Your Microbes are Super Models: Using Microbiome data to Reproducibly Predict Crohn's Disease

---

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Moira Differding

May 2017

Approved for the Division
(Biology)

_____

Anna Ritz

Acknowledgements

Many people contributed to the making of thesis.

I can't thank my family and friends enough. I appreciate you all more than I can put into words. Thank you for making me take breaks, helping me relax, and listening when I needed it.

Thank you to Anna, who was an amazing help and the best advisor I could've asked for the whole way through. Your help with python and debugging my scripts was indispensible, and I appreciated your guidance whenever I got stuck on something.

Thank you to Kristin, for being a great boss and sending interesting tech projects my way during work, and for encouraging me to learn more R code which I use so much now. To Chester, who helped me sort out the taxa plots in R, and who designed the wonderful R thesis template that worked so seamlessly. To Jay, for teaching the microbiome seminar that pulled me into this topic. And finally...

To the enigmatic Gevers et al., whose wonderful data made this thesis possible.

Table of Contents

List of Figures

Abstract

Every healthy person's gut is inhabited by up to a 100 trillion microbial life forms. Together, these diverse colonies of bacteria, fungi, and viruses form a unit known as the human microbiome, and without them, we might never have evolved. The microbiome is thought to actively keep our bodies healthy in a variety of ways, including promoting the activation of immune cells to fight infection and regulating our metabolisms. Dysbiosis of the microbiome is strongly associated with Crohn's Disease, which causes ulcers, chronic gastrointestinal inflammation, and deregulation of the immune system.

In an effort to characterize the effects CD has on the microbiome, a 2014 study by Gevers et al. sequenced the microbiomes of 447 pediatric patients newly diagnosed with Crohn's and 221 non-IBD controls. While Gevers et al.'s data was made available online, the model they use to predict CD with their data is not, leaving a gap in reproducibility. This thesis demonstrates that creating a fully reproducible microbiome model to predict CD is not only possible, but can also be done using free bioinformatics software on a typical laptop. The model created here accuractely predicts disease state at a rate of ~82.9 percent on average. In creating this model, this thesis demonstrates one of many practical applications that the intersections of the microbiome, Crohn's Disease, and bioinformatics can have on public health.

Chapter 1

Introduction

1.1 You've got bugs

Every healthy person's gut is inhabited by up to 100 trillion microbial life forms (Ley, Peterson, & Gordon, 2006). Together, these diverse colonies of bacteria, fungi, and viruses form a unit known as the human microbiome, and without them, we might never have evolved (Dethlefsen, McFall-Ngai, & Relman, 2007; Kinross, Darzi, & Nicholson, 2011). Microbial cells outnumber human cells 10 to 1, while their number of genes exceed ours at least 100 to 1 (S. R. Gill et al., 2006). Having co-evolved with us, they can play roles as commensals, where the microbe benefits without significantly affecting the host, symbiotes, where both microbe and host benefit from each other, and pathogens, where they cause harm or disease in the host (Dethlefsen et al., 2007). Classifying our gut bacteria into these three roles is often difficult, as they can behave differently depending on the hosts' health and a variety of other environmental factors (Dethlefsen et al., 2007).

An easy example to demonstrate this balance is *Clostridium difficile* (C. difficile), a bacteria whose infections were associated with 29,000 deaths in the US in 2011 (Leffler & Lamont, 2015). C. difficile infection symptoms can range from frequent diarrhea to colitis, or inflammation of the colon (Leffler & Lamont, 2015). A young to middle-aged healthy person is generally protected from C. difficile infections by their gut microbiomes and immune system (Leffler & Lamont, 2015). C. difficile is a major obstacle in hospital environments, however, as they contain a much larger proportion of people whose microbiomes and immune systems have been disrupted, either by antibiotics, illness, or wounds, and are more susceptible to C. difficle infections (Leffler & Lamont, 2015). Further, C. difficile have been found to colonize infants without causing any adverse effects, demonstrating that its presence doesn't always guarantee illness (Leffler & Lamont, 2015). The gastrointestinal (gut) microbiome in particular is especially enriched in bacteria, and is thought to actively keep our bodies healthy in a variety of ways (Ley et al., 2006; Turnbaugh et al., 2007). Recent studies suggest it plays a critical role in human health, including regulating our metabolic energy

efficiency, clearing bacterial infections via hematopoesis, and relieving inflammation (Donohoe et al., 2011; Khosravi et al., 2014; Mazmanian, Round, & Kasper, 2008). While the specific taxa, or different kinds of bacteria, making up different healthy people's gut microbiomes can vary significantly, the functions these bacteria perform within them are thought to remain consistent (Donohoe et al., 2011). When these baseline functions become imbalanced, a microbiome is said to be in dysbiosis, which can have significant effects on the host.

Dysbiosis of the gut microbiome is postively associated with a number of diseases, including obesity, asthma, and liver disease (Arrieta et al., 2015; Schnabl & Brenner, 2014; Turnbaugh et al., 2009). While restoration of the microbiome with fecal transplants can ameliorate recurrent *Clostridium difficile* infections, rectifying diseases involving relatively unknown etiological factors like those listed above or Crohn's Disease (CD) isn't as simple (Colman & Rubin, 2014; Youngster et al., 2014). There is no one factor that induces CD (Gevers et al., 2014). Current research suggests a tangled combination of environmental, genetic, and microbial elements are responsible for the disease, but the specifics of these and their interactions with each other remain unclear (Gevers et al., 2014). Attempts to elucidate these relationships are further complicated by the lack of a sufficient CD animal model to test on (Antoniou et al., 2016). IBDs like CD are well-suited for gut microbiome studies because of the recognizable dysbiosis they cause (Manichanh, Borruel, Casellas, & Guarner, 2012). While certain taxa have been associated with CD in a variety of different cohort studies, these links aren't strong enough to predict if a patient has CD with a chance much greater than flipping a coin (Gevers et al., 2014).

Recent advances in meta'omic technology, including the ability to quickly sequence the genes belonging to microbiota, now enable further investigation into the etiology of CD, however, and researchers are optimistic in designing new treatment plans based on this knowledge (Manichanh et al., 2012). These new plans could include fecal transplants as for C. difficile, specific dietary recommendations, or even help deciding which currently prescribed medications would work best with each patient (Colman & Rubin, 2014; Gevers et al., 2014; Manichanh et al., 2012). Exploring the extensive interactions between the microbiome and CD using bioinformatics could elucidate new correlations to guide future studies, and so these relationships are the focus of this thesis, as seen in Figure 1.1.

1.1.1 Goals of this Thesis

The goals of this thesis were two-fold: 1) building a model to predict CD using microbiome data and bioinformatic programs, and 2) assessing the reproducibility of an open-access microbiome paper's results. Included are the procedures used in an attempt to replicate those detailed in Gevers et al.'s 2014 paper titled "The Treatment-Naive Microbiome in New-Onset Crohn's Disease" and my own downstream analyses performed on their unprocessed data to build a model predicting the disease state of their samples (Gevers et al., 2014). Because Gevers et al. processed their

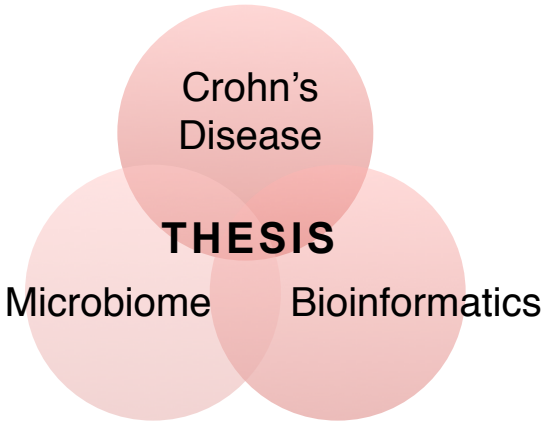


Figure 1.1: This thesis explores the relationships between bioinformatics, Crohn's Disease, and the microbiome, and is visually represented in the above venn diagram.

Illumina sequence data using Qiita and then made it open-access, their pipeline steps, parameters, and data at each stage are available for download on the Qiita website under study IDs #1939 and #1998 (Robbins-Pianka, 2015). Their data's availability allowed me to verify that the data I processed looked identical to theirs, and so the differences in our models do not stem from processing the samples differently.

Gevers et al.'s data was chosen for its open-access data, treatment naive population, and sequencing depth, as summarized in Figure 1.2. The RISK cohort[1] that Gevers et al. sampled from is one of the largest CD microbiome cohorts, and is especially unique in that only pediatric patients who had not been treated for CD before were sampled (Gevers et al., 2014). Further detail on the composition of this study is located in the Results chapter (Gevers et al., 2014).

1.2 From one genome to millions: Bioinformatics and the Microbiome

While work on the microbiome dates back to at least 1956, recent advances in sequencing technology over the last decade have made studying it much cheaper, opening the field up to further investigation (J. Gregory Caporaso et al., 2011; Sayers et al., 2011). As a result, the number of microbiome related studies has increased exponentially over the last decade, as seen in the histogram in Figure 1.3 (Sayers et al., 2011). In 2001, when the first sequenced human genome was published, its estimated cost was $300,000,000, and it took 251 researchers working on it years to complete (Pushkarev, Neff, & Quake, 2009). By 2009, however, the use of new sequencing technologies allowed the cost to decrease to $48,000 with 3 authors working together (Pushkarev et al., 2009). This enormous decrease in costs allowed more researchers to start microbiome studies (J. Gregory Caporaso et al., 2011). One of

[1] RISK = Risk Stratification Study

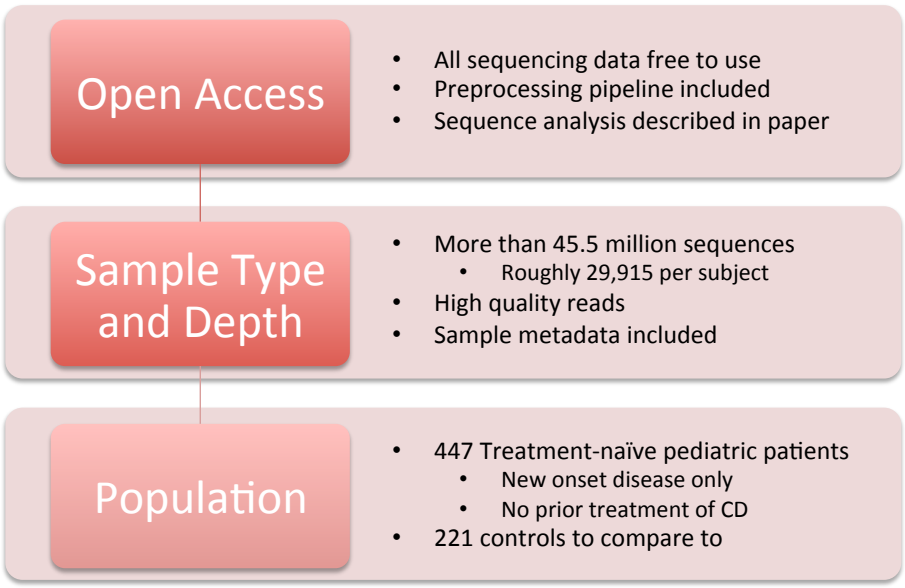


Figure 1.2: A summary of the main reasons that Gevers et al.'s treatment naive population paper was chosen for this thesis (Gevers et al., 2014).

the most commonly used sequencing platforms today is the Illumina MiSeq v2, which can produce 4.5 billion basepairs in 24 hours for only $1015 (Glenn, 2011). The DNA sequences are only produced in approximately 300 bp chunks, however, and a massive amount of computational resources are needed to process the data (Glenn, 2011). In order to keep up with this sudden deluge of data, a surge of bioinformatics programs entered development (Glenn, 2011).

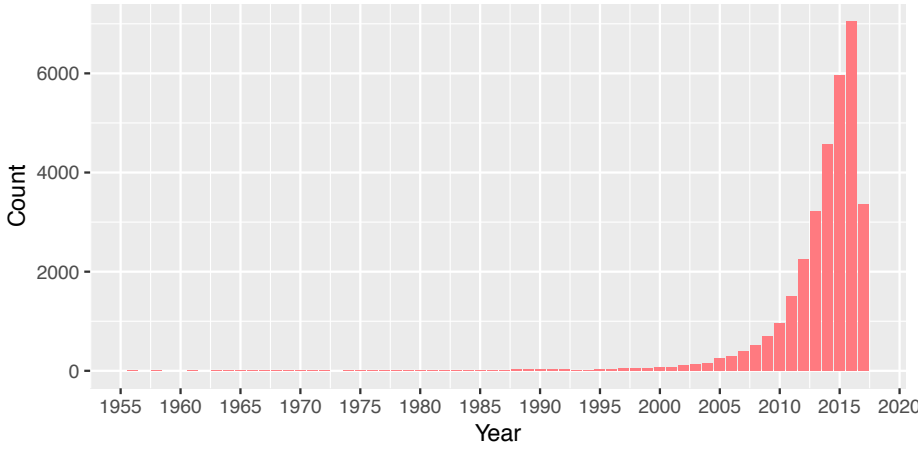


Figure 1.3: The number of Pubmed studies on microbiomes by year (Sayers et al., 2011).

As a result of this intensified sampling of the microbiome, methodologies regarding data analysis have rapidly evolved, but not always in a unified way (Gerber, 2014). While researchers tend to follow a select few sequencing protocols, dozens of software packages

and databases for platforms like Python and R are invariably mixed and used for data analysis, generating hundreds of unique data collection and analysis combinations (Noecker, McNally, Eng, & Borenstein, 2017; Weiss et al., 2016). Amongst this complicated affair, a simpler solution emerged: publishing the raw data (J. Peterson et al., 2009). By publishing sequences and the data pipelines used to process them, microbiome analyses are becoming increasingly reproducible (Ravel & Wommack, 2014).

While a variety of factors still affect the sequencing data itself, extensive supplementary methods published alongside results help alleviate this. Additionally, a multitude of studies focus on helping researchers account for these hidden variables, including potential biases introduced by sample type, storage, processing time, bacterial blooms, contaminants, and primers (Nsubuga et al., 2004). These specifics fall outside the scope of this paper, but one prime example of these biases are fecal samples (Nsubuga et al., 2004). One methodological study found that stool consistency had a significant effect on the types of bacteria found within it (Vandeputte et al., 2015). Most studies don't publish the consistency of the stool samples used, so it is impossible to account for this bias, especially in papers with hundreds of samples (Vandeputte et al., 2015). Additionally, more studies raise serious concerns about the validity of benchmarking fecal bacterial taxa for clinical applications (Nsubuga et al., 2004). For example, Hepatic encephalopathy (HE), a disease involving loss of brain function due to liver malfunction, is linked to gut microbiome dysbiosis, but fecal samples cannot distinguish between patients with the disease and those with unreleated cirrhosis (Bajaj et al., 2012). More generally, detection of specific bacterial enterotypes is associated with fecal water content (Vandeputte et al., 2015).

