterile spatula and then transferred into 10 ml of 1X Phosphate Buffer Solution (pH 7.4) in a 35 ml Oak Ridge

tube and centrifuged for 5 minutes at 6,500 x *g*. The pellet was resuspended with 5 ml of PBS and then heat-shocked at 56˚ C. for 10 minutes to kill any remaining vegetative cells, then incubated on ice for 10 min. This suspension was again centrifuged at 6,500 x *g*. , after which the pellet was resuspended in 9.0 mL PBS and stored at -70˚ C. until further use.

Spores were titrated via serial dilution IN PBS and plating onto pre-reduced (i.e. anaerobic) TFA plates (Merrigan et al., 2010), then incubated overnight in an anaerobic GasPak chamber at 37˚ C.

Sampling

As at time point 1 all mice were equivalent to negative control mice, it was deemed unnecessary to sample 30 mice (i.e. 6 from each of the 5 treatment groups). Instead, 12 mice were sampled at timepoint 1, prior to the beginning of the experiment. The subsequent timepoints were at -3 d (pre-infection), just prior to infection, 24 h post-infection, 4 d post infection and 9 d post-infection (this last timepoint was empirically determined as when clinical signs showed the mice had active CDAD (i.e. they relapsed).

At each sampling timepoint, mice were euthanized humanely in a CO_2 chamber. Intestinal tract samples were then surgically removed from each of 6 sampling points (Fig. 2). The sampling points comprised the distal ileum (defined as the last 3 cm of the ileum), the cecum, the tip of the cecum, the proximal colon (the region nearest the cecum containing liquid digesta), the mid-colon (the mid-portion containing the first-formed, soft, ‘pre-fecal’ pellets), and the distal colon (defined as the last 2 cm including fully formed fecal pellets and the rectum). Due to the low numbers of mice being sampled and

the need to keep experimental groups separate, randomization of sampling was not required.

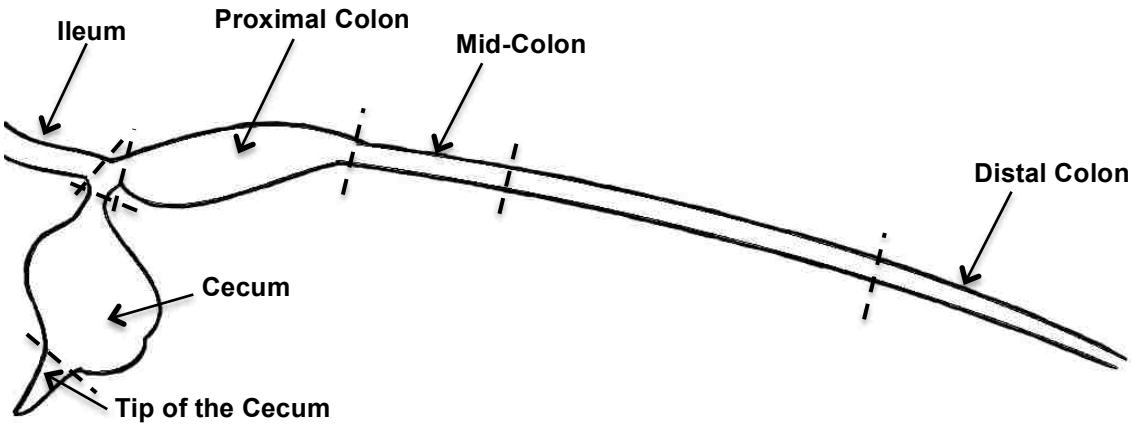


Figure 2: Sampling sites along the mouse lower intestine.

The luminal contents of surgically removed samples were recovered by gentle squeezing into sterile microcentrifuge tubes, which were kept on ice during collection then stored at -70 ˚C prior to downstream processing. All samples from within each mouse were considered to be linked and therefore were processed for microbiome DNA recovery, PCR amplification and sequencing together. The processing order of the samples was randomized using a random number table (Rand Corporation, 2001).

Microbiome DNA Recovery, Amplification and Sequencing.

Microbiome DNA was recovered using a protocol adapted from Apajalahti *et al*. (Apajalahti et al., 1998) that was shown to provide highly effective recovery of bacterial DNA from GI tract samples. Digesta samples were placed into sterile Oak Ridge tubes containing 10 mL of sterile wash buffer (0.5 M sodium phosphate [pH 8.0], 0.1% Tween-80) and washed 4 times as follows. Samples were vortexed briefly before being shaken at high speed on a reciprocating shaker for 10 min. Next, samples were centrifuged at

30,000 x *g* for 15 min at room temperature, after which the supernatant was removed by aspiration and the samples resuspended in 10 ml of wash buffer. Following the final wash step and centrifugation, the samples were resuspended in 3 ml of Qiagen Buffer B1 (50 mM sodium EDTA, 50 mM Tris base [pH 8.0], 0.5% Tween-20, 0.5% Triton X-100; Qiagen, Valencia, CA) to which RNase A was added to a final concentration of 200 μg/ml, then stored at -70˚C to initiate the 5 freeze-thaw cycles that facilitate bacterial cell lysis. The samples were thawed and refrozen a total of 5 times by being placed in a water bath at 40˚C for 15 min, then placed back at -70 ˚C for at least 1 h before being thawed again. Following the final thaw, 50 μL of lysozyme (200 mg/ml) and 90 μL of proteinase K (20 mg/ml) were added. Samples were then incubated in a water bath at 37˚C for 45 min, after which 1 mL of Qiagen B2 buffer (3 M guanidine HCl, 20% Tween-20) was added. The samples were incubated in a water bath at 50˚C for 45 min, then centrifuged for 10 min at 5,000 x *g* at 4˚C. The supernatant was transferred to a sterile microcentrifuge tube and vortexed for 10 sec. At this point, the Qiagen Genomic Tip 20G protocol was followed precisely to elute microbiome DNA, except that 1 extra 70% ethanol wash was performed. The dried samples were resuspended in 50 ml of TE (10 mM Tris [pH 8.0], 1 mM EDTA) and then the eluted DNA was centrifuged at maximum speed using a microfuge (Eppendorf 5415C) for 15 minutes to remove any contaminating particulates. The supernatant, which contained purified DNA was quantified using a nanophotometer (Implen P 300, Implen, Inc., Westlake Village, CA).

To compare microbiome composition among samples, partial 16/18S rRNA gene sequences encompassing regions V4 & V5 were PCR amplified from the microbiome DNA (25 ng) using barcoded versions of the highly conserved primers 536f and 907r

(Holben et al., 2004), which span variable regions V4 & V5. Where samples were not of sufficient concentration to provide 25 ng for PCR, 3 µl of purified microbiome DNA was used as template. The resulting 16S-sized amplicons were gel purified using the Qiagen Gel Purification kit per manufacturer's instructions. Purified DNA was quantified, multiplexed and sequenced at the UC Davis Genome Center DNA Technologies Core (http://dnatech.genomecenter.ucdavis.edu/) The DNA was sequenced using Illumina Miseq technology (Illumina; http://www.illumina.com/, San Diego, CA, USA).

Data Analysis and Statistics.

Identification of OTUs and core diversity

Sequences were received as paired-end reads and joined using fast-q join (https://code.google.com/p/ea-utils/).ddswwwwwq11). After changing the header conformation of the sequences and compiling the sequences into one file using fna format, the sequences were processed using the Quantitative Insights Into Microbial Ecology (QIIME) pipeline (Caporaso et al., 2010). A combination of both closed and open reference was used to identify sequences to the genus level with UClust using a 97% similarity cut-off (Caporaso et al., 2010). The number of taxa (i.e. OTUs) were identified for each sample. To generate diversity plots, a 4,300 rarefaction cut-off was set to retain an appropriate number of samples to support downstream analyses. Alpha diversity plots were run using QIIME, and PD whole tree, Chao 1 and observed OTU numbers were all calculated and plotted. Unifrac beta diversity was also calculated in QIIME using the same rarefaction cut-off.

Feature Selection and LDA

A taxa summary table was built containing a row for each sample (one for each locational sampling point for each mouse) and a column for each genus identified. The entries in the matrix were the proportion of reads in the sample represented by the associated genus. Any reads that were unclassified at the genus level were removed from further consideration.

Clostridiaceae and Peptostreptococcaceae were identified as families of taxa of which *C. difficile* was possibly a member (Yutin and Galperin, 2013). All sequences in these families were identified and processed using the Ribosomal Database Project (RDP) Sequence Match (SeqMatch (Cole et al., 2009)) to the species level to determine the amount of *C. difficile* in each treatment at any timepoint . These numbers were then used downstream for feature selection, statistical and co-occurrence analyses.

Due to individual variation between animals, we found it difficult to distinguish between treatment types using typical beta diversity measures such as Unifrac ((Lozupone and Knight, 2005) supplementary Fig. S1); however, we were able to distinguish between treatments and sampling locations using a novel bioinformatics workflow developed in support of this work. This approach employs a feature selection algorithm known as floating search (in this case the important features being identified are the genera (Pudil et al., 1994)) combined with a classification step that involved the use of Linear Discriminant Analysis (LDA) applied to the selected genera to predict the location from which the sample was taken. For visualization purposes, a computational voting process identified the relevant genera across all instances of feature selection (see below).

This approach presented a challenge for the analysis. Because we used a classifier (LDA) to determine whether microbial community composition could be used to discriminate between sample locations, cross-validation was required to provide confidence in the results. Thus, each of the feature selection and locational classification steps were performed within a cross-fold validation process. Each fold involved withholding data from the sample sites within a single mouse, then testing the classifier's ability to predict the location from which the withheld samples were drawn. The feature selection should be performed for each fold in a cross-fold validation process in a two-step process like this (feature selection followed by classification), just as the training phase is performed during each fold. This meant that, potentially, a different subset of genera might have been identified during each fold of the cross-fold-validation process, leading to the problem of which genera to use for visualization. To overcome this, we employed a computational voting process (described below) to identify which genera to use during visualization in the LDA scatter plots. The more often a given genus was selected as being important for discrimination during the cross-validation process, the more likely it was ultimately used for visualizing the results.

Another challenge was identification of the proper number of features (genera) to utilize in the analyses. In machine learning, this is often accomplished by examining classifier performance across different numbers of features, and choosing the number of features that provide the best classifier performance. The observation of classifiers operating best at a specific number of features is known as the 'peaking phenomenon' (Trunk, 1979). This also helps prevent over-fitting (the classifier becoming overly sensitive to nuances in the data). The approach we chose for identifying the number of

genera, or dimensions, to use was to run the classifier on the datasets (treatments at each timepoint) at various numbers of dimensions (from 5 to 30 genera) with various levels of 'pre-pruning.' Pre-pruning involved the removal of genera if they were not present in at least 1 sample (equivalent to no pruning), and in 3%, 5%, 8%, and 16% of all samples. This provided 15 accuracies at each number-of-dimensions tested for each classifier. A different classifier was used for each LDA scatter plot. These accuracies were visualized in a boxplot format (Figs. S2 – S4).

To remove human bias from the process, we determined the appropriate number of features by choosing the number with good performance (classifier median accuracy), low variation (accuracies at that number of features tend to have low variance), and a larger number of dimensions. This was accomplished using a Pareto-front-analysis ((Hwang and Masud, 1979) Figs S5 – S7). All data points on the Pareto front are of equivalent multi-objective quality. A representative number of dimensions was chosen for visualization from the set on the Pareto front and presented in Fig. 4.

This strategy also facilitated the voting process. The genera chosen in each of the folds of the leave-one mouse-out cross-validation runs (because samples from the same mouse were considered linked) for each of the timepoint datasets and for each of the five pre-pruning levels contributed to the voting tallies. Those with the highest normalized tallies were used for visualization purposes. By 'normalized' it is meant that each vote was weighted according to its achieved accuracy. That is, a genus chosen during a fold that achieved a greater accuracy was given more weight than one that did not perform as well. A schematic diagram of the bioinformatic and computational workflow is presented

in Supplementary Fig. S8, while example code listings for generation of LDA plots and cross validation, respectively, are presented in Supplementary Fig. S7 in Chapter 3.

Distribution of Peptoclostridium across all treatments

C. difficile has been reclassified recently as *Peptoclostridium difficile* and is found under that nomenclature within the RDP, as well as the Green Genes database (DeSantis et al., 2006). We used the QIIME interface to subsample all Peptostreptococcaceae sequences and submitted them to RDP via Sequence Match (SeqMatch) for identification at the species level.

Peptoclostridium and Clostridium Group XI sequences were graphed using Excel to demonstrate the distribution of sequences across all treatments for all locations (Fig. 5 and 6). In addition, because it was possible that the Clostridium Group XI sequences contained *Peptoclostridium* sequences, we submitted 498 *Peptoclostridium* sequences to the European Bioinformatics Institute (EMBL-EBI; http://www.ebi.ac.uk/). These were aligned using MUltiple Sequence Comparison by Log-Expectation (MUSCLE (Edgar, 2004)) and then a consensus sequence was made using MView (Brown et al., 1998). The consensus sequence was used to identify sequences within the Clostridium Group XI that were *Peptoclostridium* sequences.

Core Microbiome Determination:

The core microbiome for each of the main treatments for each timepoint was determined to the genus level and used to generate pie-chart plots (Fig. 9). A genus

was considered part of the core microbiome if all samples for that treatment or location from all animals contained that genus.

Analysis of mean proportion of reads by sample location and treatment

The average proportion of reads for the 15 most abundant taxa as well as for *Peptoclostridium* and Clostridium Group XI, were calculated using Excel. These were graphed for each treatment as well by sampling site (Fig. 6, 7, 10 – 13, some data not shown).

Results:

*Clinical signs in animals and treatment effects on major phyla***:**

Only animals that were challenged with *C. difficile* showed clinical signs of illness. The animals challenged with the antibiotic cocktail alone had soft stools and moved more slowly than usual for approximately 24 h after being challenged with *C. difficile*, but this was the extent of clinical signs seen in those mice. Mice that were given vancomycin after being challenged with *C. difficile* all became sick within 24 hours after the vancomycin treatment was discontinued, as evidenced by the presence of diarrhea in all cages. All mice in the group that had been treated with vancomycin and then discontinued had soft stools but most ate and drank normally. One mouse became moribund during the vancomycin treatment and had to be euthanized. Another died overnight after discontinuing the vancomycin.

The graph below (Fig. 3) can be considered a general diagram of how different treatments affect the intestinal microbiome on a general (i.e. low-resolution) level during the course of this experiment. The three major phyla found in the intestinal microbiome

are Firmicutes, Bacteroidetes and Proteobacteria, of which Proteobacteria are normally not seen in great numbers as indicated by the negative controls (Fig. 3). As the experiment began, Proteobacteria levels were barely above the limit of detection. The administration of clindamycin stimulated a short period of outgrowth of Proteobacteria, while the relative proportions of Firmicutes and Bacteroidetes decreased. The proportion of Firmicutes dropped to below 50% of the proportion observed in the control animals. Challenging the mice with *C. difficile* with no antibiotic pre-treatment did not produce an increase in Proteobacteria. In mice challenged with *C. difficile* while being treated with antibiotics, the course of the experiment appears much the same as in those that weren't challenged – except that some mice exhibited clinical signs of CDAD. Finally, the administration of vancomycin, (first measured 4 days into a 7-day treatment regimen) produced a sustained increase in the relative proportion of Proteobacteria detected within the intestinal microbiome. As vancomycin treatment was discontinued, and overt clinical signs of disease were observed in the mice, the relative proportion of Firmicutes and Proteobacteria decreased, while that of Bacteroidetes rose.

