Anaerobic Microbes and Communities In the Context of Soil and the Equine Digestive Tract

Amy Biddle

University of Massachusetts - Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2

Part of the Microbiology Commons

Recommended Citation

https://scholarworks.umass.edu/dissertations_2/2

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
ANAEROBIC MICROBES AND COMMUNITIES IN THE CONTEXT OF SOIL
AND THE EQUINE DIGESTIVE TRACT

A Dissertation Presented

By

AMY SANDERS BIDDLE

Submitted to the Graduate School of the
University of Massachusetts, Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2014

Department of Microbiology
ANAEROBIC MICROBES AND COMMUNITIES IN THE CONTEXT OF SOIL
AND THE EQUINE DIGESTIVE TRACT

A Dissertation Presented

By

AMY SANDERS BIDDLE

Approved as to style and content by:

______________________________
Jeffrey L. Blanchard, Chair

______________________________
Susan Leschine, Member

______________________________
Samuel P. Hazen, Member

______________________________
Kristen M. DeAngelis, Member

______________________________
John M. Lopes, Department Head
Department of Microbiology
DEDICATION

To all of the horse owners who are anxiously waiting for the vet to arrive tonight.
ACKNOWLEDGMENTS

I wish to thank my advisor, Jeffrey L. Blanchard, for his friendship and guidance and for fostering a supportive lab environment in which to do science; Susan Leschine for her sound advice and uncanny ability to appear at odd hours when things are falling apart; members of my committee for their time and helpful suggestions; members of the Blanchard and Leschine Labs for all of their assistance and input, especially Kelly Haas, Supratim Mukherjee, Maddie Coppi, Tom Warnick, Mary Hagen, Tara Mahendrarajah, Josephine Aiyeku, Sonia Filipczak, and Ava Bakhtyari; Grace Pold and members of the DeAngelis Lab for their technical expertise, snacks and general good cheer; Lucy Stewart and the members of the Holden Lab for their help with GC and the taxonomy project; Ruthie, Lorelle, Sue and Maryanne who keep the Microbiology Department running so smoothly; Carlos Gradil and Sam Black for their assistance and encouragement for the horse gut work; Manfred Auer and Danielle Jorgens for giving me a nudge at the very beginning.

Special thanks go to friends and family who have cheered me onward through each step of this process, in particular Suzy Biddle, Chris Madigan, Marilyn Miller, Gale Christensen, Jody Goodrich, Libby Volckening, and Charles Malloch.
ABSTRACT

ANAEROBIC MICROBES AND COMMUNITIES IN THE CONTEXT OF SOIL AND THE EQUINE DIGESTIVE TRACT

FEBRUARY 2014

AMY SANDERS BIDDLE, B.S., UNIVERSITY OF NEW HAMPSHIRE

M.Ed., ANTIOCH NEW ENGLAND

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Jeffrey L. Blanchard

Soil and herbivore gut environments present different challenges to plant degrading bacteria in terms of nutrient availability, fluctuations in moisture, pH and temperature, and temporal constraints, however complex communities of microbes in each serve similar roles in hydrolyzing and fermenting the diverse components of plant biomass.

This dissertation describes four projects with the underlying purpose to further understand the structure and functioning of anaerobic plant degrading communities.

(1) A three-year microcosm experiment using enrichment and serial transfers to reduce the diversity of a complex soil community over time, tested the hypothesis that changes in community structure would be consistent across replicate samples and enabled the detection and isolation of persistent community members. (2) A second project to track the changes in bacterial community structure and function that occur during starch induced lactate acidosis in horses identified specific microbes that could be implicated in the recovery and/or resistance to these community changes. (3) Genomic data was used to
compare clostridial species inhabiting the gut (belonging to the Lachnospiraceae, and Ruminococcaceae) and those that are free-living (belonging to the Clostridiaceae) to identify metabolic strategies that could enable specialization to a host associated or free-living lifestyle. (4) The genomic sequence analysis of *Clostridium indolis*, a member of the *C. saccharolyticum* species group, provided insights into the genetic potential of this poorly described taxa which will drive hypotheses regarding its metabolic and ecological activities and help to resolve distinctions between closely related taxa in this taxonomically confusing clade within the Lachnospiraceae.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Acknowledgements</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>viii</td>
</tr>
</tbody>
</table>

## 1. Introduction

1.1 Microbes in plant rich anaerobic environments .......................................................... 1
1.2 Plant degradation in soil ............................................................................................... 1
1.3 Plant degradation in the herbivore gut ........................................................................ 3
1.4 The genomic basis of plant degradation in soil and gut microbes ............................ 4
1.5 The genome analysis of the *Clostridium saccharolyticum* species group ............ 5

## 2. Early Establishment and Persistence of Microcosm Community Diversity Through Three Years of Serial Transfers

2.1 Abstract ...................................................................................................................... 7
2.2 Introduction ............................................................................................................... 8
2.3 Materials and methods ............................................................................................ 10
2.4 Results .................................................................................................................... 15
2.5 Discussion ............................................................................................................... 25
2.6 Conclusion ............................................................................................................... 29

## 3. An *In Vitro* Model of the Horse Gut Microbiome Enables Identification of Lactate-Utilizing Bacteria That Differentially Respond to Starch Induction

3.1 Abstract .................................................................................................................. 31
3.2 Introduction ............................................................................................................. 32
3.3 Materials and methods ............................................................................................ 35
3.4 Results .................................................................................................................... 38
3.5 Discussion ............................................................................................................... 47
3.6 Conclusion ............................................................................................................... 51
H. ACTIVITIES OF ENZYMES THAT DIFFER BETWEEN THE LACHNOSPIRACEAE, CLOSTRIDIACEAE, AND RUMINOCOCCACEAE ................................................................. 99

I. SUBSTRATE AFFINITIES OF CARBOHYDRATE BINDING MODULES THAT DIFFER BETWEEN THE LACHNOSPIRACEAE, CLOSTRIDIACEAE, AND RUMINOCOCCACEAE ........................................... 100

REFERENCES ................................................................................................................ 101
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Sequence metrics of five microcosm samples at nine transfer points</td>
<td>17</td>
</tr>
<tr>
<td>2. Sequence metrics of sequencing quality assessment</td>
<td>24</td>
</tr>
<tr>
<td>3. Sequence metrics of horse gut samples</td>
<td>41</td>
</tr>
<tr>
<td>4. Classification and general features of <em>Clostridium indolis</em> DSM 755</td>
<td>71</td>
</tr>
<tr>
<td>5. Project information</td>
<td>72</td>
</tr>
<tr>
<td>6. Nucleotide content and gene count levels of the genome of <em>C. indolis</em> DSM 755</td>
<td>74</td>
</tr>
<tr>
<td>7. Number of genes in <em>C. indolis</em> DSM 755 associated with the 25 general COG functional categories</td>
<td>75</td>
</tr>
<tr>
<td>8. Number of genes in <em>C. indolis</em> DSM 755 not found in near-relatives associated with the 25 general COG functional categories</td>
<td>76</td>
</tr>
<tr>
<td>9. Selected carbohydrate active genes in the <em>C. indolis</em> DSM 755 genome</td>
<td>77</td>
</tr>
<tr>
<td>10. Selection of <em>C. indolis</em> DSM 755 genes related to citrate utilization</td>
<td>78</td>
</tr>
<tr>
<td>11. Selection of <em>C. indolis</em> DSM 755 genes related to nitrogen fixation</td>
<td>79</td>
</tr>
<tr>
<td>12. Selection of <em>C. indolis</em> DSM 755 genes related to lactate utilization</td>
<td>81</td>
</tr>
<tr>
<td>13. Selection of <em>C. indolis</em> DSM 755 genes related to degradation of aromatics</td>
<td>82</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Short chain fatty acid metabolites in microcosms over time</td>
<td>16</td>
</tr>
<tr>
<td>2.</td>
<td>Rarefaction curve of microcosm sequencing data</td>
<td>17</td>
</tr>
<tr>
<td>3.</td>
<td>Richness in microcosms over time</td>
<td>18</td>
</tr>
<tr>
<td>4.</td>
<td>Alpha diversity of microcosms over time</td>
<td>19</td>
</tr>
<tr>
<td>5.</td>
<td>Beta diversity of microcosms by sample and transfer</td>
<td>19</td>
</tr>
<tr>
<td>6.</td>
<td>Distribution of microcosm sequences by phyla</td>
<td>20</td>
</tr>
<tr>
<td>7.</td>
<td>Distribution of Proteobacteria sequences by OTU</td>
<td>21</td>
</tr>
<tr>
<td>8.</td>
<td>Distribution of microcosm sequences by OTU, across all transfers</td>
<td>22</td>
</tr>
<tr>
<td>9.</td>
<td>Distribution of shared sequences in all replicates at Transfer 62</td>
<td>23</td>
</tr>
<tr>
<td>10.</td>
<td>pH of cultures over time by horse</td>
<td>39</td>
</tr>
<tr>
<td>11.</td>
<td>Short chain fatty acid metabolites over time by horse</td>
<td>40</td>
</tr>
<tr>
<td>12.</td>
<td>Distribution of horse gut sequences by phyla</td>
<td>42</td>
</tr>
<tr>
<td>13.</td>
<td>Distribution of horse gut Firmicute sequences by family</td>
<td>43</td>
</tr>
<tr>
<td>14.</td>
<td>Distribution of most abundant horse gut OTUs at time 48</td>
<td>44</td>
</tr>
<tr>
<td>15.</td>
<td>Change in most abundant horse gut Veillonellaceae OTUs over time</td>
<td>45</td>
</tr>
<tr>
<td>16.</td>
<td>Change in single, abundant horse gut Lactobacillaceae OTU over time</td>
<td>46</td>
</tr>
<tr>
<td>17.</td>
<td>Change in abundant horse gut Streptococcaceae OTUs over time</td>
<td>47</td>
</tr>
<tr>
<td>18.</td>
<td>16S Neighbor-joining tree of the representative taxa used in this analysis</td>
<td>59</td>
</tr>
<tr>
<td>19.</td>
<td>Proportion of members of each family that are gut associated</td>
<td>60</td>
</tr>
<tr>
<td>20.</td>
<td>Comparison of Carbohydrate-Active enZymes (CAZy) with respect to abundance in the genomes of Lachnospiraceae, Clostridiaceae, and Ruminococcaceae</td>
<td>61</td>
</tr>
</tbody>
</table>
21. Glycoside hydrolase families that differ in abundance ..................................................61
22. Glycoside hydrolase enzymes that differ in abundance ...............................................62
23. Carbohydrate-binding module (CBM) families that differ in abundance ....................63
24. Average abundance of phosphotransferase system (PTS) and ATP-binding cassette (ABC) transporter genes ..........................................................64
25. Average abundance of sugar transport genes ...............................................................65
26. Average abundance of genes in each Ecocyc degradation pathway ..............................66
27. Carbohydrate and Polymeric Compound Degradation pathways ................................67
28. Phylogenetic tree of 16S rRNA genes of the Clostridium saccharolyticum species group .................................................................................................................69
29. Distribution of ABC and PTS transporters in the genomes of C. indolis and close relatives .........................................................................................................77
30. Citrate utilization genes in C. indolis ...........................................................................79
CHAPTER 1
INTRODUCTION

1.1 Microbes in plant rich anaerobic environments

Plant biomass is a composite of fibrils and sheets of cellulose, hemicellulose, pectin, lignin, waxes, and proteins forming a complex network that provides support for plants while resisting attack from bacteria and fungi. While soil and herbivore gut environments present different challenges to plant degrading bacteria in terms of nutrient availability, fluctuations in moisture, pH and temperature, and temporal constraints, complex communities of microbes in each serve similar roles in hydrolyzing and fermenting the diverse components of plant biomass. The mechanisms of plant degradation by microbial communities are fundamental to carbon flow dynamics in anaerobic environments, and underlie diverse research areas including climate change, gut microbe-host interactions, soil formation, waste management, and biofuels development.

1.2 Plant degradation in soil

Three quarters of the carbon in terrestrial ecosystems is found as organic matter in soils, mostly of plant origin [1]. Anaerobic degradation accounts for as much as 10% of cellulose hydrolysis in soils through transformations mediated by microbes [2]. Soil microbes convert lignocellulose to dissolved organic matter, recalcitrant aggregates, or gases such as CO₂ and CH₄ that are released into the atmosphere [3]. Soil microbes can be challenged by nitrogen limitation, and/or wide fluctuations in temperature [4], moisture levels [5], and redox conditions [6], but under normal conditions there is little
mixing of the substrate, favoring the persistence of slow growing bacteria. Aerobic fungi and bacteria begin the decomposition process of fresh leaf litter, but as layers build over time and become saturated with water, the rate of oxygen diffusion is soon insufficient to maintain aerobic conditions, resulting in a thin oxic layer restricted to the surface [7]. Mature forest soil, where layers of leaf litter have accumulated over decades, is an ideal source for discovering anaerobic plant degraders because anoxic conditions prevail as leaf layers become compacted and waterlogged.

Soil is one of the most diverse microbial habitats, harboring thousands of species in a single gram [8]. While estimates of soil species diversity are quite high in comparison to other environments [9], phylogenetic diversity has been found to be surprisingly low [10], suggesting the presence of many closely related species with similar functions that may be adapted to slightly different niches [11]. Efforts to understand the physiological ecology of plant degrading microbial communities in soil has been hampered by the complexity of these communities. Forest soil organisms such as *Clostridium phytofermentans* [12] have been isolated and studied for biofuels development, but the process of isolation of single bacterial strains is laborious. The estimated 1% of soil microbes that are culturable have exacting nutritional requirements that are likely met in natural environments through associations with other organisms. In fact, cellulose breakdown is hastened by cocultures of cellulolytic and noncellulolytic microbes [13–17]. Reports suggest that noncellulolytic microbes may help to maintain a favorable pH, consume oxygen or inhibitory metabolites [14], cooperate through crossfeeding of nitrogenous compounds [17], or remove inhibitors to degradation [18].
Chapter 2 of this dissertation describes a three-year microcosm experiment using enrichment and serial transfers to reduce the diversity of a complex soil community over time. This strategy tested the hypothesis that changes in community structure over time would be consistent across replicate samples, and enabled the detection of persistent community members.

1.3 Plant degradation in the herbivore gut

The gastrointestinal tract harbors a dense and diverse community of microbes with bacterial numbers exceeding $10^{14}$ per gram [19]. 70-80% of total gut volume is devoted to fermentation in herbivores, while this volume is much less in omnivores and carnivores (17% in humans and 14% in dogs) [20,21]. In horses, microbial fermentation supplies the host with over 70% of its energy needs through the production of volatile fatty acids and microbial proteins [22].

The biochemistry of plant degradation is essentially the same between soils and the gut, but the processivity in the digestive systems of animals is inherently inefficient, as the transit time of material through the gut constrains the extent of breakdown. Horses, with an average transit time of 30 hours, digest 50% of the dry matter they consume, while giant pandas, with an average transit time of 8 hours, digests only 20% of dry matter [23]. While the transit time through the digestive system imposes temporal constraints, each compartment of the gut is a remarkably stable environment in terms of temperature and moisture levels, although pH fluctuates over time between meals [24]. This stability can be disrupted by dietary and/or metabolic changes with drastic consequences for the
microbial community and its activity. In horses, a sudden influx of dietary starch, favoring the growth of lactic acid bacteria in the hindgut, can precipitate lactic acidosis, leading to colic and/or laminitis.

Chapter 3 describes an experiment to track the changes in bacterial community structure and function that occur during starch induced lactate acidosis in horses and identify specific microbes that could be implicated in the recovery and/or resistance to these dramatic changes.

1.4 The genomic basis of plant degradation in soil and gut microbes

While the capacity to break down lignocellulose is a widespread characteristic in the microbial world, the majority of cellulolytic anaerobic bacteria are clostridia, mainly from families within the Order Clostridiales. Cellulolytic activity in this group is concentrated within three families, the Lachnospiraceae, and Ruminococcaceae (clusters III and XIVa as described by Collins [25]). Not surprisingly, cultured species from these groups have been isolated from environments where plant degradation is actively occurring such as soils, compost, guts, and bioreactors. The abundance of clostridia species varies by environment with gut samples being especially enriched. Soils can harbor fewer than 5% Firmicutes [26], while this phyla comprises between 50–80% of the taxa in the core human gut microbiota [27,28] and more than 84% of the active fraction [29]. The majority of the Firmicutes in the human gut is estimated to be clostridial [28–30], while estimates of clostridia in the horse gut are closer to 70% [31–33].
Chapter 4 of this dissertation describes work to use genomic data to distinguish between clostridial species inhabiting the gut (belonging to the Lachnospiraceae, and Ruminococcaceae) from those that are free-living (belonging to the Clostridiaceae).

1.5 The genome analysis of the *Clostridium saccharolyticum* species group

Efforts to isolate bacteria from the experimental microcosms discussed in chapter 2 have resulted in the discovery of novel species as well as new strains closely related to those already in the literature. One particularly abundant family, the Lachnospiraceae, yielded several isolates closely related to *Clostridium saccharolyticum*, a taxa that has gained attention because its saccharolytic capacity has been shown to be syntrophic with the cellulolytic activity of *Bacteroides cellulosolvens* in coculture, enabling the conversion of cellulose to ethanol in a single step [16,34]. The *C. saccharolyticum* species group is a poorly described and taxonomically confusing clade. This group includes *C. indolis*, *C. sphenoides*, *C. methoxybenzovorans*, *C. celerecrescens*, and *Desulfotomaculum guttoideum*, none of which are well studied, despite interesting characteristics. Members of this group such as *C. celerecrescens* are cellulolytic [35], and others are known to degrade unusual substrates such as methylated aromatic compounds (C. *methoxybenzovorans*) [36], and the insecticide lindane (*C. sphenoides*) [37]. Seven members of this group (including two microcosm isolates) were targeted for whole genome sequencing in a proposal for the Community Sequencing Project of the Joint Genome Institute (JGI) (http://www.jgi.gov) to provide insight into the genetic potential of these taxa that could then direct experimental efforts to understand the physiology and ecology of this species group.
Chapter 5 of this dissertation describes the genomic sequence analysis of *Clostridium indolis*, a member of the *C. saccharolyticum* species group, a poorly described and taxonomically confusing clade within the Lachnospiraceae.
CHAPTER 2
EARLY ESTABLISHMENT AND PERSISTENCE OF MICROCOSM COMMUNITY DIVERSITY THROUGH THREE YEARS OF SERIAL TRANSFERS

2.1 Abstract

Substrate utilization in nature results from metabolic and spatial interactions between microbes that are more easily teased apart in simplified communities. While these can be established by intentionally combining bacteria of interest, it is not clear how stable resulting communities would be over time. Here we used a strategy of enrichment and serial transfers to reduce the diversity of a complex soil community, to enable the detection of persistent community members, and to discover whether changes in community structure over time would be consistent across replicate samples. Five replicate microcosms containing switchgrass as the primary carbon source were inoculated with forest soil.