To avoid issues related to fecal samples, some studies choose to sequence biopsy samples instead, which are more stable and are thought to be more representative of patient microbiomes (Gevers et al., 2014). Intuitively, this makes sense, as the fecal samples only represent taxa that are leaving the gut, and not necessarily what is still there. Biopsies are more invasive, however, and there are risks involved with collecting them from the gastrointestinal (GI) tract, which are detailed in the next section (Shergill et al., 2015). As medical doctors usually require a biopsy sample taken during a colonoscopy to diagnose CD, however, obtaining them for CD microbiome studies is not overly difficult (Dignass et al., 2010). Further, because CD most commonly affects the colon, where biopsies are taken during colonoscopies, these samples represent a snapshot of the microbiome in the area that CD affects the most (Dignass et al., 2010). The larger obstacle then is obtaining biopsies from control patients to compare them to; in Gevers et al.'s data, for instance, there are many more CD than control samples, which can imbalance downstream analyses if not accounted for (Gevers et al., 2014). The control samples in Gevers et al.'s data are from patients who initially displayed IBD symptoms, including diarrhea and abdominal pain, but received a negative IBD diagnosis after a colonoscopy and biopsy found no signs of GI inflammation (Gevers et al., 2014).

1.3 Crohn's Disease and Diagnostics

Crohn's disease (CD) causes chronic gastrointestinal inflammation and deregulation of the immune system (Dignass et al., 2010). Its characteristic ulcers manifest most commonly in the ileal lumen and colon, and while treatable, there is no cure (Dignass et al., 2010). Diagnoses have increased worldwide over the last few decades, with the highest rates in Europe at 322 per 100,000 persons and in North America at 319 per 100,000 persons (Laass, Roggenbuck, & Conrad, 2014). A variety of diagnostic methods are required to confirm the presence and severity of the disease (Shergill et al., 2015). For an approximate visual aid of the relevant gut anatomy, please refer to the diagram in Figure 1.4.

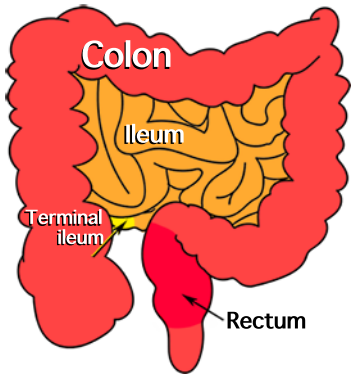


Figure 1.4: A simplified diagram of the intestinal anatomy most affected in Crohn's Disease. The small intestine, or ileum (orange), feeds into the large intestine, or colon (red), at the terminal ileum (yellow) using a valve (not pictured). In this thesis, samples from both the terminal ileum and rectum (in dark pink) are used.

Endoscopy with biopsy is the most common method for diagnosing and monitoring Crohn's (Dignass et al., 2010). Considered safe for all ages and recommended as the first diagnostic tool when colitis is suspected, gastroenterologists usually require one to rule out other GI diseases (Dignass et al., 2010; Terheggen et al., 2008). While clinical, radiological, and serological techniques are also used to diagnose Crohn's, endoscopy is the most reliable for specific disease diagnosis (Shergill et al., 2015). Despite the benefits, endoscopies are invasive and typically require anesthetization (Shergill et al., 2015). The most concerning risks include GI perforation, bleeding, infection, and sedation-induced cardiopulmonary events (Deas Jr & Sinsel, 2014). Patients with electromagnetic implants such as pacemakers must be constantly monitored during the procedure, as the endoscopy can interfere with their electromagnetic rhythms (A. C. Travis, Pievsky, & Saltzman, 2012). While colonscopies allow detailed examination of the patients' illeal lumen and colon and give the opportunity to take biopsies, they require the patient to flush out their intestines with an electrolyte solution first (Jalanka et al., 2014). This flush can reportedly reduce the patient's microbial load by 31-fold, and while it usually recovers to baseline levels within 14 days post-procedure, lasting differences in bacterial abundances within taxa such as *Proteobacteria* and *Clostridium* can linger (Jalanka et al., 2014).

1.3.1 The Crohn's Microbiome

Dysbiosis of the microbiome is strongly associated with CD (Gevers et al., 2014). Microbial dysbiosis means that there is some sort of imbalance in a bacterial community, and does not indicate any specific taxa or patterns, only that there is a disruption from normal behavior or composition (H. Morgan Xochitl C., 2012). Because current clinical disease indices were not designed to work with microbiome data, however, a number of studies have attempted to build new, microbiome-based diagnostic measures (Gevers et al., 2014). A simple logistical problem is partly to blame for the difficulty in modeling CD: studies attempting to characterize the microbiomes of patients with diseases that tend to decrease bacterial diversity can have trouble collecting a large enough sample size or sequencing depth (Gevers et al., 2014). Finding a willing cohort of treatment naive CD patients large enough to get statistically significant results is already difficult, and because patients experience the disease in different subsections of the intestine at vastly different severities, deciding on a single sampling location can be tricky (Dignass et al., 2010). As these sites have different healthy base microbiomes to begin with, it's difficult to compare them, and even more difficult to draw generalized conclusions about the disease state from them (Laass et al., 2014). Additionally, a variety of treatments patients take to combat the disease and allow for a reasonable quality-of-life include antibiotics, steroids, and immunotherapies, which can dramatically alter the microbiome and skew results of what a diseased Crohn's microbiome looks like (Dignass et al., 2010; Gevers et al., 2014). It's not always possible for studies to take samples during both patient's Crohn's flareups and relatively undisturbed periods either, and so microbial abundancy differences between these often cannot easily be accounted for (Gevers et al., 2014). Despite these difficulties, however, new studies are continuing to build and improve models demonstrating associations between CD and the microbiome (Gevers et al., 2014).

1.4 DNA Sequencing and Processing Pipelines

As the vast majority of bacteria cannot be cultured, gut microbiome studies commonly use either 16S rRNA amplicon sequencing (16S) or whole genome shotgun sequencing (WGS) (Hamady & Knight, 2009). Most microbiome studies utilize 16S analysis to group sequenced bacteria, as every bacterial genome contains a species-specific 16S gene that functions as a barcode, allowing for accurate identification of taxa while avoiding the costs of WGS (Hamady & Knight, 2009). While 16S sequencing typically focuses on what bacteria are present in the gut and in what abundance relative to other bacteria, WGS gives a snapshot of what the total gene content in the community is like (H. Morgan Xochitl C., 2012). While WGS generally produces vastly more sequence data than 16S, the cost is much greater, especially if it is performed on hundreds of samples, and it can require more computational resources to process (H. Morgan Xochitl C., 2012). Both methods require different bioinformatic tools to infer these conclusions from the raw sequencing data, but Figure 1.5 summarizes the biggest

differences between the two methods in a basic pipeline.

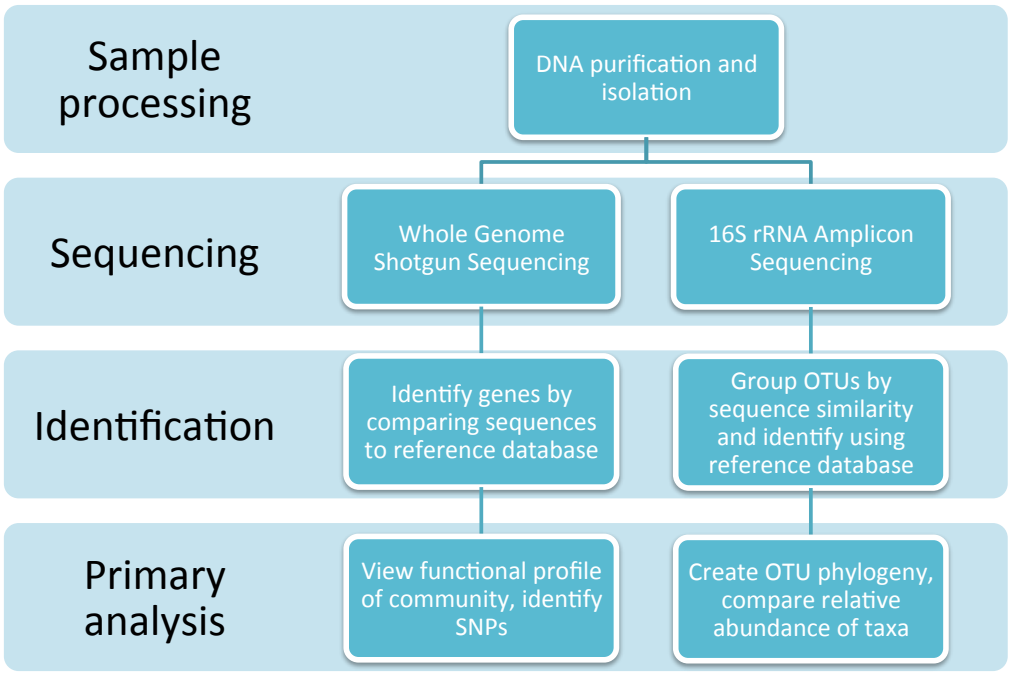


Figure 1.5: A pipeline summarizing the basic steps involved in Whole genome shotgun sequencing (WGS) on the left and 16s rRNA amplicon sequencing (16S) on the right. The Primary analysis step shows the most reliable data that can be inferred for each method, but is not exhaustive (Huttenhower and Morgan, 2012).

WGS more accurately predicts the potential roles the gut microbiota play within their host, as the short sequences it produces can be directly aligned and matched to genes in reference databases (Hamady & Knight, 2009). For 16S amplicon sequencing, gene content can only be inferred from gut taxa with a matching full sequence or from other closely related taxa (Hamady & Knight, 2009). Because it doesn't amplify DNA using PCR, however, less abundant taxa are less likely to be represented in the pool of sequenced DNA unless a sufficient number of reads are reached (K. Chen & Pachter, 2005). 16S, however, can avoid this read bias by amplifying the same gene using universal primers (Hamady & Knight, 2009). In order to be cost-effective, however, 16S sequencing is usually performed on a few lesser-conserved regions of the full ~1500 bp 16S gene, potentially allowing for some primer bias (K. Chen & Pachter, 2005). Gevers et al. performed mostly 16S amplicon sequencing in their 2014 paper, so 16S will be the focus for this thesis.

1.4.1 16S Analysis

For 16S analysis, researchers select a subset of the 16S gene, amplify it using PCR, shotgun sequence it, and then identify taxa by comparing these sequences to reference genomes, as shown in Figure 1.5 (Hamady & Knight, 2009). While shotgun 16S cannot

accurately classify bacteria at the species level, it allows researchers to group bacteria into Operational Taxonomic Units (OTUs), which are typically assigned at 97% similarity (Noecker et al., 2017). The bacterial 16S gene contains nine hypervariable regions (V1-V9) ranging from about 30-100 base pairs long that are involved in the secondary structure of the small ribosomal subunit, as seen in Figure 1.6 (Gray, Sankoff, & Cedergren, 1984; Lee & Gutell, 2012).

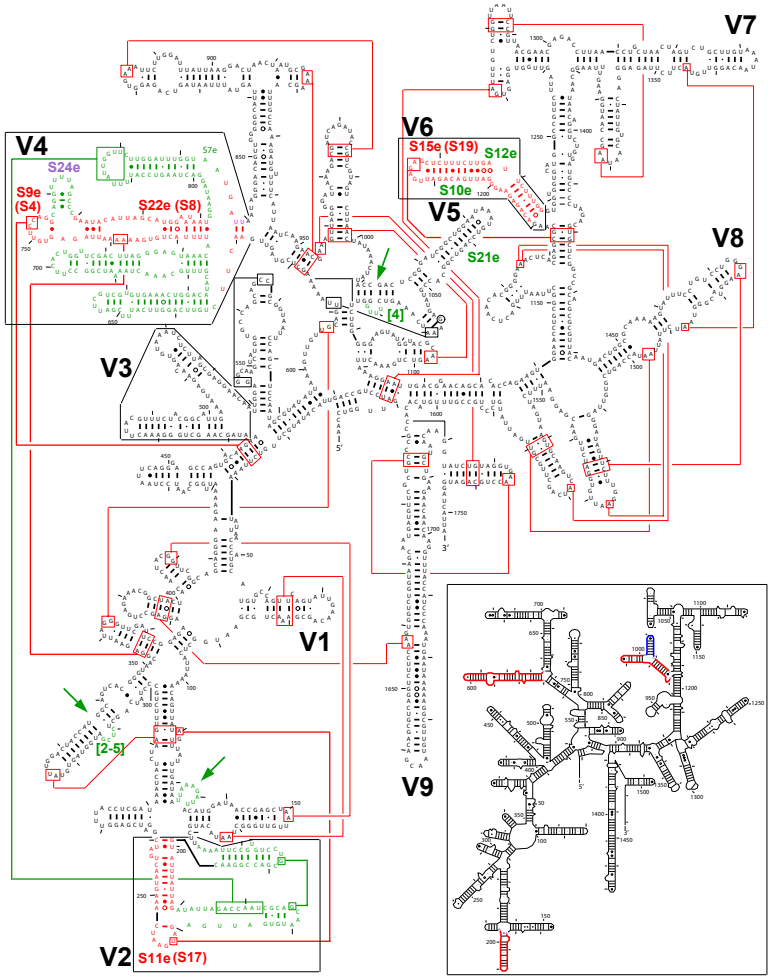


Figure 1.6: A visualization of the 16S rRNA nucleotide sequence and structure, with hypervariable regions labeled V1-V9 (Lee and Guttell, 2011). Image © 2011, Guttell and Lee. Used with permission.

The degree of conservation varies widely between hypervariable regions, with more conserved regions correlating to higher-level taxonomy and less conserved regions to lower levels, such as genus and species (Yang, Wang, & Qian, 2016). While the entire 16S sequence allows for comparison of all hypervariable regions, at approximately 1500 base pairs long it can be prohibitively expensive for studies seeking to identify or characterize diverse bacterial communities (Yang et al., 2016). These studies commonly utilize the Illumina platform, which produces reads at rate 50-fold and 12,000-fold less expensive than 454 pyrosequencing and Sanger sequencing, respectively

(Bartram, Lynch, Stearns, Moreno-Hagelsieb, & Neufeld, 2011). Although cheaper and allowing for deeper community coverage, Illumina sequencing only produces reads 75-150 bp long, and has no established protocol for reliably assembling the full gene in community samples (Burke & Darling, 2016). Multiple hypervariable regions can be assembled from a single Illumina run, however, making them ideal targets for the platform (Burke & Darling, 2016). While 16S hypervariable regions can vary dramatically between bacteria, the 16S gene as a whole maintains greater length homogeneity than its Eukaryotic counterpart, which can make aligning them easier (Van de Peer, Chapelle, & De Wachter, 1996). In addition to this, the 16S gene contains highly conserved sequences between hypervariable regions, enabling the design of universal primers that can reliably produce the same sections of the 16S sequence across different species (Weisburg, Barns, Pelletier, & Lane, 1991). While these regions cannot produce a full taxa estimate from Domain to Species as accurately as the full gene, the unique features of each region enable the accurate prediction of different taxonomic components (Yang et al., 2016). Many community studies select semi-conserved hypervariable regions like the V4 for these analyses, which can provide resolution at the phylum level as accurately as the full 16S gene (Yang et al., 2016). Hypervariable regions with the lowest conservation, like the V1 and V2 regions, are used to resolve species and genus level taxonomy when the higher order taxonomy is already known (Yang et al., 2016). While 16S analysis is a powerful tool for bacterial taxonomic studies, it struggles to differentiate between closely related species (Větrovsky & Baldrian, 2013). In the families Enterobacteriaceae, Clostridiaceae, and Peptostreptococcaceae, species can share up to 99% sequence similarity across the full 16S gene (Jovel et al., 2016). As a result, the V4 sequences can differ by only a few nucleotides, leaving reference databases unable to reliably classify these bacteria at the genus or species levels (Jovel et al., 2016). By limiting 16S analysis to select hypervariable regions, these studies can fail to observe differences in closely related taxa and group them into single taxonomic units, therefore underestimating the total diversity of the sample (Větrovsky & Baldrian, 2013). Furthermore, bacterial genomes can house multiple 16S genes, with the V1, V2, and V6 regions containing the greatest intraspecies diversity (Coenye & Vandamme, 2003; Van de Peer et al., 1996). While not the most precise method of classifying bacterial species, analysis of the 16S hypervariable regions remains one of the most useful tools available to bacterial community studies (Jovel et al., 2016). The vast majority of microbiome analyses include 16S sequencing data, but the ways they process them and the models they include them in vary dramatically (Differding, 2017).