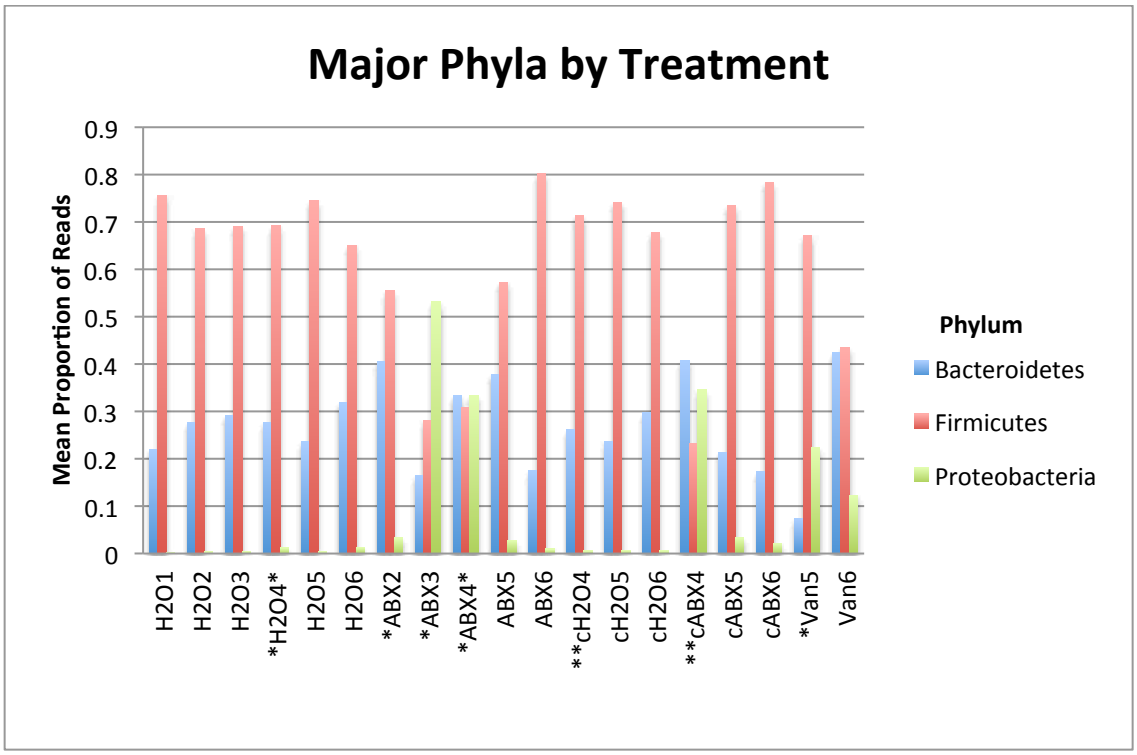


Figure 3: Mean proportional differences in major phyla as a function of treatment. *Denotes antibiotic administration **Denotes *C. difficile* challenge. H_2O4 and ABX4 occur at the time of challenge, but the mice were not exposed (as indicated by bracketing asterisks), although they continued to be handled. H_2O: Negative controls (water or saline in place of treatments); ABX: Antibiotic controls; cH_2O: Challenged controls; cABX: Challenged with antibiotic treatment prior to *C. difficile* challenge; Van5: Vancomycin given for 4 days, starting 24 h after *C. difficile* challenge; Van6: Relapse after course of Vancomycin terminated.

Alpha and Beta Diversity:

Three different indices of alpha diversity gave similar results (Chao1 is depicted in Fig. 4, see also Supplementary Table 1). There were significant difference in diversity due to antibiotic treatment at times 2, 3, and 4, and between the controls and the vancomycin treated mice at time 5. Although there was no significant difference between groups with respect to the *C. difficile* challenge, there was a suggested trend among the antibiotic treatments in that the cABX mice appeared to have a slightly higher alpha

diversity than those from the antibiotic treatment group that was not challenged with *C. difficile*.

Unifrac was applied to the dataset with a sequence cut-off (so-called, rarefaction) of 4300 sequences; however, Unifrac was unable to distinguish between all treatments. (Supplementary Fig. 1). Because of this, we used feature selection with LDA analysis to determine whether there were differences between treatments within timepoints.

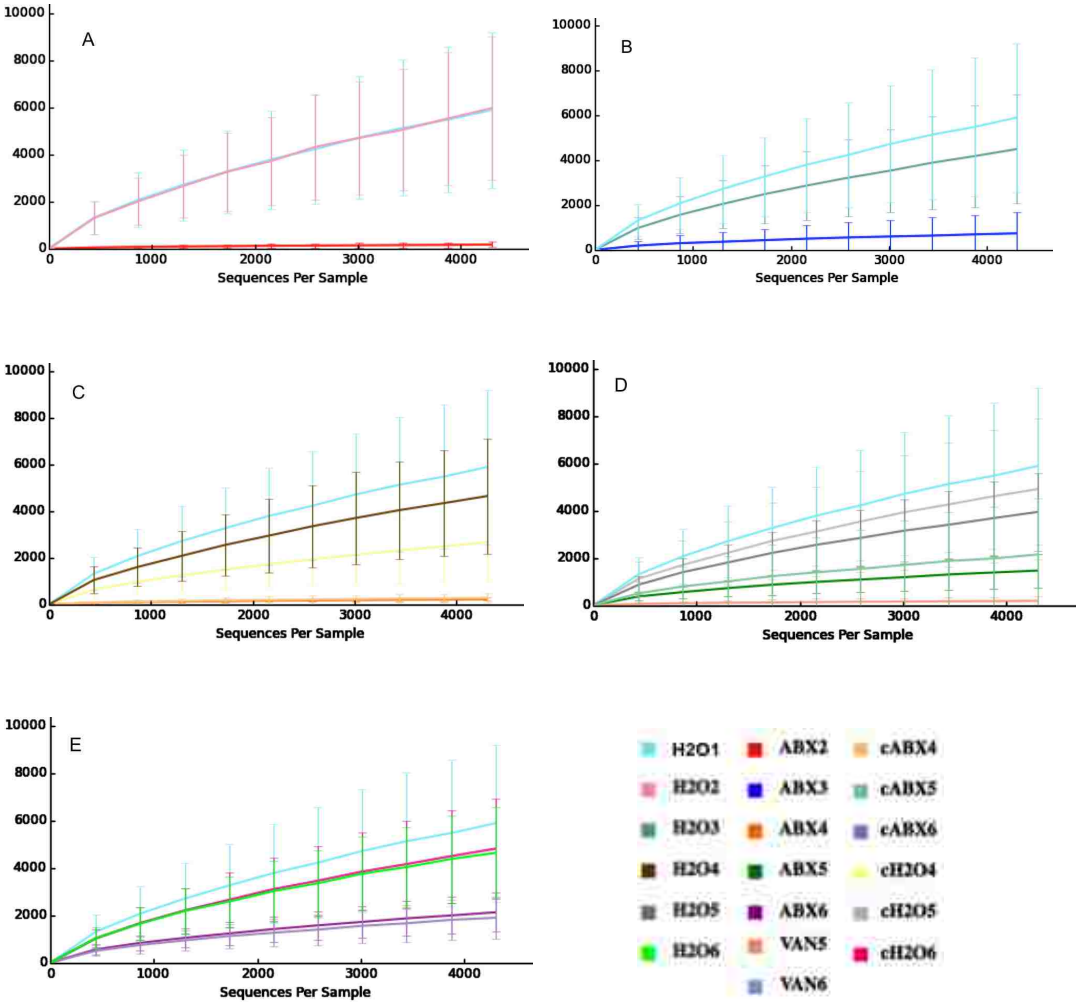


Figure 4: Chao1 alpha diversity plots, showing alpha diversity for samples from different points in time. The plot for the Time1 (H2O1) controls is overlaid on each plot to enable comparisons. A: Timepoint 2, B: Timepoint 3, C: Timepoint 4, D: Timepoint 5, E: Timepoint 6. H2O: water treatment (negative controls); ABX: antibiotic treatment; Van5: vancomycin treatment; Van6: halting of vancomycin treatment (relapse).

Feature Selection and LDA:

Feature Selection distinguished clearly between groups within timepoints and LDA plots (Fig. 5) show clear separation between treatment groups, except in the case of timepoint 4 and timepoint 6 (Fig. 5E).

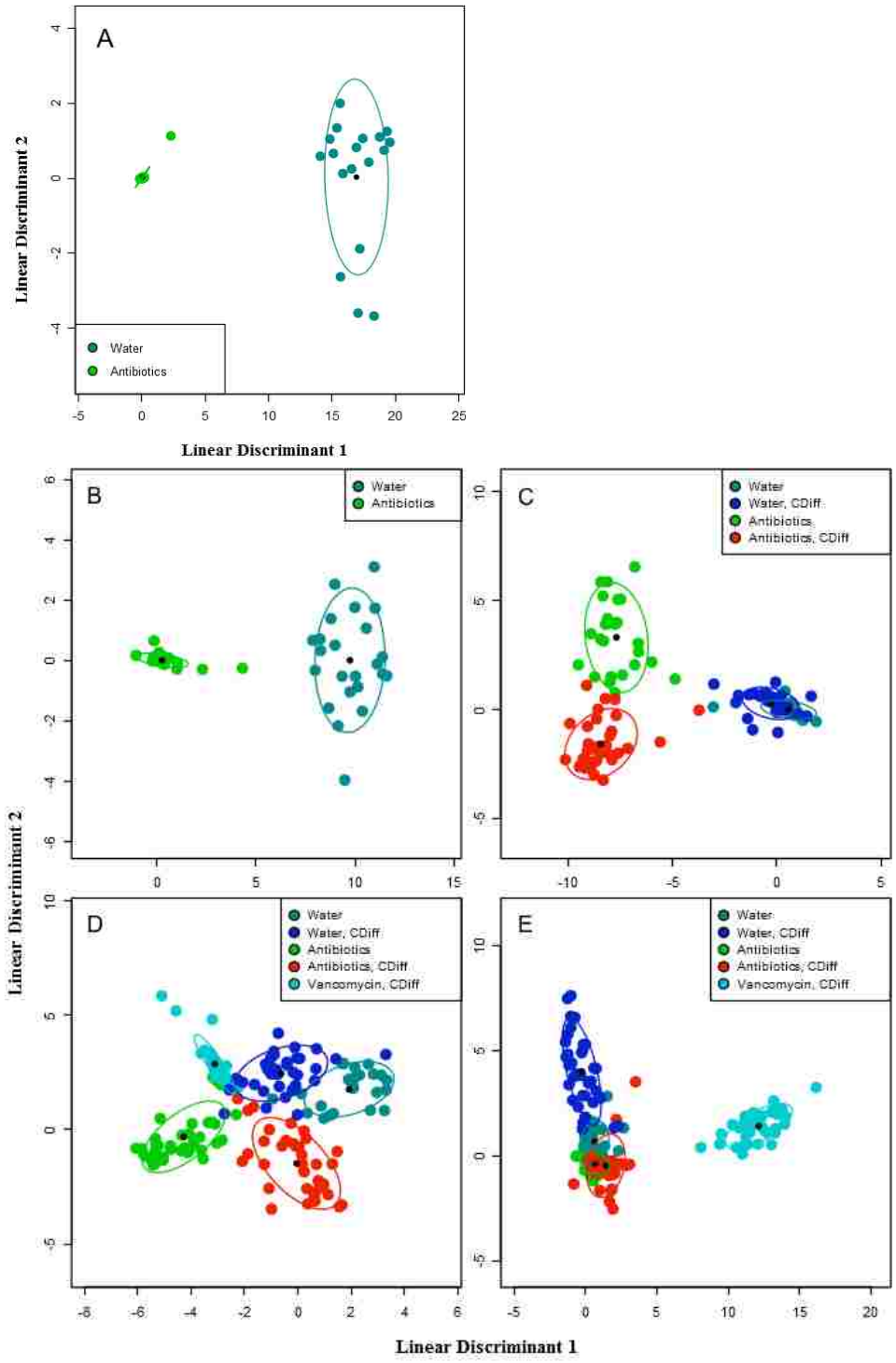


Figure 5: Feature selection combined with LDA showing that within different sampling timepoints, treatments can be distinguished from one another. The plots were made using vote-determined genera. The first accuracies listed used vote-determined genera, while accuracies in parentheses were for genera identified using 'floating search within each fold'. Black dots represent the centroid for each cluster and ellipses indicate 1 standard deviation. Cross-validation accuracies: A: Timepoint 2 accuracies: 83.33%(74.07%), 10 taxa; B. Timepoint 3 accuracies: 93.1% (74.14%), 12 taxa; C.

Timepoint 4 accuracies: 81.51%(75.63%), 17 taxa; D. Timepoint 5 accuracies: 80.24%(68.86%),18 taxa; E. Timepoint 6 accuracies: 73.56%(81.03%), 18 taxa

Distribution mapping of C. difficile *prevalence across all samples, treatments and timepoints*

C. difficile has been reclassified recently as *Peptoclostridium difficile* and is found under that nomenclature within the RDP, as well as the Green Genes database (DeSantis et al., 2006); however, due to the widespread use of the older designation, "*Clostridium difficile*" in the experimental and medical literature, the older term is used herein. Although good experimental technique was used to select a clonal colony to plate for spores, we found that we had a mixture of results, with some sequences being identified as *Peptoclostridium difficile*, but the majority were identified as being in Clostridium Group XI (the group that contains *C. difficile*). While all identified members of *Peptoclostridium difficile* were limited to the mice that had been experimentally treated with *C. difficile*, the Clostridium Group XI were not (Figs. 6 and 7). We therefore used the *Peptoclostridium difficile* sequences to create a consensus sequence. Unfortunately, the best consensus was only at 97% identity for the entire alignment, limiting our ability to use the sequence to identify *Peptoclostridium difficile* sequences that might be among those contained in the Clostridium Group XI file.

Using this approach, distribution mapping of *C. difficile* (*Peptoclostridium difficile*) shows that it was detected only in mice that were challenged with *C. difficile*. Furthermore, in this set of experiments it was detected only in the mice pretreated with antibiotics. Although a supershedder state has been reported previously in some other facilities for mice not pre-treated with antibiotics (Buffie et al., 2012; Lawley et al.,

2009), we did not find this here. The proportion of OTUs is highest in the mice that showed the most severe clinical signs – those that had received vancomycin and were allowed to relapse.