Community structure over time was determined using deep sequencing to track changes in bacterial groups across sixty-two serial transfers over a period of three years. After an initial, dramatic reduction in bacterial diversity, the five replicates were not significantly different over time (ANOVA p < 0.05). After sixty-two transfers 83-93% of reads were shared between replicates with a total of 265 shared OTUs (out of 92,346 OTUs). All of these except for one classified into the Firmicutes. Of those that classified to the family level, 82-91.5% of the reads were found in the Enterococcaceae, Clostridiaceae, Lachnospiraceae, and Ruminococcaceae (137 shared OTUs). OTUs that differed between samples came from the same families. The most abundant OTUs include novel species
and new strains of the biofuels catalyst, *Clostridium phytofermentans*. Metabolic outputs were consistent between replicates over time.

The consistency of metabolic products and similarity of community profiles between replicates suggest selection for closely related taxa that may have similar metabolic capacity and/or cooperative interactions. The persistence of taxa in a community context suggests the potential of this method for developing stable communities from complex environments such as soil, bioreactors, and the gut.

### 2.2 Introduction

Soil harbors thousands of species in a single gram [8], representing enormous taxonomic and metabolic diversity. As revealed by 16S rRNA surveys, an estimated 99% of microbial diversity is missed using traditional culture techniques [38–40]. A recent analysis of 16S surveys suggested that while estimates of soil species diversity are quite high in comparison to other environments [9], phylogenetic diversity at higher taxonomic levels is surprisingly low [10]. This suggests the presence of many closely related species with similar functions that may occupy redundant or slightly different niches [11].

Traditional techniques of continuous culture in chemostats or serial transfers for enrichment have been commonly used to select for specific metabolic traits such as substrate utilization or stress tolerance, thus simplifying the process of isolating bacteria
of interest from environmental samples. How communities change during enrichment is a question with broad implications for understanding the response of community composition and function to disturbance, and developing model systems to study the dynamics of microbial communities that underlie ecosystem processes. In particular, the simplification of complex plant degrading communities from the gut, soil, or bioreactors offers a useful approach for teasing out spatial and metabolic interactions that evade understanding due to the complexity of the whole community. Understanding the interactions between plant degrading microbes in these environments could provide insight into the mechanisms of degradation at the community level and strategies for improving its efficiency for waste treatment, digestive physiology and biofuels production. Establishing simplified communities that resist drastic changes in membership could assist in the development of mixed cultures for a wide range of applications, from bioreactors to probiotic production.

While the composition of microbial communities has been shown to be sensitive to disturbance regardless of the taxonomic level described or the method of measurement employed, these changes may not affect ecosystem processing rates due to functional redundancy (species that perform similar ecological roles) [41]. In cellulose-rich microcosms inoculated with defined communities of microbes, functional redundancy has been demonstrated to maintain both biodiversity and ecosystem function [42]. In one study, efforts to link membership to function using a stable isotope probe pointed to ten taxa in the Clostridiaceae and thirteen taxa in Enterobacteriaceae as the most active
fermenters of glucose, suggesting the presence of many closely related and functionally redundant taxa in the earthworm gut community [43].

In this study, five communities were established from one forest soil sample and transferred weekly, then monthly into minimal media with switchgrass as the primary carbon source. Here we demonstrate that the selective pressure of enrichment and serial transfer resulted in a rapid reduction of diversity that was consistent between the five replicates over a period of three years (sixty-two transfers). Deep sequencing of the 16S rRNA gene and analysis of metabolic outputs revealed that the resulting five communities, while not identical, were strikingly similar with respect to membership, relative abundance, and metabolism.

2.3 Materials and methods

2.3.1 Soil sampling

A soil sample was collected on 9/21/09 from a mixed hardwood forest in Montague, Massachusetts at a depth of 6-10 cm, and stored at 4 °C overnight prior to inoculation. A thick slurry was prepared by the addition of sterile water followed by vortexing.

2.3.2 Inoculation and maintenance of communities

The anaerobic techniques of Hungate [7] were used in all media preparation and transfers. Microcosms were established by inoculating five 25 ml samples of modified MS media and 10% salt solution with 1 g of soil slurry. Cultures were incubated at 30 °C, and vortexed 2-3 times during the incubation period. Each sample was transferred
(2%) into fresh media every seven days for a total of 33 transfers, then every 3-4 weeks into 10 ml of the above media thereafter. Samples were collected from each microcosm at each transfer for DNA extraction and metabolic analysis. Sequencing was performed for all replicates at Time 0 and Transfers 2, 10, 20, 30, 40, 49, 55, and 62, except for replicates 1 and 2 from Transfer 55.

Modified MS media (adapted from [8,9]) was prepared as follows: For 1.0 L: KH$_2$PO$_4$, 1.04 g; K$_2$HPO$_4$, 1.11 g; NaHCO$_3$, 2.50 g; NH$_4$Cl, 0.40 g; L-cysteine HCl, 0.50 g; resazurin 0.15%, 1.20 ml, switchgrass (see below for preparation), 6.00 g. The pH was adjusted to 7.0 using 6N KOH. Switchgrass was a homogenate of chopped stem and leaf material grown and harvested as hay at the University of Massachusetts, South Deerfield farm facility in 2009. The switchgrass was further dried in a 50 °C oven for two days, and ball milled at 200 Hz for 2 minutes. Salt solution was prepared as follows: for 100 ml: MgCl$_2$×6H$_2$O, 1.0 g; CaCl$_2$×2H$_2$O, 0.15 g, FeSO$_4$×7H$_2$O, 0.00125g; autoclaved for 20 minutes.

**DNA extraction, amplification, and sequencing**

DNA was extracted from each sample as described elsewhere [6,44] with the following modifications. Samples were subjected to two extractions in Matrix E bead tubes (MP Biomedicals, Solon, OH) containing 5% CTAB in 1 M NaCl and 0.25 M phosphate buffer (pH 8), phenol: chloroform: isopropyl alcohol (25:24:1), and 0.1 M ammonium aluminum sulfate, followed by separation with 24:1 chloroform: isoamyl alcohol. DNA was then precipitated with 30% polyethylene glycol 6000 in 1.6 M NaCl, washed with
70% ethanol, and resuspended in 10 mM Tris, pH 8. Replicate extractions were pooled and purified using MoBio PowerClean DNA clean-up kit (MoBio, Carlsbad, CA) and quantified using Quant-iT PicoGreen assay (Invitrogen, Carlsbad, CA).

Amplification of the V4 region of the 16S rRNA gene and attachment of Illumina adaptors and barcodes for multiplexing samples (reverse read only) was done in triplicate as described elsewhere [45]. Briefly, genomic DNA was amplified with universal forward 515F (5’- Illumina adapter- Forward primer pad- Forward primer linker-GTGCCAGCMGCCGCGGTAA-3’) and universal reverse 806R (5’- Illumina adapter-Golay barcode- Reverse primer pad- Reverse primer linker-GGACTACHVGGGTWTCTAAT-3’). Each PCR reaction mixture contained 10 ng of genomic DNA, 2.5 ul 10X buffer, 2.0 ul MgCl₂ (25 mM), 2.0 ul dNTP (2.5 mM each), 5.0 mM (each) forward and reverse primers, 1.25 ul (25 ug) BSA (Roche, Indianapolis, IN), 0.25 ul (1.25 U) Ex Taq (TaKaRa, Japan), and molecular grade water to reach a volume of 25 ul. PCR was performed with 3 min of initial denaturation at 94 °C followed by 30 cycles of the following program (denaturation, 94 °C for 45 sec; annealing, 50 °C for 30 sec; and extension, 72 °C for 45 sec) followed by a final extension at 72 °C for 7 min. PCR products were cleaned using Qiagen MinElute kit (Qiagen, Valencia, CA) as directed and quantified using the Quant-iT PicoGreen assay (Invitrogen, Carlsbad, CA). Equal concentrations were pooled and sequenced using the Illumina MiSeq platform at the Dana Farber Cancer Institute, Molecular Biology Core Facilities (Cambridge, MA).
2.3.3 Sequence analysis

Sequences were demultiplexed and trimmed of bar codes and primer sequences, then filtered for quality. Reverse and forward reads were assembled into contigs using default parameters in FLASH [46]. Sequence processing was done in QIIME using the following workflow: Reads were aligned using default parameters (PyNAST) [47], operational taxonomic units (OTUs) were picked at the 97% similarity threshold using QIIME [48] default parameters. Chimeric sequences were identified using Chimera Slayer [49] and removed, taxonomic assignments were made against the most recent Greengenes database (October, 2012) [50]. Rarefaction curves, alpha diversity (Shannon and Simpson), richness (observed species and Chao1) and Good’s coverage were calculated using default parameters in QIIME [48]. The Shapiro-Wilk test was used to confirm the normal distribution of alpha diversity values, and ANOVA was used to indicate differences between samples and across transfer points using the statistical package R [51].

Beta diversity estimates (Bray-Curtis and weighted Unifrac) were made at a sampling depth of 40,000 in QIIME [48]. Ordination of the Bray-Curtis and weighted Unifrac matrices was done using non-metric multidimensional scaling (NMDS) and principle components analysis (PCoA) respectively. 2-D plots were visualized in R [51]. Sequence data has been submitted to the NCBI Sequence Read Archive (SRA), Accession number: SRP029387.
2.3.4 Metabolic outputs

Concentrations of acetate, ethanol, lactate, formate, propionate, and butyrate were measured from centrifuged, filtered culture supernatant by HPLC using a BioRad Aminex HPX-87H column (Bio-Rad, Hercules, CA) with 0.005 M H₂SO₄ as the running buffer.

2.3.5 Isolation of microcosm bacteria

Eight strains were isolated from microcosms at Transfer 43. Prior to DNA extraction, culture purity was confirmed by microscopy and taxonomy assigned by amplifying the 16S rRNA gene using universal bacterial primers: 27F (5’-AGAGTTTGATCCTGGCTCAG-3’) and 1492R (5’-ACGGCTACCTTGTTACGACTT-3’) (Integrated DNA Technology, Coralville, IA). Using the taq PCR Core Kit (Qiagen, Valencia, CA), each PCR reaction contained a single colony or 2.0 ul of culture, 5.0 ul 10X buffer, 2.0 ul MgCl₂ (25 mM), 1.0 ul dNTP (2.5 mM each), 5.0 mM (each) forward and reverse primers, 0.25 ul Taq polymerase and molecular grade water to reach a volume of 50.0 ul. PCR was performed with 5 min of initial denaturation at 95 °C followed by 30 cycles of the following program (denaturation, 94 °C for 1 min; annealing, 53 °C for 1 min; and extension, 72 °C for 2 min) followed by a final extension at 72 °C for 10 min. PCR products were cleaned using Qiagen PCR Purification kit (Qiagen, Valencia, CA) as directed and sequenced at the University of Massachusetts Genomics Resource Lab (Amherst, MA) using an ABI 3130XL Sequencer (Beckman Coulter, Danvers, MA). General taxonomy assignments based on full length 16S rRNA sequences were done using the Ribosomal Database Project (RDP) classifier tool [52,53].
Finer level taxonomic assignments were made by constructing neighbor-joining phylogenetic trees in MEGA [54] using 16S rRNA gene sequences from the isolates and all type species from the order Clostridiales downloaded from RDP [52].

2.3.6 Deep sequencing of pooled isolates

A sample containing equal concentrations of genomic DNA (5 ng/ul) from eight bacteria isolated from the microcosm cultures was amplified and sequenced using the Illumina Miseq platform. Sequence data was analyzed as described above. Local BLASTN [55] was used to run full-length isolate 16S rRNA sequences against a database of the representative microcosm sequences created by QIIME [48] during OTU picking to determine which OTU mapped to each isolate.

2.4 Results

2.4.1 Metabolic outputs were consistent between replicates

Measurements of fermentation products (propionate, ethanol, butyrate, acetate, formate, and lactate) at fourteen time points via HPLC revealed that levels of lactate, formate, and propionate dropped below detectable levels after Transfer 20. Concomitantly, the levels of acetate and butyrate increased. There was a leveling off in the production of all metabolites after Transfer 30, with increasing variability between replicates in the production of acetate and butyrate over time. (Fig. 1).

2.4.2 Sequencing metrics of microcosm samples

As shown in Table 1, there were over 4.5 million sequences and 99,558 OTUs in the dataset following quality filtering and initial OTU picking in QIIME [48]. Chimera
removal using Chimera Slayer [49] left more than 4.2 million sequences and 92,346 OTUs for further analysis. Average read length was 254 bp.

Figure 1. Short chain fatty acid metabolites in microcosms over time. Concentration (mM) of acetate, butyrate, lactate, and propionate measured by HPLC at thirteen transfer points. Error bars show standard deviation of five samples.
Table 1. Sequence metrics of five microcosm samples at nine transfer points.

<table>
<thead>
<tr>
<th></th>
<th>Before Chimera Check</th>
<th>After Chimera Check</th>
<th>Chimera Checked &gt; 1.5% Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of sequences</td>
<td>4,554,420</td>
<td>4,204,943</td>
<td>3,328,037</td>
</tr>
<tr>
<td>Sequences per sample (mean)</td>
<td>114,734.5</td>
<td>105,123.575</td>
<td>83,200.925</td>
</tr>
<tr>
<td>Sequences per sample (min)</td>
<td>49,400</td>
<td>45,989</td>
<td>37</td>
</tr>
<tr>
<td>Number of OTUs</td>
<td>99,558</td>
<td>92,346</td>
<td>18</td>
</tr>
</tbody>
</table>

Total number of sequences, sequences per sample, and OTUs before and after chimera detection using Chimera Slayer[49].

Figure 2. Rarefaction curves of microcosm sequencing data.
Sequences per sample vs. observed species at a sampling depth of 40,000 sequences calculated and plotted in QIIME [48].
Rarefaction curves of the number of sequences per sample by observed species (Fig. 2) indicated a slowing down of new species at the minimum sequence depth of 40,000, but not a clear leveling off. The rarefaction curves were based at a low sampling depth due to one sample with a low number of sequences. Since all other samples had over twice this depth, and Good’s Coverage exceeds 94% for all samples we are confident that our sequencing efforts have captured most of the diversity in these samples.

2.4.3 Diversity is reduced early, and then stabilizes over time

Estimators of observed species (Fig. 3a), and Chao1 (Fig. 3b) indicated that species richness fell dramatically between Time 0 and Transfer 2, and then leveled off. Alpha diversity as estimated by Shannon (Fig. 4a) and Simpson Indices (Fig. 4b) showed a similar pattern, with the highest diversity at Time 0, followed by a sharp decline and leveling off. Based on ANOVA tests ($\alpha=0.05$) there were no statistical differences between replicates or across transfers after Time 0 for all four diversity measures.

Figure 3. Richness of microcosms over time.
Observed species (a) and Chao1 (b) calculated in QIIME [48]. Error bars show standard deviation of five samples.
Figure 4. Alpha diversity in microcosms over time.
Shannon Index (a) and Simpson Index (b) calculated in QIIME [48]. Error bars show standard deviation of five samples.

Beta diversity plots showed that after Time 0 there was no clear pattern in distance between replicates or across transfers as shown by weighted Unifrac (Fig. 5a and b), although there were two loose clusters dividing the earlier transfers (below 20) from the later transfers (30 and above). Bray-Curtis dissimilarity plots showed the same pattern (not shown).

Figure 5. Beta diversity of microcosms by sample and transfer.
Principle component (PCoA) plots of weighted Unifrac analysis colored by Sample (a) and Transfer (b). The circled region highlights Transfers 2-20.
2.4.4 Simplified community structure is consistent between replicates across transfers

The microbial community of the original sample was dominated by three major phyla: *Acidobacteria*, *Proteobacteria*, and *Verrucomicrobia*, with a total of 10 phyla at an abundance of greater than 1.5%, comprising 93% of the sequences (Fig. 6). The *Firmicutes* at Time 0 were at an abundance of 0.47%. There was an immediate and dramatic reduction in the number of phyla across replicates from Time 0 to Transfer 2 that was sustained throughout the time course, with only the *Firmicutes* and the *Proteobacteria* present at abundances greater than 1.5% at and beyond Transfer 2.

![Figure 6. Distribution of microcosm sequences by phyla.](image)
Relative abundance of sequences for each operational taxonomic unit found at or greater than 1.5% at each transfer point by phyla.
2.4.5 **Taxa level community composition shifts dramatically early, then stabilizes**

Consistent with the dramatic changes seen in phyla level abundance, there were enormous shifts in the identity and distribution of OTUs between Time 0 and Transfer 2. At Time 0 there were twelve OTUs with abundance greater than 1.5%, all belonging to the *Acidobacteria, Proteobacteria*, and *Verrucomicrobia* phyla. Together these OTUs comprised only 33-42% of the total sequence abundance at Time 0, with the majority of OTUs found at less than 1.5% abundance each.

Of the major phyla present at Time 0, only the Proteobacteria were present above 1.5% at Transfer 2, however the identity of the OTUs changed dramatically (Fig. 7).

*Figure 7. Distribution of Proteobacteria sequences by OTU.*
Relative abundance of Proteobacteria sequences for each operational taxonomic unit found at or greater than 1.5% at each transfer point.
The two Proteobacteria OTUs found at Time 0 in abundances greater than 1.5% (Gamma1 and Gamma2) did not persist in Transfer 2 and beyond, and the three OTUs from the Proteobacteria that were abundant after Transfer 2 were not abundant at Time 0. These three OTUs, members of the Enterobacteriaceae, persisted until Transfer 20 and then declined (Fig. 7 and 8). None of the abundant OTUs from Time 0 persisted at Transfer 2 at greater than 1.5% abundance. Additionally, none of the eight OTUs at 1.5% abundance or greater at Transfer 2 were found at greater than 0.09% in the Time 0 dataset.

Figure 8. Distribution of microcosm sequences by OTU, across all transfers. Relative abundance of sequences for each operational taxonomic unit found at or greater than 1.5% at each transfer point. Data is missing for Transfer 55, replicates 1 and 2.
Beyond Transfer 2, eighteen OTUs from five families were found at abundances greater than 1.5% (Fig. 8) including the Enterobacteriaceae discussed above. The four other families that contained abundant OTUs were the Enterococcaceae, Lachnospiraceae, Clostridiaceae, and Ruminococcaceae (Fig. 8). The phylogenetic relationships of these OTUs to selected type species is shown (Appendix A). By Transfer 62, ten OTUs were present in the sequence data at abundances greater than 1.5%, three of which were common to all five replicates (Fig. 9). Two OTUs, Entcoc1 and Lachno6 were found at greater than 1.5% abundance in every replicate and time point after Time 0. Entcoc1, Lachno6, and Lachno4 comprised between 48-67% of the total sequence abundance in all

Figure 9. Distribution of shared sequences in all replicates at Transfer 62.
Relative abundance of three most abundant operational taxonomic units (OTUs) shared between all samples as well as all shared OTUs for Transfer 62.
replicates at Transfer 62 (Fig. 9). The five replicates at this time point were remarkably similar, sharing 265 OTUs that together comprised 83-93% of the reads (Fig. 9).