1.5 Machine Learning for Taxa Classification

Machine learning simply means that a computer builds the prediction model using a given set of variables, or predictors, without explicit instructions from the programmer detailing every step (Touw et al., 2013). It does so by identifying trends or patterns within the data and then using these to predict the outcome of interest chosen by the

programmer (Breiman, 2001). Supervised machine learning means that the computer is given the actual outcome for a subset of the data that it uses to determine which variables are the best predictors of the designated outcome (Breiman, 2001). Consider this example: Say you give a computer a dataset consisting of hundreds of oranges and apples, using type of fruit as the outcome and the color and shape of each fruit as predictors. You tell the computer the type of fruit for a randomly selected 2/3 subset of the data (training data), and then ask it to choose the variables that accurately predict the type of fruit. The computer determines that color is the best predictor, despite not being told this by the programmer, and tests this theory on the remaining 1/3 of the dataset (test data). The final model built by the training data is evaluated by how well it performs classification on the test data.

In this thesis, I use random forests to predict a categorical variable (disease state -either CD or no) as the outcome. Random Forests are supervised machine learning algorithms capable of quickly and accurately classifying noisy data (Touw et al., 2013). They're well suited to 16S microbiome data for this reason, as even the taxa of healthy patients can vary hugely (Touw et al., 2013). Random forests can also predict continuous variables using regression, but these will not be considered here. Random Forests perform well on noisy data, or data with many variables that don't correlate to the class, without excess overfitting because they utilize randomness in two ways: random observation sampling and random variable selection (Breiman, 2001). By training the model's classification trees on different subsets of the training population using different variables each time, the random forest model can determine which variables are the most important in classifying the outcome across different subsets of the population (Breiman, 2001). The chosen variables from the training set are then validated against the test data in order to measure their performance (Breiman, 2001).

1.5.1 A Random Forest Example

To understand how a random forest model works, consider the following toy example. Imagine you have a dataset consisting of 100 patients. Half of them have contracted a new disease called "Poochophobia", which embeds an instantaneous phobia of dogs upon infection. Researchers have determined that the disease is likely the result of an environmental trigger and a combination of genetic factors, and so have sequenced a number of genes associated with altered behavior and performed a survey of lifestyle choices of the patients. You need to make a Random Forest model to find what variables are most associated with the disease so researchers know which gene to investigate. The predictors here are smoking (yes/no), pet dog (yes/no), if they live within 1 km of an old toxic waste dump (yes/no), and the presence of gene1, gene2..., gene49, and gene50 (all yes/no). The outcome, or target variable, you want to predict is if the patient has the disease or not. A tree is a graph of nodes connected by branches, with all of the data starting in a single node and different subsets of that data in the following nodes. A simple binary classification tree is a structure that

takes your data as the root of the tree, or parent node, chooses a variable to predict the class of each sample, then splits the data into two final branches, or terminal nodes, representing either class (Loh, 2011). To visualize what this structure looks like, look at the annotated tree in Figure 1.7, which describes each step the tree makes. If a tree can't classify all of the data using one variable, it splits the data into one two nodes:one terminal node containing only one class, and one child node containing both classes, which are then grown until all of the data is classified into terminal nodes (Loh, 2011). A model containing multiple trees is called a forest. Classification trees can be tuned using a variety of parameters, two including index measures to split nodes and mtrys (Loh, 2011). Before running the model, you decide the number of classification trees (ntree) you want to use and the number of variables (mtry) you randomly select from to sort the data in each tree (Breiman, 2001). Random forests use multiple trees, where the final classification for each sample is made by averaging the individual assignments from each tree where it appeared. Having an excess of trees is crucial to the model's success, as you want each sample and variable to be used at least one time (Breiman, 2001). If they're not, you could miss finding associations within your data (Breiman, 2001).

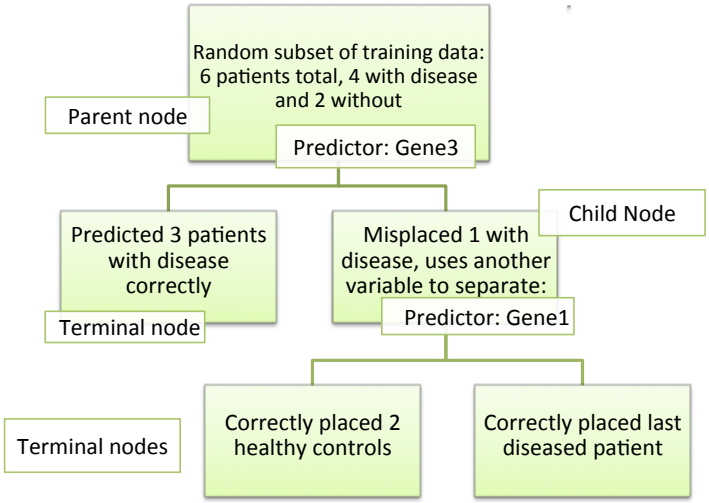


Figure 1.7: A single tree generated by the poochophobia data random forest model using 1000 trees and 1 variable predictor per node as parameters. Gene3 correctly predicts some of the observations (patients), but the ones it cannot proceed into a new node and are then split by a new, randomly chosen variable (Breiman, 2001).

For your model, you decide to run a random forest algorithm in `R` with all the variables, generating 1000 trees. Every tree is built by randomly sampling your training data with replacement 1000 times, which are put in the parent nodes. Instead of choosing from all the variables for each decision, a smaller, randomly selected subset is used each time. One variable is chosen from this pool for its ability to accurately split the observations (here, the patients) (Breiman, 2001). After every tree is made in this fashion, each tree gets a "vote" on what class the samples it contains belong to (Breiman, 2001). These votes are averaged across all the trees, and the class with

the highest average number of votes is assigned to each sample. After the training data is processed, the model decides the best way to use these to predict the disease. To see if these are accurate, it then uses the associations between the variables and classes it made from the training data on unlabeled test data. It then assigns the test data classes using trees as in the training data, and checks how accurate it was at the end. The error rate can then be used to determine how well the model performed (Breiman, 2001).

Further details of random forests are included in the methods chapter.

Chapter 2

Methods and Results

2.1 Methods

The methods in this chapter describe how a model was built to predict CD using Gevers et al.'s microbiome data and to assess the reproducibility of their results. Their data was grouped into three categories by sample type to build three models, including the terminal ileum, rectum, and fecal samples. The exact processing steps and scripts used by Gevers et al. on their data can be viewed on their Qiita studies listed in the introduction, which are the same as the ones performed in this thesis. The purpose of these processing steps are generally explained in the following sections. For general definitions of some of the terms and techniques used in this thesis, please see the Appendix 1. The analysis was performed using the R program language and scripts, which will be made available in a github repository.

2.1.1 From DNA to Data

Microbiome researchers are increasingly publishing their raw data and analyses on open-source databases, giving these studies the potential to be highly reproducible (Ravel & Wommack, 2014). There are hundreds of free online tools available, most commonly available as `Python` scripts and `R` packages, to analyze microbiome sequences. The following sections give a brief description of the basic steps used to process sequencing data in one commonly used pipeline and to create the random forest model (J. G. Caporaso et al., 2010).

Reference Databases

A reference database is essentially an online library full of annotated sequence files, either full genomes or fragments such as 16S regions, that can be used to help researchers determine what taxa or genes their experimental data contains (McDonald

et al., 2012). Annotated reference database sequences are compared to experimental sequences in order to assign gene content or taxa (McDonald et al., 2012). For many 16S community studies, BLAST searching experimental 16S data against GenBank's stored 16S references is not sufficient for taxa assignment, as over half of GenBank's 16S sequences are only annotated to the Domain level (McDonald et al., 2012). To fill this gap, researchers have collaborated to construct and curate databases dedicated to storing and annotating 16S reference genes (McDonald et al., 2012). Two of the most commonly used for microbiome analysis are Greengenes and Silva, which are both accessible online (McDonald et al., 2012; Quast et al., 2012). Greengenes is managed by a handfull of researchers internationally, including the University of Colorado, University of Queensland, and Second Genome Inc., and contains only 16S rRNA sequences. Currently, Greengenes only offers downloads of its reference sequences for researchers to utilize (McDonald et al., 2012). Silva, which is managed by the Max Planck Institute for Marine Microbiology and Jacobs University in Germany, maintains a wider spectrum of rRNA sequences for all domains (McDonald et al., 2012). Gevers et al. used Greengenes to process their sequences, as did this thesis (Gevers et al., 2014; McDonald et al., 2012).

QIIME (Quantitative Insights Into Microbial Ecology):

Qiime is a powerful open-source software that processes raw DNA sequencing data and outputs a broad range of microbial analyses (J. G. Caporaso et al., 2010). Qiime runs the most commonly used data processing pipeline for microbiome analyses, and can installed on any computer using the command line or terminal (J. G. Caporaso et al., 2010). In 2010, Qiime's managers published an initial paper demonstrating its uses, which has since been cited 7,154 times according to a Google Scholar search of papers in April 2017. Some of this popularity can be attributed to its availablity as both an installable program and online server called Qiita, which allows users to run Qiime analyses on data they upload or source from other studies using Qiita (Robbins-Pianka, 2015). The Qiita server is essentially a more limited online installation of Qiime, and runs the same Qiime commands just as the command line installation does; the only difference is that each processing step in Qiita is visualized in a flowchart style, and not all of the newer Qiime commands are included (Robbins-Pianka, 2015). For the purposes of this section, they will not be differentiated between.
Gevers et al. processed their data according to Qiime's recommended pipeline, which can be summarized in three steps:

Filtering: Demultiplexing and quality filtering raw sequencing data

Grouping: Pick & group OTUs at 97% similarity using the Greengenes database

Analysis: Peform diversity analyses, make PCoA plots, etc.

Gevers et al. used the Illumina MiSeq platform to target the V4 hypervariable region on the 16S gene for their purified fecal DNA samples (Gevers et al., 2014). The Illumina MiSeq produced paired end reads of 175 bp long, which Qiime then sorts

through to remove primers, dispose of the reads exceeding a specified error threshold, and combines reads that have overlapping segments of about 97 bp long (Gevers et al., 2014). These quality filtered reads now undergo the second step of OTU picking. Gevers et al. used closed-reference OTU picking, which means that Qiime takes the reads and compares them to the reference sequences in Greengenes (J. G. Caporaso et al., 2010; Gevers et al., 2014). Reads that don't match a reference sequence are thrown out, so closed-reference OTU picking cannot identify and include new taxa in downstream analyses (J. G. Caporaso et al., 2010).

2.1.2 Random Forest Modeling

Random Forests were used to construct an algorithm to predict disease state in Crohn's patients. In order to explain the different parts of a random forest model and how it can be used for prediction, we'll go back to the toy example of poochophobia from the Introduction.

After running a random forest to see which variables affect the model the most on our training set of data (66 randomly selected patients), we can make a variable importance plot, which shows the relative effect, or weight, that each variable or feature in the data had on the model's prediction ability (Breiman, 2001). The variable importance plot generated by the poochophobia algorithm, seen in Figure 2.1, shows the importance that both genes (gene3 and gene1) and other features had on the poochophobia random forest model (like if the patient lived close to a toxic waste site: toxic_waste).

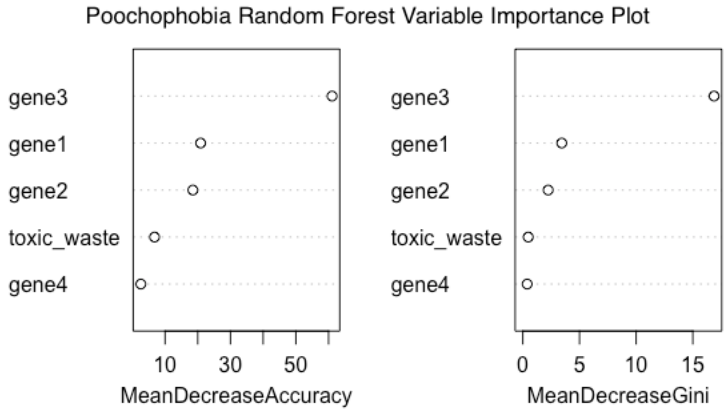


Figure 2.1: The variable importance plot showing both the Mean Decrease in Accuracy (MDA) and Mean Decrease Gini (MDG) generated by the randomForest R package using all variables in the poochophobia dataset (Breiman 2001).

The Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG) seen in Figure 2.1 are indices that account for different accuracy measures (Breiman, 2001). Both unitless measures are specific to the random forest model they're generated for, so comparing the values of different models' variables doesn't give any meaningful

info. They are both calculated using their performance on the test data, which is also called the out-of-bag (OOB) data (Breiman, 2001).

The MDA for each variable is a direct measure of how a variable's presence affects a tree's prediction accuracy (Strobl, Boulesteix, Zeileis, & Hothorn, 2007). After the OOB error is calculated for each tree the variable is in, its values are randomly permuted across patients, and if the re-calculated OOB goes up, it is considered important (Breiman, 2001). In Figure 1.7 from the introduction, for example, if patients were randomly assigned either "yes" or "no" for gene3 and the error rate of the model went up when averaged across all trees, then it would earn a higher MDA. Essentially, it is scored by the average overall accuracy of each tree, and is adjusted to fit the forest after averaging by accounting for the variance in OOB changes.