The presence and proportions of *C. difficile* in mice was found to be quite variable between compartments, treatments and timepoints (Fig. 8), but both the Ileum and the tip of the cecum were found to have the highest mean proportion of reads. Small amounts of *C. difficile* were found in the one mouse that became moribund (49.1) while on vancomycin. In that mouse, *C. difficile* was detected in the tip of the cecum, as well as the cecum with some in the mid-colon.

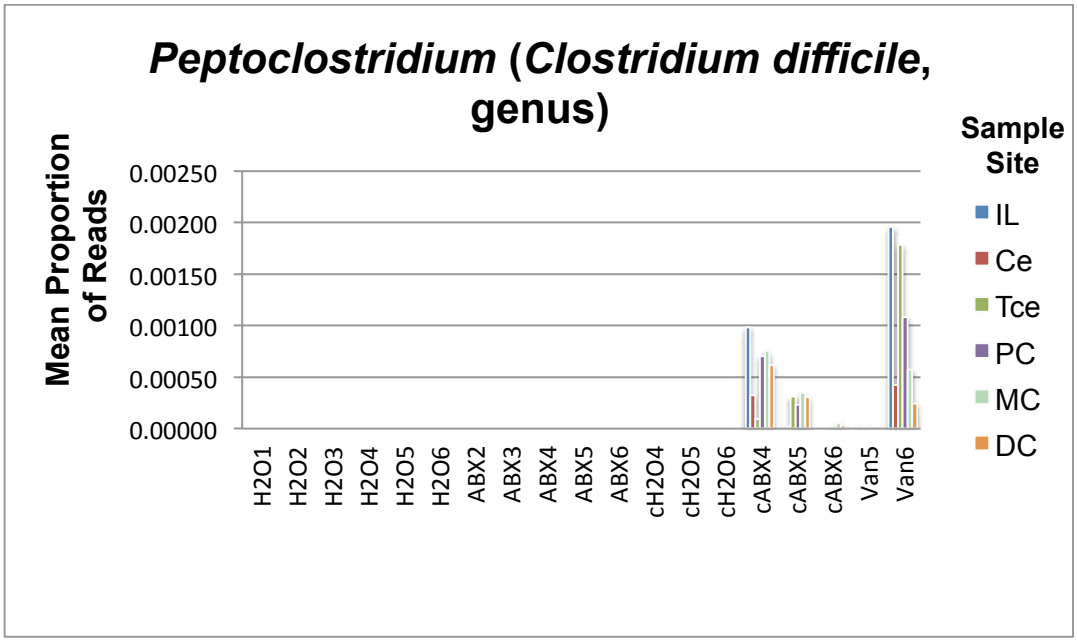


Figure 6: Distribution of *Peptoclostridium difficile* reads across all sampling times/treatments

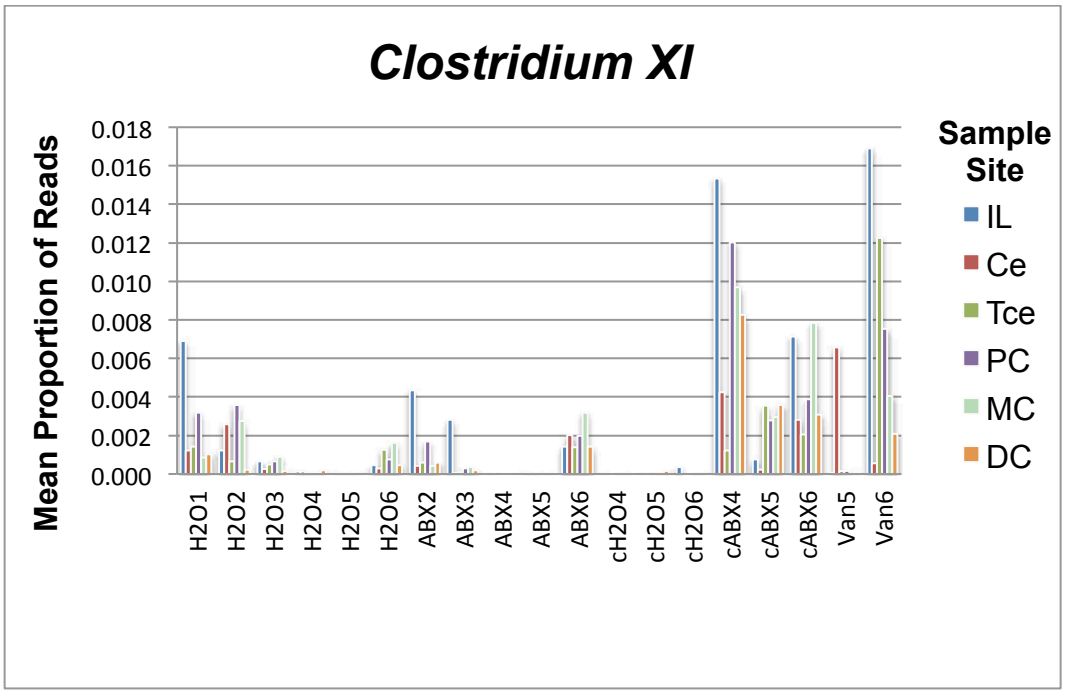


Figure 7: Distribution of *Clostridium* Group XI showing the more broad dispersal across treatments that implies that other members of Clostridium Group XI are present in addition to *C. difficile*.

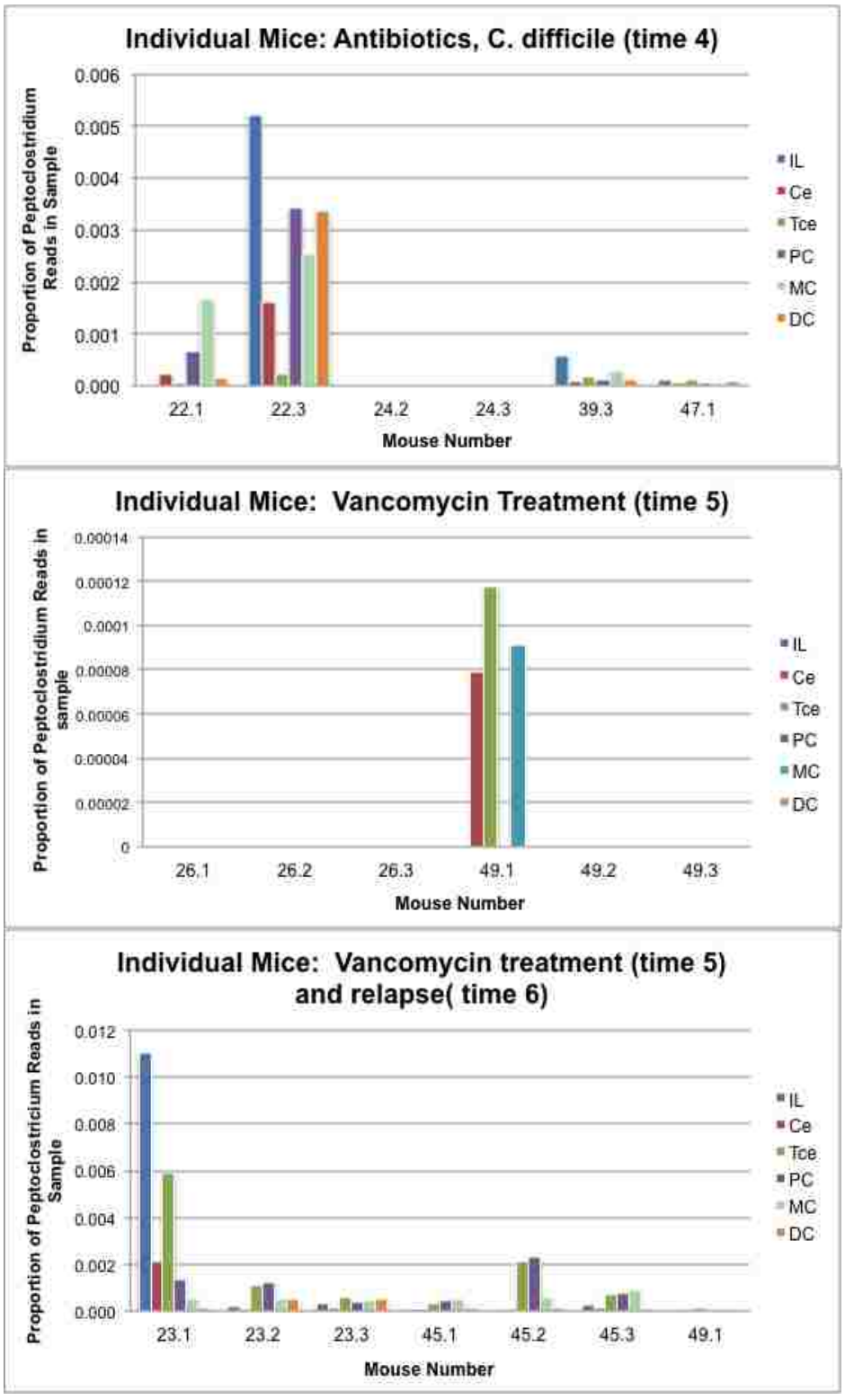


Figure 8: Proportion of *Peptoclostridium* reads for individual mice from times 4, 5, and 6 showing variation of *Peptoclostridium* in individuals. Top: Mice at timepoint 4, previously treated with antibiotics, showing variable detection of *Peptoclostridium*

between individual mice. Middle: *Peptoclostridium* was only detected in one mouse during vancomycin treatment (49.1). Bottom: Mice at timepoint 6 (during relapse). *Peptoclostridium* was detected in all mice, but variability between individuals can be seen. Mouse 49.1 (from timepoint5, during vancomycin treatment) is shown on the right to give an indication of the difference in the amount of *Peptoclostridium* found during vancomycin treatment and that found during relapse

Core microbiome analysis:

The core microbiome at the genus level was determined for all treatments by timepoint (Fig. 9). *Lactobacillus, Parabacteroides, Bacteroides* and *Escherichia* were the predominant taxa observed as a function of the different treatments administered.

The core plots provide a simple method for following key genera as they change with the experimental regime over time. Notably, antibiotic treatment causes a loss in core microbiome diversity. The antibiotic cocktail, as well as the vancomycin treatment appeared to cause a short-term increase in the proportion of *Lactobacillus* within the intestine, which rapidly diminished in favor of *Parabacteroides* along with *Escherichia*. Over the course of the experiment, there appeared to be a shift in the core microbiome of mice that were treated with antibiotics to include a greater proportion of *Bacteroides* along with *Parabacteroides* and *Dorea*. Clindamycin, while sustaining the decrease in core diversity, did not cause a similar proportional increase in *Lactobacillus*. Changes in the control samples (Fig. 9, top row) appeared to reflect the initiation of experimental manipulations (Time 2, top row) with a commensurate increase in core diversity that is not reflected in the alpha diversity plots (Fig. 1A). By timepoint 3, the core community for the controls reverted to looking much the same as at the start of the experiment (timepoint 1). One unique aspect of the control cores was that they were enriched in the proportion of *Turicibacter* relative to the other treatments.

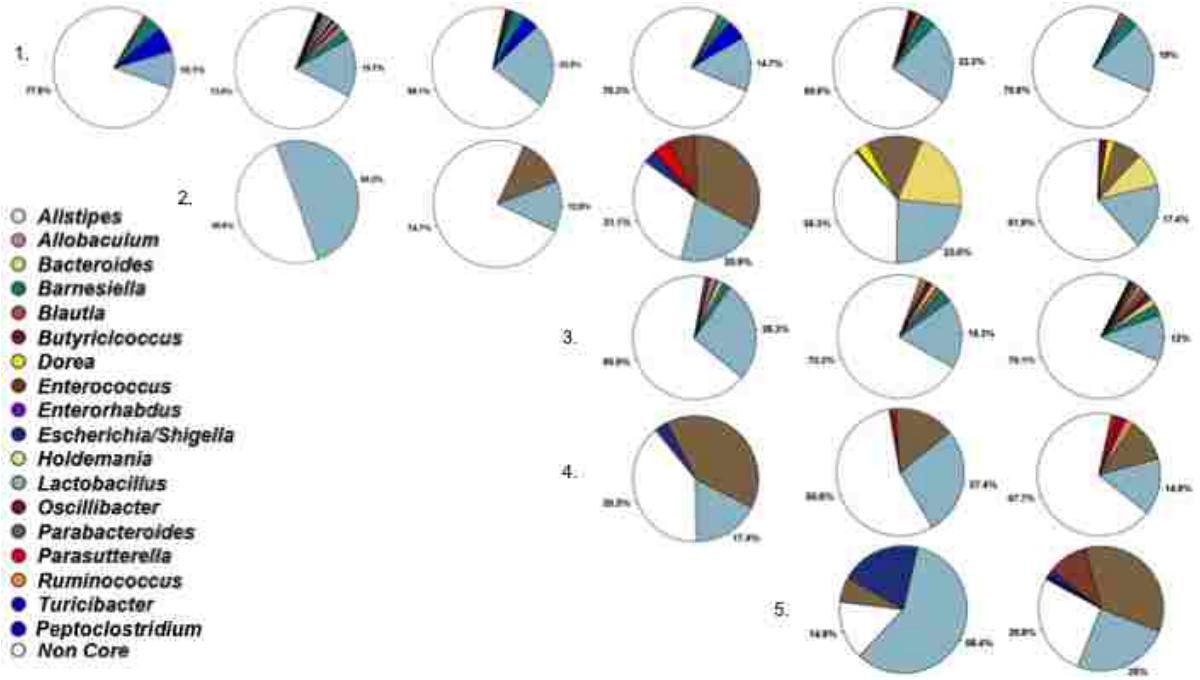


Figure 9: Core plots of treatments. Genera represented in any core plot are present in all samples for that treatment at that timepoint. From Left to Right, columns are: timepoint1 (control), timepoint 2 (antibiotic treatment), timepoint 3 (clindamycin treatment), timepoint 4 (*C. difficile* challenge), timepoint 5 (vancomycin treatment), and timepoint 6 (Relapse (vancomycin removed). From top to bottom, rows represent treatments: 1. Negative controls (water only)), 2. antibiotic treatments only, 3. *C. difficile* challenge, 4. *C. difficile* challenge combined with antibiotic treatment, 5. Vancomycin treatment and relapse (far right bottom). The dark blue/purple in the relapse treatment is *Peptoclostridium*, the genus that contains *C. difficile* alongside *Escherichia*.

Interestingly, the core plots for the *C. difficile* treated controls (no antibiotics) looked much like the core plot for the control at timepoint 2, with many of the same taxa present in both, but the *C. difficile* treated controls also exhibited an increased proportion of *Parabacteroides*. We had also expected to detect *C. difficile*, but saw no evidence of its presence in the *C. difficile* control core plots or samples. The timepoint 2 control also contained a small proportion of *Blautia*, which is a member of the *Clostridium* XIV

Group, which contains many taxa that are considered beneficial to the host, while *Blautia* didn't appear in the core microbiome of the *C. difficile* controls.

The core communities from mice challenged with *C. difficile* after having been given antibiotics were less diverse than those that were exposed to antibiotics. There was a modestly increased proportion of *Escherichia* along with *Parabacteroides*, which progressed to a core community dominated by *Parabacteoides* and *Parasutterella*.