2.4.6 Sequencing metrics of eight pooled isolates

The number of total OTUs even at Transfer 62 was surprisingly high (21,255). To test the ability of our sequencing and analyses strategies to accurately assess OTU assignments, genomic DNA was pooled from the eight isolated strains and the resulting mixture was sequenced using the Illumina MiSeq platform as above for the microcosms. As shown in Table 2, there were over 334,000 sequences and 5,900 OTUs in the dataset following quality filtering and initial OTU picking in QIIME [48]. Average read length was 254 bp. Chimera removal using Chimera Slayer [49] removed approximately 1,000 sequences, leaving 4,131 OTUs at a similarity threshold of 97% for further analysis.

Table 2. Sequence metrics of sequencing quality assessment.

<table>
<thead>
<tr>
<th></th>
<th>Before Chimera Check</th>
<th>After Chimera Check</th>
<th>Isolate Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of</td>
<td>334,881</td>
<td>333,086</td>
<td>305,356</td>
</tr>
<tr>
<td>sequences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of OTUs</td>
<td>5,918</td>
<td>4,131</td>
<td>8</td>
</tr>
</tbody>
</table>

Total number of sequences and OTUs before and after chimera detection using Chimera Slayer [49].

The sequence identities of the eight, pooled isolates were matched against the OTU sequences in the dataset using local BLASTN to verify recovery of the known sequences. Percent identity to each OTU representative sequence ranged from 97% to 100%, and the coverage was 100%. Phylogenetic affinities between the isolate sequences, OTU representatives, and selected type species verify of relationships (Appendix A).
The eight isolates corresponded to the top eight OTUs and represented 92% of the total reads. The remaining reads all classified as Firmicutes.

2.5 Discussion

This research suggests that for plant degrading anaerobic bacteria from soil, a dramatic change in microbial community composition occurs during the initial process of culture enrichment and serial transfers, but this drastic change is followed by relative stability, in a process that is highly replicable. Additionally, this strategy enables the discovery of novel microbes in a community context that might otherwise evade detection due to low abundance in the natural environment.

2.5.1 Dramatic change early during the time course

The soil sample used to inoculate the microcosm cultures at Time 0 was comprised of three abundant phyla (Acidobacteria, Proteobacteria, and Verrucomicrobia), with over half of the sequences spread between several less abundant phyla, notably the Actinobacteria, Bacteroidetes, Planctomycetes, and Chloroflexi). At Time 0 in this study, as in many other soil surveys, the Firmicutes was a relatively rare group. The distribution of bacterial phyla found at Time 0 in this study mirrors reports from 16S rRNA surveys for a variety of soil samples [56–58].

The rapid and dramatic change in bacterial community structure we observed between Time 0 and Transfer 2 has been observed in other microcosm experiments [59,60]. All eighteen OTUs present at abundances of greater than 1.5% after Transfer 2 were present,
but extremely rare in Time 0, and the Time 0 and Transfer 2 cultures had no OTU in common at abundances greater than 1.5%.

Further changes in abundance, notably the rise and decline of Enterobacteriaceae taxa, occurred over a period of 30 transfers, indicating that the community that emerged from the initial shift was resistant to further change. By Transfer 62, 83-93% of the total sequence abundance was shared between all samples. Three OTUs, Lachno4, Entcoc1 and Lachno6 comprised between 48-67% of these sequences.

2.5.2 Stability of microcosm communities over time

Throughout the time course of this experiment, four bacterial families, Enterococcaceae, Lachnospiraceae, Clostridiaceae, and Ruminococcaceae were abundant and persistent in all samples (Fig. 7). There were 21,256 OTUs at Transfer 62, however only 265 were common across all five replicates. This suggests the possibility that many OTUs are spurious, as many of the OTUs that are not shared are at low abundance. We thus sequenced a pooled sample of eight isolates from the microcosms in order to assess our ability to distinguish real OTUs from artifacts.

We were able to recover the eight isolate sequences from the 5,900 called OTUs in the sequencing assessment dataset at abundances greater than 1.5%. The relative abundances of the eight isolates were not equal indicating primer/ PCR bias, differences in DNA recovery and/or variation in rRNA gene copy numbers (S5). Errors in PCR and sequencing can generate artifacts resulting in overestimation of diversity [61,62].
Differentiating real sequences from artifacts presents a challenge for researchers seeking to estimate OTU abundance and diversity accurately, and is an active topic of discussion in the field of microbial ecology.

Following published recommendations [61], we used a 97% threshold in our OTU picking method and focused our analysis on more abundant OTUs, and those that were shared between samples and over time. Despite the risk of ignoring important rare taxa, we chose a more conservative abundance cutoff (1.5%) based on the resolution of our sequence assessment experiment. As sequencing technology improves it will be possible to determine the number of OTUs in a complex sample with greater accuracy and confidence.

2.5.3 Reproducibility

Our microcosm enrichments produced simplified, replicate communities from a complex environmental sample. A similar result was found from adapted consortia from tropical soil also using switchgrass as a carbon source in which a few dominant Firmicutes OTUs and an abundance of rare taxa emerged after 8 weeks of incubation [63]. Stable, replicable consortia of plant degraders have also been produced in intentional communities of up to eight species [42]. We know of no study that demonstrates community stability and replicability over a time course as long as three years, and suggest that these results demonstrate the robustness of this strategy for the long term cultivation of bacterial communities in industrial contexts such as bioreactors or probiotic production. Since the microcosm experimental design in no way models...
conditions in the natural soil environment, we make no speculations about whether the observed patterns of diversity reflect ecologically relevant trends.

2.5.4 Potential functional redundancy among the four Firmicutes families

While the four bacterial families, Enterococcaceae, Lachnospiraceae, Clostridiaceae, and Ruminococcaceae were abundant and persistent in all samples the distribution of OTUs within each family differed by replicate (Fig. 8). We suggest that these closely related taxa may be fulfilling similar ecological roles in the microcosm environment. While it is unclear whether these groups form stable associations in nature, we hypothesize that they were active in the utilization of plant material in our experimental system, and served roles similar to those described in previous studies [2,64,65]. The Lachnospiraceae, Ruminococcaceae and some Clostridiaceae are well known gut symbionts that are specialized to degrade complex plant material [66–68]. We suspect that taxa from these groups, specifically OTUs closely related to known cellulose degraders *Clostridium phytofermentans*, *Clostridium sporosphaeroides*, and *Clostridium celerecrescens* (Appendix A), to be involved in solubilizing complex plant polysaccharides to oligosaccharides, making them available for noncellulolytic, fermentative taxa such as those related to *Clostridium xylanovorans* and *Enterococcus sp.* In addition to being common gut symbionts, Enterococcaceae are ubiquitous in soils, freshwater and marine sediments, watersheds, and vegetation (reviewed in [69]). They are capable of fermenting a wide variety of carbon sources producing short chain fatty acids such as acetate, lactate, and butyrate, but they are not cellulolytic [69,70]. There is an extensive body of literature describing the pathogenicity of *Enterococcus sp.* [70] and
their utility as indicators of fecal contamination in drinking water [69,71]. 

Enterococcus is used in cocultures, producing bacteriocins in the production of fermentation products such as sausages [72] and cheese [73], but relatively little is known about their ecology and how they interact with other taxa in the environment.

We are especially intrigued by the persistent co-occurrence of OTUs Entcoc1, Lachno4, and Lachno6, as these are closely related to bacteria that are commonly found in the gut. Given the high abundance of the Entcoc1, an OTU related to a group of well described lactic acid bacteria, we were surprised to detect low amounts of lactic acid in the culture supernatant. We suggest that lactate utilizers in the community were using this resource to produce butyrate and acetate as has been described in gut environments [74,75]. Elucidating the roles and interactions between members of these and other members of our microcosm communities in the degradation of plant material is a subject of continued research.

2.6 Conclusions

Enrichment and serial transfers have been traditionally used to aid in the isolation of bacteria from environmental samples, but the specific changes communities undergo, the stability of communities across transfers, and the replicability of such strategies over a long time course have not been described. This research suggests that enrichment and serial transfers do not generate communities that are strongly divergent in terms of bacterial community diversity and membership or metabolic outputs. While community membership over time may differ slightly between replicate samples, the differences
involve closely related species, thus reflect functional redundancy and not divergent communities or metabolic activity. Additionally, this strategy represents an effective way to detect and select for novel taxa that might be missed in shorter-term enrichments and isolation protocols from complex communities. These have implications for the maintenance of long-term stable communities in bioreactors, and other industrial processes utilizing mixed communities.
3.1 Abstract

Laminitis is a chronic, crippling disease triggered by the sudden influx of dietary starch. Starch reaches the hindgut resulting in enrichment of lactic acid bacteria, lactate accumulation, and acidification of the gut contents. Bacterial products enter the bloodstream and precipitate systemic inflammation. Hindgut lactate levels are normally low because specific bacterial groups convert lactate to short chain fatty acids. Why this mechanism fails when lactate levels rapidly rise, and why some hindgut communities can recover is unknown.

Fecal samples from three adult horses eating identical diets provided bacterial communities for this in vitro study. Triplicate microcosms of fecal slurries were enriched with lactate and/or starch. Metabolic products (short chain fatty acids, headspace gases, and hydrogen sulfide) were measured and microbial community compositions determined using Illumina 16S rRNA sequencing over 12-hour intervals.

We report that patterns of change in short chain fatty acid levels and pH in our in vitro system are similar to those seen in in vivo laminitis induction models. Community differences between microcosms with disparate abilities to clear excess lactate suggest profiles conferring resistance of starch-induction conditions. Where lactate levels recover following starch induction conditions, propionate and acetate levels rise correspondingly
and taxa related to *Megasphaera elsdenii* reach levels exceeding 70% relative abundance. In lactate and control cultures, taxa related to *Veillonella montpellierensis* are enriched as lactate levels fall. Understanding these community differences and factors promoting the growth of specific lactate utilizing taxa may be useful to prevent acidosis under starch-induction conditions.

### 3.2 Introduction

#### 3.2.1 The equine digestive tract

Horses are hindgut fermenters, adapted to grazing continually on marginal forages that change seasonally, thus slowly [76]. The hindgut (caecum and colon) comprises roughly two thirds of the volume of the equine digestive tract [77]. Here complex plant material is fermented by microbes to short chain fatty acids (SCFA) such as acetate, propionate, and butyrate, which provide 60-70% of the daily energy needs of the horse [22,78]. Rapid dietary change and modern feeding practices of 2-3 meals a day of starch-based concentrate and/or fructans from rich pasture can disrupt normal fermentation in the hindgut, causing lactic acidosis, and colic [79–82], and predisposing animals to bouts of laminitis [83–85].

#### 3.2.2 Laminitis and lactic acidosis

Laminitis is a chronic, crippling disease, accounting for 15% of all lameness in horses in the United States, with over 27% unable to return to normal work, and 4.7% mortality [86]. It is characterized by weakened adhesion and eventual detachment of the distal phalynx from the lamellae of the inner hoof wall resulting in permanent rotation of the
coffin bone and severe pain. Factors released into the bloodstream by bacteria in the gut during lactic acidosis are thought to serve as triggers for dietary laminitis [87], however the molecular mechanisms underlying induction are unknown.

Surveys of equine hindgut bacteria using culture based methods [88] and 16S rRNA gene sequencing [31] have detected a diverse community of novel microbes dominated by Firmicutes, with Bacteroidetes, Proteobacteria, and Verrucomicrobia as other major phyla. Studies have detected a greater proportion of fibrolytic bacteria than starch and lactate utilizing bacteria in the cecum than in the colon [89], reflecting a substrate content normally low in starch and soluble carbohydrates due to the action of endogenous enzymes and absorption of nutrients in the small intestine.

Experimental in vivo models of laminitis induction using starch gruel or oligosaccharide [90] have revealed changes in hindgut microbiota during the developmental stage (24-36 hours post induction), correlated with a drop in pH and an increase in lactic acid concentration [91]. Lactic acid bacteria, specifically members of the Streptococcus bovis/equinus group (now renamed Streptococcus lutetiensis [92]), have been implicated as major producers of lactic acid, rapidly increasing in numbers as lactic acid levels rise and the pH drops. At the lowest caecal pH (4-4.5) and levels of lactic acid reaching 1000 umol/g caecal fluid, acid sensitive fibrolytic and gram negative bacteria die off, while Lactobacilli sp. and Mitzuokella sp. increase [91]. By 32-36 hr, hindgut lactate levels and pH approach normal levels in most horses [91]. In other experiments of starch induction,
blood D-lactate levels peaked at 20-24 hr, then declined and disappeared by 36-40 hr [93].

3.2.3 Lactate utilizing bacteria

Lactate levels in the hindgut are normally low due to the activity of lactate utilizing bacteria. It is unclear why this mechanism fails during conditions of starch induction. While studies of lactate producers have pointed to specific taxa that proliferate during the developmental stage of laminitis [91], little is known about how the abundance of lactate utilizing bacteria changes over the same time course, which lactate utilizers survive the drop in pH, and which lactate utilizers are active in the later stages to bring lactate concentrations back to normal levels.

3.2.4 Experimental summary

In this study we used fecal samples collected from 3 healthy, adult horses eating an identical pasture based diet in an \textit{in vitro} model system to track bacterial metabolites and community shifts over time in response to enrichment with starch and/or lactate. Patterns of changes in lactate concentration and pH were similar to those reported in published \textit{in vivo} studies [79]. Illumina 16s rRNA amplicon sequencing was used to track changes in hindgut microbiota over the time course, specifically identifying lactate utilizing taxa and bacterial groups that proliferated as lactate levels drop. Additionally, we identified community differences in cultures lacking the ability to clear excess lactate, which may lead to further insight into why some horses are resistant to starch induction, and point to bacteria with the potential to attenuate or prevent lactic acidosis.
3.3 Materials and Methods

3.3.1 Sampling and in vitro enrichments

Fecal samples were manually collected from the midrectum of three Morgan geldings cohoused at the University of Massachusetts Hadley Farm and fed identical hay based diets. None of the animals had received antibiotics, anthelmintics, or other medications for at least three months prior to sampling. Samples were protected from oxygen exposure during collection by inverting the glove around each as it was removed, and placing each immediately in a container evacuated of oxygen. Samples were kept as close to 39°C as possible in a heated, insulated container during transport and the following steps of dilution and inoculation. In an anaerobic chamber, samples were diluted to make a 10% slurry in anaerobic dilution media prepared as described by Bryant [94]. This slurry was used to inoculate basal media described by de Carvalho [95] at 2.5% which was enriched with either 1% soluble starch, 50 mM sodium lactate, or both 1% soluble starch and 50 mM sodium lactate. Control cultures were diluted at 2.5% of basal media with no enrichment. Cultures were kept under anaerobic conditions, incubated at 39°C and sampled every 12 hours for metabolite analysis and DNA extraction. Sampling and enrichments were subsequently repeated for the same horses, inoculant concentrations, culture and incubation conditions in 125 ml serum bottles (Wheaton, Millville, NJ) fitted with stoppers that enable headspace gas collection and analysis.

3.3.2 Metabolite measurements

Short chain fatty acids (acetate, lactate, butyrate, succinate, formate, and propionate) were measured for samples taken over the time course using high performance liquid
chromatography (HPLC) (Shimadzu, Japan) with an Aminex HPX-87H column (Bio-
Rad, Hercules, CA). Headspace gases were analyzed using a gas chromatograph
(Shimadzu GC-8A, Shimadzu, Japan) fitted with a HayeSep DB column 100/120
(Bandera, TX). Hydrogen sulfide levels were measured using the methylene blue assay as
in Cline [96] modified for culture samples as follows: at each time point, 1.0 ml of
culture was removed from each serum bottle via syringe, transferred into sealed vials
containing an equal volume of 1.2% degassed zinc acetate solution, and stored at 4°C
until all samples were collected. To a 1.0 ml subsample, 62.5 ul of 7% sodium hydroxide
was added. Following a 15 min incubation at room temperature, 187.5 ul of 0.1% N,N’-
dimethyl-p-phenylenediamine and 187.5 ul 10 mM iron (III) chloride were added,
stoppered immediately, and incubated for 20 minutes at room temperature. The resulting
suspension was spun down and the absorbance of the supernatant was measured at 670
nm in comparison to standard solutions (0 - .55mM sodium sulfide, and uninoculated
media controls. pH was determined using EMD colorphast pH strips (Fischer Scientific,
Pittsburgh, PA).

3.3.3 DNA extraction, amplification, and sequencing

DNA was extracted from each sample as described elsewhere [44] with the following
modifications. Samples were subjected to two extractions in Matrix E bead tubes
containing 5% CTAB in 1 M NaCl and .25M phosphate buffer (pH 8), phenol:
chloroform: isopropyl alcohol (25:24:1), and 0.1 M ammonium aluminium sulfate,
followed by separation with 24:1 chloroform: isoamyl alcohol. DNA was then
precipitated with 30% polyethylene glycol 6000 in 1.6 M NaCl, washed with 70%
ethanol, and resuspended in 10 mM Tris, pH 8. Replicate extractions were pooled and purified using MoBio PowerClean DNA clean-up kit (MoBio, Carlsbad, Ca) and quantified using Quant-iT PicoGreen assay (Invitrogen, Carlsbad, Ca).

Amplification of the V4 region of the 16S rRNA gene and attachment of Illumina adaptors and barcodes for multiplexing samples (reverse read only) was done in triplicate as described elsewhere [45]. Briefly, genomic DNA was amplified with universal forward 515F (5’- Illumina adapter- Forward primer pad- Forward primer linker-GTGCCAGCMGCCGCGGTAA-3’) and universal reverse 806R (5’- Illumina adapter-Golay barcode- Reverse primer pad- Reverse primer linker-GGACTACHVGGGTWTCTAAT-3’). Each PCR reaction mixture contained 10 ng of genomic DNA, 2.5 ul 10X buffer, 2.0 ul MgCl$_2$ (25 mM), 2.0 ul dNTP (2.5 mM each), 5.0 mM (each) forward and reverse primers, 1.25 ul (25 ug) BSA (Roche, Indianapolis, IN), 0.25 ul (1.25 U) Ex Taq (TaKaRa, Japan), and molecular grade water to reach a volume of 25 ul. PCR was performed with 3 min of initial denaturation at 94°C followed by 30 cycles of the following program (denaturation, 94 °C for 45 sec; annealing, 50 °C for 30 sec; and extension, 72 °C for 45 sec) followed by a final extension at 72 °C for 7 min. PCR products were cleaned using Qiagen MinElute kit (Qiagen, Valencia, Ca) as directed and quantified using the Quant-iT PicoGreen assay (Invitrogen, Carlsbad, Ca). Equal concentrations were pooled and sequenced using the Illumina MiSeq platform at the Dana Farber Cancer Institute, Molecular Biology Core Facilities (Cambridge, MA).