In contrast, the MDG looks at how often a variable, like gene1 in Figure 2.2, is chosen as the best variable to correctly split the observations into new child nodes, and is not tree-dependent (Breiman, 2001). The MDG is scaled to fit the forest by looking at how many times the variable appeared in each random pool of variables used for splitting and was not used, and so is also referred to as the mean decrease in node purity, because that variables' removal results in more classes being incorrectly split, or less pure, at that node (Touw et al., 2013). In the context of Figure 2.2, if gene1 and gene3 were only present in the variable pools they were chosen in, they would both have the gini score. The final MDG values are further adjusted by comparing the accuracy of the variable in splitting nodes to the accuracy obtained by randomly classifying cases based on the class proportions of the samples in the tree (Breiman, 2001). So if you start with 5 infected and 5 healthy patients in a tree, the gini score reflects how often a variable like gene3 will accurately split the patients at a node compared to randomly assigning patients healthy or infected classes by flipping a coin. A visual example of this is shown in Figure 2.2.

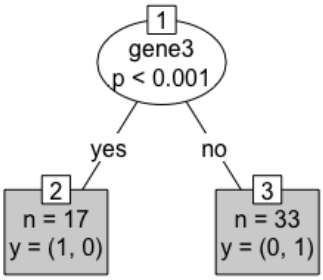


Figure 2.2: A node-splitting scheme showing how MDG values would be calculated for gene3 if the proportion (p) of the test data patients with gene3 was <0.001 and it was included in 50 test data variable splitting pools. Because gene3 is chosen above other variables to correctly assign 17 patients, and $0.34 > 0.001$, gene3 would have a high MDG (Hothorn et al., 2006).

This is done for all trees' nodes that the variable appears in and is averaged, with the MDG value representing how much your model will decrease in accuracy when that variable is removed (Breiman, 2001). So the MDG is a measure of global variable importance, taking into account both the ability of a variable to split nodes with the

lowest impurity across all trees in the model in addition to a single tree basis, whereas the MDA is more of a local measure, as its value can only be assessed on a per tree basis and then averaged. Because it takes into account all trees, MDG is used more often to pick variables in biological models (Touw et al., 2013). The bias of MDG is often reduced by the same measures used to normalize microbiome data, like rarefying (described in the following section), which is used reduce uneven OTU clustering due to differing sequence counts across samples (Touw et al., 2013).

The Random Forest model determined that the presence of gene3 has the largest effect on both MDA and MDG, as it has the highest variable importance in Figure 2.1, but because it alone cannot accurately predict all patients disease state, we want to include gene1 and gene2, the next most important variables, in our model as they might account for this discrepancy (Breiman, 2001).

To get a stable random forest prediction model, you'd want to account for the potential effects that variables could have on model accuracy due to associations with other variables or on specific samples (Touw et al., 2013).To ensure you've included enough trees to account for these interactions, you can plot your OOB error, as in Figure 2.3 for the poochophobia model. In the `randomForest` package, 1/3 of the samples from the data you supply it are randomly chosen as test data to validate your model for each tree.

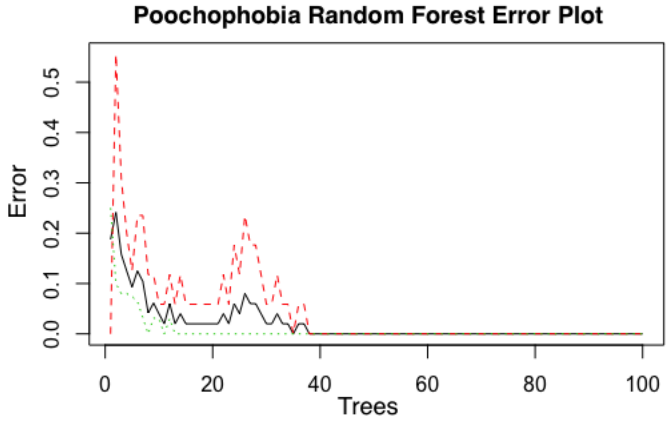


Figure 2.3: The poochophobia random forest model's error plot, where the lines represent the error rates of different accuracy measures, including the OOB total error (black), and the individual class error rates for healthy (red) and sick (green). The minimum number of trees you'd want to include in future models would be around 50, when the error rate has leveled off (Breiman 2001).

After the plot stabilizes around 50 trees, adding more trees to this model is unlikely to improve the model's accuracy and will just take more time to run. Seeing an error rate of 0 for a giant should raise alarm bells, however (Touw et al., 2013). It means that your model has either found a perfect way to predict class, or, more likely, that you've overfit your model and it will predict poorly on new data (Strobl et al., 2007). The poochophobia model is a case of the former, as the dummy dataset was designed so that gene3 was a highly important variable.

Rarefying Data Microbiome sequencing typically produces different numbers of sequences for multiple samples run at once due to the nature of the technology used and quality control techniques (Noecker et al., 2017). For example, this can mean in a single sequencing reaction, one patient's sample got 200 sequences, while another got 500. The amount of sequences per sample is often referred to as sequencing depth, so for the two patients above, we would say that the first patient had a lower sequencing depth than the second. Rarefying is used to minimize the range of, or normalize, uneven sequencing depths across samples. To rarefy data, you set a threshold level of minimum sequencing depth required to be included, and then randomly pick a specified number of sequences without replacement from each collected sample so that all samples include about the same number of sequences(Noecker et al., 2017). The reasoning is that samples with more sequences have more chances to pick up unique OTUs, so an observed increase could be solely due to sequencing depth (Noecker et al., 2017). The downside to rarefying is that while you may avoid grouping samples based on sequencing depth, by discarding the pieces of your data that weren't initially deep enough or randomly selected, you are lowering the potential diversity of every sample, and could be masking an actual difference of OTU diversity (McMurdie & Holmes, 2014). Rarefaction curves are created by thresholding and subsampling at different sequencing depths, and then plotting the results to observed the induced changes in diversity. In the rarefaction curve shown in Figure 2.4, all of the treatment naive samples are subsampled at the depths and threshold minimums marked by the confidence intervals and plotted against the alpha diversity measure Observed OTUs, which is a measure reflecting the frequency of unique OTUs.

While the American Gut Project study found that rarefying from 1,000 to 2,000 sequences did not significantly affect their clustering results, in cases like Figure 2.4 where your lowest sequencing depth is about 150, you dramatically lower your statistical power to detect a difference in diversity (Koren et al., 2012). On the other hand, if your rarefaction curve looked flatter like the antibiotic-free biopsy data plotted in Figure 2.5, you might not be worried (Noecker et al., 2017).

Biostatistians have also argued against it, and a variety of different measures have been created to managing uneven sampling depths, but there is no perfect technique, so the one used will depend on the nature of the study (Koren et al., 2012; McMurdie & Holmes, 2014). The ideal rarefaction curve would show all samples reaching an asympotote before the sequencing depth of the lowest group, indicating that nearly all the diversity of the sample has been accounted for (Noecker et al., 2017). A slope of 0 would require extremely deep sampling, however, and are generally not financially feasible (Koren et al., 2012). Another commonly used normalization technique is relative abundance, which divides each OTU's count by the total counts in that sample (McMurdie & Holmes, 2014).

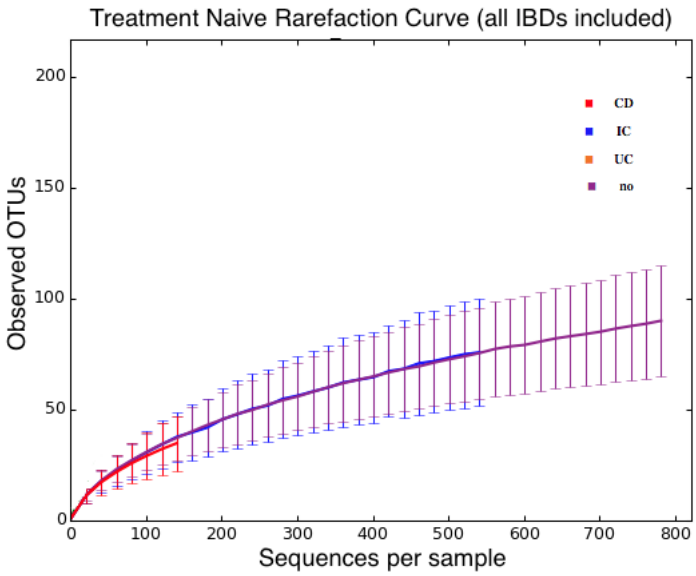


Figure 2.4: A rarefaction curve showing the observed OTUs, an alpha diversity measure showing how diverse the OTUs within a single sample are, against the total number of sequences per sample for CD (red), control (purple), and colitis (blue, covering orange) patients for all treatment naive data. Because the number of control sequences extend far beyond the CD samples, it's possible that any diversity meausure we calculate for these samples is only due to different sequencing depths, but by rarefying to the CD sample depth of about 150 sequences, you would lose much of the diversity to the right (Caporaso et al. 2010; Noecker et al. 2017).

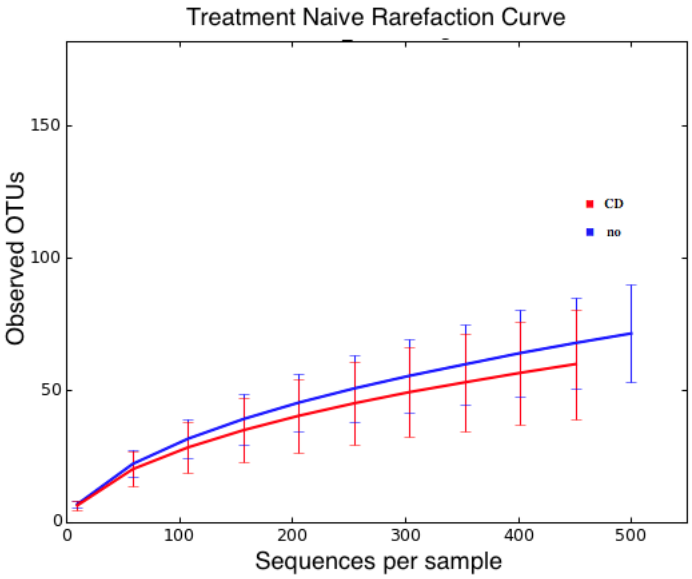
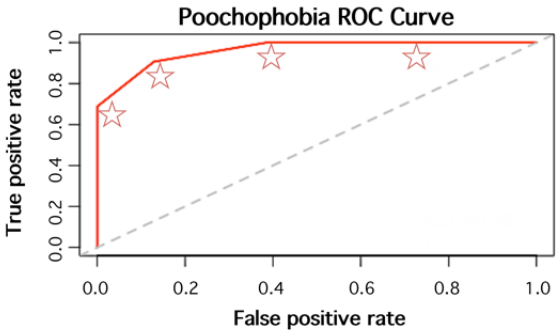


Figure 2.5: A rarefaction curve showing the observed OTUs, an alpha diversity measure showing how diverse the OTUs within a single sample are, against the total number of sequences per sample for CD (red) and control (blue) patients for the antibiotic-free, treatment naive biopsy sample data. If the control sequences' depth ended at 500, we would probably not be losing much diversity by rarefying to a depth of 450 (Caporaso et al. 2010; Noecker et al. 2017).

ROC Curves

A Receiver Operating Characteristic (ROC) curve is a plot showing how well a binary classification model performed (Bradley, 1997). The x-axis is the false positive rate (FPR) and the y-axis is the true positive rate (TPR), with both ranging in value from 0 to 1 (Bradley, 1997).

By definiton, the TPR is the number of correctly identified positive cases (TN) over the total number of identified positive cases, and is represented the equation: $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ (Bradley, 1997). The FPR is the number of cases incorrectly marked positive over the total number of samples labeled negative, and is written $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$ (Bradley, 1997). In the ROC curve plotted for the poochophobia example in Figure 2.6, a correctly identified positive case means an infected patient is classified as infected, while the healthy control samples are negative cases (Poynard et al., 2007). The TPR is also known as the sensitivity of a model, which refers to its ability to correctly identify infected people; a model that prioritizes catching every infected person will have a high TPR to catch all the positives (Poynard et al., 2007). Conversely, the FPR is also known as 1 minus the specificity, which is higher when a model prioritizes catching all infected at the cost of more false positives (Poynard et al., 2007). When plotted together, the area under the curve (AUC) represents the probability of favoring a true positive over a false positive.



Patient#	Sick?	RF P(Sick)	FPR	TPR	Cutoff
50	yes	0.86	0	0.687	0.86
35	no	0.58	0.129	0.906	0.58
6	no	0.34	0.387	1	0.34
12	no	0.08	1	1	0.08

Figure 2.6: An ROC curve generated by the Poochophobia random forest model with an area under the curve (AUC) value of 0.961, indicating that the model can correctly predict all of the infected cases to a high degree without incurring too many false positives. The curve was plotted using the Poochophobia random forest model to test its prediction ability on 4 patient's data (stars), with the table of values used to calculate the ROC curve below (Breiman 2001; Sing et al. 2005).

To generate an ROC curve for a random forest model, the TPR and FPR are calculated for each classified observation by the model, ranked in order of probability, and then plotted as a point on the graph (Bradley, 1997). A model with 100% sensitivity and

100% specificity will look like Figure 2.7, with the curve going straight up and then right across the top of the graph.

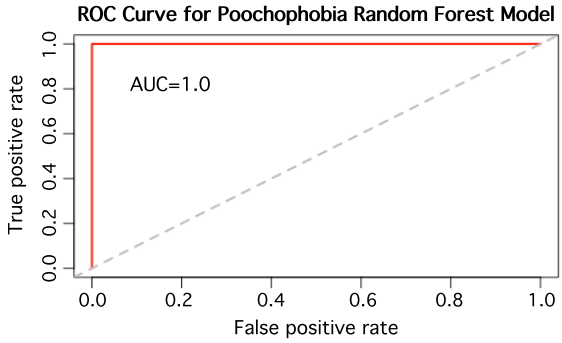


Figure 2.7: The hypothetical ROC curve that would be generated if the Poochophobia random forest model predicted all infected classes accurately and never incurred false positives. The AUC of this curve is 1 (Hajian-Tilaki 2013; Sing et al. 2005).

An ROC curve like this will have an area under the curve of 1.0, signifying that it's a perfect model for predicting your given outcome with the variables you supplied it (Bradley, 1997). The area under the curve (AUC) is an easy way of quantifying the performance of a binary model like a classification random forest (Poynard et al., 2007). If a model has an AUC of 0.5, it essentially means your model predicts an outcome no better than simply flipping a coin (Bradley, 1997). To understand how ROC curves and AUC's are used for prediction of new cases, consider the following scenario. You are an epidemiologist trying to make a model to predict the chance that a person will develop Poochophobia. It is non-infectious, and you are supplied with the diagnoses and gene data of all the infected and a large set of controls who live in the same area. You end up with an ROC curve that looks exactly like the one in Figure 2.6, and now need to decide what point on that ROC curve you want to take as the cutoff point to predict if someone will get infected. The ideal cutoff will change depending on the nature of disease: if people who develop it suddenly die within 24 hours, you'd want the highest sensitivity possible to make sure you didn't miss anyone, but a lot of people who wouldn't get the disease will also be pre-treated as a result (Hajian-Tilaki, 2013). If it's a very minor disease, like a small cough for a week, you'd probably want a lower false positive rate, so that you don't waste too many resources treating people who won't get the disease (Hajian-Tilaki, 2013).