Treatment with vancomycin following challenge with *C. difficile* exhibited predominance by *Lactobacillus* and *Escherichia,* both of which subsequently diminished in favor of an increased proportion of *Enterococcus*.

Biogeographical Analysis

Distribution graphs (by treatment and location) were made for several taxa including *C. difficile* (Figures 5 and 6). Only some of these suggested differences in distributional patterns associated with *C. difficile*, and these are presented below in Figs. 10-14. Of all of these taxa, *Lactobacillus* was found in the greatest proportional abundance, which is of interest given that bacteria in this genus supposedly afford protection against *C. difficile* infection (Lawley et al., 2012; Schubert et al., 2015). The distribution graph for *Lactobacillus* (Fig. 10) shows that normally (i.e.in mice not challenged with antibiotics) *Lactobacillus* is primarily found in the ileum, with the next most preferred site being the mid-colon. When mice were administered antibiotics, there was a general increase in the proportion of *Lactobacillus* across the rest of the compartments of the lower gastrointestinal tract. This is most apparent at Timepoint 2 following treatment with the antibiotic cocktail (ABX2), as well as at timepoint 5 during

vancomycin treatment (Van5). At these timepoints, the distribution pattern for *Lactobacillus* proportional abundance increased, with it being detected at higher levels in all locations of the intestine. Interestingly, during relapse (timepoint 6), the distribution of *Lactobacillus* proportional abundance reverted to being similar to that found normally in control mice (Fig. 10).

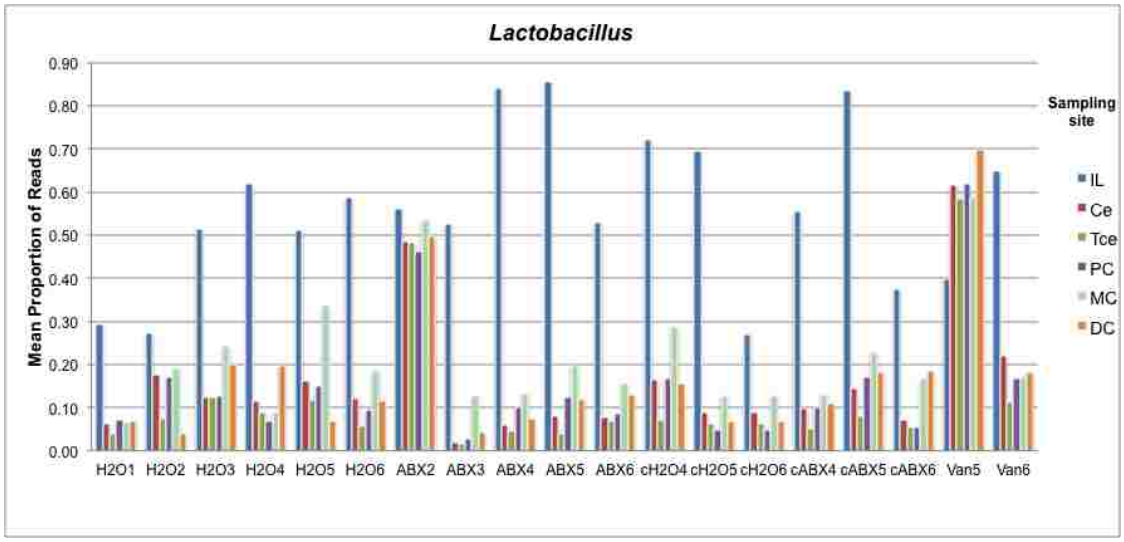


Figure 10: Distribution of *Lactobacillus* by treatment and location

Bacteroidetes are also an important group of bacteria in the gut, but haven't been shown to have an influence on the outcome of infection by *C. difficile*. The genus *Parabacteroides* was found in high proportion in many of the samples in this experiment, particularly in mice that were treated with antibiotics (Fig. 11). Where proportionally abundant, *Parabacteroides* had a fairly even biogeographical distribution across all of the sampling sites. Following cessation of antibiotic treatment, the distribution patterns became less even, with *Parabacteroides* being more prevalent in the mid colon and distal colon. Antibiotic treatment appeared to be the primary reason for increased levels of *Parabacteroides*, given that *C. difficile* challenged mice exhibited a similar profile to

those that were not challenged. Interestingly, when compared to the negative controls, *Parabacteroides* did reflect the effects of *C. difficile* challenge in the absence of antibiotic treatment. Mice that were challenged without antibiotics showed an increase in the proportion of *Parabacteroides* compared to the negative controls (Fig. 11). In this case, instead of the fairly even profile seen for other treatments involving antibiotics, the profiles were uneven, making it difficult to determine where *Parabacteroides* were found in the greatest proportional numbers in those mice.

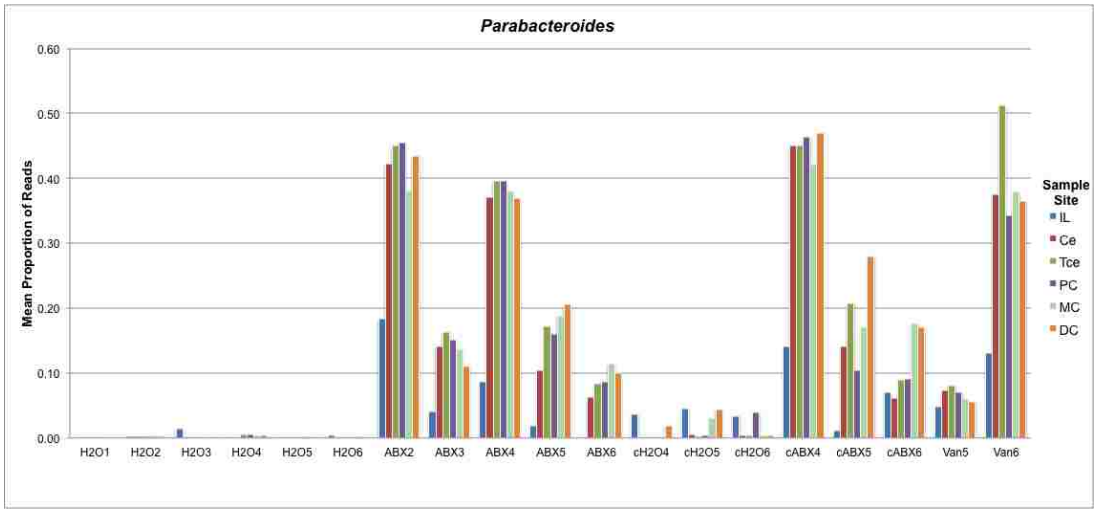


Figure 11: Distribution of *Parabacteroides* by treatment and location

The genus *Bacteroides* also showed differences in distribution patterns due to *C. difficile* challenge alone; i.e. with no antibiotics (Fig 12). Antibiotic controls that weren't challenged were in decreased proportion in *Bacteroides* in the first 24 h after being given clindamycin. Twenty-four hours later they had rebounded and high proportions of *Bacteroides* were found throughout the lower intestine except for the distal ileum, with the highest proportions found in the mid colon and distal colon. Mice challenged with *C. difficile* showed a decrease in the proportion of *Bacteroides* associated with clindamycin treatment, but the levels never reached those of the mice that weren't challenged. Despite

this, the highest proportions of *Bacteroides* in the *C. difficile* challenged mice were in the colon (Fig. 12).

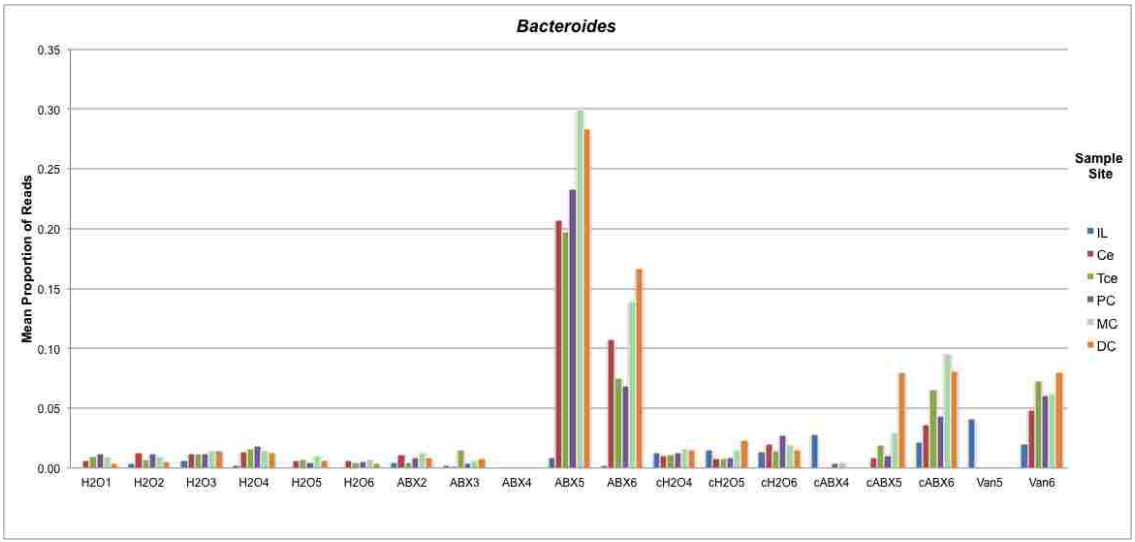


Figure 12: Distribution of *Bacteroides* by treatment and location

The genus *Alistipes* (Fig. 13) is a Bacteroidete that may protect against *C. difficile* infection (Schubert et al., 2015). The observed proportional distribution of this genus is paradoxical to that notion in that challenge with *C. difficile* is associated with changes in where the highest proportions of *Alistipes* were found. Prior to *C. difficile* challenge and in unchallenged controls, the highest proportions of *Alistipes* were found in the proximal colon. By contrast, following challenge with *C. difficile*, the highest proportions were observed in the mid-colon. In antibiotic treated mice, *Alistipes* proportions dropped below the level of detection and then rebounded. The highest proportions of *Alistipes* were found at timepoint 3 and were the most proportionally abundant in the mid colon and distal colon. Upon treatment with clindamycin, the proportion of *Alistipes* dropped below the level of detection. By contrast, in *C. difficile* challenged mice, *Alistipes* remained at fairly normal levels (compared to negative controls), although the highest

proportions within the colon were found in the mid-colon in these mice (representing a change in predominant location) (Fig. 13).

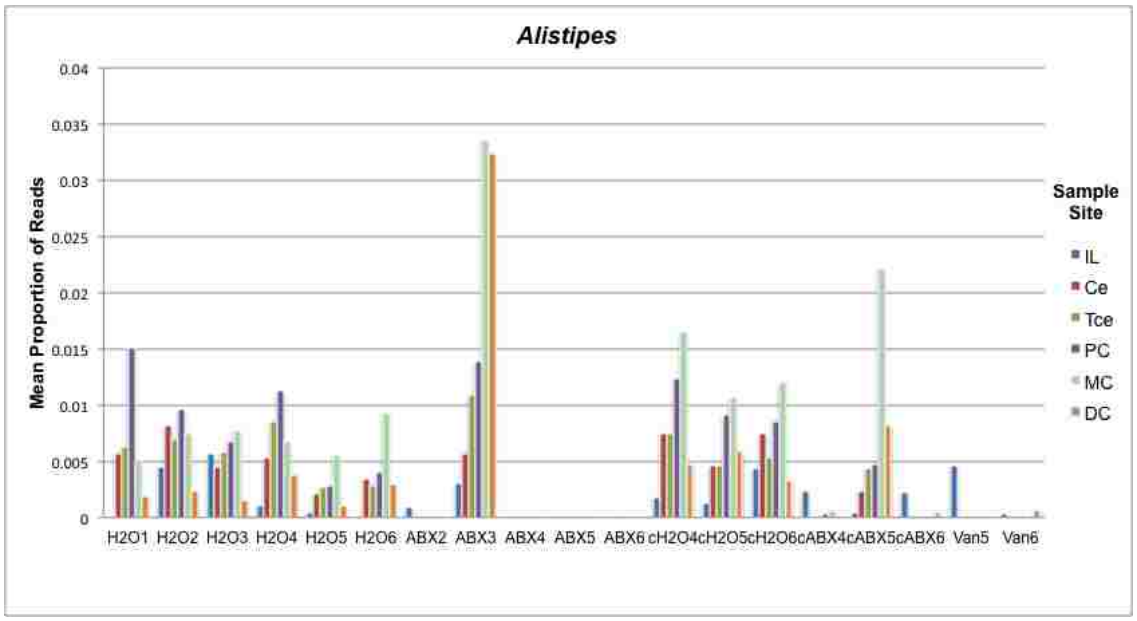


Figure 13: Distribution of *Alistipes* by treatment and location

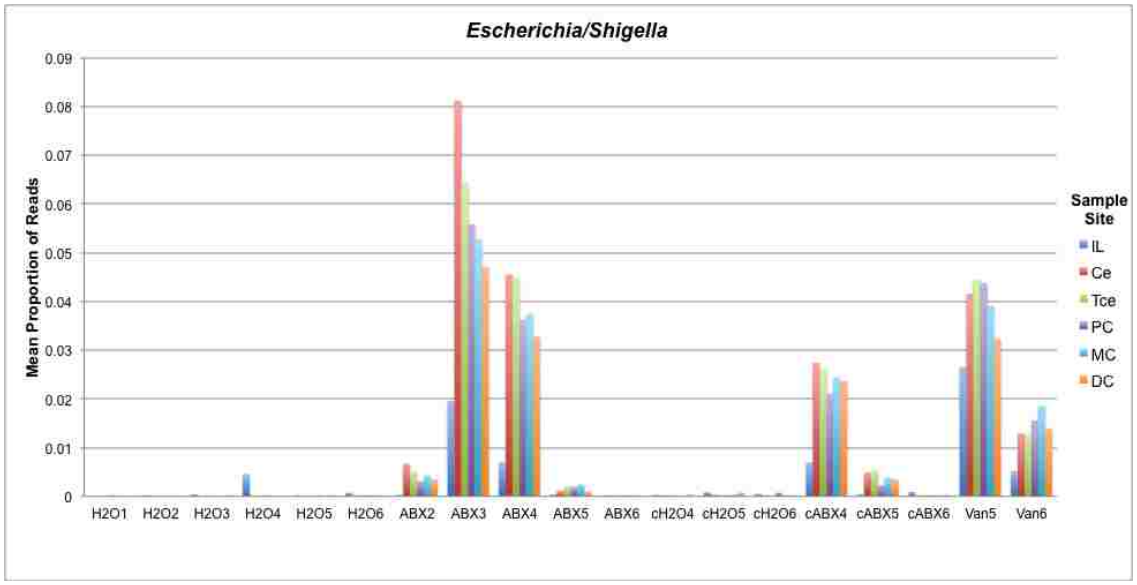


Figure 14: Distribution of *Escherichia* by treatment and location

The genus *Escherichia* was normally present in low abundance within the intestine (Fig. 14, H2O1 – H2O6). Increased numbers of *Escherichia* were detected after

antibiotic administration. While *Escherichia* has been positively associated with overt CDAD (Lawley et al. 2012; Schubert et al., 2015), in this experiment it was barely above the limit of detection in mice that were challenged with *C. difficile* but not pretreated with antibiotics.