3.3.4 Sequence analysis

Sequences were demultiplexed and trimmed of bar codes and primer sequences using
FastQC [97], then filtered for quality and reverse and forward reads were assembled into contigs using FLASH [46]. Sequence processing was done in QIIME [48] using the following workflow: Reads were aligned using default parameters (PyNAST) [47], operational taxonomic units (otus) were picked at the 97% similarity threshold using the subsampled open-reference option, chimeric sequences were identified using ChimeraSlayer [49] and removed, taxonomic assignments were made against the most recent greengenes database (October, 2012) [50]. Sequence data has been submitted to the NCBI Sequence Read Archive (SRA), Accession number: SRP028582.

3.4 Results

3.4.1 Patterns of changes in metabolites

Despite variation in the pH response between the three horses, in all of the starch and starch/lactate enrichments, the pH dropped to below 6 by hour 12, and reached levels between 4 and 5 by hour 18 (Fig. 10). These values paralleled the peak in lactate levels for these cultures over the same time course (Fig. 11). The control and lactate enriched treatment groups showed an initial increase in lactate followed by a rapid decline as the lactate was used up. In the starch enriched cultures, lactate levels peaked by hour 18, exceeding 100 mM in cultures enriched with both starch and lactate. Where lactate levels dropped over time, there was a corresponding increase in acetate, propionate, and butyrate.

The three horse cultures differed dramatically in the ability to attenuate accumulated lactate. The lactate was reduced to below detectable limits in the control and lactate
Figure 10. pH of cultures over time by horse.
P\(\text{H}\) of cultures from each horse and culture condition measured at 6 hour intervals.

Enriched treatment groups of all three horse cultures, however the peak was higher for horse 3 cultures and took longer to drop. In the starch and starch/lactate enrichments the differences were more dramatic. Lactate persisted at maximum levels in horse 3 cultures over the full time course while dropping for horse 1 and 2 cultures by hour 36.

Headspace gases (hydrogen and methane) and hydrogen sulfide levels measured at 12 hour intervals did not show consistent differences between treatment and control conditions due to variation between horses especially for the starch and starch/lactate cultures (Appendix B and C).
Figure 11. Short chain fatty acid metabolites over time by horse.
Concentration (mM) of acetate, butyrate, lactate, and propionate measured by high performance liquid chromatography from each horse and culture condition at times 0, 6, 18, 36, and 48 hours.

3.4.2 Sequence metrics

As shown in Table 3, there were over 6 million sequences and 41,000 OTUs in the dataset following quality filtering and initial OTU picking in QIIME [48]. Chimera removal using Chimera Slayer [49] left more than 2.5 million sequences and 32,000 OTUs for further analysis. Average read length was 254 bp.

Rarefaction curves of the number of sequences per sample by observed species (Appendix D) indicated a leveling off in terms of new species at the minimum sequence
depth of 9685. Since 83% of the samples had over twice this depth, and Good’s coverage at the depth of the smallest library (9685) was 89% (Appendix E), we are confident that our sequencing efforts have captured most of the diversity in these samples.

Table 3. Sequence metrics of horse gut samples.

<table>
<thead>
<tr>
<th></th>
<th>Before Chimera Check</th>
<th>After Chimera Check</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of sequences</td>
<td>6,043,194</td>
<td>2,576,535</td>
</tr>
<tr>
<td>Sequences per sample (mean)</td>
<td>125,900</td>
<td>53,678</td>
</tr>
<tr>
<td>Sequences per sample (min)</td>
<td>51,890</td>
<td>9,685</td>
</tr>
<tr>
<td>Number of OTUs</td>
<td>41,055</td>
<td>32,563</td>
</tr>
</tbody>
</table>

Total number of sequences, sequences per sample, and OTUs before and after chimera detection using Chimera Slayer.

3.4.3 Relative abundance at the phylum level

Taxonomic assignments at the phylum level (Fig. 12) showed that the Firmicutes were the most abundant group by time 48 in all cultures regardless of enrichment conditions, with a corresponding decline in all other major groups, namely the Verrucomicrobia, Spirochaetes, Proteobacteria, and Bacteroidetes, however at this taxonomic level, clear differences between horse and treatment groups were not apparent.

3.4.4 Relative abundance at the family level for Firmicutes

Differences between horse cultures and treatment conditions were evident at the family level, specifically for the most abundant phyla, the Firmicutes. The Veillonellaceae, a family with known lactate utilizing species [98,99], increased in abundance in all cultures in the control and lactate treatment groups. In the starch and starch/lactate enrichments there were dramatic differences between horse 3 cultures and those of horse 1 and 2.
Figure 12. Distribution of horse gut sequences by phyla.
Relative abundance of sequences for each operational taxonomic unit found at or greater than 1% at each time point for each culture condition and horse sample by phyla.

(Fig. 13). By time 48, the Veillonellaceae dropped to less than 5% in horse 3 cultures, while in horse 1 and 2 cultures this group was the highest in abundance, making up more than 70% of the total sequences (Fig. 4). The most abundant family for the starch and starch/lactate horse 3 cultures at time 48 were the Lactobacillaceae, making up greater than 40% of the total sequences at this time point.

The Streptococcaceae reached a peak in abundance by 24 hours in the starch and starch/lactate conditions for all horse cultures, after which the abundance dropped. In horse 1 and 3 cultures this decrease was accompanied by an increase in the Lactobacillaceae,
Figure 13. Distribution of horse gut Firmicute sequences by family.
Relative abundance of Firmicutes sequences found at or greater than 1% by family at each time point for each culture condition and horse sample.

however in horse 2 the abundance of Lactobacillaceae remained relatively low even as the Streptococcaceae declined over time.

3.4.5 Distribution of abundant OTUs
Identification of specific taxa that were most abundant by hour 48 (Fig. 14 and Appendix F) suggested community differences between horse cultures of specific interest in light of differences in lactate utilization and attentuation. There was a striking difference in OTU abundance and distribution between horse 3 cultures and those of horse 1 and 2 in all treatment conditions at hour 48. While the most abundant sequence
Figure 14. Distribution of most abundant horse gut OTUs at time 48.
Relative abundance of OTUs found at greater than 5% in each culture condition by horse. OTUs are identified by best BLAST match, and percent sequence similarity is given.

(between 50-60% relative abundance, with 98% identity to *Veillonella montpellierensis*) under control and lactate conditions for horse 3 was also found in horse 1 and 2 cultures, it reached a lower abundance in both. While horse 3 control and lactate enriched cultures were able to attenuate lactate, it persisted in high concentrations in starch and starch/lactate enrichments in which no member of the Veillonellaceae was highly abundant.

A different OTU, related to *Megasphaera elsdenii* (100% identity) was the most abundant OTU in horse 1 and 2 cultures, especially in the starch and starch and lactate
enrichments, reaching relative abundances of over 70%. While this OTU was present in single numbers at Time 0 in all three horse samples, and persisted in horse 3 in low numbers under all culture conditions (Fig. 15), it never reached abundances of greater than 0.40% at any condition or time point.

**Figure 15. Change in most abundant horse gut Veillonellaceae OTUs over time.**
Relative abundance of most abundant Veillonellaceae OTUs over time in each culture condition by horse. OTUs are identified by best BLAST match, and percent sequence similarity is given.

The most abundant OTU in horse 3 starch and starch and lactate cultures was related to *Lactobacillus equi* (100% sequence identity) and represented 97.7% of the Lactobacillaceae sequences across the dataset. This OTU was present in all three horse
Figure 16. Change in single, abundant horse gut Lactobacillaceae OTU over time. Relative abundance of the single, dominant Lactobacillus OTU over time for each culture condition by horse.

cultures, reaching abundances of 20% and greater in horse 1 in starch and starch and lactate enrichments at time 36, but then falling to 10% or less by time 48 (Fig. 16).

Instead of being dominated by one or two abundant OTUs, the distribution of the Streptococcaceae taxa was more dispersed, with the two most abundant OTUs (related to Streptococcus equi and Streptococcus infantarius (with 100% and 99% sequence identities respectively) together accounting for less than 50% of the sequences in this group (Fig. 17).
Figure 17. Change in abundant horse gut Streptococcaceae OTUs over time. Relative abundance of dominant Streptococcus OTUs over time for each culture condition by horse. OTUs are identified by best BLAST match, and percent sequence similarity is given.

3.5 Discussion

3.5.1 Our in vitro system captures elements observed in vivo

The accumulation of lactate has long been recognized as an early event in the microbial response to carbohydrate overload leading to colic and laminitis in horses [79]. In the in vitro system described herein, we were able to simulate many key aspects of starch induced conditions that have been reported elsewhere for in vivo experiments. The extent and timing of pH changes (Fig.10) and fluctuations of SCFA levels (Fig. 11), especially
with respect to lactate, acetate, and propionate are similar to those reported for *in vivo*
starch induction studies [79,91].

No study thus far has tracked microbial changes over the time course of starch induction
in horses using 16S rRNA deep sequencing as described here, however, the changes that
we observed in community composition are similar to what has been reported in equine
and ruminant culture and probe based studies [81,100]. Specifically, in the first 24 hours
following enrichment, we noted an increase in Firmicutes, especially members of the
Streptococcaceae, coinciding with a decrease in fibrolytic groups (Ruminococcaceae and
Lachnospiraceae), followed by an increase in the Lactobacillaceae as the abundance of
Streptococcaceae falls.

Despite a sample size of three horses, we were able to observe between-horse
differences in the ability to attenuate accumulated lactate as has been described in
previous studies [79,91].

Certainly the microbial dynamics reported here may not reflect the actual community
compositions of the caecum and large intestine of horses who recover or resist conditions
of lactic acidosis, however the elucidation of endogenous lactate utilizers that thrive
under conditions of low pH following starch enrichment in our *in vitro* model may
provide insight into microbial mechanisms of resistance and/or recovery.
3.5.2 Lactate producing bacteria proliferate following starch induction

While we observed the pattern of increase in *Streptococcus* species during the first 24 hours of starch induction followed by an increase in *Lactobacilli* species noted in other studies [81,91] we expected the *Streptococcus lutetiensis* group (formerly *Streptococcus bovis* [92]) to be more highly represented in the starch and starch/lactate cultures, as it has been identified as the major lactate producer in *in vivo* starch induction studies [93,101]. Our data indicates a high abundance of the family Streptococcaceae in the starch and starch/lactate cultures, but fails to resolve subtle differences between species. We recognize that basing taxonomic assignments on a single region (V4) of the 16S gene is inherently limited [102,103] for all bacterial groups since 500-700 bases are recommended for species resolution [104], and suspect that taxa within the *Streptococcus lutetiensis* group may be especially difficult to distinguish due to their close phylogenetic relationships [92,105]. Sequencing with longer reads would clarify the species composition of this family in our starch and starch/lactate enrichments.

3.5.3 Members of the Veillonellaceae are the most abundant lactate utilizing bacteria

While studies have pointed to *Streptococcus lutetiensis* and *Lactobacillus* sp as major lactic acid producers [93], relatively little attention has been paid to bacteria in the horse gut that actively utilize lactic acid. Studies of lactic acidosis in ruminants [106,107] have identified specific genera, namely *Megasphaera*, *Veillonella*, *Selenomonas*, *Propionibacterium*, and *Anaerovibrio* as key lactate utilizers. In fact, strains of *Megasphaera elsdenii* have been shown to be effective in preventing lactic acidosis in cattle and are under development as probiotic therapies [108–110].
Using deep sequencing of the 16S rRNA gene of fecal communities challenged with starch, lactate, or both in an *in vitro* model of starch induction, we report here changes in the abundance of specific microbes associated with the reduction of lactic acid. As we observed at time 0 in this study, 16S rRNA surveys of the equine gut microbiome have shown that under normal conditions, the Veillonellaceae (known lactate utilizers) comprise 1% or less of the total bacterial abundance [31,111,112]. One probe-based *in vivo* study did not see a difference in the abundance of Veillonellaceae in response to dietary change despite an increase in lactate levels [101]. It is unclear why the Veillonellaceae in that study did not increase in abundance as lactate accumulated, or which specific taxa were present in those horses.

In our study we observed that one particular taxa closely related to *Megasphaera elsdenii* was highly abundant in all starch and starch/lactate cultures in which lactate accumulated and was attenuated. This taxa was present in very low abundance in cultures in which lactate persisted. It is unclear why this taxa fails to proliferate in any of the horse 3 cultures while reaching such high abundances under all conditions in horse 1 and 2 cultures. A second OTU related to *Veillonella montpellierensis* was highly abundant in control and lactate enrichments specifically in horse 2 and 3 cultures, but did not thrive in the starch or starch/lactate enrichments, suggesting that other factors such as pH or competitive interactions exert selective pressure under starch and starch/lactate enrichment conditions.
While our data indicates a relationship between the presence of the *Megasphaera elsdenii* OTU and the reduction of lactate, conclusive evidence that this taxa is responsible for reducing lactate levels will require further study. It is possible that other community members with lactate utilizing capabilities are playing active roles as well. Understanding the factors stimulating or preventing the proliferation of lactate utilizers in the horse gut microbiome could provide valuable information about why some horses are more sensitive to starch induction, and the microbial basis behind mechanisms of resistance.

### 3.6 Conclusion

A robust *in vitro* model for starch induced laminitis in horses as described here could provide a convenient and cost effective means to understand the microbial dynamics underlying colic and laminitis, and test hypotheses for ways to prevent or interrupt the progress of these equine diseases. Specific taxa in the family Veillonellaceae were highly abundant in starch-enriched cultures that were able to attenuate lactate. These could provide useful insights into mechanisms of recovery or resistance, and could be valuable, individually or in consortia, as probiotics to prevent starch induced colic and laminitis.
CHAPTER 4
UNTANGLING THE GENETIC BASIS OF FIBROLYTIC SPECIALIZATION
BY LACHNOSPIRACEAE AND RUMINOCOCCACEAE IN DIVERSE GUT
COMMUNITIES

4.1 Abstract
The Lachnospiraceae and Ruminococcaceae are two of the most abundant families from
the order Clostridiales found in the mammalian gut environment, and have been
associated with the maintenance of gut health. While they are both diverse groups, they
share a common role as active plant degraders. By comparing the genomes of the
Lachnospiraceae and Ruminococcaceae with the Clostridiaceae, a more commonly free-
living group, we identify key carbohydrate-active enzymes, sugar transport mechanisms,
and metabolic pathways that distinguish these two commensal groups as specialists for
the degradation of complex plant material.

4.2. Introduction

4.2.1 Taxonomic Revision of the Clostridiales is a Work in Progress
Classically, the genus Clostridium was described as comprising spore-forming, non-
sulfate reducing obligate anaerobic bacteria with a gram-positive cell wall. Over the
years, the classification and naming of new species based on phenotypic traits has led to
much confusion about the relationships between taxa in this and related groups. In fact,
many species classified as Clostridium are more closely related to members of other genera
than to the type species, Clostridium butyricum [25,113]. Prior to 2009, the order
Clostridiales was divided into eight families, many of which were recognized at the time
to be paraphyletic [114]. The most recent taxonomic revision of the Phylum Firmicutes in
Bergey’s Manual of Systematic Bacteriology [115] divided the Clostridiales into ten named families. An additional nine families were identified as Incertae sedis (Latin for uncertain placement) in an effort to regroup species found to fall outside of the named families.

In the past decades, surveys of 16S rRNA gene sequence diversity have led to the identification of thousands of novel taxa and a new appreciation for the overwhelming diversity of the microbial world. At the same time, knowledge about the roles of new species in their environments has remained scanty [116]. For diverse groups with confusing taxonomic structure, such as the clostridia, linking the phylogeny of novel, uncultured taxa to possible ecological and/or physiological roles requires extensive prior knowledge and time-consuming literature review.

4.2.2 Lachnospiraceae and Ruminococcaceae are active members of the gut environment

Abundance estimates based on 16S rRNA surveys suggest that Firmicutes comprise between 50–80% of the taxa in the core human gut microbiota [27,28], and more than 84% of the active fraction [29]. Lachnospiraceae and Ruminococcaceae are the most abundant Firmicute families in gut environments, accounting for roughly 50% and 30% of phylotypes respectively [27,30]. Lachnospiraceae such as *Eubacterium rectale, Eubacterium ventriosum, Coprococcus sp.* and *Roseburia sp.* have been associated with the production of butyrate necessary for the health of colonic epithelial tissue [117,118], and have been shown to be depleted in inflammatory bowel disease [119]. Unusual polysaccharide binding and degradation strategies have been described in *Ruminococcus flavefaciens*.
[64,120], while another Ruminococcaceae, *Faecalibacterium prausnitzii*, has been shown to be depleted in Crohn’s disease [121].

4.2.3 The complexity of plant material poses challenges for bacterial decomposition

Plant biomass is a fibrous composite of fibrils and sheets of cellulose, hemicellulose, lignin, waxes, pectin, and proteins forming a complex network that provides support for the plant while resisting attack from bacteria and fungi. Due to the size and complexity of the substrate, bacterial glycoside hydrolases (GH) are generally produced extracellularly. In anoxic environments such as the gut, bacteria utilize complexed, multienzyme catalytic systems found on the cell surface or in organelles called cellulosomes. These complexes are modular in design and often include one or more carbohydrate-binding modules (CBM) that attach to the substrate enabling easy access [2,65]. Surveys of plant degradation in the rumen show that bacteria that can degrade easily available substrates colonize plant material first, and that these communities are replaced by others capable of degrading more recalcitrant substrates such as cellulose [66]. Non-adherent *Bacteroides* sp. and *Bifidobacterium* sp. have been shown to outcompete gram-positive bacteria (such as Firmicutes) for easily hydrolysable starch [122,123], while Lachnospiraceae and Ruminococcaceae persist in fibrolytic communities and are uniquely suited to degrade a wide variety of recalcitrant substrates [66].
4.2.4 Genomic clues to fibrolytic function in gut environments

With progress of the Human Microbiome Project and other efforts to understand the complexity of microbial communities living on and inside of humans, an increasing number of sequenced genomes and datasets have been released [124]. Here we use genomic data to describe two families of Clostridiales that are highly abundant in the human gut microbiota, the Lachnospiraceae and the Ruminococcaceae. By comparing the distribution and abundance of carbohydrate-active enzymes and transporters, and the differences in key metabolic pathways present in the genomes of each group, we reveal genetic components supporting plant degradation by these fibrolytic specialists, and provide clues to help distinguish gut microbes from their primarily free-living relatives in the Clostridiaceae.