2.2 Results

All of the following results are entirely reproducible, and the code or scripts used to make each will be available in a github repository. All OTU processing and analyses were performed using Qiime, while both Qiime and R were used to make the figures. For the following analyses, the treatment naive data was filtered to include only samples that met the standards described in section 3.1.

2.2.1 Treatment Naive Microbiome Composition

In order to perform further downstream analyses, the OTUs chosen during closed reference OTU-picking in Qiime were first visualized at higher taxonomic levels as a last, approximate quality check. For the compositional plotting, the mean relative abundance of each phylum's OTUs was calculated by summing each sample's OTU counts by phylum and then dividing by the total number of OTU counts, so that each sample's relative abundances would sum to 1. A visual representation of the relative abundance of a single OTU in one sample can also be represented by the following equation, where `RA` is the Relative Abundance of OTU `i (O[i])` in sample `s`, given `n` unique OTUs in a sample.

$$\mathrm{RA[s, i]} = \frac{\mathrm{O[i]}}{\mathrm{O[1]+O[2]+...O[n]}}$$

Following this, the mean relative abundances for a phylum of all CD patients could be represented by this next equation, where `MRA` is the Mean Relative Abundance of a single phylum `i (P[i])` in all `CD` samples, given `n` unique Phyla in a sample.

$$\mathrm{MRA[CD, i]} = \frac{\mathrm{P[i]}}{\mathrm{P[1]+P[2]+...P[n]}}.$$

The mean relative abundances were calculated at the phylum taxonomic level and plotted for a variety of different sample characteristics below in Figure 2.8.The phylum level was chosen both because it is the rank that V4 sequencing can predict most accurately and because it is the most visually accessible (Yang et al., 2016).

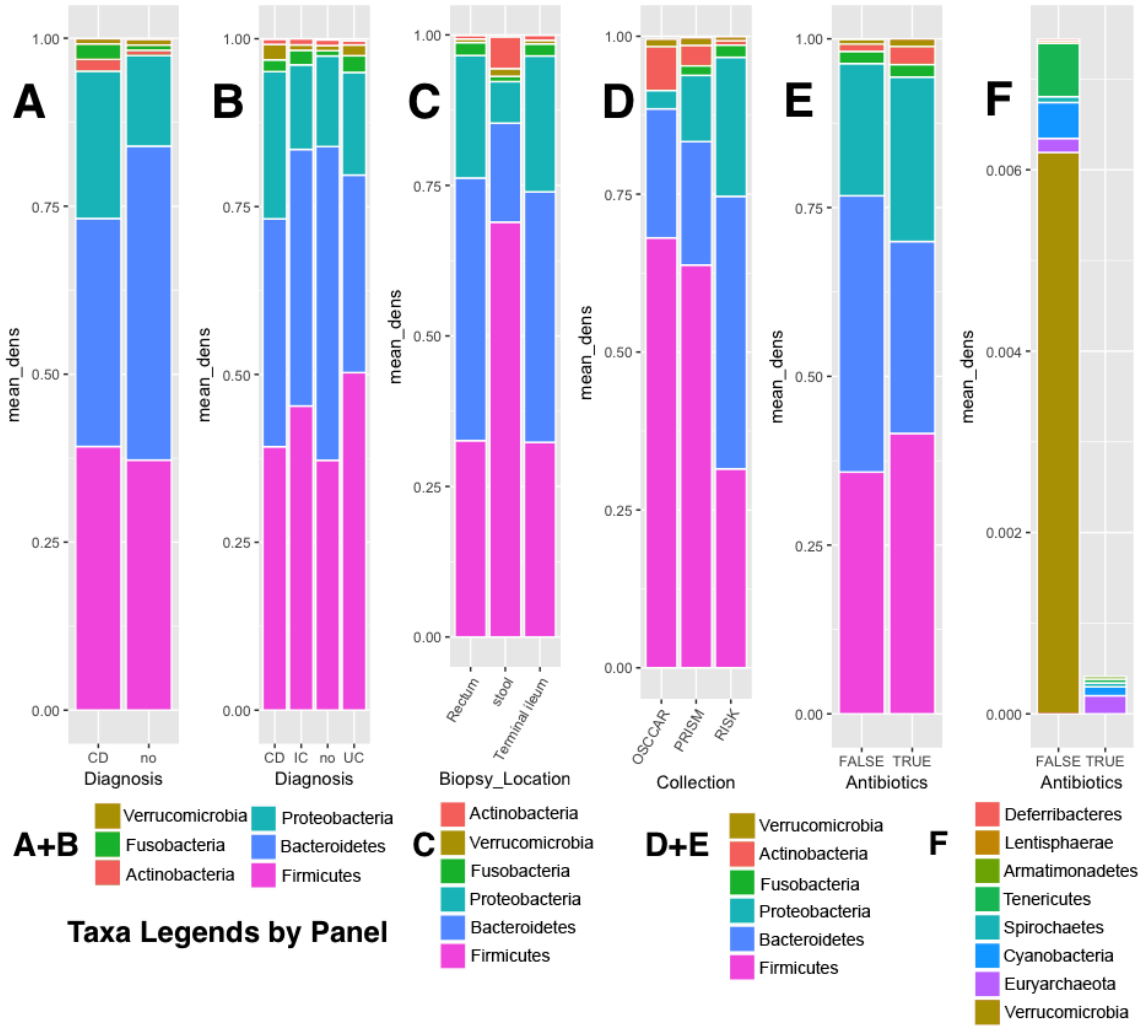


Figure 2.8: A barplot showing the mean relative abundance of the most common phyla from the treatment naive microbiome samples. All OTU counts were collapsed to phylum level and normalized by calculating the mean relative abundance (MRA) for each phylum by sample characteristic. In order, the plots include the MRA of A) antibiotic-free biopsy samples in both CD and control RISK patients that were used in the rest of the downstream analyses, B) all samples in panel A, but expanded to include samples with Colitis diagnoses, C) all treatment naive samples sorted by biopsy location, D) the RISK cohort compared to the OSCCAR and PRISM cohorts also sequenced by Gevers et al., E) all treatment naive samples grouped by antibiotic use, where FALSE means no recent antibiotic usage, and F) all antibiotic use as in E, but where the MRA's of the phyla are less than 1 percent (but greater than 0.000001 percent) of the total abundance (Gevers et al. 2014). Phyla with MRAs of less than 1 percent were removed from panels A-E.

In panel A of Figure 2.8, the mean phylum relative abundance was calculated for the RISK cohort biopsy samples of CD and control patients who had no recent history of antibiotics. While the Firmicutes, in pink, don't appear to differ too much between healthy and CD samples, there are much larger differences between the remaining phyla, including Bacteroidetes (sky blue, 12.8 percent difference) and

Proteobacteria (turqoise, 8.4 percent difference). This difference in phyla we observe isn't suprising, as other studies have reported that although the taxa of the microbiome vary, the overall metabolic functions they contribute to usually remain consistent (Consortium & others, 2012). While the CD/no diagnosis plots of panel A look like they have significantly different microbiota, it could simply be due to this natural variance in microbiomes across individuals (Consortium & others, 2012). Beyond this variation, however, other characteristics of the samples could also be contributing to the differences in phyla composition of these groups. To help understand possible differences we see in the CD/no phylum plot in panel A above, the phyla composition of all treatment naive microbiome samples were plotted for different attributes in panels B-F of Figure 2.8.

In the expanded diagnoses plots in panel B of Figure 2.8, all 4 diagnoses look distinct from each other. Although the MRAs of Firmicutes (pink) from CD and control samples look fairly similar and all IBD diagnoses have a similar MRA of Actinobacteria (green), the MRA of Proteobacteria in CD samples is not shared with the other diagnoses. Despite UC, IC, and CD being clustered together under the umbrella term IBD, there are marked differences in how they are categorized and how they effect those diagnosed with them (Laass et al., 2014). It makes sense then that they might also affect the microbiome differently, and as these samples make up only a small part of Gevers et al.'s data, IC and UC samples were excluded (Gevers et al., 2014).

In the sample type plot in panel C, the stool samples are sandwiched between the terminal ileum and biopsy samples. The fecal sample plot is radically different from the biopsy plots, with an enormous difference between the relative abundance of Firmicutes and Bacteroidetes. You don't need to subtract the MRAs of these to infer that, as it's quite apparent just by looking at them. This makes sense, as gut biopy samples are thought to be more representative of the actual gut environment than fecal samples, which are mainly representative of what's leaving the gut microbiome (Gevers et al., 2014).

In the Collection plot in panel D, the phyla for each of the cohorts that Gevers et al. used in some of their analyses are plotted (Gevers et al., 2014). The differences between the mean abundances of Firmicutes (pink), Bacteroidetes (sky blue), and Proteobacteria (turqoise) are much more pronounced than in panel A of Figure 2.8. The RISK (Risk Initiative) cohort only includes pediatric treatment naive patients, while the OSCCAR (Ocean State Crohn's and Colitis Registry) and PRISM (Prospective Registry in IBD Study at Massachusetts general hospital) cohort samples include a much larger age range and are not all new-onset or treatment naive (Gevers et al., 2014). Because the microbiome has been demonstrated to change considerably as we age or take medications for different illnesses, as seen in panel D of Figure 2.8 the OSCCAR and PRISM cohorts were excluded from the data analysis in this thesis (Yatsunenko et al., 2012).

Gevers et al. also excluded patients who had recently used antibiotics from their analyses (Gevers et al., 2014). When we limit the antibiotic plot in panel E of Figure 2.8 and plot only the phyla averaging less than 1 percent in panel F (minus the

phyla approaching 0 percent abundance once averaged, which are still too small to see even in this reduced plot), you can see that you lose a huge amount of diversity in the less abundant phyla between antibiotic users. The recent-antibiotic (TRUE) samples of panel F lack many of the low-abundance taxa that the antibiotic-free (FALSE) samples display. The biggest visual difference observed in these samples is between the MRAs of Verrucomicrobia (yellow), but an additional 18 phyla were removed from these plots as they were of differing magnitudes and thus too small to easily compare. These differences have been described by a number of studies looking at the effect of antibiotics on the microbiome, and is usually why antibiotic users are excluded from models not focusing on antibiotic use (Gerber, 2014).

Filtered Samples

A graphic summarizing the samples used in this thesis can be seen in Figure 2.9.

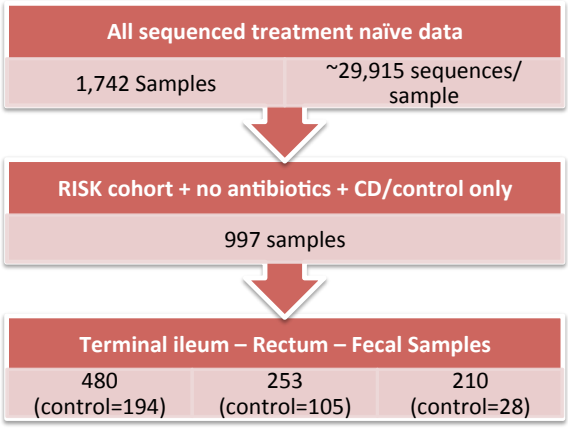


Figure 2.9: A flowchart summarizing why samples were excluded from the data analysis in this thesis (Morgan et al. 2012).

The discarded cohorts OSCCAR and PRISM were excluded because a number of their patients were adults without treatment naive, new-onset disease. Additionally, they included patients diagnosed with other IBDs, including Ileal and Ulcerative Colitis (IC and UC), and were more limited by geographic region, as they are both New England based studies (Morgan et al., 2012).

2.2.2 Random Forest Disease State Models

Random forest models were created to evaluate the individual abilities of terminal ileum, rectum, and stool samples to accurately predict if a patient had Crohn's. Samples that had taken antibiotics recently were excluded from the model, as were the PRISM and OSCCAR samples for reasons described in the previous section (Gevers et al., 2014).

In order to stabilize the prediction ability of the model and reduce model run times, an `R` package titled `VSURF` was used to select the most important variables across multiple random forest runs (Genuer, Poggi, & Tuleau-Malot, 2015). Briefly, `VSURF` makes 20 random forests with 10,000 trees each, chooses the taxa that consistently have the highest permutation variable importance, and then runs 10 addtional models twice using these threshold step variables to filter out any additional poorly predicting taxa (Genuer et al., 2015).

The following code was the `R` input used for `VSURF` to select taxa.

```
Equation 1: VSURF(x=training.data, y=training.data$diagnosis, ntree =
10000, nfor.thres = 20, nfor.interp = 10, nfor.pred = 10)
```

After the most important variables are selected, a regular random forest is constructed using Breiman and Cutler's algorithm, as seen in the code below titled `Equation 2` (Liaw & Wiener, 2002). The formula specifies that it will try to predict the diagnosis of the samples as a function of the taxa selected by `VSURF`. The training data is about 2/3 of the treatment naive data, while the other 1/3 was set aside to later test each model's accuracy using an ROC curve, unless stated otherwise. This forest grows 80,000 trees, assesses the gini variable importance of the variables used `(importance=TRUE)`, and calculates the proximity, or similiarity, `(proximity=TRUE)` of the samples included in the training data for each class. By comparing the samples within the same diagnosis to each other, we can assess if any samples included in either group are outliers that could potentially skew the model (Liaw & Wiener, 2002).

The following code was the `R` input used for `randomForest` to predict disease state (diagnosis) as a function of the variables chosen by `VSURF` (VSURF.variables).

```
Equation 2: randomForest(formula="diagnosis ~ VSURF.variables", data=
training.data, ntree=80000, importance=T, keep.forest=TRUE, proximity=TRUE)
```

The ntree value for this model was chosen by running the same model dozens of times, but growing an excessive number of trees (80,000) and then plotting the OOB error multiple times, as shown in Figure 2.10.

The error rate stabilizes around 18,000 trees in this plot, but for a few other models it stabilized only around 25,000, and so a higher limit of 30,000 was chosen. For all 200 models included in this thesis, the error plot was saved and checked for behavior differing from the plot in Figure 2.10. While a select few showed bumps after 30,000 like those around 17,000 in Figure 2.10, the accuracy of the models was not greatly affected. Although reducing the biopsy models to 30,000 trees does not greatly affect their error rates, the trees in this thesis all use 80,000 for consistency, as the fecal sample models required the full 80,000 to stabilize.

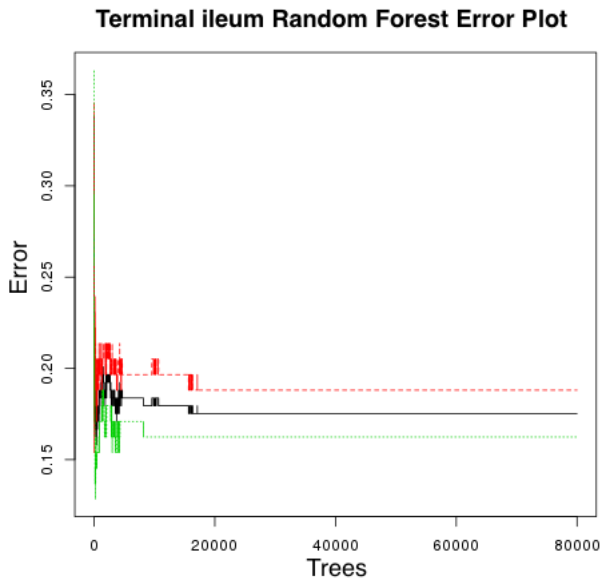


Figure 2.10: A representative error plot from a terminal ileum random forest model growing 80,000 trees. The total OOB error is shown in black, the CD error rate is in red, and the control error rate is in green (Breiman, 2001).