Discussion:

Following the development of mouse models for CDAD that correspond closely with the course of the disease in humans, there have been many excellent studies examining *C. difficile* colonization and its potential interactions with the intestinal microbiome (Koenigsknecht et al., 2015; Lawley et al., 2012; Schubert et al., 2015; Sun et al., 2011). In the current study, for the first time, the interactions of *C. difficile* with the intestinal microbiome are explored at the genus level along the landscape of the lower intestinal tract. In addition to examining the effect of *C. difficile* on the intestinal microbiome during overt disease, the effects of *C. difficile* in the absence of perturbation (antibiotics) were examined, as well as during antibiotic treatment and relapse. This approach was both promising and rewarding in that it examined how *C. difficile* might behave under a variety of conditions, providing considerable insight into its etiology.

CDAD has been associated with reduced alpha-diversity in the intestinal microbiome (Buffie et al., 2012; Young and Schmidt, 2004), which is most likely due to its relationship to treatment with antibiotics in its host. Here, we show that while the antibiotic regimen significantly decreased the alpha diversity of the intestinal microbiome, the introduction of *C. difficile* alone did not significantly lower alpha-diversity (Fig. 4, Supplementary Table 1). This observation lends substantive credibility

to the notion that the course of antibiotic treatment itself altered the intestinal microbiome, resulting in a decrease in diversity in those prior studies. Our findings are also concordant with prior work suggesting that *C. difficile* takes advantage of the disturbance of the microbiome brought on by antibiotic treatment in order to invade the intestine. In this etiology, timing may play an important role, as *C. difficile* and other pathogens appear to be dependent on the series of successional changes that take place as the intestinal microbiome recovers from perturbation (Fig. 8, see also David et al., 2015; Schubert et al., 2015)

Another goal of this study was to determine whether *C. difficile* had an effect on the intestinal microbiome without prior disturbance due to antibiotic treatments. To do this, negative control samples were compared to *C. difficile* control samples. There have been prior reports in the literature suggesting that a "super shedder state" is created when mice are exposed to *C. difficile* without antibiotic treatment (Buffie et al., 2012; Lawley et al., 2009), but this did not occur in this study, as spores were never recovered by plating from these samples (data not shown). However, reports of the super shedder state, as well as other disease states with mouse models appear to be dependent on differences between animal facilities (Buffie et al., 2012; Lawley and Young, 2013).

Unifrac was not able to discriminate between the negative control and the *C. difficile* control at time 4 (Supp. Fig. 1). By contrast, while the LDA plot associated with feature selection appeared to have failed to separate the two treatments (Fig. 5C), the cross validation values obtained (81.51% and 75.63% for cross validation inside and outside of feature selection, respectively) indicate that this approach was, in fact, able to discriminate between all of the time 4 treatments including the negative control and the

C. difficile control. Feature selection and LDA done at later time points confirmed that feature selection successfully discriminated between the two control treatments, indicating that were indeed different from one another.

Interestingly, in this study, *C. difficile* itself was never found in the *C. difficile* controls (Fig. 6), although it may have been present as a member of the Clostridium XI Group. We attempted to try and create a 100% identity consensus sequence using the *C. difficile* sequences in order to query Clostridium Group XI sequences for potential *C. difficile* candidates to obtain a better distribution map, but this was unsuccessful when we barely achieved a 90% identity consensus. The poor success rate developing a consensus was likely due to the fact that *C. difficile* genome contains between 10 and 12 rRNA operons, most of which appear to be functional (van Eijk et al., 2015). Sequence differences between these operons, including their respective 16S rRNA genes were an obvious explanation for our lack of success, as well as probably being the reason why identification of recovered *C. difficile* sequences was divided between *Peptoclostridium* and *Clostridium XI* at the genus level.

Core plot analysis not only helped in selecting genera of interest for distribution maps from among 300+ genera, but also gave pertinent information at the genus level on changes in 'major players' as different experimental treatments were applied (Fig. 8). For instance, the *C. difficile* control core communities appear to be enriched in *Parabacteroides* as compared to the negative control core communities. *Parabacteroides* are able to digest resistant starches (making them likely members of the mouse intestinal microbiome (Flint et al., 2012), but the role they play in *C. difficile* infection when antibiotics are not present is not clear and may reward future investigation(s).

Additionally, *Parabacteroides* and *Bacteroides* have been found to harbor resistance genes to a variety of antibiotics (Nakano et al., 2011), which perhaps explains their ability to survive in the face of multiple rounds of antibiotic treatments. While *Bacteroides* did not fare well with respect to the vancomycin treatments, *Parabacteroides* was one of the few community members to thrive during this treatment, along with *Escherichia* and *Lactobacillus*.

The core plots also present an interesting picture of what appear to be "waves" of succession in microbiome composition, as a variety of taxa increase in proportional abundance (i.e. "bloom") and then decrease in abundance, only to be followed by "blooms" of other taxa. For example, in following the antibiotic treatments over time (Fig. 8, row 2), a "bloom" of *Lactobacillus* occured within 24 hours after the administration of antibiotics. This swiftly declined again and was followed by a "bloom" of *Parabacteroides* at time 4 (which can be seen within the core microbiome at time three). A third "bloom" followed, with *Lactobacillus* and *Bacteroides* in approximately equal proportions, which declined by time 6.

Both *Escherichia* and *Lactobacillus* have been associated with CDAD, but *Lactobacillus* has been suggested to be protective against *C. difficile* infection, while *Escherichia* has been shown to facilitate infection (Lawley et al., 2012; Schubert et al., 2015). In the current study, high levels of *Escherichia* were detected in the intestine following antibiotic treatment and just before challenge with *C. difficile* (Fig. 8 and 14). This appears to affirm their role as facilitating *C. difficile* infection. *Lactobacillus*, on the other hand, also increased proportionally following antibiotic treatment and was much more prevalent than *Escherichia* (Figs. 8 and 9). However, *Lactobacillus* only increased

in prevalence following the first treatment with the antibiotic mixture, not following clindamycin treatment (Time 3). This means that it was present in normal amounts prior to challenge with *C. difficile*. The role that *Lactobacillus* might play in protection against CDAD is also called into doubt by its rather puzzling presence during vancomycin treatment and prior to relapse (Fig. 9). One interpretation is that *Lactobacillus* only provides protection prior to colonization by *C. difficile* and that once *C. difficile* is established, *Lactobacillus* cannot not prevent its overgrowth (i.e. range expansion).

There are a couple possible scenarios for how this might work. A recent study has postulated that the microbiome works as a consortium to provide colonization resistance to *C. difficile* (Schubert et al., 2015). In this scenario, vital members of the protective consortium are depleted by continued antibiotic use and even though one or more members may be present in increased proportions, they can't protect against *C. difficile* invasion, leading to recurrent disease as more antibiotics are used to control CDAD in the host. In another scenario, different members of the genus *Lactobacillus* form a mixed species group in which some taxa are protective and others are not (and may even facilitate invasion). In this case, antibiotics select for specific species of *Lactobacillus* that are protective leaving the host with other species of Lactobacillus that don't provide protection against *C. difficile* invasion and/or relapse. This is supported by other work done in our lab on how *Lactobacillus* species change during invasion of probiotic species in yogurt (supplementary Fig. S9). The second scenario is also supported in part by particular changes illustrated in the distribution map that show the biogeographical distribution of *Lactobacillus* throughout the experiment (Fig. 9). Those data show that *Lactobacillus* is located in high proportions within the ileum prior to

antibiotic use. At timepoint 2 *Lactobacillus* had rapidly expanded it range to be found in large quantities throughout the lower intestine. By time 3, the *Lactobacillus* population has decreased to less than normal proportions, with the highest proportion again in the ileum. *Lactobacillus* does not appear to be much affected by clindamycin, (administered prior to timepoint 3), but responded to vancomycin treatment by again expanding its range to include the entire lower intestinal tract.

Another objective of this research was to investigate whether *C. difficile* was sequestered somewhere in the lower intestinal tract during treatment with vancomycin, leading to recurrent disease from this reservoir as opposed to simply reinfecting the host *de novo* following the end of the antibiotic treatment. We found some evidence supporting both alternatives in this study. Of the 6 mice sampled during vancomycin treatment, 5 appeared to have cleared *C. difficile*, or if present, it was below the limit of detection (Fig. S8, middle panel). However, one mouse did have *C. difficile* present in detectable abundance and actually became morbid during the vancomycin treatment. The *C. difficile* was, however, present in very low abundance, leading to a question of whether it was cause of the mouse being ill. Considering the biogeographical distribution maps for *C. difficile*, it can be seen that while *C. difficile* is found in a variety of locations within the lower intestine, it is primarily found in the ileum in infected mice that are not being concurrently treated with antibiotics (Fig. 3, points ABX4 and Van6). In the one mouse that was infected during vancomycin treatment, the *C. difficile* was located primarily in the cecum and tip of the cecum (the latter equivalent to the appendix in the mouse), although a smaller amount was detected in the mid-colon as well. While this is certainly not overwhelming support for *C. difficile* sequestration while under vancomycin

treatment, it is interesting to note that 1 mouse out of 6 is consistent with the recurrence rate of 20% given in the literature (Ananthakrishnan, 2011). Thus, the mouse results mirror human disease with 5 mice appearing to clear the disease and then getting reinfected, while 1 mouse contained sequestered *C. difficile,* which expanded its distribution once the antibiotic was removed.

This is probable, considering that the intestinal microbiome is an ecosystem with all of the complexity and 'checks and balances' thereof. A major disturbance of the system can lead to opportunities both for pathogens and possibly probiotic bacteria, thus facilitating disease or preventing it. One challenge with this hypothesis is that every individual appears to have a different microbiome, creating intersubject variation that may lead to high levels of variation during experimental work or interfere with patient care during clinical work. For instance in this study, *Alistipes* was found in mice that had *C. difficile* (Fig. 13). Conversely, in another study it was not present in any animals that had CDAD and therefore was considered to be protective against infection with *C. difficile* (Schubert et al., 2015). Thus, the interpretation of any study as it applies to specific taxa and their roles in disease etiology should be received with caution, just as in any other ecosystem study specific to that system. Nonetheless, the current study and others can be read with cautious optimism as different methods of examining the effects of invasion on the intestinal microbiome substantiate information about the system as a whole.

Here, we have shown the value of using feature selection to discriminate between taxa and combined that with analysis of the proportional abundance and distribution of specific taxa to differentiate between locations within the intestine and provide insight

into how *C. difficile* infection progresses, as well as gaining insight into how *C. difficile* might survive antibiotic treatment within the host in order to cause recurrent disease. In addition, we have provided evidence that *C. difficile* is associated with changes to the intestinal microbiome even in the absence of antibiotic treatments (Figs. 5 and 6). By examining individual taxa, such as *Lactobacillus*, *Escherichia* and *Parabacteroides*, mechanisms for how their abundance and distribution change during antibiotic treatment and *C. difficile* infection, as well as possible mechanisms for changes in the intestinal microbiome during the course of CDAD have been suggested. This is important in searching for taxa that interact directly with *C. difficile* to facilitate or protect against disease. It is only by examining how members of the intestinal microbiome interact with medical treatment and nutrition as well as invaders, pathogenic or otherwise, that we can begin to develop new treatment regimes and therapeutics based on the microbiome itself.

Literature Cited:

Ackermann, M., Stecher, B., Freed, N.E., Songhet, P., Hardt, W.D., and Doebeli, M. (2008). Self-destructive cooperation mediated by phenotypic noise. Nature *454*, 987-990.

Ananthakrishnan, A.N. (2011). Clostridium difficile infection: epidemiology, risk factors and management. Nat. Rev. Gastroenterol. Hepatol. *8*, 17-26.

Apajalahti, J.H., Sarkilahti, L.K., Maki, B.R., Heikkinen, J.P., Nurminen, P.H., and Holben, W.E. (1998). Effective recovery of bacterial DNA and percent-guanine-plus-cytosine-based analysis of community structure in the gastrointestinal tract of broiler chickens. Appl. Environ.Microbiol. *64*, 4084-4088.

Bartlett, J.G. (1979). Antibiotic-Associated Pseudomembranous Colitis. Reviews of Infectious Diseases *1*, 530-539.

Bollinger, R.R., Barbas, A.S., Bush, E.L., Lin, S.S., and Parker, W. (2007). Biofilms in the large bowel suggest an apparent function of the human vermiform appendix. J Theor Biol *249*, 826-831.

Brown, N.P., Leroy, C., and Sander, C. (1998). MView: a web-compatible database search or multiple alignment viewer. Bioinformatics *14*, 380-381.

Buffie, C.G., Jarchum, I., Equinda, M., Lipuma, L., Gobourne, A., Viale, A., Ubeda, C., Xavier, J., and Pamer, E.G. (2012). Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to Clostridium difficile-induced colitis. Infect. Immun. *80*, 62-73.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I*., et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Meth. *7*, 335-336.

Chen, X., Katchar, K., Goldsmith, J.D., Nanthakumar, N., Cheknis, A., Gerding, D.N., and Kelly, C.P. (2008). A mouse model of Clostridium difficile-associated disease. Gastroenterology *135*, 1984-1992.

Clements, A.C., Magalhaes, R.J., Tatem, A.J., Paterson, D.L., and Riley, T.V. (2010). Clostridium difficile PCR ribotype 027: assessing the risks of further worldwide spread. Lancet Infect. Dis. *10*, 395-404.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M*., et al.* (2009). The

Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. *37*, D141-145.

Cuddington, K., and Hastings, A. (2004). Invasive engineers. Ecol. Model. *178*, 335-347.

David, L.A., Weil, A., Ryan, E.T., Calderwood, S.B., Harris, J.B., Chowdhury, F., Begum, Y., Qadri, F., LaRocque, R.C., and Turnbaugh, P.J. (2015). Gut microbial succession follows acute secretory diarrhea in humans. mBio *6*, e00381-00315.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ.Microbiol. *72*, 5069-5072.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

Flint, H.J., Scott, K.P., Louis, P., and Duncan, S.H. (2012). The role of the gut microbiota in nutrition and health. Nat. Rev. Gastroenterol. Hepatol. *9*, 577-589.

Holben, W.E., Feris, K.P., Kettunen, A., and Apajalahti, J.H. (2004). GC fractionation enhances microbial community diversity assessment and detection of minority populations of bacteria by denaturing gradient gel electrophoresis. Appl. Environ.Microbiol. *70*, 2263-2270.

Hwang, C.L., and Masud, A.S.M. (1979). Multiple objective decision making, methods and applications: a state-of-the-art survey (Springer-Verlag).

Karatayev, A.Y., Boltovskoy, D., Padilla, D.K., and Burlakova, L.E. (2007). The invasive bivalves dreissena polymorpha and limnoperna fortunei: parallels, contrasts, potential spread and invasion impacts. J. Shellfish Res. *26*, 205-213.