4.3 Materials and methods

4.3.1 Phylogenetic arrangement of Lachnospiraceae, Clostridiaceae, and Ruminococcaceae

Sixty-eight high quality (greater than 1,200 bases) 16S rRNA sequences of representative taxa from each family in the most recent taxonomic revision of the Phylum Firmicutes in Bergey’s Manual of Systematic Bacteriology [115] were downloaded from the Ribosomal Database Project website [52], and aligned using MUSCLE [125]. A maximum likelihood tree was created from these sequences with 500 bootstrap replicates using Escherichia/Shigella coli (ATCC 11775T; X80725) as the outgroup. Tree building was accomplished using the Tamura_Nei model [126] with default parameters in the program MEGA [54]. Tree visualization was carried out using Figtree [127].
4.3.2 Habitat association by group

Isolation site and habitat preference (gut vs. non-gut) was determined for listed members of each group from either the IMG (Integrated Microbial Genomes) metadata table [128] and published reports of isolation or 16S surveys. Pathogenic organisms were treated as non-gut residents if they were not reported to be part of the normal gut flora of a mammalian species. Logistic regression was performed to test whether each group was more likely to be gut associated, and the Wald test was used to test whether group assignment significantly predicted gut association.

4.3.3 Comparative analysis of carbohydrate-active enzymes

The numbers of carbohydrate-active genes and gene families were tabulated for each member of the Lachnospiraceae, Clostridiaceae, and Ruminococcaceae found in the Carbohydrate-Active enZymes Database (CAZy) [129] (Appendix G). GH and CBM families represented in greater than 50% of total taxa (more than 15 of 31) and a difference of greater than two times in average abundance between any two groups were chosen for further analysis. All statistical analyses were done in R [51] using the package MASS [130]. Gene counts for each family and individual genes were tested for normality using the Shapiro-Wilk test [131]. Due to overdispersion and zero values across the data set, we did not transform the data [132]. Instead, we compared the fits of the negative binomial and Poisson regression models for the gene counts for each family. Clostridiaceae was used as a reference. Significance was estimated from the model with the best fit as determined by the p-values reported for likelihood ratio tests comparing both models.
Genomes were further searched for enzymes from each significant GH and CBM family using the genome comparison tool in the Integrated Microbial Genomes system [128]. Enzymes with significantly different abundances per group were identified by the best fit between negative binomial and Poisson regression models as described above for each GH and CBM family. Enzyme functions were assigned using UniprotKB [133].

4.3.4 Comparative analysis of sugar transport genes

Genomes from each member of the Lachnospiraceae, Clostridiaceae, and Ruminococcaceae found in the IMG database [128] were compared in terms of numbers of carbohydrate transport genes. Specifically, genomes were searched for 63 PTS (phosphotransferase system) genes, and 137 ABC (Adenosine triphosphate-binding cassette) transporter genes identified by KEGG orthology [134]. Gene counts for each taxa were tabulated and the difference between groups for each category of transporter genes was tested using the best fit between negative binomial or Poisson regression models as determined by likelihood ratio test ($p$-value < 0.05) as described earlier. Individual genes were chosen for further analysis if they had an average number of copies per taxa of at least 1% in any group. Significant differences between counts for these genes between families were estimated as described above for each category.

4.3.5 Comparative analysis of metabolic pathways

Metabolic pathways characterized in Ecocyc [135] for each member of the Lachnospiraceae, Clostridiaceae, and Ruminococcaceae in the Biocyc database system [136] (Appendix G) were tabulated for each pathway class. Significantly different pathway classes
between the three families were identified using either one-way ANOVA (for data with a normal distribution based on the Shapiro-Wilks test) [131] or the best fit between negative binomial or Poisson regression models as described above.

For carbohydrate degradation pathways showing a difference between families, the percentage of genomes in each family containing the pathway was determined to highlight differences between families for these functions.

4.4 Results and Discussion

4.4.1 Phylogenetic arrangement of Lachnospiraceae, Clostridiaceae, and Ruminococcaceae

The topology of the 16S rRNA neighbor-joining tree for the representative taxa chosen for this study confirms the clustering of the Lachnospiraceae, Clostridiaceae, and Ruminococcaceae into distinct clades (Figure 18) with the exception of Clostridium sporosphaeroides DSM 1294, which is identified as a Lachnospiraceae in Bergey’s newest revision [115] yet clearly clusters with the Ruminococcaceae.
Figure 18. 16S Neighbor-joining tree of the representative taxa used in this analysis. Bootstrap values are given for 500 replicates. (blue) Lachnospiraceae, (red) Clostridiaceae, (green) Ruminococcaceae.
4.4.2 Habitat Association by Group

The logit model concluded that the three groups differ significantly in their likelihood to be gut associated (Figure 19). The Wald test predicted that the Ruminococcaceae and Lachnospiraceae were more likely than the Clostridiaceae to be gut associated ($X^2 = 17.5$, df = 2, P (> $X^2$) = 0.00016).

![Figure 19. Proportion of members of each family that are gut associated.](image)

The Clostridiaceae are less likely to be gut associated based on logistical regression model predictions (p-value < 0.05) tested using the Wald test ($X^2 = 17.5$, df = 2, P (> $X^2$) = 0.00016).

4.4.3 Comparative Analysis of Carbohydrate-Active Enzymes

The Lachnospiraceae, Clostridiaceae, and Ruminococcaceae differ with respect to the average numbers of carbohydrate-active genes and gene families, particularly the glycoside hydrolases (GH) and carbohydrate-binding modules (CBM) (Figure 20), which are more abundant and more diverse in the Lachnospiraceae and Ruminococcaceae.

A closer look at the average numbers of genes per GH family reveals that significant differences exist between these groups for thirteen GH families most of which include enzymes used to degrade complex plant polymers. With the exception of GH1, all of these families are more highly represented in the Lachnospiraceae, and
Figure 20. Comparison of Carbohydrate-Active enZymes (CAZy) with respect to abundance in the genomes of Lachnospiraceae, Clostridiaceae, and Ruminococcaceae. (a) Average numbers of CAZy families per group. (b) Average numbers of CAZy genes per group. CAZy enzyme classes indicated as follows: GH = glycoside hydrolase, GTF = glucosyltransferase, CE = carbohydrate esterase, PL = polysaccharide lyase. (*) Groups showing significant differences in either one-way ANOVA testing (for data with a normal distribution based on the Shapiro-Wilks test) or the best fit between negative binomial and Poisson models as described above.

Ruminococcaceae (Figure 21). GH2, GH3, GH43, and GH51 are associated with cleaving pectin and hemicellulose sidechains, GH5 and GH9 contain cellulases, GH13 and GH31 consist of starch-degrading alpha-glucosidases, while GH10 includes xylanases. GH94 contains phosphorylases that cleave beta-glycosidic bonds in cellobiose, celloextrin and chitobiose.

Figure 21. Glycoside hydrolase families that differ in abundance. Families of glycoside hydrolases that differ in abundance between the genomes of Lachnospiraceae, Clostridiaceae, and Ruminococcaceae based on best fit between negative binomial or Poisson regression models (Clostridiaceae as reference) as determined by likelihood ratio test (p-value < 0.05).
When GH enzyme families are broken into individual enzymes (Figure 22; Activities are given in Appendix H), differences between microbial families related to specific plant degradation processes are revealed. The enzyme that differs significantly between the genomes of Lachnospiraceae, Clostridiaceae, and Ruminococcaceae in the GH1 family is a beta-glucosidase (EC: 3.2.1.21), which is found in almost every GH family. Considering other GH families, the Lachnospiraceae and Ruminococcaceae have higher numbers of genes equipped to degrade a wide variety of polysaccharides. The Ruminococcaceae are enriched in endo-1, 4-beta-xylanase and cellulase genes, while both groups have higher numbers of alpha-glucosidases and both alpha and beta-galactosidases. Thus, members of these microbial families are better equipped to cleave the cellulose and hemicellulose components of plant material.

Figure 22. Glycoside hydrolase enzymes that differ in abundance. Glycoside hydrolase enzymes that differ significantly between the genomes of Lachnospiraceae, Clostridiaceae, and Ruminococcaceae based on best fit between negative binomial or Poisson regression models (Clostridiaceae as reference) as determined by likelihood ratio test (p-value < 0.05).
Carbohydrate-binding modules that differ significantly between the Lachnospiraceae, Clostridiaceae, and Ruminococcaceae include CBM6, CBM22, and CBM48 (Figure 23; Activities are given in Appendix I). All are significantly enriched in the Ruminococcaceae which has been shown to have unusual substrate binding capabilities [120]. CBM6 binds to both cellulose and hemicellulose components of plant material, while CBM22 binds primarily to xylan, and CBM48 is associated with glycogen.

4.4.4 Comparative Analysis of Transporter Proteins

Genomes of the Clostridiaceae contained more PTS (phosphotransferase) genes than the Lachnospiraceae or Ruminococcaceae, while the later two groups were more highly enriched in ABC (ATP-binding cassette) genes (Figure 24).

PTS systems transport a wide variety of mono- and disaccharides, especially hexoses such as glucose [137]. In PTS transport, substrates are phosphorylated upon entry, which makes their subsequent metabolism efficient, and also provides a means for regulation.
and preferential sugar utilization via catabolite repression [138]. Thus, PTS transport enables bacteria living in carbohydrate-limited environments, such as soils and sediments, to efficiently utilize and compete for substrates as they become available.

ABC transporters, on the other hand, tend to carry oligosaccharides, and have less preference for hexoses [139,140]. Oligosaccharide import is energetically favorable because it enables the conservation of the energy of hydrolysis intracellularly. Regulation of ABC transporters is less well studied than for PTS; however, they are thought to be controlled by proteins acting to block specific domains [141]. The abundance of ATP transporters in the Lachnospiraceae and Ruminococcaceae is consistent with their capacity to utilize complex plant material, and transport degradation products of various sizes and compositions. Since carbon is not limited and is present as a range of complex

![Figure 24. Average abundance of phosphotransferase system (PTS) and ATP-binding cassette (ABC) transporter genes.](image)

A comparison of Lachnospiraceae, Clostridiaceae, and Ruminococcaceae genomes from the IMG database [128] (*) Genes showing significant differences between groups (Clostridiaceae as reference) in best fit between negative binomial or Poisson regression models as determined by the likelihood ratio test ($p$-value < 0.05).
polymers in the gut environment, the ability to utilize many different substrates may be more advantageous than the efficient intake of a preferred carbon source.

On the single gene level, the average abundance of sugar transport genes in each genome of Lachnospiraceae, Clostridiaceae, and Ruminococcaceae demonstrates the preference of the Clostridiaceae for simple hexoses such as glucose and cellobiose, and the wider range of substrates, including pentoses, transported by the Lachnospiraceae and Ruminococcaceae (Figure 25).

**Figure 25. Average abundance of sugar transport genes.**
A comparison of each genome of Lachnospiraceae, Clostridiaceae, and Ruminococcaceae. (*) Genes showing significant differences between groups (Clostridiaceae as reference) in best fit between negative binomial or Poisson regression models as determined by the likelihood ratio test ($p$-value < 0.05).

4.4.5 Comparative Analysis of Metabolic Pathways

Four metabolic pathways were found to differ in terms of average number of genes between the Lachnospiraceae, Clostridiaceae, and Ruminococcaceae, namely alcohol
degradation, carbohydrate degradation, polymeric compound degradation, and generation of precursor metabolites and energy (Figure 26). Considering degradation pathway classes for carbohydrates and polymeric compounds, ten specific pathways were identified as differing between the Lachnospiraceae, Clostridiaceae, and Ruminococcaceae (Figure 27) revealing the capacity to break down a full range of plant-derived substrates including cellulose, hemicellulose, and starch.

![Figure 26. Average abundance of genes in each Ecocyc degradation pathway.](image)

A comparison between Lachnospiraceae, Clostridiaceae, and Ruminococcaceae genomes. (*) Pathways showing significant differences between groups in either one-way ANOVA testing (for normal data) \( (p\text{-value} < 0.05) \) or the best fit between negative binomial or Poisson regression models as determined by the likelihood ratio test \( (p\text{-value} < 0.05) \).

4.5 Conclusion

Genome comparisons of the carbohydrate-active enzymes, transporters, and metabolic pathways of the Lachnospiraceae and Ruminococcaceae in comparison with the Clostridiaceae as described here reveal these groups to be more highly specialized for the degradation of complex plant material.
Figure 27. Carbohydrate and Polymeric Compound Degradation pathways. Percentages for pathways found to be significantly different between Lachnospiraceae, Clostridiaceae, and Ruminococcaceae (indicated by (*) in Figure 9). Roman numerals reference specific pathway designations in Ecocyc [135].

In gut environments the ability to degrade cellulose and hemicellulose components of plant material enables members of the Lachnospiraceae and Ruminococcaceae to decompose substrates that are indigestible by the host. These compounds are then fermented and converted into short chain fatty acids (mainly acetate, butyrate, and propionate) that can be absorbed and used for energy by the host.
CHAPTER 5
THE COMPLETE GENOME SEQUENCE OF CLOSTRIDIUM INDOLIS

5.1 Abstract

*Clostridium indolis* DSM 755 is a bacteria commonly found in soils and feces of birds and mammals. Despite its prevalence, little is known about the ecology or physiology of this species, however close relatives, *C. saccharolyticum* and *C. hathewayi*, have demonstrated interesting metabolic potentials related to plant degradation and human health. The genome of *C. indolis* DSM 755 reveals an abundance of genes in functional groups associated with the transport and utilization of carbohydrates, as well as citrate utilization and nitrogen fixation. Our genome analysis suggests hypotheses to be tested in future culture based work to better understand the physiology of this poorly described species.

5.2 Introduction

The *C. saccharolyticum* species group is a poorly described and taxonomically confusing clade in the Lachnospiraceae, a family within the Clostridiales that includes members of clostridia, cluster XIVa [25]. This group includes *C. indolis*, *C. sphenoides*, *C. methoxybenzovorans*, *C. celerecrescens*, and *Desulfotomaculum guttoideum*, none of which are well studied (Figure 28). *C. saccharolyticum* has gained attention because its saccharolytic capacity was shown to be syntrophic with the cellulolytic activity of *Bacteroides cellulosolvens* in coculture, enabling the conversion of cellulose to ethanol in a single step [16,34]. Members of this group such as *C. celerecrescens* are themselves
cellulolytic [35] and others are known to degrade unusual substrates such as methylated aromatic compounds (C. methoxybenzovorans) [36], and the insecticide lindane (C. sphenoides) [37]. C. indolis was targeted for whole genome sequencing to provide insight into the genetic potential of this taxa that could then direct experimental efforts to understand its physiology and ecology.

Figure 28. Phylogenetic tree of 16S rRNA genes of the Clostridium saccharolyticum species group. The strains and their corresponding NCBI accession numbers (and, when applicable, draft sequence coordinates) for 16S rRNA genes are: Desulfotomaculum guttoideum strain DSM 4024\textsuperscript{T}, Y11568; C. sphenoides ATCC 19403\textsuperscript{T}, AB075772; C. celerecrescens DSM 5628\textsuperscript{T}, X71848; C. indolis DSM 755\textsuperscript{T}, Pending release by JGI: 1620643-1622056; C. methoxybenzovorans SR3, AF067965; C. saccharolyticum WM1\textsuperscript{T}, NC_014376:18567-20085; C. algidixylanolyticum SPL73\textsuperscript{T}, AF092549; C. hathewayi DSM 13479\textsuperscript{T}, ADLN0000000:202-1639; C. phytofermentans ISDg\textsuperscript{T}, CP000885:15754-17276; C. jejuense HY-35-12, (AY494606.2); C. xylanovorans HESP1\textsuperscript{T}, AF116920; C. phytofermentans ISDg, CP000885:15754-17276; and C. stercorarium ATCC 35414\textsuperscript{T}, CP003992:856992-858513. The tree uses sequences aligned by MUSCLE, and was inferred using the Neighbor-Joining method [142]. The optimal tree with the sum of branch length = 0.44337019 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [143]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method [144] and are in the units of the number of base substitutions per site. Evolutionary analyses were conducted in MEGA5 [54]. Bacillus subtilis DSM 10\textsuperscript{T} (AJ276351) was used as an outgroup.
5.3 Organism information

The general features of *Clostridium indolis* DSM 755 are listed in Table 4. *C. indolis* DSM 755 was originally named for its ability to hydrolyze tryptophan to indole, pyruvate, and ammonia [145] in the classic Indole Test used to distinguish bacterial species. It has been isolated from soil [146], feces [147], and clinical samples from infections [148]. Despite its prevalence, *C. indolis* is not well characterized, and there are conflicting reports about its physiology. It is described as a sulfate reducer with the ability to ferment some simple sugars, pectin, pectate, mannitol, and galacturonate, and convert pyruvate to acetate, formate, ethanol, and butyrate [115]. According to this source, neither lactate nor citrate are utilized, however other studies demonstrate that fecal isolates closely related to *C. indolis* may utilize lactate [74], and that the type strain DSM 755 utilizes citrate [149]. It is unclear whether *C. indolis* is able to make use of a wider range of sugars or break down complex carbohydrates, however growth is reported to be stimulated by fermentable carbohydrates [115].

5.4 Genome sequencing information

5.4.1 Genome project history

The genome was selected based on the relatedness of *C. indolis* DSM 755 to *C. saccharolyticum*, an organism with interesting saccharolytic and syntrophic properties. The genome sequence was completed on May 2, 2013, and presented for public access on June 3, 2013. Quality assurance and annotation done by DOE Joint Genome Institute (JGI) as described below. Table 5 presents a summary of the project information and its association with MIGS version 2.0 compliance [150]
Table 4. Classification and general features of *Clostridium indolis* DSM 755.