Terminal Ileum Biopsy model For the Terminal ileum model, a total of 480 samples, 194 control and 286 CD, were included from the treatment naive data. The following procedure was repeated 50 times, resulting in 50 models, ROC curves, and AUC values. A training data set was made by randomly sampling 117 controls and 117 CD samples without replacement, while the test data set contained the remaining 234 samples, which included 77 controls and 246 CD samples. The training set was constructed so that it contained an even ratio of CD and control samples, and was limited by the number of control samples. Including all of the leftover CD samples in the test set and having an uneven CD/no ratio does not change the model, as it only makes predictions for these classes.

`VSURF` was then run on the training data to calculate and select the variables with the highest local variable importance. The variables are ranked in order of importance, so the most important variable is listed first. The taxa chosen the most frequently by `VSURF` in order of importance, and therefore position in the model, up to the sixth most important variable are summarized in Figure 2.11. After six taxa are selected by `VSURF`, no remaining taxa are repeated in more than 3 models, and so are excluded in the figure for brevity.

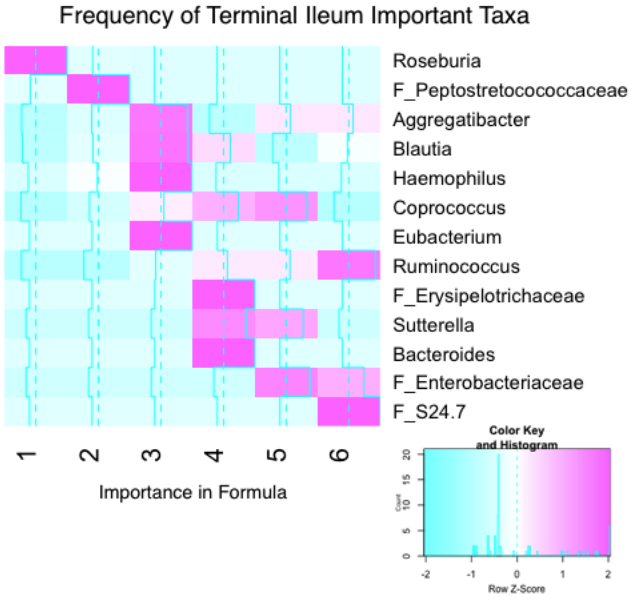


Figure 2.11: A heatmap showing the top six most frequently chosen variables in decreasing order of importance by VSURF for the terminal ileum random forest models. Roseburia was chosen as the most important taxa for 49 models, and so was the first variable in the VSURF output. Taxa were colored by Z-score, a normalizing measure describing how many standard deviations away from the mean frequency the taxa was, with pink indicating that a taxa was chosen more often on average and blue that it was chosen less (Genuer et al., 2015).

A random forest model was then generated for each model using the variables selected by `VSURF`, and was used to predict the diagnosis of the samples in the test data. An ROC curve and its AUC were calculated for each model, with the average model ROC and AUC plotted in Figure 2.12.

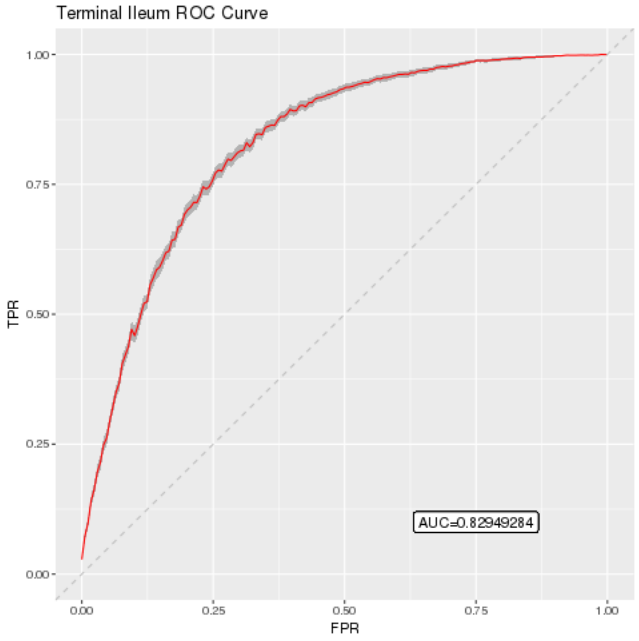


Figure 2.12: The average ROC curve and calculated AUC for the 50 terminal ileum random forest models. The averaged model is plotted by True positive rate (TPR) and False positive rate (FPR), the shaded gray area is the 95 percent confidence interval for the mean, and the dashed line through the center diagonal shows how a model with a 50 percent accuracy rate would be plotted, equivalent to flipping a coin (Sing et al., 2005).

The terminal ileum model performed the best of the sample types at predicting disease state, with a 0.83 AUC value on average. It was slightly less accurate than Gevers et al.'s model, which used regression and 5-fold cross-validation on all normalized genus level taxa for a mean AUC of 0.85 (Gevers et al., 2014). For further description of the model differences, please refer to the Results.

Rectum Biopsy model The procedures above were repeated to generate 50 random forest models, ROC curves, and AUC values for rectum biopsy samples. For the rectum data, a total of 253 samples were used, consisting of 148 CD and 105 control samples. For the training data, 63 CD and 63 control samples were included, while the rest were put in the test data. The taxa chosen the most frequently by `VSURF` for the recutm model in order of importance, and therefore position in the model, up to the sixth most important variable are summarized in Figure 2.13.

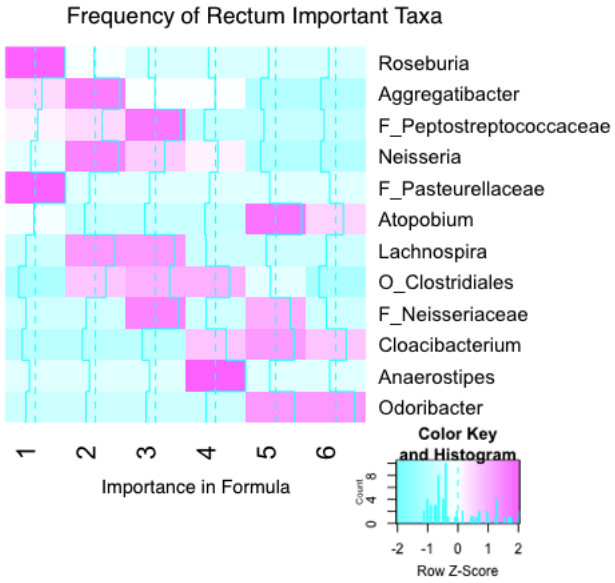


Figure 2.13: A heatmap showing the top six most frequently chosen variables in decreasing order of importance by VSURF for the rectum biopsy random forest models. Roseburia was chosen as the most important taxa for 32 models, and so was the first variable in the VSURF output for those. Taxa were colored by Z-score, a normalizing measure describing how many standard deviations away from the mean frequency the taxa was, with pink indicating that a taxa was chosen more often on average and blue that it was chosen less (Genuer et al., 2015).

An ROC curve and its AUC were calculated using each rectum random forest model, with the average model ROC and AUC plotted in Figure 2.14.

The rectum model performed below the terminal ileum model at predicting disease state, with a 0.81 AUC value on average. It was more accurate than Gevers et al.'s rectum model, however, which had a mean AUC of 0.78 (Gevers et al., 2014). While AUCs are sometimes referred to as the percent accuracy of a model, it is unwise to do so, as they are only an accuracy measure used to compare true and false positive rates of a model (Bradley, 1997). By removing the term AUC when discussing the model, you are distancing the fact that the accuracy of your model is that achieved only when it is performed on test data samples, and may not necessarily reflect the model's actual predictive ability on the entire population (Bradley, 1997). While the random forest rectum model could have a higher accuracy than Gevers et al.'s model, this would depend on a number of factors, which are explained in the next chapter. To

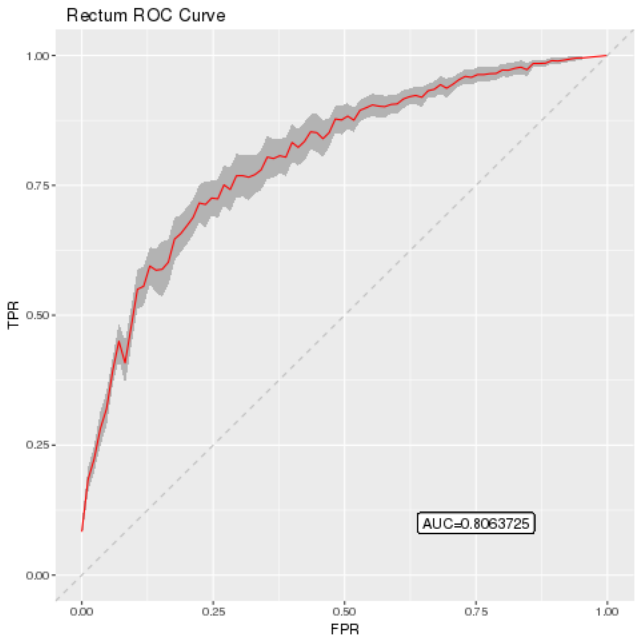


Figure 2.14: The average ROC curve and calculated AUC for the 50 rectum biopsy random forest models. The averaged model is plotted by True positive rate (TPR) and False positive rate (FPR), the shaded gray area is the 95 percent confidence interval for the mean, and the dashed line through the center diagonal shows how a model with a 50 percent accuracy rate would be plotted, equivalent to flipping a coin (Sing et al., 2005).

simplify discussion of the predictive models in this thesis, however, when accuracy or predictive ability are mentioned in the context of the random forest models, it refers only to the accuracy of the model on the RISK cohort population.

Fecal model The procedures above were repeated to generate 100 random forest models, ROC curves, and AUC values for rectum biopsy samples. More models were generated for fecal samples to obtain a more stable mean value, as these models appeared to be relatively poor predictors of disease state and ranged from AUC's of 0.2 to 0.8. To ensure that the first 50 runs didn't trend towards an unusually low or high AUC from random sampling, another 50 runs were added. For the fecal data, a total of 210 samples were used, consisting of 182 CD and 28 control samples. For the training data, 17 CD and 17 control samples were included, while the rest were put in the test data. No taxa were chosen by `VSURF` more than 20 percent of the time as the most important, so no heatmap was generated to summarize them. Unlike the ileum and rectum models, where `VSURF` chose between 3-11 taxa, the number of fecal taxa chosen ranged only from 1-7. An average ROC curve and AUC were calculated using all 100 fecal random forest models, and are plotted in Figure 2.15.

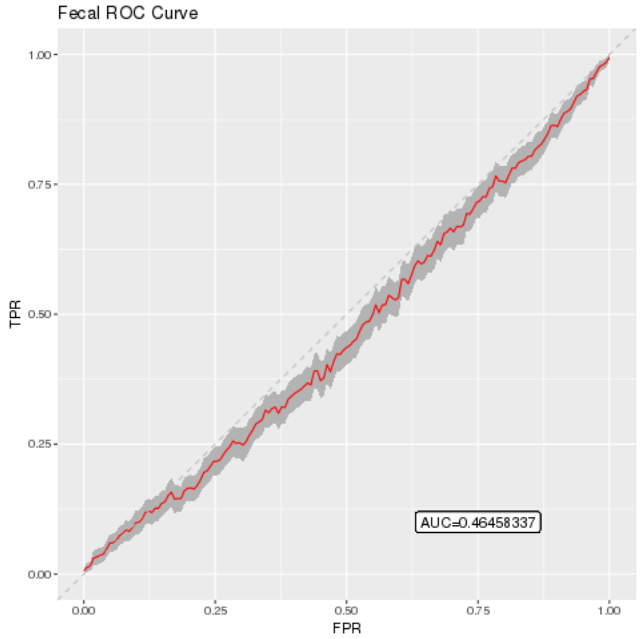


Figure 2.15: The average ROC curve and calculated AUC for the 100 fecal sample random forest models. The averaged model is plotted by True positive rate (TPR) and False positive rate (FPR), the shaded gray area is the 95 percent confidence interval for the mean, and the dashed line through the center diagonal shows a model with a 50 percent accuracy rate, equivalent to flipping a coin. Because the average model has an AUC of under 0.5, it is worse than a coin flip at predicting CD diagnosis (Sing et al., 2005).

The fecal model here performed the worst at predicting disease state, with a 0.46 AUC value on average. It was much less accurate than Gevers et al.'s model, which had a mean AUC of 0.66 (Gevers et al., 2014). Like Gevers et al.'s models, however, both biopsy models significantly outperformed the fecal model, with the terminal ileum biopsies having the best predictors. Both the small sample size and type of sample for the fecal model were suspected to be two reasons behind these models' poor accuracy.

Fecal model outlier plots To investigate potential reasons behind the fecal model's poor prediction ability, the outlier plots for each model were computed and compared to the others, which is reflected in the two representative outlier plots shown in Figure 2.16.

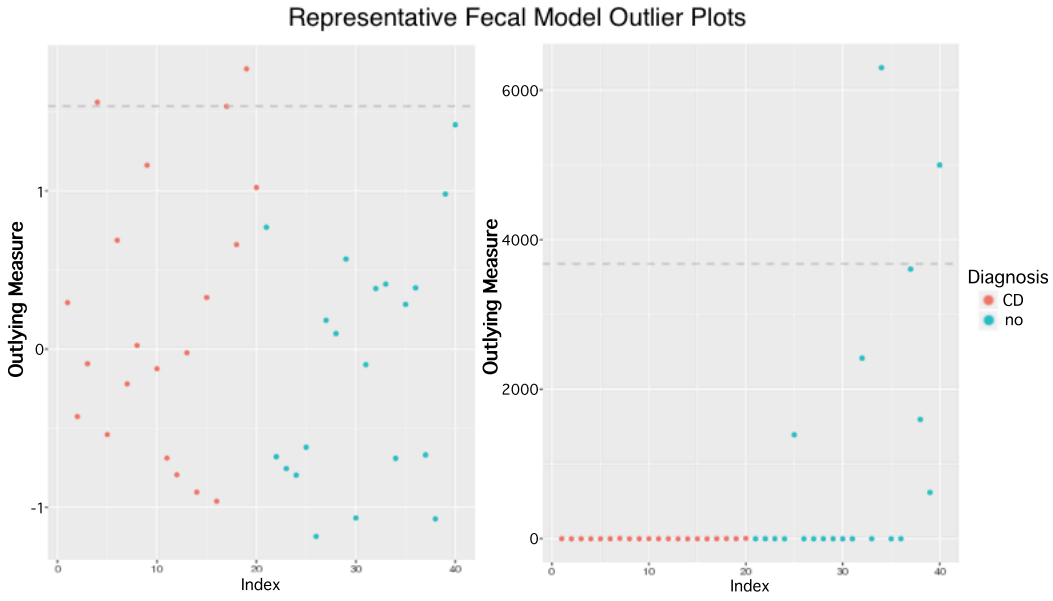


Figure 2.16: Two representative outlier plots for the random forest fecal models, where outlier measure for each sample represents how different a sample is from the others in its disease state, and index indicates the order each sample appears in on the plot. Red samples are from the CD training data, while blue controls. An outlier measure above 10 can indicate a sample is considerably different than the others in its class, while measures between -1 and 1 indicate very little difference (Breiman 2004).