Knodler, L.A., Vallance, B.A., Celli, J., Winfree, S., Hansen, B., Montero, M., and Steele-Mortimer, O. (2010). Dissemination of invasive Salmonella via bacterial-induced extrusion of mucosal epithelia. Proc Natl Acad Sci U S A *107*, 17733-17738.

Koenigsknecht, M.J., Theriot, C.M., Bergin, I.L., Schumacher, C.A., Schloss, P.D., and Young, V.B. (2015). Dynamics and establishment of Clostridium difficile infection in the murine gastrointestinal tract. Infect. Immun. *83*, 934-941.

Laurin, M., Everett, M.L., and Parker, W. (2011). The cecal appendix: one more immune component with a function disturbed by post-industrial culture. Anat. Rec. (Hoboken) *294*, 567-579.

Lawley, T.D., Clare, S., Walker, A.W., Goulding, D., Stabler, R.A., Croucher, N., Mastroeni, P., Scott, P., Raisen, C., Mottram, L.*, et al.* (2009). Antibiotic treatment of

clostridium difficile carrier mice triggers a supershedder state, spore-mediated transmission, and severe disease in immunocompromised hosts. Infect. Immun. *77*, 3661-3669.

Lawley, T.D., Clare, S., Walker, A.W., Stares, M.D., Connor, T.R., Raisen, C., Goulding, D., Rad, R., Schreiber, F., Brandt, C.*, et al.* (2012). Targeted restoration of the intestinal microbiota with a simple, defined bacteriotherapy resolves relapsing Clostridium difficile disease in mice. PLoS pathog. *8*, e1002995.

Lawley, T.D., and Young, V.B. (2013). Murine models to study Clostridium difficile infection and transmission. Anaerobe *24*, 94-97.

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ.Microbiol. *71*, 8228-8235.

Merrigan, M., Venugopal, A., Mallozzi, M., Roxas, B., Viswanathan, V.K., Johnson, S., Gerding, D.N., and Vedantam, G. (2010). Human hypervirulent Clostridium difficile strains exhibit increased sporulation as well as robust toxin production. Journal of bacteriology *192*, 4904-4911.

Nakano, V., Nascimento e Silva, A., Merino, V.R., Wexler, H.M., and Avila-Campos, M.J. (2011). Antimicrobial resistance and prevalence of resistance genes in intestinal Bacteroidales strains. Clinics *66*, 543-547.

Nolte, A.W. (2011). Dispersal in the course of an invasion. Mol. Ecol. *20*, 1803-1804.

Pudil, P., Novovicova, J., and Kittler, J. (1994). Floating Search Methods in Feature-Selection. Pattern Recogn. Lett. *15*, 1119-1125.

Rand Corporation (2001). A Million Random Digits with 100,000 Normal Deviates (Glencoe, IL USA: Rand Corporation), pp. B1 - 8.

Reeves, A.E., Theriot, C.M., Bergin, I.L., Huffnagle, G.B., Schloss, P.D., and Young, V.B. (2011). The interplay between microbiome dynamics and pathogen dynamics in a murine model of Clostridium difficile Infection. Gut microbes *2*, 145-158.

Schubert, A.M., Sinani, H., and Schloss, P.D. (2015). Antibiotic-Induced Alterations of the Murine Gut Microbiota and Subsequent Effects on Colonization Resistance against Clostridium difficile. mBio *6*.

Shin, B.M., Kuak, E.Y., Yoo, H.M., Kim, E.C., Lee, K., Kang, J.O., Whang, D.H., and Shin, J.H. (2008). Multicentre study of the prevalence of toxigenic Clostridium difficile in Korea: results of a retrospective study 2000-2005. J. Med. Microbiol. *57*, 697-701.

Sun, X., Wang, H., Zhang, Y., Chen, K., Davis, B., and Feng, H. (2011). Mouse relapse model of Clostridium difficile infection. Infect. Immun. *79*, 2856-2864.

Tae, C.H., Jung, S.A., Song, H.J., Kim, S.E., Choi, H.J., Lee, M., Hwang, Y., Kim, H., and Lee, K. (2009). The first case of antibiotic-associated colitis by Clostridium difficile PCR ribotype 027 in Korea. J. Korean Med. Sci. *24*, 520-524.

Trunk, G.V. (1979). A problem of dimensionality: A simple example. IEEE Trans. Pattern Anal. Mach. Intell. *1*, 306 - 307.

van Eijk, E., Anvar, S.Y., Browne, H.P., Leung, W.Y., Frank, J., Schmitz, A.M., Roberts, A.P., and Smits, W.K. (2015). Complete genome sequence of the Clostridium difficile laboratory strain 630Deltaerm reveals differences from strain 630, including translocation of the mobile element CTn5. BMC genomics *16*, 31.

Viswanathan, V.K., Mallozzi, M.J., and Vedantam, G. (2011). Clostridium difficile infection: An overview of the disease and its pathogenesis, epidemiology and interventions. Gut Microbes *1*, 234-242.

Weiss, B., Kleinkauf, N., Eckmanns, T., an der Heiden, M., Neumann, M., Michels, H., and Jansen, A. (2009). Risk factors related to a hospital-associated cluster of Clostridium difficile PCR ribotype 027 infections in Germany During 2007. Infect. Control. Hosp. Epidemiol. *30*, 282-284.

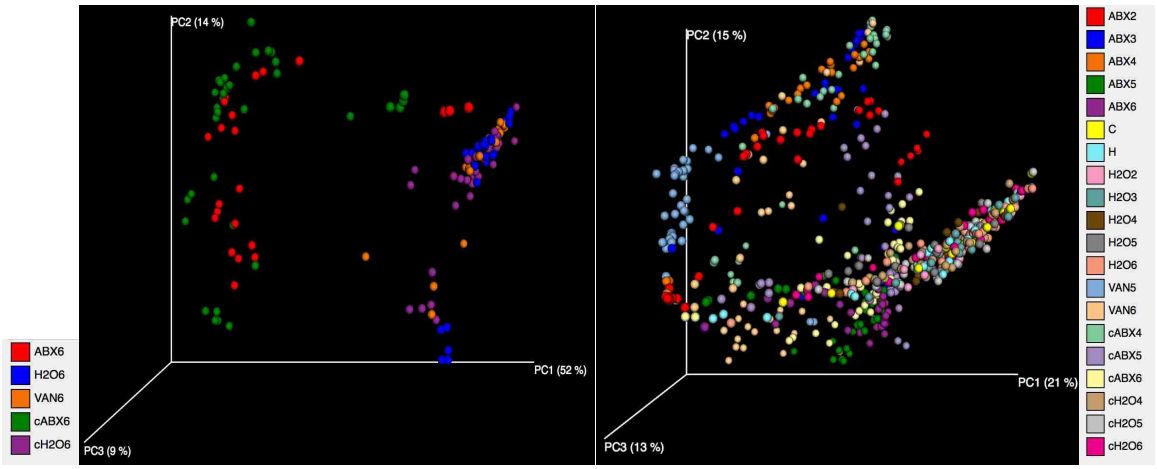
Young, V.B., and Schmidt, T.M. (2004). Antibiotic-associated diarrhea accompanied by large-scale alterations in the composition of the fecal microbiota. J. Clin. Microbiol. *42*, 1203-1206.

Yutin, N., and Galperin, M.Y. (2013). A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. Environ. Microbiol. *15*, 2631-2641.

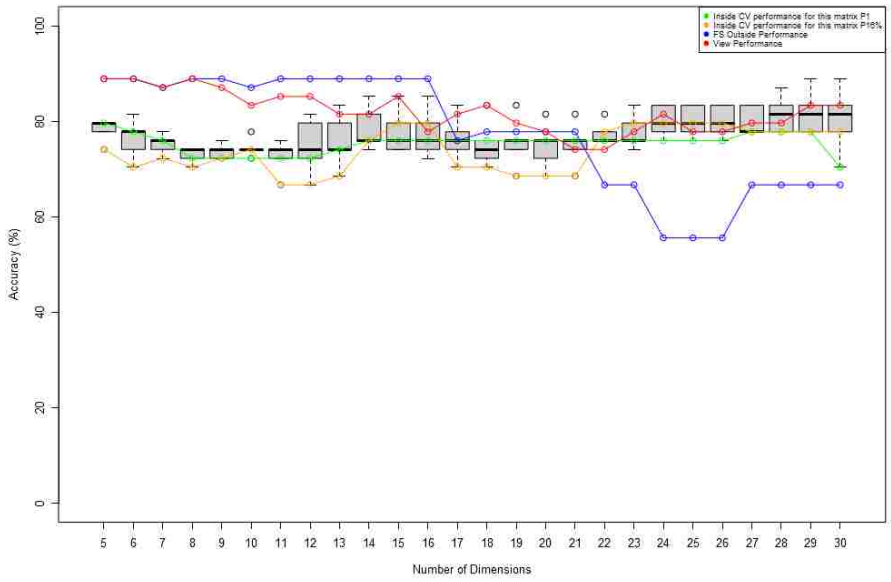
Supplementary Figures and Tables:

Supplementary Table 1: Table of PD_Whole_Tree, Chao1 and observed species diversity indices and averages.