<table>
<thead>
<tr>
<th>MIGS ID</th>
<th>Property</th>
<th>Term</th>
<th>Evidence Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Current classification</td>
<td>Domain <em>Bacteria</em></td>
<td>TAS [151]AM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phylum <em>Firmicutes</em></td>
<td>TAS [115]AM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Class <em>Clostridia</em></td>
<td>TAS [113]AM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Order <em>Clostridiales</em></td>
<td>TAS [113,115]AM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Family <em>Lachnospiraceae</em></td>
<td>TAS [115]AM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Genus <em>Clostridium</em></td>
<td>TAS [25,152]AM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Species <em>Clostridium indolis</em></td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td></td>
<td>Strain DSM 755</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gram stain</td>
<td>Negative</td>
<td></td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>Cell shape</td>
<td>Rod</td>
<td></td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>Motility</td>
<td>Motile</td>
<td></td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>Sporulation</td>
<td>Terminal, spherical spores</td>
<td></td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>Temperature range</td>
<td>Mesophilic</td>
<td></td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>Optimum temperature</td>
<td>37°C</td>
<td></td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>Carbon sources</td>
<td>Glucose, lactose, sucrose, mannitol, pectin, pyruvate, others</td>
<td></td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>Terminal electron receptor</td>
<td>Sulfate</td>
<td></td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>Indole test</td>
<td>Positive</td>
<td></td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>MIGS-6</td>
<td>Habitat</td>
<td>Isolated from soil, feces, wounds</td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>MIGS-6.3</td>
<td>Salinity</td>
<td>Inhibited by 6.5% NaCl</td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>MIGS-22</td>
<td>Oxygen</td>
<td>Anaerobic</td>
<td>TAS [145,146]AM</td>
</tr>
<tr>
<td>MIGS-15</td>
<td>Biotic relationship</td>
<td>Free living and host associated</td>
<td>TAS [146,147]AM</td>
</tr>
<tr>
<td>MIGS-14</td>
<td>Pathogenicity</td>
<td>No</td>
<td>NAS</td>
</tr>
<tr>
<td>MIGS-4</td>
<td>Geographic location</td>
<td>Soil, feces</td>
<td>TAS [146,147]AM</td>
</tr>
</tbody>
</table>

**Evidence Codes** - IDA: Inferred from Direct Assay; TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from [http://www.geneontology.org/GO.evidence.shtml](http://www.geneontology.org/GO.evidence.shtml) of the Gene Ontology project [153].

5.4.2 Growth conditions and DNA isolation

*C. indolis* DSM 755 was cultivated anaerobically on GS-2 media as described elsewhere [12]. DNA for sequencing was extracted using the DNA Isolation Bacterial Protocol available through the JGI ([http://www.jgi.doe.gov](http://www.jgi.doe.gov)). The quality of DNA extracted was assessed by gel electrophoresis and NanoDrop (ThermoScientific, Wilmington, DE) according to the JGI recommendations, and the quantity was measured using the Quant-iT™ Picogreen assay kit (Invitrogen, Carlsbad, CA) as directed.
Table 5. Project information.

<table>
<thead>
<tr>
<th>MIGS ID</th>
<th>Property</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIGS-31</td>
<td>Finishing quality</td>
<td>Improved Draft</td>
</tr>
<tr>
<td>MIGS-28</td>
<td>Libraries used</td>
<td>Shotgun and long insert mate pair (Illumina)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMRTbell™ (PacBio)</td>
</tr>
<tr>
<td>MIGS-29</td>
<td>Sequencing platforms</td>
<td>Illumina and PacBio</td>
</tr>
<tr>
<td>MIGS-31.2</td>
<td>Fold coverage</td>
<td>759.7X (Illumina)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>51.6X (PacBio)</td>
</tr>
<tr>
<td>MIGS-30</td>
<td>Assemblers</td>
<td>Velvet, AllpathsLG</td>
</tr>
<tr>
<td>MIGS-32</td>
<td>Gene calling method</td>
<td>Prodigal, GenePRIMP</td>
</tr>
<tr>
<td>Genome Database release</td>
<td></td>
<td>June 3, 2013 (IMB)</td>
</tr>
<tr>
<td>Genbank ID</td>
<td></td>
<td>Pending release by JGI</td>
</tr>
<tr>
<td>Genbank Date of Release</td>
<td></td>
<td>Pending release by JGI</td>
</tr>
<tr>
<td>GOLD ID</td>
<td></td>
<td>Gi22434</td>
</tr>
<tr>
<td>Project relevance</td>
<td></td>
<td>Anaerobic plant degradation</td>
</tr>
</tbody>
</table>

5.4.3 Genome sequencing and assembly

The draft genome of *Clostridium indolis* was generated at the DOE Joint genome Institute (JGI) using a hybrid of the Illumina and Pacific Biosciences (PacBio) technologies. An Illumina std shotgun library and long insert mate pair library was constructed and sequenced using the Illumina HiSeq 2000 platform [154]. 16,165,490 reads totaling 2,424.8 Mb were generated from the std shotgun and 26,787,478 reads totaling 2,437.7 Mb were generated from the long insert mate pair library. A Pacbio SMRTbellTM library was constructed and sequenced on the PacBio RS platform. 99,448 raw PacBio reads yielded 118,743 adapter trimmed and quality filtered subreads totaling 330.2 Mb. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov. All raw Illumina sequence data was passed through DUK, a filtering program developed at JGI, which removes known Illumina sequencing and library preparation artifacts [155]. Filtered Illumina and PacBio reads were assembled using AllpathsLG (PrepareAllpathsInputs: PHRED 64=1 PLOIDY=1 FRAG COVERAGE=50 JUMP COVERAGE=25; RunAllpath- sLG: THREADS=8
The final draft assembly contained 1 contig in 1 scaffold. The total size of the genome is 6.4 Mb. The final assembly is based on 2,424.6 Mb of Illumina Std PE, 2,437.6 Mb of Illumina CLIP PE and 330.2 Mb of PacBio post filtered data, which provides an average 759.7X Illumina coverage and 51.6X PacBio coverage of the genome, respectively.

5.4.4 Genome annotation

Genes were identified using Prodigal [157], followed by a round of manual curation using GenePRIMP [5] for finished genomes and Draft genomes in fewer than 10 scaffolds. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database, UniProt, TIGRfam, Pfam, KEGG, COG, and InterPro databases. The tRNAscanSE tool [158] was used to find tRNA genes, whereas ribosomal RNA genes were found by searches against models of the ribosomal RNA genes built from SILVA [159]. Other non-coding RNAs such as the RNA components of the protein secretion complex and the RNase P were identified by searching the genome for the corresponding Rfam profiles using INFERNAL [160]. Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG) platform [128] developed by the Joint Genome Institute, Walnut Creek, CA, USA [161].
5.5 Genome properties

The genome of *C. indolis* DSM 755 consists of a 6,383,701 bp circular chromosome with GC content of 44.93% (Table 6). Of the 5,903 genes predicted, 5,802 were protein-coding genes, and 101 RNAs; 170 pseudogenes were also identified. 81.21% of genes were assigned with a putative function with the remaining annotated as hypothetical proteins. The genome summary and distribution of genes into COGs functional categories are listed in Tables 7 and 8.

**Table 6. Nucleotide content and gene count levels of the genome of *C. indolis* DSM 755.**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Genome (total)</th>
<th>% of total&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (bp)</td>
<td>6,383,701</td>
<td></td>
</tr>
<tr>
<td>G+C content (bp)</td>
<td>2,868,247</td>
<td>44.93</td>
</tr>
<tr>
<td>Coding region (bp)</td>
<td>5,688,007</td>
<td>89.10</td>
</tr>
<tr>
<td>Total genes&lt;sup&gt;b&lt;/sup&gt;</td>
<td>5,903</td>
<td>100.00</td>
</tr>
<tr>
<td>RNA genes</td>
<td>101</td>
<td>1.71</td>
</tr>
<tr>
<td>Protein-coding genes</td>
<td>5,802</td>
<td>98.02</td>
</tr>
<tr>
<td>Protein-coding with function pred.</td>
<td>4,794</td>
<td>81.21</td>
</tr>
<tr>
<td>Genes in paralog clusters</td>
<td>4,527</td>
<td>76.69</td>
</tr>
<tr>
<td>Genes assigned to COGs</td>
<td>4,643</td>
<td>78.65</td>
</tr>
<tr>
<td>Genes with signal peptides</td>
<td>421</td>
<td>7.13</td>
</tr>
<tr>
<td>Genes with transmembrane helices</td>
<td>1,494</td>
<td>25.31</td>
</tr>
<tr>
<td>Paralogous groups</td>
<td>4,527</td>
<td>76.69</td>
</tr>
</tbody>
</table>

<sup>a</sup> The total is based on either the size of the genome in base pairs or the total number of protein coding genes in the annotated genome.

<sup>b</sup> Also includes 170 pseudogenes.

The genomes of *C. indolis* and its near relatives (*C. saccharolyticum, C. hathewayi, and C. phytofermentans*) have similar numbers of genes in each of the 25 broad COG categories (not shown), however differences exist in the type and distribution of genes in specific functional groups (Table 8), particularly those related to COG categories (G) Carbohydrate transport and metabolism, (C) Energy production and conversion, and (Q) Secondary metabolites biosynthesis, transport and catabolism.
Table 7. Number of genes in C. indolis DSM 755 associated with the 25 general COG functional categories

<table>
<thead>
<tr>
<th>Code</th>
<th>Value</th>
<th>% of total</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>184</td>
<td>3.57</td>
<td>Translation</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>RNA processing and modification</td>
</tr>
<tr>
<td>K</td>
<td>531</td>
<td>10.3</td>
<td>Transcription</td>
</tr>
<tr>
<td>L</td>
<td>191</td>
<td>3.71</td>
<td>Replication, recombination and repair</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.02</td>
<td>Chromatin structure and dynamics</td>
</tr>
<tr>
<td>D</td>
<td>28</td>
<td>0.54</td>
<td>Cell cycle control, mitosis and meiosis</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>0</td>
<td>Nuclear structure</td>
</tr>
<tr>
<td>V</td>
<td>107</td>
<td>2.08</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>T</td>
<td>335</td>
<td>6.5</td>
<td>Signal transduction mechanisms</td>
</tr>
<tr>
<td>M</td>
<td>235</td>
<td>4.56</td>
<td>Cell wall/membrane biogenesis</td>
</tr>
<tr>
<td>N</td>
<td>70</td>
<td>1.36</td>
<td>Cell motility</td>
</tr>
<tr>
<td>Z</td>
<td>0</td>
<td>0</td>
<td>Cytoskeleton</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>0</td>
<td>Extracellular structures</td>
</tr>
<tr>
<td>U</td>
<td>41</td>
<td>0.8</td>
<td>Intracellular trafficking and secretion</td>
</tr>
<tr>
<td>O</td>
<td>124</td>
<td>2.41</td>
<td>Posttranslational modification, protein turnover, chaperones</td>
</tr>
<tr>
<td>C</td>
<td>261</td>
<td>5.06</td>
<td>Energy production and conversion</td>
</tr>
<tr>
<td>G</td>
<td>910</td>
<td>17.65</td>
<td>Carbohydrate transport and metabolism</td>
</tr>
<tr>
<td>E</td>
<td>493</td>
<td>9.56</td>
<td>Amino acid transport and metabolism</td>
</tr>
<tr>
<td>F</td>
<td>110</td>
<td>2.13</td>
<td>Nucleotide transport and metabolism</td>
</tr>
<tr>
<td>H</td>
<td>153</td>
<td>2.97</td>
<td>Coenzyme transport and metabolism</td>
</tr>
<tr>
<td>I</td>
<td>77</td>
<td>1.49</td>
<td>Lipid transport and metabolism</td>
</tr>
<tr>
<td>P</td>
<td>325</td>
<td>6.3</td>
<td>Inorganic ion transport and metabolism</td>
</tr>
<tr>
<td>Q</td>
<td>70</td>
<td>1.36</td>
<td>Secondary metabolites biosynthesis, transport and catabolism</td>
</tr>
<tr>
<td>R</td>
<td>590</td>
<td>11.45</td>
<td>General function prediction only</td>
</tr>
<tr>
<td>S</td>
<td>319</td>
<td>6.19</td>
<td>Function unknown</td>
</tr>
<tr>
<td>-</td>
<td>1260</td>
<td>21.35</td>
<td>Not in COGs</td>
</tr>
</tbody>
</table>

a) The total is based on the total number of protein coding genes in the annotated genome.

5.5.1 Carbohydrate transport and metabolism

Plant biomass is a complex composite of fibrils and sheets of cellulose, hemicellulose, waxes, pectin, proteins, and lignin. Bacteria from soil and the gut generally possess a variety of genes to degrade and transport the diversity of substrates encountered in these plant-rich environments. The genome of C. indolis includes 910 genes (17.65% of total protein coding genes) in this COG group including glycoside hydrolases with the potential to degrade complex carbohydrates including starch, cellulose, and chitin (Table 9), as well as an abundance of carbohydrate transporters (Figure 29).
Table 8. Number of genes in *C. indolis* DSM 755 not found in near-relatives associated with the 25 general COG functional categories*

<table>
<thead>
<tr>
<th>Code</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>4</td>
<td>Translation</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>RNA processing and modification</td>
</tr>
<tr>
<td>K</td>
<td>5</td>
<td>Transcription</td>
</tr>
<tr>
<td>L</td>
<td>9</td>
<td>Replication, recombination and repair</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>Chromatin structure and dynamics</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>Cell cycle control, mitosis and meiosis</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>Nuclear structure</td>
</tr>
<tr>
<td>V</td>
<td>1</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>T</td>
<td>2</td>
<td>Signal transduction mechanisms</td>
</tr>
<tr>
<td>M</td>
<td>8</td>
<td>Cell wall/membrane biogenesis</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>Cell motility</td>
</tr>
<tr>
<td>Z</td>
<td>0</td>
<td>Cytoskeleton</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>Extracellular structures</td>
</tr>
<tr>
<td>U</td>
<td>1</td>
<td>Intracellular trafficking and secretion</td>
</tr>
<tr>
<td>O</td>
<td>10</td>
<td>Posttranslational modification, protein turnover, chaperones</td>
</tr>
<tr>
<td>C</td>
<td>28</td>
<td>Energy production and conversion</td>
</tr>
<tr>
<td>G</td>
<td>6</td>
<td>Carbohydrate transport and metabolism</td>
</tr>
<tr>
<td>E</td>
<td>8</td>
<td>Amino acid transport and metabolism</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>Nucleotide transport and metabolism</td>
</tr>
<tr>
<td>H</td>
<td>11</td>
<td>Coenzyme transport and metabolism</td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td>Lipid transport and metabolism</td>
</tr>
<tr>
<td>P</td>
<td>11</td>
<td>Inorganic ion transport and metabolism</td>
</tr>
<tr>
<td>Q</td>
<td>10</td>
<td>Secondary metabolites biosynthesis, transport and catabolism</td>
</tr>
<tr>
<td>R</td>
<td>18</td>
<td>General function prediction only</td>
</tr>
<tr>
<td>S</td>
<td>21</td>
<td>Function unknown</td>
</tr>
</tbody>
</table>

* Number of genes from a set of 158 genes not found in near relatives (*C. saccharolyticum, C. phytofermentans, C. hathewayi*) associated with the 25 general COG functional categories.

Almost 8% of the protein-coding genes in the genome of *C. indolis* were found to be associated with carbohydrate transport, represented by two main strategies. ABC (ATP binding cassette) transporters tend to carry oligosaccharides, and have less affinity for hexoses [139,140], while PTS (phosphotransferase system) transporters carry many different mono- and disaccharides, especially hexoses [137]. PTS systems provide a means of regulation via catabolite repression [138], and are thought to enable bacteria living in carbohydrate-limited environments to more efficiently utilize and compete for substrates [138]. Both *C. indolis* and its near relatives are more highly enriched in ABC
Table 9. Selected carbohydrate active genes in the *C. indolis* DSM 755 genome.

<table>
<thead>
<tr>
<th>Gene Count</th>
<th>Gene Product</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>Beta-glucosidase (GH-1)</td>
<td>EC:3.2.1.86</td>
</tr>
<tr>
<td>8</td>
<td>Beta-galactosidase/ beta-glucuronidase (GH-2)</td>
<td>EC:3.2.1.23</td>
</tr>
<tr>
<td>7</td>
<td>Beta-glucosidase/ related glucosidases (GH-3)</td>
<td>EC:3.2.1.21</td>
</tr>
<tr>
<td>14</td>
<td>Alpha-galactosidases/ 6-phospho-beta-glicosidases (GH-4)</td>
<td>EC:3.2.1.122</td>
</tr>
<tr>
<td>2</td>
<td>Cellulase, endogluconase (GH-5)</td>
<td>EC:3.2.1.14</td>
</tr>
<tr>
<td>14</td>
<td>Alpha-amylase</td>
<td>EC:3.2.1.10</td>
</tr>
<tr>
<td>8</td>
<td>Beta-xylosidase (GH 39)</td>
<td>EC:3.2.1.37</td>
</tr>
<tr>
<td>2</td>
<td>Chitinase (GH 18)</td>
<td>EC:3.2.1.14</td>
</tr>
</tbody>
</table>

than PTS transporters (Fig 29), however nearly a third of *C. indolis* and *C. saccharolyticum* transporters are PTS genes, suggesting a preference for hexoses, as well as an adaptation to more marginal environments. *C. indolis* also possesses ten genes associated with all

Figure 29. Distribution of ABC and PTS transporters in the genomes of *C. indolis* and close relatives.
Shown by (a) Number of COGS, and (b) Percentage of genes in the genome.
three components of the TRAP-type C4-dicarboxylate transport system, which transports C4-dicarboxylates such as formate, succinate, and malate [162], as well as six putative malate dehydrogenases and two putative succinate dehydrogenases suggesting that \textit{C. indolis} may have the potential to utilize both of these short chain fatty acids.

5.5.2 Energy production and conversion

The genome of \textit{C. indolis} contains 261 genes in COG category (C) Energy production and conversion, 28 of which are not found in the near relatives analyzed, including genes for citrate utilization (Table 10 and Figure 30) and nitrogen fixation (Table 11).

Table 10. Selection of \textit{C. indolis} DSM 755 genes related to citrate utilization.

<table>
<thead>
<tr>
<th>Locus Tag</th>
<th>Gene Product</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>*K401DRAFT_2892</td>
<td>holo-ACP synthase (CitX)</td>
<td>EC:2.7.7.61</td>
</tr>
<tr>
<td>*K401DRAFT_2893</td>
<td>citrate lyase acyl carrier (CitD)</td>
<td>EC:4.1.3.6</td>
</tr>
<tr>
<td>*K401DRAFT_2894</td>
<td>citrate lyase beta subunit (CitE)</td>
<td>EC:4.1.3.6</td>
</tr>
<tr>
<td>*K401DRAFT_2895</td>
<td>citrate lyase alpha subunit (CitF)</td>
<td>EC:4.1.3.6</td>
</tr>
<tr>
<td>*K401DRAFT_2896</td>
<td>triphosphoribosyl-dephospho-CoA synthase (CitG)</td>
<td>EC:2.7.8.25</td>
</tr>
<tr>
<td>*K401DRAFT_2897</td>
<td>citrate (pro3S)-lyase ligase (CitC)</td>
<td>EC:6.2.1.22</td>
</tr>
<tr>
<td>K401DRAFT_2898</td>
<td>response regulator, CheY-like receiver domain, winged helix DNA binding domain</td>
<td>–</td>
</tr>
<tr>
<td>*K401DRAFT_2899</td>
<td>signal transduction histidine kinase</td>
<td>–</td>
</tr>
<tr>
<td>*K401DRAFT_2900</td>
<td>citrate transporter, CITMHS family</td>
<td>KO:K03303 TC:LCTP</td>
</tr>
</tbody>
</table>

* Genes without homologs in near relatives (\textit{C. saccharolyticum}, \textit{C. phytofermentans}, \textit{C. hathewayi}).
Figure 30. Citrate utilization genes in C. indolis.
A single gene cluster on K401DRAFT_scaffold0000.1.1, including citCDEFGX, the citrate transporter CitMHS, and a putative two-component system.