The proximity measures for randomForest models are calculated by determining how many times two samples from the same class are put in the same terminal node of a tree, and then divided by the number of trees included in the model (Breiman & Cutler, 2004). So if one model sorts CD and control patients correctly 100 percent of the time in their terminal nodes, all of the samples' proximities would be 1. To compute the outlier measure, the median proximities and the mean absolute deviations, or the mean difference between each sample and the median, are computed for each disease state. Let `P=[P[1],P[2],...P[n]]` be the proximities for the `n` samples. The median proximity is subtracted from each proximity measure `(P[i])` and then divided by the mean absolute deviation for each sample, which is what is plotted on the y-axis in Figure 2.16. This measure can be represented as the following formula, where n=number of samples in 1 class :

To calculate `P[i]` for every sample in n, do: $\text{OutlyingMeasure[i]} = \frac{\text{P[i]} - \text{Median(P[n])}}{\text{MeanAbsoluteDeviation(P[n])}}$

Samples with outlying measures between -1 and 1 are considered close to others in their class, while Breiman and Cutler reccommend that samples with measures above 10 should be considered very carefully (Breiman & Cutler, 2004). In Figure 2.16, the

outlier plot on the left indicates a fairly even sample distribution, while the plot at right shows enormous differences between the control samples. We can infer from the rightmost outlier plot that either the normal variation displayed by microbiomes isn't being accounted for due to sample size, that the fecal microbiota aren't representative of the gut microbiome, both of the former are true, or that there is another confounding variable.

Comparing Disease State Classification Models

To quickly compare this thesis' models to each other and to Gevers et al.'s models, please refer to the summary table in Figure 2.17. For an explanation of how each value was obtained, please read the individual model and summary sections above. A more thorough comparison of Gevers et al.'s models to the random forest models is included in the next chapter. The following section will focus on the differences between the random forest models produced for this thesis.

	This Thesis			Gevers et al.		
Model	Thesis AUC	Thesis Runs	Important Taxa	Gevers runs	Gevers AUC	Important Taxa
Terminal Ileum	0.83	50	See next table (SNT)	5	0.85	SNT
Rectum	0.81	50	SNT	5	0.78	SNT
Fecal	0.46	100	Inconsistent	5	0.66	SNT

Figure 2.17: Comparative models at a glance.

Biopsy vs. Fecal Models Fecal models performed much worse than biopsy models at predicting disease state, as summarized in the table of AUC values in Figure 2.17. Their AUC of only 0.46 means that they can predict disease state about as accurately as a coin flip, which would correspond to an AUC of 0.5 (Hajian-Tilaki, 2013). Previous studies have shown that fecal samples are poor substitutes for biopsies when trying to characterize the gut microbiome (P. B. Eckburg et al., 2005). After visualizing the taxa composition plot comparing biopsy and fecal samples in panel C of Figure 2.8, this huge difference in model prediction ability doesn't come as much of a surprise. While the biopsy models shared most of their important taxa as seen in the heatmaps in Figure 2.11 and Figure 2.13, no single taxa was given the same importance ranking more than 20 times out of the 100 times it was run. A fecal taxa heatmap didn't make any sense to plot, as even the top 3 most important taxa would include dozens of genera repeated only a few times. Additionally, because the taxa chosen for these models results in an AUC of about 0.5, they aren't good predictors of diease state anyways (Gevers et al., 2014; Hajian-Tilaki, 2013). While some studies have used fecal data to model CD with some success, they usually either sample from non-treatment-naive

patients with established CD diagnoses, or they compare fecal sample composition over multiple collection points to see the progression of taxa (Gevers et al., 2014; Shaw et al., 2016). The poor prediction ability of the fecal model compared to the biopsy models again supports the sentiment that fecal samples should only be viewed as a meausure of what taxa the host is shedding from their microbiome, and should not be used as a proxy for biopsy samples alone (Shaw et al., 2016).

Biopsy model comparison The biopsy models' AUCs differed only by 0.02, with average ROC curves that don't differ much beyond the 95 percent confidence intervals. This relatedness extends to the taxa that are used the most frequently in either model, which is reflected in the table showing the most frequently selected taxa in Figure 2.18.

Model	Taxa #1	Taxa #2	Taxa #3	Taxa #4	Taxa #5	Taxa #6
Ileum	Roseburia (49)	F_Peptostrepto-coccaceae (42)	Aggregatibacter (11)	F_Erysipelotrichaceae (9)	Coprococcus (6)	Ruminococcus (9)
Rectum	Roseburia (32)	Neisseria / Aggregatibacter (11)	O_Clostridiales / F_Peptostrepto-coccaceae (8)	O_Clostridiales (8)	Atopobium / Cloacibacterium (4)	No taxa chosen > 3 times

Figure 2.18: A table summarizing the differences between the most frequently selected taxa in the terminal ileum and rectum biopsy random forest models. The taxa are ordered by the number of times they appear in each of the 50 models, which is included in parentheses.

For discussion of individual taxa, their potential roles in CD, and how they compare to the taxa selected by Gevers et al.'s models, please see the next chapter. When comparing these models, it is important to keep in mind that for these ROC curves, the 95 percent confidence intervals are not indicative of how accurate their corresponding random forest models are (Bradley, 1997). This is not a feature unique to these random forest models or results, but is important to note when interpreting them. The following description of the confidence intervals is not necessary to understanding these results or models, but may be of interest anyways. They are built using point-wise, or empirical, bootstrapped ROC values from the average model, and then plotted over the mean curve(Sing, Sander, Beerenwinkel, & Lengauer, 2005). They are more useful for visualizing how consistent the models are on average, and should not be used to infer any associations by comparing models other than to see if one was dramatically less consistent than another (Fawcett, 2006). Similarly, Gevers et al.'s confidence intervals were constructed using the normal distribution, and so were not built the same as these were (Gevers et al., 2014). The only data values that are directly based on performance are the AUC values, which tell us the rectum model performs only slightly below the terminal ileum model (Hajian-Tilaki, 2013). Although terminal ileum biopsy models are not a clinically feasible method for monitoring the progression of CD over time, as they carry risks and aren't as important to collect after a diagnosis is made, the ileum model shows great potential in being used to help make the initial

diagnosis (Gevers et al., 2014; Shaw et al., 2016). Additionally, because obtaining a rectum biopsy is much easier than a terminal ileum biopsy and the rectum model's AUC is only 0.02 less than the ileum models, this shows that rectum samples do have the potential to be used to use in clinic to monitor disease progression (Fawcett, 2006; Gevers et al., 2014). Rectum samples are still much more invasive than examining fecal samples, but in cases of severe CD, they would be useful in understanding what exactly is going on in the patient's gut (Gevers et al., 2014).

Chapter 3

Discussion

3.1 Reproducibility of the Gevers et al. models

While Gevers et al. maintained that their results were reproducible, their disease state model could not be perfectly replicated because they did not provide the regression parameters used to run their equations or the exact number and nature of samples used. They write that the Python package they used to run it was the "LogisticRegression" module of the scikit-learn package, but do not say if they used the default parameters of the package or not (Gevers et al., 2014; Hsu, Chang, Lin, & others, 2003). Further, after the data has been filtered to exclude samples with the undesired characteristics that they list (antibiotic users, non-RISK cohorts, specific biopsy types), the resulting numbers of samples do not match the numbers they list in the descriptions of the ROC curves (Gevers et al., 2014). For instance, they write that they use 425 terminal ileum biopsies in their ileum ROC curve. After the previously listed filters are applied to the samples listed in their supplemental table, however, you end up with 441 samples, while filtering from their raw data gives you 480 samples (Gevers et al., 2014). There is no other mention in the paper or supplementals where the other samples went. In addition to the disease state prediction models, Gevers et al. also created a random forest model to correlate a clinical disease index with microbiome data (Gevers et al., 2014). The Pediatric Crohn's Disease Activity Index (PCDAI) is scored from 1-60 points, with less than 10 indicating very mild symptoms or remission, and 60 indicating severe symptoms (Dignass et al., 2010). Points are added by clinicians based on a variety of criteria, including symptoms, histological tests, and colonoscopy results (Dignass et al., 2010). RISK patients' PCDAI scores were recorded at the time of diagnosis, as well as at the 6- and 12-month appointments for a smaller suubset of patients who consented (Gevers et al., 2014). The model produced using random forests to predict 6-month PCDAI scores by Gevers et al. could not be replicated because they only published the PCDAI scores at the time of diagnosis in their supplemental table (Gevers et al., 2014). Additionally, while they had PCDAI scores at the time of diagnosis from 269 patients, only 13 subjects had scores of less than 10. This is significant because

they divide their test data into two classifying groups–severe disease (PCDAI $\geq$ 10) and mild/remission (PCDAI<10)–and then construct the ROC curve and AUC using their random forest model's prediction for these groups (Gevers et al., 2014). Trying to generate a random forest model on data so heavily skewed toward one class can result in serious prediction biases, especially when using 5-fold cross-validation, where 1/5 of the samples are left out each time the model is run, and then averaged at the end (Breiman, 2001; C. Chen, Liaw, & Breiman, 2004). It is important to note that they trained their model on the 6-month PCDAI scores in addition to other clinical variables, including the diagnosis-PCDAI scores, taxa, and CD medications/therapies, so a bias in one variable would not necessarily hurt the model too much–it is reason for caution, however (Gevers et al., 2014). To check the predictive ability of taxa in this 5-fold cross-validated model, they compared its accuracy to one produced using permuted taxa in addition to the other variables. It is this model that would have suffered more from lack of lower PCDAI scores, and could make the taxa appear to have greater prediction ability than they would have otherwise (C. Chen et al., 2004; Gevers et al., 2014). Of course, these are only hypotheses, and without the packages and specific data used, they cannot be tested on Gevers et al.'s data. It it entirely possible that the 6-month PCDAI scores and other clinical data completely make up for the diagnosis-PCDAI scores, but without seeing the data these possibilities cannot be ruled out (C. Chen et al., 2004; Gevers et al., 2014). Despite the lengthy criticisms above of their models' reproducibility, Gevers et al.'s paper takes a huge step in the right direction for microbiome data analysis and reproducibility (Ravel & Wommack, 2014). They go beyond what many other papers provide their readers, and allowed the creation of this thesis' random forest models. The graphic in Figure 3.1 summarizes the ideal pipeline for a fully reproducible microbiome study to follow, and Gevers et al. have gotten two thirds of the way there; no small feat when the size of the dataset (over 13 gb of raw sequences alone) is considered (Gevers et al., 2014).

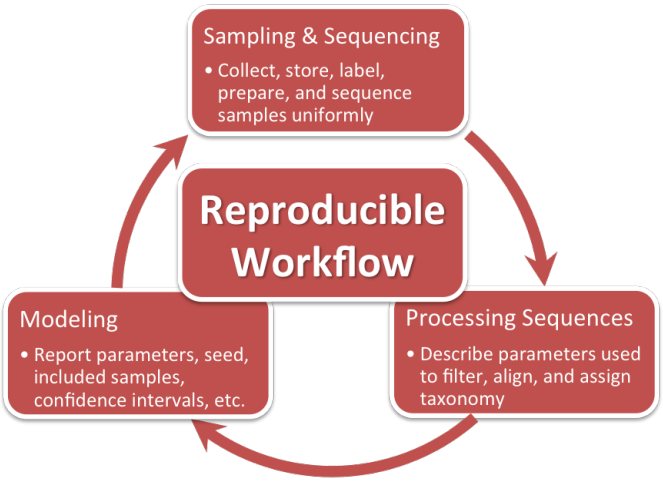


Figure 3.1: A pipeline summarizing the steps a truly reproducible microbiome study would need to follow.

A follow-up paper recently published by Gevers et al. on March 8, 2017 reflecting on their 2014 paper describes some of these issues in addition to other logistical problems, and again reiterates that these models weren't meant to be definitive diagnostic guides– simply another step to help researchers better understand Crohn's disease (Gevers et al., 2017). At the time it was published, it contained the largest gut microbiome dataset for both control and CD patients, and has helped guide new wet lab studies since (Gevers et al., 2014, 2017). In the sections below, a more detailed account of these studies and further comparison of the random forest and Gevers et al.'s models are given.

3.2 Random Forest Regression and Missing Data

Comparing models built using different methods can be tricky (Fawcett, 2006). Even limiting your models to random forests, you can still produce both binary classification trees, like the models generated in this thesis, or regression trees, like the ones generated by Gevers et al. for their PCDAI outcome prediction model (Breiman, 2001). Regression trees output the average value calculated by each tree for a continuous variable, like PCDAI score, instead of voting on a class for each test sample like in classification (C. Chen et al., 2004). Training a random forest on continuous values instead of binary classes imposes a few key differences in how the forest can be applied (Sing et al., 2005). Instead of simply judging the performance of the model based on it's ability to correctly assign a discrete label to test samples, you need to artificially create a cutoff to divide the predicted regression values into two classes, or you cannot plot the ROC curve (Fawcett, 2006; Sing et al., 2005).

Gevers et al. defined their regression PCDAI classes by using scores as a cutoff measure, so patients with PCDAI scores of less than 10 were put in class 1, and the rest of the patients into class 2 (Gevers et al., 2014). This cutoff introduces the potential class bias issue mentioned earlier, as only 13 samples fit into the first class (Gevers et al., 2014). Because this enforced regression cutoff is more subjective than disease state, this adds another layer of uncertainty to the model (Gevers et al., 2014). PCDAI scores aren't diagnoses, but rather clinical measures designed to help doctors track the progress of a patient's CD and alter their treatments accordingly (Hyams et al., 2005). While a PCDAI score of less than 10 is generally interpreted as inactive disease, one study using ROC curves to find the best cutoff value found that a score of 30 was the best at dividing severe and milder disease activity (Hyams et al., 2005). Further, Gevers et al.'s disease outcome model only achieved a 67 percent accuracy rating using both clinical data and taxa as variables (Gevers et al., 2014). This is an improvement over the 52.9 percent they got by training on clinical data alone, but they state that one of the most important variables in accurately predicting high 6-month PCDAI (scores ≥ 10) was, in fact, the PCDAI score at time of diagnosis (Gevers et al., 2014). While they admit that this difference isn't spectacular, they follow it with, "the

performance gain driven by the microbiome is a direct and unbiased demonstration of the utility of microbiome features for predicting clinical outcomes"(Gevers et al., 2014). This is a strong claim, and one that can't be directly tested without the 6- or 12-month PCDAI scores. To be fair, their recent paper rectifies this somewhat by tempering their model's relative importance in saying that they hope it can be used as an additional diagnostic tool, not a powerful tool in and of itself (Gevers et al., 2017).