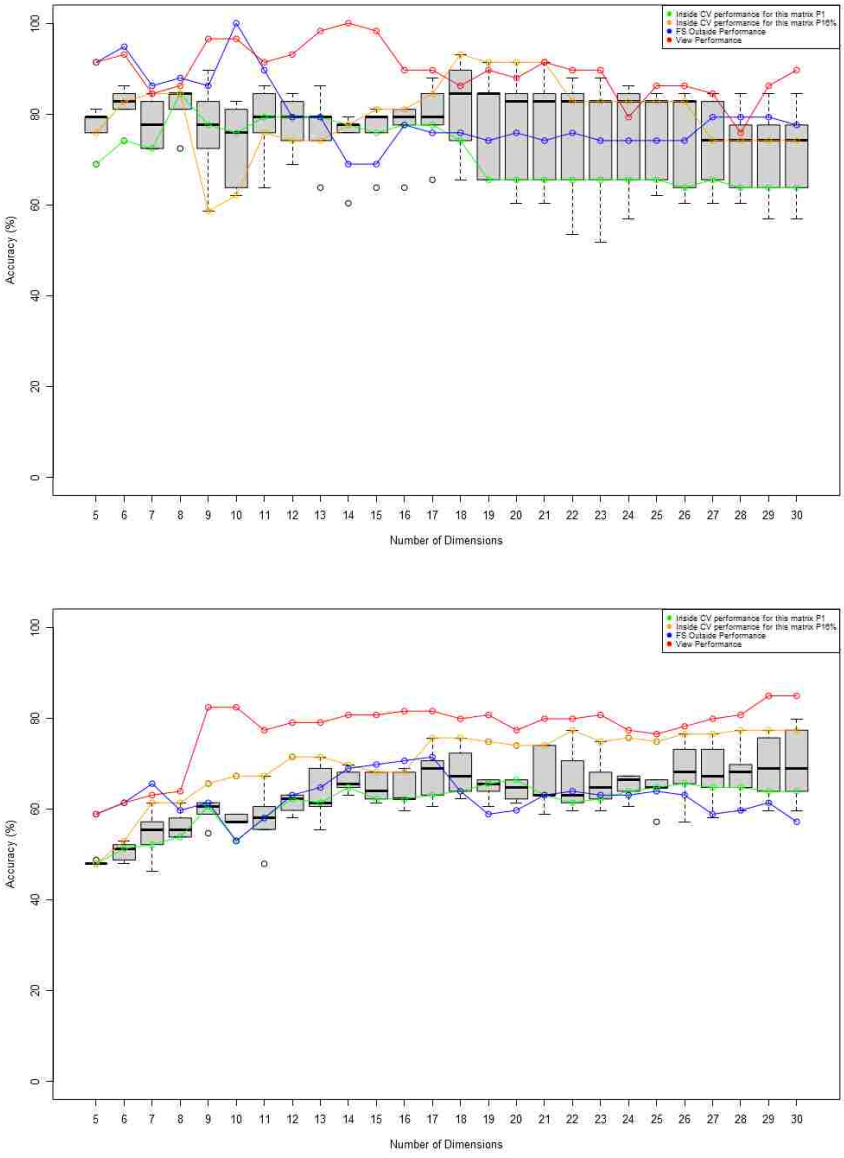
Treat_Time	Seqs/Sample	PD_whole_tree Ave.	PD_whole_tree Err.	chao1 Ave.	chao1 Err.	observed_species Ave.	observed_species Err.
ABX2	10.0	1.735	0.616	5.814	4.196	3.825	1.603
ABX2	439.0	7.816	4.384	56.170	46.011	28.117	20.360
ABX2	868.0	9.903	5.731	79.778	55.055	40.058	30.040
ABX2	1297.0	11.400	6.510	100.929	72.893	48.653	35.888
ABX2	1726.0	12.515	6.982	112.019	70.584	55.725	40.525
ABX2	2155.0	13.573	7.379	129.236	76.789	62.294	43.370
ABX2	2584.0	14.407	7.715	138.046	84.176	68.306	46.667
ABX2	3013.0	15.270	8.174	148.701	99.654	73.731	49.809
ABX2	3442.0	15.994	8.334	158.745	101.157	78.544	52.169
ABX2	3871.0	16.627	8.647	167.429	101.377	82.819	54.320
ABX2	4300.0	17.315	8.742	182.596	113.158	87.600	56.781
ABX3	10.0	2.676	0.652	11.880	8.415	5.512	1.648
ABX3	439.0	14.378	8.527	198.468	221.772	68.197	49.031
ABX3	868.0	19.281	12.752	307.733	351.176	105.691	84.510
ABX3	1297.0	22.831	15.793	370.079	427.780	135.809	112.967
ABX3	1726.0	25.755	18.635	439.692	503.923	162.485	140.963
ABX3	2155.0	28.316	21.239	506.350	624.209	186.470	166.677
ABX3	2584.0	30.733	23.448	562.468	686.422	209.264	192.002
ABX3	3013.0	32.801	25.631	605.208	744.249	231.258	217.045
ABX3	3442.0	34.559	27.403	644.754	802.821	249.697	237.650
ABX3	3871.0	36.615	29.257	699.965	865.939	268.991	258.370
ABX3	4300.0	38.293	31.079	745.285	937.869	288.185	280.009
ABX4	10.0	2.521	0.586	9.390	3.593	5.271	1.361
ABX4	439.0	9.399	1.976	74.988	20.987	41.583	11.976
ABX4	868.0	11.376	2.248	107.571	29.088	57.825	16.159
ABX4	1297.0	12.806	2.667	130.786	38.112	70.629	19.747
ABX4	1726.0	13.779	2.894	146.050	45.345	79.683	22.108
ABX4	2155.0	14.812	3.084	168.681	49.213	88.629	24.054
ABX4	2584.0	15.447	3.284	180.443	54.275	96.267	26.078
ABX4	3013.0	16.287	3.495	192.184	55.397	103.246	27.811
ABX4	3442.0	16.949	3.660	207.489	61.738	109.862	29.103
ABX4	3871.0	17.399	3.711	216.588	61.259	115.338	30.394
ABX4	4300.0	18.100	3.917	225.051	64.756	121.779	31.531
ABX5	10.0	3.122	0.597	22.501	9.911	7.811	1.354
ABX5	439.0	22.077	6.931	386.498	176.938	129.275	50.683
ABX5	868.0	30.360	9.727	571.552	258.548	206.161	84.523
ABX5	1297.0	36.512	11.858	737.278	340.565	269.519	113.170
ABX5	1726.0	41.647	13.800	880.851	417.616	326.350	140.651
ABX5	2155.0	46.007	15.469	999.751	477.085	376.783	164.082
ABX5	2584.0	49.697	16.698	1099.968	531.564	421.433	183.954
ABX5	3013.0	53.255	18.155	1201.101	574.658	464.419	204.691
ABX5	3442.0	56.613	19.483	1316.589	630.464	507.344	223.939
ABX5	3871.0	59.617	20.486	1398.793	684.082	546.364	241.938
ABX5	4300.0	62.362	21.499	1480.462	716.122	583.175	259.777
ABX6	10.0	2.982	0.475	29.384	10.190	8.414	1.188
ABX6	439.0	26.824	6.789	572.432	218.240	166.981	52.894
ABX6	868.0	38.250	10.393	849.265	319.092	274.125	92.510
ABX6	1297.0	46.671	13.019	1060.980	397.758	362.672	124.854
ABX6	1726.0	53.555	15.005	1245.880	475.398	440.569	153.518
ABX6	2155.0	59.783	17.059	1431.842	538.090	513.081	180.564
ABX6	2584.0	65.272	18.518	1587.943	590.046	580.764	204.032
ABX6	3013.0	70.203	20.190	1734.126	654.280	643.308	228.299
ABX6	3442.0	74.897	21.603	1884.094	726.373	703.000	250.933
ABX6	3871.0	78.974	22.829	2006.164	779.901	758.108	271.996
ABX6	4300.0	83.023	24.030	2140.281	821.639	811.539	291.026
C	10.0	3.607	0.483	44.900	11.996	9.367	1.188
C	439.0	48.163	10.941	1706.121	729.146	285.942	73.883
C	868.0	74.524	18.203	2725.631	1160.761	512.575	140.836
C	1297.0	95.070	24.524	3540.258	1391.270	712.650	204.551
C	1726.0	113.133	29.614	4193.735	1730.099	901.538	265.020
C	2155.0	129.239	34.263	4813.687	1974.608	1078.683	322.260
C	2584.0	143.789	38.582	5423.859	2173.980	1249.229	378.126
C	3013.0	157.563	42.765	6014.184	2433.353	1412.475	433.858
C	3442.0	170.438	46.682	6584.883	2663.306	1570.525	487.192
C	3871.0	182.322	50.364	7082.003	2873.496	1723.021	541.309
C	4300.0	193.477	53.978	7525.560	3037.425	1870.625	592.715
H	10.0	3.319	0.395	37.600	13.112	8.856	1.426
H	439.0	41.694	12.945	1324.079	719.294	250.152	88.998
H	868.0	63.438	21.667	2082.016	1164.385	437.119	169.974
H	1297.0	80.599	28.636	2714.694	1496.538	604.596	244.796
H	1726.0	95.448	34.732	3271.057	1766.531	759.737	315.395
H	2155.0	108.720	40.406	3795.281	2087.058	904.563	384.483
H	2584.0	120.459	45.408	4234.560	2324.326	1040.433	449.294
H	3013.0	131.994	50.083	4724.071	2600.649	1176.856	512.234
H	3442.0	142.288	54.647	5138.849	2892.097	1302.374	574.565
H	3871.0	151.987	58.846	5481.060	3079.037	1426.130	635.995
H	4300.0	161.654	63.255	5897.378	3315.762	1547.896	695.824
H2O2	10.0	3.414	0.356	38.233	12.641	9.072	1.007
H2O2	439.0	44.302	10.478	1311.507	674.430	254.356	75.322
H2O2	868.0	67.372	18.176	2019.507	1003.119	443.467	147.779
H2O2	1297.0	85.598	24.485	2652.153	1341.228	609.656	215.435
H2O2	1726.0	100.531	30.031	3270.689	1680.956	760.706	280.455
H2O2	2155.0	114.877	35.370	3724.365	1865.624	906.589	343.354
H2O2	2584.0	127.636	39.937	4325.524	2216.247	1046.294	402.485
H2O2	3013.0	138.477	44.217	4711.268	2404.276	1172.156	458.820
H2O2	3442.0	149.978	48.418	5061.781	2563.566	1301.372	516.916
H2O2	3871.0	159.720	52.787	5533.480	2842.174	1423.067	575.086
H2O2	4300.0	170.003	56.853	5973.498	3053.021	1546.361	630.602
H2O3	10.0	3.281	0.497	34.202	13.107	8.587	1.466
H2O3	439.0	37.419	10.405	982.699	500.542	215.739	77.527
H2O3	868.0	56.096	17.020	1570.566	820.111	372.874	143.583
H2O3	1297.0	70.707	22.369	2046.803	1085.183	511.030	204.642
H2O3	1726.0	82.936	26.862	2483.358	1287.862	636.448	258.367
H2O3	2155.0	93.827	31.367	2866.880	1527.196	753.039	314.870
H2O3	2584.0	103.862	35.412	3221.130	1722.289	863.409	364.548
H2O3	3013.0	112.997	38.868	3536.724	1861.136	971.174	414.564
H2O3	3442.0	121.853	42.341	3891.142	2081.698	1073.252	461.743
H2O3	3871.0	129.445	45.816	4182.573	2264.147	1169.426	510.998
H2O3	4300.0	137.599	48.678	4494.060	2419.305	1266.174	553.586
H2O4	10.0	3.391	0.491	34.793	12.399	8.765	1.174
H2O4	439.0	40.080	10.564	1054.304	569.440	225.883	70.526
H2O4	868.0	59.835	17.137	1609.812	832.028	385.652	130.738
H2O4	1297.0	74.969	22.514	2084.756	1081.101	524.970	188.024
H2O4	1726.0	88.475	27.435	2554.307	1290.323	657.674	241.691
H2O4	2155.0	100.556	31.970	2952.559	1566.029	778.230	294.889
H2O4	2584.0	111.034	35.444	3357.610	1766.406	889.996	340.254
H2O4	3013.0	120.931	39.386	3714.638	1972.293	998.922	390.575
H2O4	3442.0	130.085	42.916	4049.435	2073.835	1105.257	435.649
H2O4	3871.0	138.766	46.016	4345.520	2276.838	1206.891	481.804
H2O4	4300.0	146.947	49.237	4651.138	2456.174	1303.687	525.806
H2O5	10.0	3.134	0.475	32.028	10.839	8.513	1.522
H2O5	439.0	36.154	8.908	882.138	375.121	211.761	64.225
H2O5	868.0	54.765	14.168	1414.084	592.769	362.157	114.384
H2O5	1297.0	68.836	18.914	1815.335	775.776	491.596	161.838
H2O5	1726.0	80.950	22.491	2220.349	911.456	611.839	203.776
H2O5	2155.0	91.443	26.252	2565.005	1051.039	721.887	245.465
H2O5	2584.0	101.082	29.164	2856.059	1180.033	825.309	284.612
H2O5	3013.0	109.796	32.291	3168.813	1306.876	923.617	323.625
H2O5	3442.0	118.063	35.159	3416.452	1429.812	1018.243	359.654
H2O5	3871.0	126.209	37.720	3693.577	1554.814	1107.830	394.143
H2O5	4300.0	132.771	39.963	3956.652	1649.213	1195.226	426.280
H2O6	10.0	3.246	0.348	36.285	12.091	8.833	1.304
H2O6	439.0	38.225	9.857	1050.160	442.897	231.148	70.627
H2O6	868.0	57.878	15.723	1660.926	684.756	399.989	128.409
H2O6	1297.0	73.244	21.086	2197.579	945.448	547.844	183.345
H2O6	1726.0	86.412	25.314	2591.586	1071.772	684.130	231.927
H2O6	2155.0	97.854	29.216	3033.211	1289.965	809.811	280.126
H2O6	2584.0	108.234	32.998	3364.227	1406.480	926.759	325.336
H2O6	3013.0	118.299	36.115	3767.888	1562.723	1044.489	368.743
H2O6	3442.0	127.149	39.525	4047.441	1691.000	1151.267	413.036
H2O6	3871.0	135.902	42.404	4388.372	1839.330	1256.574	452.074
H2O6	4300.0	144.048	45.118	4646.900	1913.519	1358.726	491.370
VAN5	10.0	2.260	0.499	6.893	3.801	4.158	1.227
VAN5	439.0	8.554	4.619	73.262	74.370	35.306	23.428
VAN5	868.0	10.656	6.322	101.397	94.970	50.206	36.351
VAN5	1297.0	12.132	7.811	121.651	109.470	61.733	47.509
VAN5	1726.0	13.364	8.982	134.825	125.444	71.303	57.714
VAN5	2155.0	14.416	10.111	150.374	143.603	80.064	67.008
VAN5	2584.0	15.343	10.991	164.931	153.599	88.186	75.614
VAN5	3013.0	15.942	11.434	169.296	149.999	94.086	80.499
VAN5	3442.0	16.601	11.981	182.365	160.902	100.025	86.319
VAN5	3871.0	17.290	12.844	189.804	168.891	107.144	95.342
VAN5	4300.0	17.915	13.224	202.598	175.273	112.317	99.271
VAN6	10.0	2.744	0.546	10.426	3.338	5.731	1.069
VAN6	439.0	13.362	3.204	129.111	41.984	57.206	14.466
VAN6	868.0	17.232	4.185	186.602	61.875	84.586	22.554
VAN6	1297.0	19.821	4.743	231.747	76.118	105.922	28.976
VAN6	1726.0	21.922	5.417	262.718	83.125	124.167	34.770
VAN6	2155.0	23.633	5.984	298.294	97.018	140.558	40.063
VAN6	2584.0	25.025	6.355	325.675	108.614	154.708	44.661
VAN6	3013.0	26.266	6.721	350.582	117.499	167.328	48.970
VAN6	3442.0	27.422	7.003	373.824	127.943	179.375	52.724
VAN6	3871.0	28.655	7.466	401.737	135.713	191.658	57.516
VAN6	4300.0	29.678	7.797	423.272	150.050	203.392	61.185
cABX4	10.0	2.298	0.660	7.326	3.609	4.594	1.031
cABX4	439.0	9.197	3.716	85.624	65.287	36.572	18.615
cABX4	868.0	11.812	5.361	126.461	89.892	53.525	29.743
cABX4	1297.0	13.832	6.683	153.989	112.955	67.258	39.018
cABX4	1726.0	15.517	7.810	184.563	139.987	79.183	48.251
cABX4	2155.0	16.862	8.729	201.722	139.869	89.808	54.900
cABX4	2584.0	18.031	9.488	222.166	159.355	99.342	62.759
cABX4	3013.0	19.081	10.244	237.526	169.057	107.653	69.316
cABX4	3442.0	20.166	10.947	259.789	180.053	116.211	74.860
cABX4	3871.0	21.035	11.611	273.023	192.268	123.778	80.603
cABX4	4300.0	21.907	12.100	287.182	201.624	131.653	86.504
cABX5	10.0	2.801	0.531	24.040	13.061	7.756	1.427
cABX5	439.0	23.374	10.614	518.040	523.974	140.553	88.250
cABX5	868.0	33.074	17.029	808.705	843.206	231.075	162.539
cABX5	1297.0	40.601	22.121	1020.015	1076.865	308.197	228.426
cABX5	1726.0	46.941	26.615	1240.332	1330.858	377.678	290.366
cABX5	2155.0	52.209	30.465	1406.527	1494.294	440.069	348.618
cABX5	2584.0	57.002	34.216	1551.132	1660.540	500.058	403.995
cABX5	3013.0	61.723	37.573	1718.694	1878.849	557.308	438.419
cABX5	3442.0	65.753	40.924	1886.025	2057.336	611.861	509.309
cABX5	3871.0	69.548	43.806	1991.128	2168.769	661.450	557.825
cABX5	4300.0	73.213	46.725	2158.758	2394.840	710.394	606.198
cABX6	10.0	2.933	0.543	28.704	11.767	8.308	1.492
cABX6	439.0	24.995	6.308	512.938	221.256	152.953	52.834
cABX6	868.0	34.868	9.462	762.251	341.653	246.767	91.028
cABX6	1297.0	42.705	11.894	955.366	426.097	327.383	124.395
cABX6	1726.0	48.852	14.066	1136.763	501.929	396.633	154.025
cABX6	2155.0	54.506	15.909	1276.810	576.936	461.844	182.551
cABX6	2584.0	59.269	17.547	1410.084	652.954	520.406	207.905
cABX6	3013.0	63.865	19.180	1568.951	715.408	577.506	233.671
cABX6	3442.0	67.985	20.520	1675.720	769.500	629.125	255.010
cABX6	3871.0	71.742	21.898	1821.444	844.324	679.906	278.638
cABX6	4300.0	75.289	23.017	1908.062	896.290	726.289	299.321
cH2O4	10.0	3.108	0.689	25.975	10.791	7.894	1.554
cH2O4	439.0	30.937	9.536	641.864	385.948	169.814	66.212
cH2O4	868.0	44.741	14.884	972.140	544.955	280.708	118.863
cH2O4	1297.0	55.179	18.947	1252.863	719.324	374.478	165.298
cH2O4	1726.0	64.200	22.564	1493.748	871.355	460.444	208.383
cH2O4	2155.0	72.262	25.591	1729.885	1014.209	540.544	248.678
cH2O4	2584.0	79.262	28.754	1936.221	1153.149	612.839	287.972
cH2O4	3013.0	85.983	31.497	2142.247	1264.980	684.731	326.543
cH2O4	3442.0	91.751	34.057	2321.875	1430.187	750.033	362.514
cH2O4	3871.0	97.513	36.676	2494.593	1505.990	813.819	396.869
cH2O4	4300.0	102.797	38.849	2665.342	1636.217	875.378	430.349
cH2O5	10.0	3.153	0.587	37.601	14.909	8.761	1.734
cH2O5	439.0	39.476	12.959	1135.086	713.447	234.961	89.455
cH2O5	868.0	59.132	20.359	1720.245	1049.695	402.625	166.032
cH2O5	1297.0	74.425	27.028	2222.201	1313.792	550.092	236.338
cH2O5	1726.0	87.707	32.693	2726.779	1631.213	686.242	304.707
cH2O5	2155.0	99.414	37.736	3125.513	1879.641	813.036	368.835
cH2O5	2584.0	110.016	42.215	3548.681	2123.318	932.128	428.530
cH2O5	3013.0	119.710	46.636	3945.967	2415.418	1046.114	489.199
cH2O5	3442.0	128.726	50.888	4273.618	2613.844	1155.633	547.441
cH2O5	3871.0	137.370	54.570	4617.253	2782.255	1261.894	603.473
cH2O5	4300.0	145.622	58.258	4921.576	2978.060	1366.142	658.353
cH2O6	10.0	3.408	0.532	37.693	11.739	9.014	1.077
cH2O6	439.0	41.922	9.397	1035.784	391.349	237.875	60.214
cH2O6	868.0	62.562	14.998	1685.618	690.921	406.356	112.933
cH2O6	1297.0	78.706	19.339	2216.342	951.535	554.658	161.010
cH2O6	1726.0	92.416	23.565	2666.134	1136.356	689.022	207.775
cH2O6	2155.0	104.623	27.183	3115.346	1343.683	817.022	252.018
cH2O6	2584.0	115.554	30.419	3470.278	1475.715	934.628	293.359
cH2O6	3013.0	125.941	33.697	3857.899	1651.793	1050.467	335.863
cH2O6	3442.0	135.251	36.753	4172.808	1841.697	1159.761	377.139
cH2O6	3871.0	144.191	39.499	4508.203	1956.465	1265.403	415.807
cH2O6	4300.0	152.750	42.596	4820.671	2102.547	1367.433	456.431