Table 11. Selection of C. indolis DSM 755 genes related to nitrogen fixation.

<table>
<thead>
<tr>
<th>Locus Tag</th>
<th>Gene Product</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>*K401DRAFT_0533</td>
<td>nitrogenase Mo-Fe protein, α and β chains</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_0534</td>
<td>nitrogenase Mo-Fe protein, α and β chains</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_0535</td>
<td>nitrogenase subunit (ATPase) (nifH)</td>
<td>Pfam00142</td>
</tr>
<tr>
<td>*K401DRAFT_0884</td>
<td>nitrogenase Mo-Fe protein, α and β chains</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_0885</td>
<td>nitrogenase Mo-Fe protein, α and β chains</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_0886</td>
<td>nitrogenase subunit (ATPase) (nifH)</td>
<td>Pfam00142</td>
</tr>
<tr>
<td>*K401DRAFT_3349</td>
<td>nitrogenase subunit (ATPase) (nifH)</td>
<td>Pfam00142</td>
</tr>
<tr>
<td>*K401DRAFT_3350</td>
<td>nitrogenase Mo-Fe protein, α and β chains</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_3351</td>
<td>nitrogenase subunit (ATPase) (nifH)</td>
<td>Pfam00142</td>
</tr>
<tr>
<td>*K401DRAFT_3874</td>
<td>nitrogenase Mo-Fe protein, α and β chains (nifD)</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_3875</td>
<td>nitrogenase Mo-Fe protein, α and β chains (nifK)</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_3876</td>
<td>nitrogenase Fe protein</td>
<td>Pfam00142</td>
</tr>
<tr>
<td>*K401DRAFT_3878</td>
<td>nitrogenase Mo-Fe protein, α and β chains (nifD)</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_3879</td>
<td>nitrogenase Mo-Fe protein, α and β chains (nifK)</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>K401DRAFT_3880</td>
<td>dinitrogenase Fe-Mo cofactor</td>
<td>Pfam02579</td>
</tr>
<tr>
<td>*K401DRAFT_3895</td>
<td>nitrogenase Mo-Fe protein, α and β chains (nifD)</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_3896</td>
<td>nitrogenase Mo-Fe protein, α and β chains (nifK)</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_5519</td>
<td>nitrogenase Mo-Fe protein, α and β chains (nifB)</td>
<td>Pfam04055</td>
</tr>
<tr>
<td>*K401DRAFT_5520</td>
<td>nitrogenase Mo-Fe protein, α and β chains (nifE)</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_5521</td>
<td>nitrogenase Mo-Fe protein (nifK)</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_5522</td>
<td>nitrogenase component 1, alpha chain</td>
<td>Pfam00148</td>
</tr>
<tr>
<td>*K401DRAFT_5525</td>
<td>nitrogenase subunit (ATPase) (nifH)</td>
<td>Pfam00142</td>
</tr>
</tbody>
</table>

* Genes without homologs in near relatives (C. saccharolyticum, C. phytofermentans, C. hathewayi). Nitrogenase genes have a common gene identifier (EC:1.18.6.1), therefore the Pfam numbers are given to distinguish between subunits.
5.5.3 Citrate utilization

Citrate is a metabolic intermediary found in all living cells. In aerobic bacteria, citrate is utilized as part of the tricarboxylic acid (TCA) cycle. In anaerobes, citrate is fermented to acetate, formate, and/or succinate. The first step is the conversion of citrate to acetate and oxaloacetate in a reaction catalyzed by citrate lyase (EC:4.1.3.6) [163]. *C. sphenoides*, a close relative of *C. indolis* that does not yet have a sequenced genome has been shown to utilize citrate [164], but there is conflicting evidence as to whether this phenotype is present in *C. indolis* [115,149]. The genome of *C. indolis* reveals a group of seven citrate genes organized in a cluster similar to operons found in other bacterial species [163,165] (Figure 30) including CitD, CitE, and CitF, the three subunits of the citrate lyase gene [163], CitG and CitX which have been shown to be necessary for citrate lyase function [165], CitMHS, a citrate transporter, and a putative two component system similar to citrate regulatory mechanisms in other bacteria [166].

5.5.4 Nitrogen Fixation

Nitrogen fixation has been observed in other clostridia [167,168] but has not been demonstrated in the *C. saccharolyticum* species group. It has been suggested that the capacity to fix nitrogen confers a selective advantage to cellulolytic microbes that live in nitrogen limited environments such as many soils [167]. The *C. indolis* genome reveals 22 nitrogenases in four gene clusters (Table 11), none of which are found in the near relatives analyzed in this study. Genes needed for the nitrogenase component proteins (*nifH*, *nifD*, and *nifK*) are present in *C. indolis*, but two of the four genes required to synthesize the nitrogenase iron-molybdenum cofactor (*nifV* and *nifN*) are absent.
5.5.5 Lactate utilization

The genome of *C. indolis* includes both D- and L-lactate dehydrogenases, which convert lactate to pyruvate. Additionally, there is a lactate transporter, suggesting the ability of *C. indolis* to utilize exogenous lactate (Table 12).

Table 12. Selection of *C. indolis* DSM 755 genes related to lactate utilization.

<table>
<thead>
<tr>
<th>Locus Tag</th>
<th>Gene Product</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>K401DRAFT_1877</td>
<td>L-lactate dehydrogenase</td>
<td>EC:1.1.1.27</td>
</tr>
<tr>
<td>K401DRAFT_5775</td>
<td>L-lactate dehydrogenase</td>
<td>EC:1.1.1.27</td>
</tr>
<tr>
<td>*K401DRAFT_3431</td>
<td>L-lactate transporter, LctP family</td>
<td>TC.LCTP</td>
</tr>
<tr>
<td>K401DRAFT_3220</td>
<td>D-lactate dehydrogenase</td>
<td>EC:1.1.1.28</td>
</tr>
</tbody>
</table>

* Genes without homologs in near relatives (*C. saccharolyticum, C. phytofermentans, C. hathewayi*).

5.5.6 Bacterial microcompartments (BMC)

Additionally, the *C. indolis* genome reveals 11 genes associated with carboxysome shell proteins used to concentrate carbon dioxide and sequester biochemical processes in two gene clusters with additional genes associated ethanolamine and propanediol utilization. Ethanolamine microcompartments have not been experimentally demonstrated in the Lachnospiraceae, and the ethanolamine-ammonia lyase gene (EC:4.2.1.7) is not found in the *C. indolis* genome, suggesting either a alternative pathway or secondary role for these BMC genes.

5.5.7 Secondary metabolites biosynthesis, transport and catabolism

Protocatechuate and other aromatics are intermediaries in the degradation of lignin in plant rich environments [169]. The genome of *C. indolis* contains two protocatechuate
dioxygenases and an aromatic hydrolase, revealing the potential for utilizing aromatic compounds (Table 13).

### Table 13. Selection of *C. indolis* DSM 755 genes related to degradation of aromatics.

<table>
<thead>
<tr>
<th>Locus Tag</th>
<th>Gene Product</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>*K401DRAFT_3571</td>
<td>Protocatechuate 3,4-dioxygenase beta subunit</td>
<td>EC:1.13.11.3</td>
</tr>
<tr>
<td>*K401DRAFT_3568</td>
<td>Protocatechuate 3,4-dioxygenase beta subunit</td>
<td>EC:1.13.11.3</td>
</tr>
<tr>
<td>*K401DRAFT_3412</td>
<td>Aromatic ring hydroxylase</td>
<td>EC:5.3.3.3, EC:4.2.1.120</td>
</tr>
</tbody>
</table>

* Genes without homologs in near relatives (*C. saccharolyticum, C. phytofermentans, C. hathewayi*).

#### 5.6 Conclusion

The genomic sequence of *C. indolis* reported here reveals the metabolic potential of this organism to utilize a wide assortment of fermentable carbohydrates and intermediates including citrate, lactate, malate, succinate, and aromatics, and points to potential ecological roles in nitrogen fixation and ethanolamine utilization. Further culture-based characterization is necessary to confirm the metabolic activity suggested by this genome analysis, and to expand the description of *C. indolis*. 
6.1 Conclusions and future directions

This dissertation describes progress made on four projects exploring the community structure and activities of anaerobic microbes from soil and gut environments.

6.1.1 Microcosmic evolution

Utilizing selective pressure to enrich for organisms with specific metabolic characteristics is a fundamental technique in microbiology, however, the application of this strategy to establish stable, replicate communities from a natural sample is a novel approach, as co-cultures of isolated strains have been traditionally employed. Given that cultured isolates are widely recognized as differing from their wild type origins, and the probability of simple pair-wise combinations occurring in nature is low, the ecological relevance of such studies could be questioned, however the enormous complexity of microbial communities in soil warrants the establishment of tractable model systems.

This research demonstrates that long-term enrichment can produce simplified, replicable communities, enabling the isolation of novel microbes, and providing a flexible model for future studies seeking to assess the effects of biotic and/or abiotic factors in a community context. Future experiments to isolate more microbes from the microcosm communities are ongoing, and will provide organisms for further sequencing and characterization.
It is interesting that the OTUs that persisted in the microcosm system for more than three years were so rare in the original soil sample. Whether their persistence was due to the ability to outcompete others for resources, establish cooperative networks with other taxa, or the adoption of another strategy is a question for further study. Metagenomic studies of isolates in pairwise combinations could inform efforts to understand the mechanisms underlying the success of the persistent OTUs, and provide insight into the metabolic strategies at work to degrade plant biomass by the adapted communities.

6.1.2 Lactate fermentation in the horse gut

By utilizing an in vitro system to model the changes in bacterial community structure and function occurring during starch induced lactic acidosis in horses, we were able to identify and isolate specific microbes that could be implicated in the recovery and/or resistance to lactate build up. Future experiments will assess the extent to which these microbes and/or microbial communities can attenuate or prevent lactic acidosis in our in vitro system modeling conditions of starch induction. We are excited by the prospect that one of our horse gut isolates could be developed as a probiotic to prevent lactic acidosis in horses.

Our research points to many additional questions regarding the utilization of lactic acid by horse gut communities. While we have demonstrated changes in community structure associated with lactate clearance, it is not clear which lactate (or starch) utilizers are most active and under what circumstances. Stable isotope probes could be used to track the flow of labeled starch and lactate through the community, pointing, in a time course, to
the microbes that utilize it most quickly and in the greatest abundance. These experiments could be coupled with tests of substrate and/or environmental preferences to identify whether specific conditions favor high lactate utilization rates. Additionally, RNA-seq could be used over the time course to measure gene expression levels indicating specific metabolic and/or competitive strategies at work to survive the stress of high lactate levels and lowered pH.

6.1.3 Comparative genomics of gut vs. free-living microbes

At the time of this writing there were 9,244 sequenced bacterial genomes available at the IMG web portal [128], and more are being released every day. While it is true that some bacterial groups are underrepresented in the database, there is still a wealth of genomic data available with which to generate and test hypotheses in silico. These results can inform and drive experimental design in the lab by providing insights into the genetic potentials of microbes of interest.

Here genome comparisons of the carbohydrate-active enzymes, transporters, and metabolic pathways between the Lachnospiraceae, Ruminococcaceae, and Clostridiaceae revealed the Lachnospiraceae and Ruminococcaceae to be more highly specialized for the degradation of complex plant material. In the gut environment, the ability to degrade both cellulose and hemicellulose components of plant biomass could enable members of these groups to utilize a wider range of substrates that are indigestible by the host. Further experimentation is needed in the lab to determine the extent to which the genetic potential of members of these groups is expressed and under what circumstances. Incorrect or
misleading gene annotation could suggest a function that does not exist. Substrate preferences, cooperative interactions with other community members, and environmental limitations are factors that influence ecosystem function without a clear genetic signal. A comparative genomics approach as described here provides insight into what molecular tools a microbe brings to the table, generating hypotheses about core genes and functions that may be associated with the commensal lifestyle.

6.1.4 Genome sequencing and characterization

_Clostridium indolis_, like many other bacteria, was isolated and characterized prior to the advent of molecular tools and the description of this taxa is scanty and conflicting. Its membership in the _C. saccharolyticum_ species group makes it interesting in terms of potential saccharolytic ability. The genome sequence of _Clostridium indolis_ reveals the potential to utilize a wider range of substrates than originally thought, including lactate, citrate, aromatics, and ethanolamine. Additionally, there is evidence that _C. indolis_ may fix nitrogen, which could give it a selective advantage in nitrogen poor soil environments, and that it may use bacterial microcompartments to sequester metabolic reactions. These results point to the need for validation experiments that could inform a more complete description of this taxa and its role in soil and gut environments.

DNA samples from four additional members of the _C. saccharolyticum_ group and two closely related microbes that were isolated from the microcosms are in the process of being sequenced. These data will provide insight into the genetic potential of the rest of
the *C. saccarolyticum* group and enable comparative genomics to better determine how they are related to each other.
APPENDIX A

PHYLOGENETIC TREES OF OTUS AND ISOLATES FROM THREE MICROCOSM FAMILIES

a

[Diagram of phylogenetic tree showing bacterial species and their relationships with branch length values]
Phylogenetic trees of OTUs and isolates from three microcosm families. To verify the match between isolate sequences (red), abundant representative OTUs (blue), and type species (black), a phylogenetic tree was constructed for each group. (a) Clostridiaceae, (b) Lachnospiraceae, (c) Ruminococcaceae. 16S rRNA sequences downloaded from the Ribosomal Protein Database [52], isolate sequences, and sequences from abundant representative OTUs were aligned in ClustalW 2.0.12 [170] using the Mobyle 1.5 web interface [171]. Sequences were trimmed, and a neighbor-joining tree [142] was made in MEGA [54] using the following parameters: 100 bootstrap iterations, Maximum Composite Likelihood method, substitutions to include transitions/transversions, uniform rates, homogeneous patterns among lineages, and complete deletions [144].
APPENDIX B

HYDROGEN SULFIDE GAS LEVELS BY HORSE

Hydrogen sulfide concentrations over time by horse.
Concentration (mM) of hydrogen sulfide measured by the Cline (methylene blue) assay from each horse and culture condition at 12 hour intervals.
**APPENDIX C**

**HEADSPACE GAS LEVELS BY HORSE**

![Graph showing headspace gas levels by horse](image)

**Headspace gas concentrations over time by horse.**
Concentration (mM) of hydrogen and methane gases measured by gas chromatography from each horse and culture condition at times 9, 20, 32, and 45 hours.
Rarefaction curves by horse.
Observed species by number of sequences per sample for each horse dataset generated using a sampling depth of 9685 (the minimum number of sequences per sample and default parameters in QIIME).
### APPENDIX E

**ALPHA DIVERSITY ESTIMATES: HORSE GUT**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Time</th>
<th>Horse</th>
<th>Reads</th>
<th>Chao1 Index</th>
<th>Obs. species</th>
<th>Shannon Index</th>
<th>Good's Cov.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 0</td>
<td>0</td>
<td>1</td>
<td>131143</td>
<td>14744.1</td>
<td>8504</td>
<td>9.6</td>
<td>97%</td>
</tr>
<tr>
<td>Starch</td>
<td>12</td>
<td>1</td>
<td>54460</td>
<td>12453.6</td>
<td>5520</td>
<td>9.4</td>
<td>94%</td>
</tr>
<tr>
<td>Lactate</td>
<td>12</td>
<td>1</td>
<td>74884</td>
<td>12035.9</td>
<td>5952</td>
<td>8.6</td>
<td>96%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>12</td>
<td>1</td>
<td>57130</td>
<td>12321.5</td>
<td>5629</td>
<td>9.3</td>
<td>94%</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>1</td>
<td>55196</td>
<td>11710.1</td>
<td>5528</td>
<td>9.5</td>
<td>94%</td>
</tr>
<tr>
<td>Starch</td>
<td>24</td>
<td>1</td>
<td>31602</td>
<td>8295.0</td>
<td>3742</td>
<td>9.1</td>
<td>93%</td>
</tr>
<tr>
<td>Control</td>
<td>24</td>
<td>1</td>
<td>56883</td>
<td>10609.1</td>
<td>4939</td>
<td>8.2</td>
<td>95%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>24</td>
<td>1</td>
<td>25002</td>
<td>7377.6</td>
<td>3160</td>
<td>8.9</td>
<td>92%</td>
</tr>
<tr>
<td>Lactate</td>
<td>24</td>
<td>1</td>
<td>62382</td>
<td>9736.8</td>
<td>4506</td>
<td>7.6</td>
<td>96%</td>
</tr>
<tr>
<td>Starch</td>
<td>36</td>
<td>1</td>
<td>45499</td>
<td>8546.2</td>
<td>3672</td>
<td>7.2</td>
<td>95%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>36</td>
<td>1</td>
<td>67107</td>
<td>8368.9</td>
<td>3735</td>
<td>5.0</td>
<td>97%</td>
</tr>
<tr>
<td>Control</td>
<td>36</td>
<td>1</td>
<td>87432</td>
<td>11093.7</td>
<td>5543</td>
<td>8.1</td>
<td>97%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>48</td>
<td>1</td>
<td>97717</td>
<td>4676.6</td>
<td>2244</td>
<td>2.1</td>
<td>99%</td>
</tr>
<tr>
<td>Lactate</td>
<td>48</td>
<td>1</td>
<td>66097</td>
<td>7949.1</td>
<td>3491</td>
<td>5.4</td>
<td>97%</td>
</tr>
<tr>
<td>Starch</td>
<td>48</td>
<td>1</td>
<td>82264</td>
<td>7082.1</td>
<td>3621</td>
<td>4.7</td>
<td>98%</td>
</tr>
<tr>
<td>Control</td>
<td>48</td>
<td>1</td>
<td>72983</td>
<td>9405.9</td>
<td>4153</td>
<td>6.5</td>
<td>97%</td>
</tr>
<tr>
<td>Time 0</td>
<td>0</td>
<td>2</td>
<td>126317</td>
<td>14301.8</td>
<td>8679</td>
<td>9.5</td>
<td>97%</td>
</tr>
<tr>
<td>Lactate</td>
<td>12</td>
<td>2</td>
<td>48086</td>
<td>10453.3</td>
<td>4993</td>
<td>9.1</td>
<td>94%</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>2</td>
<td>35210</td>
<td>9303.7</td>
<td>4207</td>
<td>8.8</td>
<td>93%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>12</td>
<td>2</td>
<td>26465</td>
<td>7229.7</td>
<td>3267</td>
<td>8.7</td>
<td>93%</td>
</tr>
<tr>
<td>Starch</td>
<td>12</td>
<td>2</td>
<td>50658</td>
<td>8722.4</td>
<td>4364</td>
<td>8.3</td>
<td>95%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>24</td>
<td>2</td>
<td>11172</td>
<td>4002.9</td>
<td>1558</td>
<td>7.0</td>
<td>91%</td>
</tr>
<tr>
<td>Lactate</td>
<td>24</td>
<td>2</td>
<td>76671</td>
<td>11587.7</td>
<td>5775</td>
<td>8.2</td>
<td>96%</td>
</tr>
<tr>
<td>Starch</td>
<td>24</td>
<td>2</td>
<td>9844</td>
<td>5160.5</td>
<td>1737</td>
<td>8.0</td>
<td>88%</td>
</tr>
<tr>
<td>Control</td>
<td>24</td>
<td>2</td>
<td>70657</td>
<td>12132.5</td>
<td>5737</td>
<td>8.5</td>
<td>95%</td>
</tr>
<tr>
<td>Control</td>
<td>36</td>
<td>2</td>
<td>26829</td>
<td>7043.4</td>
<td>3008</td>
<td>8.3</td>
<td>93%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>36</td>
<td>2</td>
<td>29086</td>
<td>4553.5</td>
<td>1701</td>
<td>4.0</td>
<td>96%</td>
</tr>
<tr>
<td>Starch</td>
<td>36</td>
<td>2</td>
<td>26266</td>
<td>4388.6</td>
<td>1938</td>
<td>5.4</td>
<td>95%</td>
</tr>
<tr>
<td>Starch</td>
<td>48</td>
<td>2</td>
<td>98241</td>
<td>6076.0</td>
<td>3311</td>
<td>4.6</td>
<td>98%</td>
</tr>
<tr>
<td>Control</td>
<td>48</td>
<td>2</td>
<td>19390</td>
<td>4969.0</td>
<td>2283</td>
<td>8.1</td>
<td>93%</td>
</tr>
<tr>
<td>Lactate</td>
<td>48</td>
<td>2</td>
<td>82176</td>
<td>5558.0</td>
<td>2506</td>
<td>3.5</td>
<td>98%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>48</td>
<td>2</td>
<td>144133</td>
<td>4536.4</td>
<td>2219</td>
<td>2.5</td>
<td>99%</td>
</tr>
</tbody>
</table>
### APPENDIX E

**ALPHA DIVERSITY ESTIMATES: HORSE GUT (CONTINUED)**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Time</th>
<th>Horse</th>
<th>Reads</th>
<th>Chao1 Index</th>
<th>Obs. species</th>
<th>Shannon Index</th>
<th>Good's Cov.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 0</td>
<td>0</td>
<td>3</td>
<td>108582</td>
<td>12506.1</td>
<td>7879</td>
<td>9.5</td>
<td>97%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>12</td>
<td>3</td>
<td>10392</td>
<td>3859.8</td>
<td>1727</td>
<td>8.1</td>
<td>90%</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>3</td>
<td>25386</td>
<td>6828.4</td>
<td>3054</td>
<td>8.3</td>
<td>93%</td>
</tr>
<tr>
<td>Starch</td>
<td>12</td>
<td>3</td>
<td>9685</td>
<td>4068.9</td>
<td>1723</td>
<td>7.7</td>
<td>89%</td>
</tr>
<tr>
<td>Lactate</td>
<td>12</td>
<td>3</td>
<td>10229</td>
<td>4201.1</td>
<td>1951</td>
<td>8.7</td>
<td>89%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>24</td>
<td>3</td>
<td>10701</td>
<td>4092.4</td>
<td>1828</td>
<td>8.3</td>
<td>90%</td>
</tr>
<tr>
<td>Control</td>
<td>24</td>
<td>3</td>
<td>42982</td>
<td>9068.7</td>
<td>3969</td>
<td>8.4</td>
<td>95%</td>
</tr>
<tr>
<td>Starch</td>
<td>24</td>
<td>3</td>
<td>13472</td>
<td>4661.1</td>
<td>2147</td>
<td>8.0</td>
<td>91%</td>
</tr>
<tr>
<td>Lactate</td>
<td>24</td>
<td>3</td>
<td>52228</td>
<td>8632.4</td>
<td>3983</td>
<td>8.0</td>
<td>96%</td>
</tr>
<tr>
<td>Lactate</td>
<td>36</td>
<td>3</td>
<td>86097</td>
<td>7136.3</td>
<td>3575</td>
<td>5.2</td>
<td>98%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>36</td>
<td>3</td>
<td>12871</td>
<td>6140.9</td>
<td>2481</td>
<td>9.0</td>
<td>88%</td>
</tr>
<tr>
<td>Control</td>
<td>36</td>
<td>3</td>
<td>81289</td>
<td>8040.5</td>
<td>3866</td>
<td>5.8</td>
<td>97%</td>
</tr>
<tr>
<td>Control</td>
<td>48</td>
<td>3</td>
<td>43180</td>
<td>5152.0</td>
<td>2527</td>
<td>4.4</td>
<td>97%</td>
</tr>
<tr>
<td>Lactate</td>
<td>48</td>
<td>3</td>
<td>68708</td>
<td>4247.8</td>
<td>2137</td>
<td>3.1</td>
<td>98%</td>
</tr>
<tr>
<td>StarchLactate</td>
<td>48</td>
<td>3</td>
<td>31334</td>
<td>6190.7</td>
<td>2805</td>
<td>6.0</td>
<td>95%</td>
</tr>
<tr>
<td>Starch</td>
<td>48</td>
<td>3</td>
<td>20417</td>
<td>4417.9</td>
<td>2102</td>
<td>5.3</td>
<td>94%</td>
</tr>
</tbody>
</table>

Alpha diversity estimates for each sample.  
Chao1, observed species, Shannon Index and Good’s coverage for each horse dataset generated using default parameters in QIIME.
### APPENDIX F

**BEST BLAST HITS FOR MOST ABUNDANT HORSE GUT OTUS AT TIME 48**

<table>
<thead>
<tr>
<th>Best BLAST hit</th>
<th>Percent similarity</th>
<th>NCBI Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Parabacteroides distasonis</em> ATCC 8503</td>
<td>85%</td>
<td>NR_074376.1</td>
</tr>
<tr>
<td><em>Streptococcus equi</em> subsp. <em>zooepidemicus</em> MGCS10565</td>
<td>100%</td>
<td>NR_102812.1</td>
</tr>
<tr>
<td><em>Desulfovibrio piger</em> ATCC 29098</td>
<td>99%</td>
<td>NR_041778.1</td>
</tr>
<tr>
<td><em>Clostridium paraputrificum</em> ATCC 25780</td>
<td>99%</td>
<td>NR_026135.1</td>
</tr>
<tr>
<td><em>Mitsuokella jalaludinii</em> M9</td>
<td>100%</td>
<td>NR_028840.1</td>
</tr>
<tr>
<td><em>Megasphaera elsdenii</em> DSM 20460</td>
<td>100%</td>
<td>NR_102980.1</td>
</tr>
<tr>
<td><em>Anaerospirobacter mobilis</em> IMSNU 40011</td>
<td>94%</td>
<td>NR_042953.1</td>
</tr>
<tr>
<td><em>Selenomonas bovis</em> strain WG</td>
<td>98%</td>
<td>NR_044111.1</td>
</tr>
<tr>
<td><em>Veillonella montpellierensis</em> ADV 281.99</td>
<td>98%</td>
<td>NR_028839.1</td>
</tr>
<tr>
<td><em>Lactobacillus equi</em> YIT 0455</td>
<td>100%</td>
<td>NR_028623.1</td>
</tr>
<tr>
<td><em>Selenomonas ruminantium</em> subsp. <em>lactilytica</em> TAM6421</td>
<td>98%</td>
<td>NR_075026.1</td>
</tr>
<tr>
<td><em>Veillonella caviae</em> PV1</td>
<td>99%</td>
<td>NR_025762.1</td>
</tr>
</tbody>
</table>

APPENDIX G

TAXA FROM EACH FAMILY USED IN COMPARATIVE ANALYSES

**Lachnospiraceae**

*Anaerostipes caccae* DSM 14662[ID]

*Blautia hansenii* VPI C7-24, DSM 20583[ID]

*Blautia hydrogenotrophica* DSM 10507[ID]

*Blautia producta* ATCC 27340[IP]

*Butyrivibrio crosstus* DSM 2876[ID]

*Butyrivibrio fibrisolvens* 16/4[C][IF]

*Butyrivibrio proteoclasticus* B316[C][IF][B]

*Catonella morbi* ATCC 51271[ID]

*Cellulosilyticum lentocellum* RHM5, DSM 5427[C][IF][B]

*Clostridium asparagiforme* DSM 15981[ID]

*Clostridium bolteae* ATCC BAA-613[ID]

*Clostridium hylemonae* DSM 15053[ID]

*Clostridium nexile* DSM 1787[ID]

*Clostridium phytofermentans* ISDg[C][IF][B]

*Clostridium saccharolyticum* WM1, DSM 2544[C][IF][B]

*Clostridium scindens* ATCC 35704[ID]

*Clostridium sporosphaeroides* DSM 1294[IP]

*Clostridium symbiosum* WAL-14163[ID][B]

*Coprococcus catus* GD/7[C][IF]

*Coprococcus comes* ATCC 27758[ID]

*Coprococcus eutactus* ATCC 27759 [ID]

*Dorea formicigenerans* ATCC 27755[ID]

*Dorea longicatena* DSM 13814[ID]

*Eubacterium cellulosolvens* 6[IP]

*Eubacterium eligens* ATCC 27750[C][IF][B]

*Eubacterium hallii* DSM 3353[ID]

*Eubacterium rectale* ATCC 33656[C][IF][B]

*Eubacterium ventriosum* ATCC 27560[ID]

*Johnsonella ignava* ATCC 51276[ID]

*Marvinbryantia formatexigens* 1-52, DSM 14469[ID][B]

*Oribacterium sinus* F0268[ID][B]

*Roseburia hominis* A2-183, DSM 16839[C][IF][B]

*Roseburia intestinalis* M50/1[C][IF][B]

*Roseburia inulinivorans* DSM 16841[ID]

*Ruminococcus gnatus* ATCC 29149[ID]

*Ruminococcus obeum* A2-162[C][IF]

*Ruminococcus torques* L2-14[C][IF]
**Clostridiaceae**

*Alkaliphilus metalliredigens* QYMF[C][IF][B]
*Alkaliphilus oremlandii* OhILAs[C][IF][B]
*Clostridium acetobutylicum* ATCC 824[C][IF][B]
*Clostridium beijerincki* NCIMB 8052[C][IF][B]
*Clostridium botulinum* BoNT/A2 Kyoto-F [C][IF][B]
*Clostridium butyricum* 5521[ID]
*Clostridium carboxidivorans* P7, DSM 15243[ID]
*Clostridium cellulovorans* 743B, ATCC 35296 [C][IF][B]
*Clostridium kluyveri* DSM 555[C][IF][B]
*Clostridium ljungdahlii* PETC, DSM 13528[C][IF][B]
*Clostridium novyi* NT[C][IF][B]
*Clostridium perfringens* ATCC 13124[C][IF][B]
*Clostridium sporogenes* ATCC 15579[ID]
*Clostridium tetani* Massachusetts E88[C][IF][B]

**Ruminococcaceae**

*Acetivibrio cellulolyticus* CD2, DSM 1870[IP]
*Anaerotruncus colihominis* DSM 17241[ID]
*Clostridium alkalicellulosi* Z-7026, DSM 17461[IP]
*Clostridium cellulolyticum* H10 [C][IF][B]
*Clostridium leptum* DSM 753[ID]
*Clostridium methylpentosum* R2, DSM 5476[ID]
*Clostridium papyrosolvens* DSM 2782[ID]
*Clostridium stercorarium* stercorarium, DSM 8532 [C][IF]
*Clostridium thermocellum* DSM 2360 [C][ID][B]
*Ethanoligenens harbinense* YUAN-3T, DSM 18485 [C][IF][B]
*Faecalibacterium prausnitzii* A2-165 [ID][B]
*Faecalibacterium prausnitzii* SL3/3 [C][IF]
*Ruminococcus albus* 7 [C][IF][B]
*Ruminococcus bromii* L2-63 [C][IF]
*Ruminococcus flavefaciens* FD-1 [ID]
*Ruminococcus gauvreauii* CCRI-16110; EF529620
*Shuttleworthia satelles* DSM 14600[ID]

[C]: CAZy database [129], [I]: IMG database[128] with status designation at time of publication: F = Finished, D = Draft, P = Permanent draft, [B]: Biocyc database [133,134]
### APPENDIX H

ACTIVITIES OF ENZYMES THAT DIFFER BETWEEN THE LACHNOSPIRACEAE, CLOSTRIDIACEAE, AND RUMINOCOCCACEAE

<table>
<thead>
<tr>
<th>EC</th>
<th>Enzyme</th>
<th>Role in Carbohydrate Degradation*</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC:3.2.1.10</td>
<td>Oligo-1,6-glucosidase</td>
<td>Hydrolysis of 1,6-α-D glucosidic linkages of oligosaccharides produced from starch or glycogen</td>
</tr>
<tr>
<td>EC:2.4.1.25</td>
<td>4-α-glucanotransferase</td>
<td>Transfer of glucose units from oligosaccharides to acceptor sugars (glucose, maltose, or maltotriose)</td>
</tr>
<tr>
<td>EC:2.4.1.18</td>
<td>1,4-α-glucan branching enzyme</td>
<td>Transfers a segment of 1,4-α-D-glucan chain to another similar glucan chain</td>
</tr>
<tr>
<td>EC:3.2.1.1</td>
<td>α-amylase</td>
<td>Endohydrolysis of 1,4-α-D-glucosid linkages in polysaccharides</td>
</tr>
<tr>
<td>EC:2.4.1.4</td>
<td>Amylosucrase</td>
<td>Addition or removal of 1,4-α-D-glucosyl linkages in starch metabolism</td>
</tr>
<tr>
<td>EC:5.4.99.16</td>
<td>Maltose α-D- glucosyltransferase</td>
<td>Conversion of maltose to α,α-trehalose</td>
</tr>
<tr>
<td>EC:3.2.1.8</td>
<td>Endo-1,4-β-xylanase</td>
<td>Endohydrolysis of 1,4-β-D-xylidosid bonds in xylans</td>
</tr>
<tr>
<td>EC:3.2.1.4</td>
<td>Cellulase</td>
<td>Endohydrolysis of 1,4-β glycosidic bonds in cellulose</td>
</tr>
<tr>
<td>EC:3.2.1.45</td>
<td>Glucosylceramidase</td>
<td>Hydrolysis of D-glucosyl-N-acylsphingosine, releasing D-glucose and N-acylsphingosine</td>
</tr>
<tr>
<td>EC:3.2.1.78</td>
<td>Mannan endo-1,4-β-mannosidase</td>
<td>Hydrolysis of 1,4-β-D-mannosidic linkages in mannan, galactomannan and glucomannan</td>
</tr>
<tr>
<td>EC:3.2.1.23</td>
<td>Beta-galactosidase</td>
<td>Hydrolysis of β-D-galactosides, releasing β-D-galactose</td>
</tr>
<tr>
<td>EC:3.2.1.31</td>
<td>Beta-glucuronidase</td>
<td>Hydrolysis of β-D-glucuronoside, releasing D-glucuronate and an alcohol</td>
</tr>
<tr>
<td>EC:3.2.1.25</td>
<td>Beta-mannosidase</td>
<td>Hydrolysis of β-D-mannosides, releasing β-D-mannose</td>
</tr>
<tr>
<td>EC:3.2.1.58</td>
<td>Glucan 1,3-β-glucosidase</td>
<td>Successive hydrolysis of 1,3-β-D-glucans, releasing β-D-glucose</td>
</tr>
<tr>
<td>EC:3.2.1.37</td>
<td>Xylan 1,4-β-xylosidase</td>
<td>Successive hydrolysis of 1,4-β-D-xylans, releasing D-xylose</td>
</tr>
<tr>
<td>EC:3.2.1.91</td>
<td>Cellulose 1,4-β-cellobiosidase</td>
<td>Hydrolysis of 1,4-β-D-glucosidic residues in cellulose, releasing cellobiose</td>
</tr>
<tr>
<td>EC:3.2.1.22</td>
<td>Alpha-galactosidase</td>
<td>Hydrolysis of terminal α-D-galactose residues, releasing α-D-galactose</td>
</tr>
<tr>
<td>EC:3.2.1.20</td>
<td>Alpha-glucosidase</td>
<td>Hydrolysis of terminal α-D-glucose residues, releasing α-D-glucose</td>
</tr>
</tbody>
</table>

*Activities as identified in available databases [129,133]
APPENDIX I

SUBSTRATE AFFINITIES OF CARBOHYDRATE BINDING MODULES THAT DIFFER BETWEEN THE LACHNOSPIRACEAE, CLOSTRIDIACEAE, AND RUMINOCOCCACEAE

<table>
<thead>
<tr>
<th>CBM</th>
<th>Substrate Affinity*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBM2</td>
<td>cellulose, chitin, xylan</td>
</tr>
<tr>
<td>CBM6</td>
<td>cellulose, xylan, (\beta)-1,3 glucan, (\beta)-1,4- glucan</td>
</tr>
<tr>
<td>CBM22</td>
<td>xylan</td>
</tr>
<tr>
<td>CBM32</td>
<td>galactose, lactose, polygalacturonic acid, LacNAc</td>
</tr>
<tr>
<td>CBM35</td>
<td>xylans, mannans, mannooligosaccharides, (\beta) -galactan</td>
</tr>
<tr>
<td>CBM48</td>
<td>glycogen</td>
</tr>
<tr>
<td>CBM50</td>
<td>chitopentaose, peptidoglycan degrading enzymes such as peptidases and amidases</td>
</tr>
</tbody>
</table>

* Activities as identified in available databases [133].
REFERENCES


155. Mingkun L, Copeland A, Han J (2011) DUK. Walnut Creek, CA, USA: JGI.