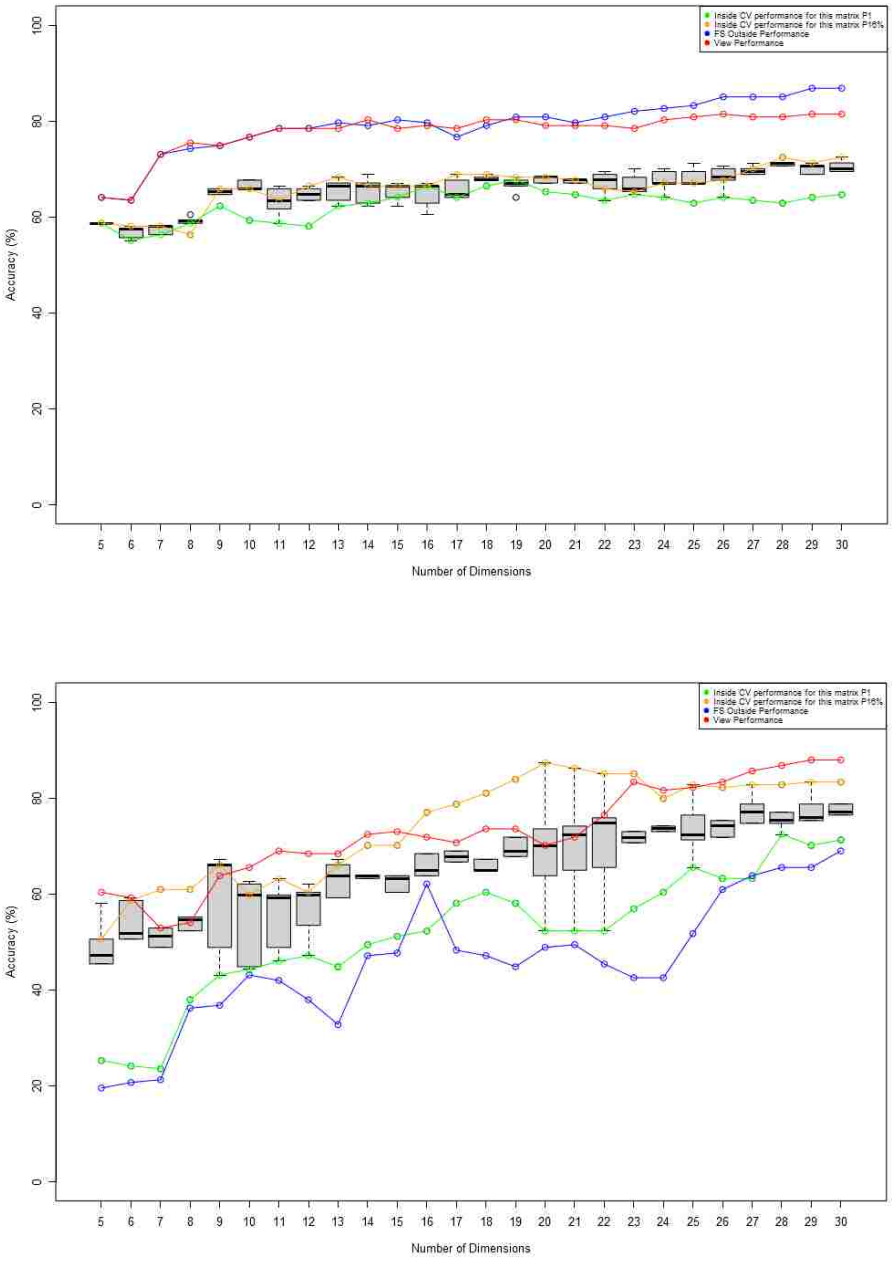
Supplementary Figure S1: Left: Unifrac for timepoint 1 and 4. Right Unifrac for total experiment.



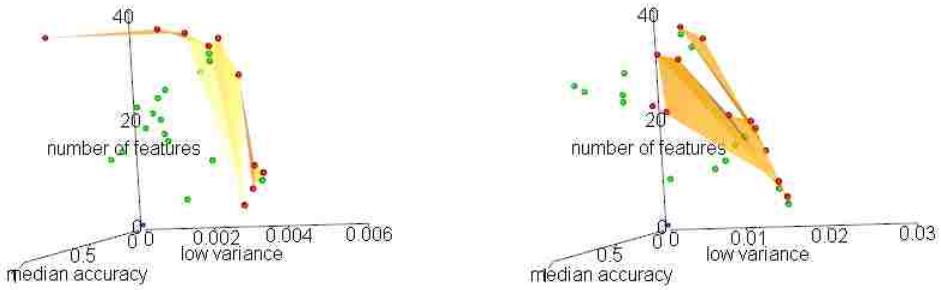
Supplementary Figure S2: Box plots showing cross-validation accuracies when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was CDF time1 and 2 filtered to 2 treatments: Water and Antibiotics. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature selection inside of cross-validation was compiled and used to select genera outside (before) cross-validation was performed.



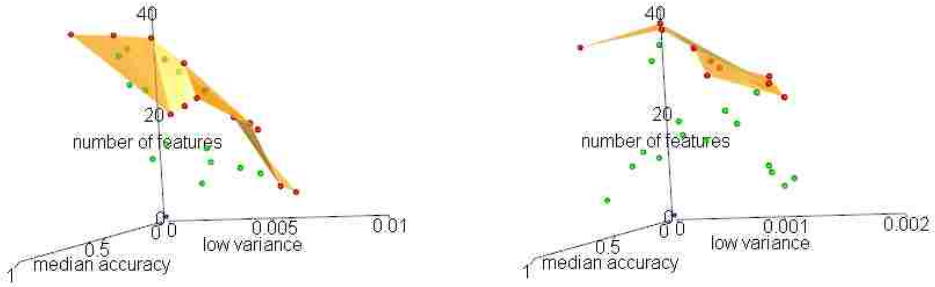
Supplementary Figure S3: Box plots showing cross-validation accuracies for 2 LDA runs when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was CDF filtered to 4 treatments: Water, Antibiotics Water, Cdiff and Antibiotics, Cdiff. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature selection inside of cross-validation was compiled and used to select genera outside (before) cross-validation was performed.



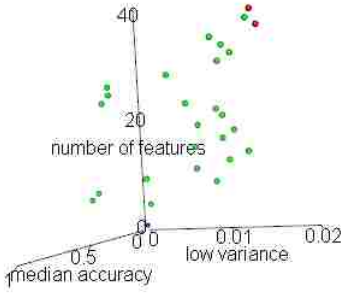
Supplementary Figure S4: Box plots showing cross-validation accuracies for 2 LDA runs when feature selection was performed inside of cross-validation to different numbers of dimensions. The base dataset used was CDF filtered to 5 treatments: Water; Antibiotics; Water Cdiff, Antibiotics, Cdiff; Vancomycin, Cdiff. The **green** line shows accuracies when using Pruning level 1 (P1) i.e. the complete dataset. The **orange** line shows accuracies using the P16% dataset (refer to METHODS). The **blue** line shows accuracies when feature selection was performed outside (before) cross-validation. The **red** line shows accuracies when data from feature selection inside of cross-validation was compiled and used to select genera outside (before) cross-validation was performed.



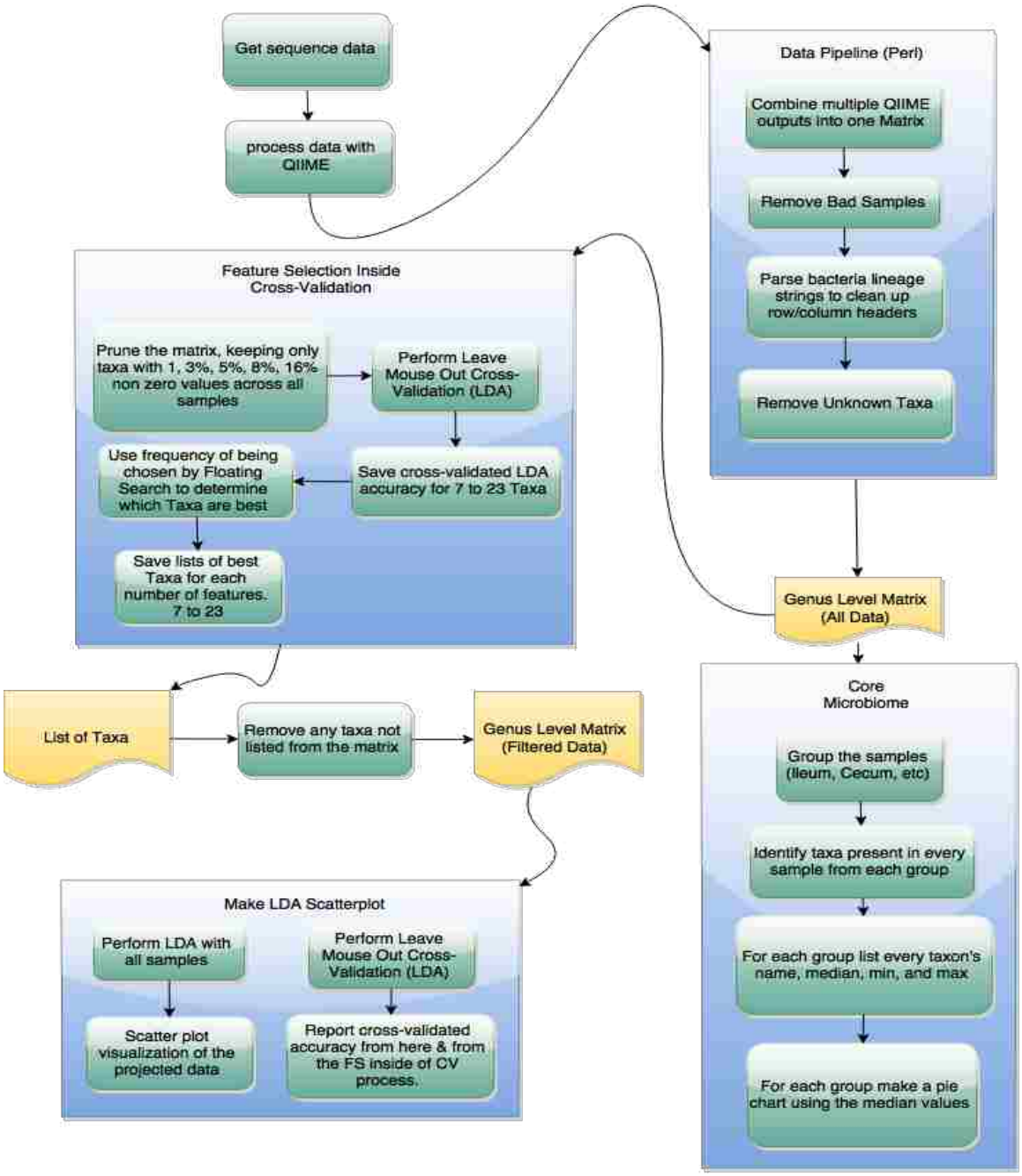
Supplementary Figure S5: Scatter plot of the 3D-Pareto frontiers when Supplementary Figures 2 and 3 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. **Green** points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.



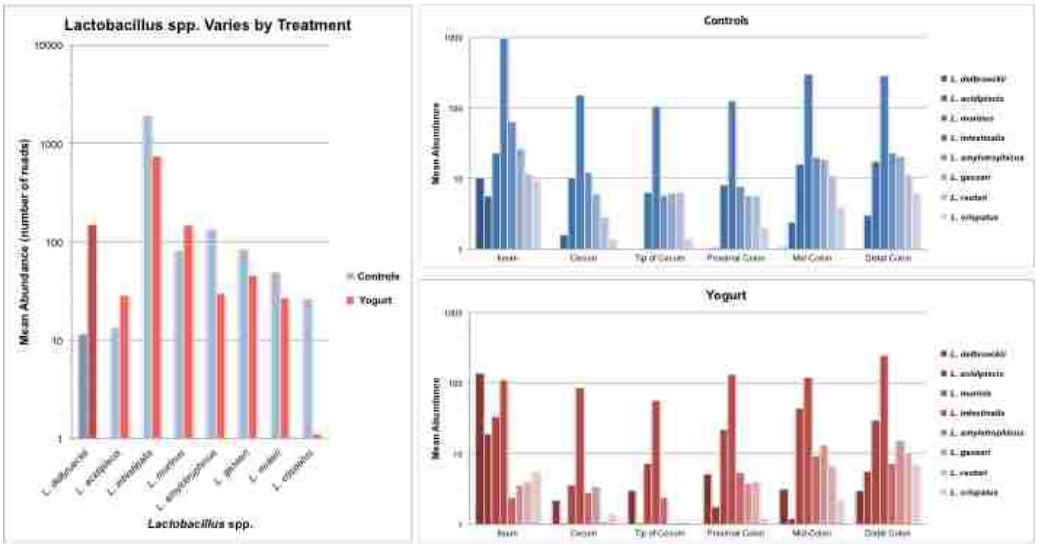
Supplementary Figure S6: Scatter plot of the 3D-Pareto frontiers when Supplementary Figures 4 and 5 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. **Green** points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.



Supplementary Figure S7: Scatter plot of the 3D-Pareto frontiers when Supplementary Figure 6 box plot data are optimized by median accuracy, lowest variance, and number of dimensions. **Green** points represent boxes that are dominated by other boxes, while red points represent boxes that are dominated by no other box, thus representing equally optimal solutions. The **orange** border is a series of triangles drawn when the red points are sorted by median accuracy and sets of 3 points are taken using a sliding window to draw n – 2 triangles. The **blue** point in the background represents the origin (0,0,0) as a frame of reference.



Supplementary Figure S8: Process flow diagram of the computational methods employed.



Supplementary Figure S9: The abundance and location of Lactobacillus spp. changes due to introduction of yogurt into the intestine. Crosshatching indicates *Lactobacillus delbrueckii*