3.3 Disease classification modeling

As no disease outcome model could be produced without the PCDAI data mentioned above, we can only directly compare the disease classification models. From this point on, random forest models refer only to the disease classification random forest models produced in this thesis, and not the disease outcome random forest model made by Gevers et al. using 6-month PCDAI scores. By plotting the ROC curves for the random forest models, we are normalizing the data so that Gevers et al.'s logistic regression output and the random forest tree votes result in the same metric: the ability to predict test data disease state (Fawcett, 2006). Because we are not using precisely the same subsets of data for our models, it is hard to tell the effect, if any, that our random sampling procedures, training/test data set partioning, or choice of statistical model had on the AUC (Fawcett, 2006; Gevers et al., 2014). Both random forest and L1 regression are known for their ability to classify sparse data without being computationally slowed by the sheer number of unrelated variables (taxa) included, so directly comparing the models' functions without having the parameters of Gevers model is meaningless (Breiman, 2001; Hsu et al., 2003). We do know, however, that their terminal ileum regression model results in an AUC value of 0.85, while the random forest model results in an AUC value of 0.83, as summarized in the previous chapter's Figure 2.17. The AUC is a popular way of representing the accuracy of many disease state models in medical diagnostics, as it represents a models' ability to discriminate true classes of the test data (R. Kumar & Indrayan, 2011). It's also bidirectional, because if you get a model with less than 0.5 AUC, then you can tell the model to switch its final predictions, resulting in a model with an AUC over 0.5; of course, this means that the functions or variables you thought would be good disease state predictors may not be (R. Kumar & Indrayan, 2011). ROC curves are also class-independent, which means they are not biased by class imbalances like random forests can be (R. Kumar & Indrayan, 2011). In their reproducible data processing, Gevers et al. attempt to minimize all of the potential biases possible for ROC curves, the foremost method using both control and biopsy samples processed in the same way (R. Kumar & Indrayan, 2011). All of the above factors emphasize that the AUCs of ROC curves are particularly well-suited to measure the prediction abilities of these models. Additionally, comparing AUCs is easy once the ROC curves are computed: the higher AUC belongs to the model that is better able to predict disease state (Hajian-Tilaki, 2013; R. Kumar & Indrayan, 2011). Following this, Gevers et al.'s

terminal ileum regression model is better at predicting disease state than the random forest mode by 0.02. Their AUCs are so close, however, that because they are both significantly above 0.5 and are not being designed as the sole diagnostic to predict CD, they will be essentially equivocal if they predict the same taxa (Hajian-Tilaki, 2013). In the table summarizing the shared taxa of either model in Figure 3.2, we see that this is true, and that they are both adept at discriminating between disease state and identifying the same important taxa (Gevers et al., 2014).

Random Forest Models	Gevers et al.
Roseburia	X
F_Peptostreptococcaceae	X
Aggregatibacter	X
F_Erysipelotrichaceae	X
Neisseria	X
O_Clostridiales	X
Ruminococcus	X
Coprococcus	X
Blautia	X
Haemophilus	X
Eubacterium	X
Sutterella	X

Figure 3.2: A summary table describing the taxa selected the most frequentely in the random forest models, with an 'X' if Gevers et al.'s paper also indicated them.

Although the taxa most implicated in Gevers et al.'s disease classification models are not listed together, the most important bacteria chosen by the random forest models were cross-checked with bacteria listed as being significant throughout the paper as a proxy (Gevers et al., 2014).

3.3.1 Taxa associated in Crohn's Disease

The most important genus chosen by the biopsy models was *Roseburia*, but all of the taxa chosen by the model have been implicated in at least one IBD study (Gevers et al., 2014; Manichanh et al., 2012; Morgan et al., 2012). Roseburia, along with others in the model such as Blautia, Ruminococcus, Odoribater, and Coprococcus, are all genera associated with producing butyrate as a metabolite (Gevers et al., 2014; Morgan et al., 2012; K. Takahashi et al., 2016). Butyrate is a fatty acid that is commonly made by gut bacteria as a by-product of fermenting undigested carbohydrates in the colon (K. Takahashi et al., 2016). It has long suspected to be a player in IBD, with a number of microbiome studies finding an association between a decrease in butyrate producing bacteria and CD or IBD (Morgan et al., 2012; K. Takahashi et al., 2016). In one study, researchers found that treating the inflamed cells of CD patients with

Butyrate reduced inflammation levels upon interaction with *E. coli* to the normal level displayed by control cells (Russo, Luciani, De Cicco, Troncone, & Ciacci, 2012). Another recent study found that butyrate induced the differentiation of regulatory immune cells in mice, which help regulate inflammatory responses (Furusawa et al., 2013). These studies, when combined with the common finding of lowered abundances of butyrate-producing bacteria in CD, point towards Butyrate as a potential therapy for CD, and implicate the dysbiosis of butyrate-producing taxa as another potential factor involved in the etiology of CD (Furusawa et al., 2013; K. Takahashi et al., 2016). As more CD and control biopsy samples are processed and more data is made open-access, models like the ones in this thesis and Gevers et al.'s will be better equipped to detect overarching patterns of differential bacteria in diseased microbiomes.

Chapter 4

Conclusion

The goal of this thesis was to generate a model that could accurately predict Crohn's disease state using microbiome data, and along the way investigate the reproducibility of Gevers et al.'s microbiome analyses. Since Crohn's is a disease that afflicts people in a number of different ways, and because the etiology is an enigmatic mixture of different genetic, environmental, and microbial factors, it can be a difficult disease to both diagnosis and treat. It can also dramatically affect a person's quality of life during flareups, and some people require surgery to resection different parts of their colon or ileum that have become too inflamed or damaged. Recent microbiome sequencing advances and cost reductions have given researchers another tool in their box when tackling Crohn's. Using microbiome data to model different aspects of the disease has already shown potential by a number of studies, including Gevers et al.'s. While they stated they hoped their models could be used by researchers to the benefit of Crohn's patients or even just understanding Crohn's itself better, not providing their model seems to impede this goal.

To try to rectify this, random forest models were used to select the most important taxa for determining disease state, and then these taxa were used to predict CD or control on unlabeled data. The random forest algorithm picked up almost all of the taxa mentioned by Gevers et al. in their paper, indicating that they performed almost at well as Gevers et al.'s regression models, which couldn't be perfectly replicated without seeing the parameters used. Additionally, while the terminal ileum biopsy random forest models scored an AUC of only 0.02 below Gevers et al.'s model, the random forest rectum model performed much better than theirs, while the fecal model performed significantly worse. These differential performances could have stemmed from the model type used. It would have been interesting to be able to directly compare the regression and random forest models to see if they got the same results using the same random sampling, but without the parameters, there's no guarantee any regression model I ran would be representative of Gevers et al.'s models. It's important to note that even running a different model type on their data got very similar results, at least for the biopsies, and that the random forest models used here are free `R` packages and don't require a separate server packed with computing power

to run. This greatly increases the models' accessibility, and could allow for an easier transition into clinical use.

Microbiome datasets themselves can take up dozens of gigabytes to store. Large databases like Qiita and EBI provide locations to store them, and at the same time give studies the opportunity to make their data public. This is a huge step forward for study reproducibility. In the current state of research, it is often said that getting funding just to replicate other studies can be incredibly difficult, as novel findings are desired more. Microbiome data doesn't require resequencing again unless issues with the data are suspected, and most of the analyses and processing pipelines used for microbiome data are free, albeit can be a little unclear on how to run them.

As microbiome datasets continue to grow, new standards are needed to regulate this data. The more accessible it is, the easier it will be for researchers to collaborate and work towards better solutions for treating both Crohn's and other diseases that have been associated with microbiome dysbiosis. Publishing all of the analyses run is almost as important as the data itself. By withholding these, it is difficult to run different models side by side and get a real feel for how they compare to each other. Although Gevers et al. did not publish their model's parameters, they published many of their other analyses, including the results from running them. This thesis shows that random forests can produce a model almost as accurate as theirs, but the parameters of this model have been explained and listed so that anyone reading this thesis could open their computer and replicate them by themselves. While the taxa implicated by my models are not anything new, as Gevers et al. and others have already written on them, they are again shown to be important in determining disease state using a different kind of model.

The more transparent studies are with their analyses in addition to their data, the easier it will be for other researchers to build upon their work. This thesis is a prime example of how difficult it can be to interpret someone's results without being able to understand the analyses used, and the scripts that will be provided as a github repository for this thesis will demonstrate how easy it can be to publish these alongside any analysis.

Appendix A

Glossary

A.1 Biological terms

Commensals: Microbe that benefits from host without significantly affecting the host.

Microbiome: All of the bacteria living as a community in the gastrointestinal tract.

Microbiome dysbiosis: an imbalance of the microbial community within the gut characterized by a disruption of normal function, metabolite production, and/or taxonomic composition

Operational Taxonomic Unit (OTU): A taxonomic unit essentially equivalent to species. Specifically means that the bacteria contained within the unit have sequences that are 97% similar to each other in this thesis.

Pathogens: Microbe causes harm or disease in the host. Can refer to a single microbe responsible for a specific disease or used more broadly as harmful bacteria.

Symbiotes: Microbe and host benefit from each other.

A.2 Model/Statistical Terms

L1-penalized regression: A regression formula that performs best on sparse data matrices, like microbiome data.

Random forest: A randomized tree-building statistical model that runs quickly and is relatively resistant to bias.

ROC curve: A curve made by plotting the TPR and 1-FPR of a model's predictions.

Variable importance: either the Mean Decrease Accuracy or Mean Decrease Gini measures, which relate to either how accurate the model is when a taxas' data is randomly permuted or how well a taxa is at splitting a node.

A.3 Abbreviations

AUC: Area Under the Curve

CD:Crohn's Disease

FPR: False Positive Rate

IBD: Irritable Bowel Disease

IC: Ileal Colitis

L1 Regression: Least absolute deviations

MDA: Mean Decrease Accuracy

MDG: Mean Decrease Gini

OTU: Operational Taxonomic Unit

PCDAI: Pediatric Crohn's Disease Activity Index

QIIME: Quantitative Insights Into Microbial Ecology

RISK: Risk initiative cohort

ROC: Receiver Operating Characteristic

TPR: True Positive Rate

UC: Ulcerative Colitis

References

Antoniou, E., Margonis, G. A., Angelou, A., Pikouli, A., Argiri, P., Karavokyros, I., ... Pikoulis, E. (2016). The tnbs-induced colitis animal model: An overview. *Annals of Medicine and Surgery*, *11*, 9–15.

Arrieta, M.-C., Stiemsma, L. T., Dimitriu, P. A., Thorson, L., Russell, S., Yurist-Doutsch, S., ... Brett Finlay, B. (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Science Translational Medicine*, *7*(307), 307ra152–307ra152. `http://doi.org/10.1126/scitranslmed.aab2271`

Bajaj, J. S., Hylemon, P. B., Ridlon, J. M., Heuman, D. M., Daita, K., White, M. B., ... Gillevet, P. M. (2012). Colonic mucosal microbiome differs from stool microbiome in cirrhosis and hepatic encephalopathy and is linked to cognition and inflammation. *Am J Physiol Gastrointest Liver Physiol*, *303*(6), G675. Retrieved from `http://ajpgi.physiology.org/content/303/6/G675.abstract`

Bartram, A. K., Lynch, M. D., Stearns, J. C., Moreno-Hagelsieb, G., & Neufeld, J. D. (2011). Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. *Applied and Environmental Microbiology*, *77*(11), 3846–3852.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., & Cutler, A. (2004). Random forests:Classification/clustering. *Department of Statistics, Berkeley*, *2*.

Burke, C. M., & Darling, A. E. (2016). A method for high precision sequencing of near full-length 16S rRNA genes on an illumina miseq. *PeerJ*, *4*, e2492.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–6.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., ... Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy*

of Sciences, *108*(Supplement 1), 4516–4522. http://doi.org/10.1073/pnas.1000080107

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, *110*.

Chen, K., & Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLOS Computational Biology*, *1s*(2), e24. http://doi.org/10.1371/journal.pcbi.0010024

Coenye, T., & Vandamme, P. (2003). Intragenomic heterogeneity between multiple 16S ribosomal rna operons in sequenced bacterial genomes. *FEMS Microbiology Letters*, *228*(1), 45–49.

Colman, R. J., & Rubin, D. T. (2014). Fecal microbiota transplantation as therapy for inflammatory bowel disease: A systematic review and meta-analysis. *Journal of Crohn's and Colitis*, *8*(12), 1569–1581. http://doi.org/10.1016/j.crohns.2014.08.006

Consortium, H. M. P., & others. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214.

Deas Jr, T., & Sinsel, L. (2014). Ensuring patient safety and optimizing efficiency during gastrointestinal endoscopy. *AORN Journal*, *99*(3), 396–406. Retrieved from http://proxy.uchicago.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=rzh&AN=104040964&site=eds-live&scope=site

Dethlefsen, L., McFall-Ngai, M., & Relman, D. A. (2007). An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature*, *449*(7164), 811–818. http://doi.org/10.1038/nature06245

Differding, M. (2017, April). 16S ribosomal rna, section: Hypervariable regions — wikipedia, the free encyclopedia. Retrieved from \url{https://en.wikipedia.org/w/index.php?title=16S_ri

Dignass, A., Assche, G. V., Lindsay, J., Lémann, M., Söderholm, J., Colombel, J., … Travis, S. (2010). The second european evidence-based consensus on the diagnosis and management of crohn's disease: Current management. *Journal of Crohn's and Colitis*, *4*(1), 28–62. http://doi.org/http://dx.doi.org/10.1016/j.crohns.2009.12.002

Donohoe, D. R., Garge, N., Zhang, X., Sun, W., O'Connell, T. M., Bunger, M. K., & Bultman, S. J. (2011). The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metabolism*, *13*(5), 517–526. http://doi.org/http://dx.doi.org/10.1016/j.cmet.2011.02.018

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., … Relman, D. A. (2005). Diversity of the human intestinal microbial flora.

Science, *308*(5728), 1635–1638.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Furusawa, Y., Obata, Y., Fukuda, S., Endo, T. A., Nakato, G., Takahashi, D., ... others. (2013). Commensal microbe-derived butyrate induces the differentiation of colonic regulatory t cells. *Nature*, *504*(7480), 446–450.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2015). VSURF: An r package for variable selection using random forests. *The R Journal*, *7*(2), 19–33.

Gerber, G. K. (2014). The dynamic microbiome. *FEBS Letters*, *588*(22), 4131–4139. http://doi.org/10.1016/j.febslet.2014.02.037

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., ... Xavier, R. J. (2014). The treatment-naive microbiome in new-onset crohn's disease. *Cell Host & Microbe*, *15*(3), 382–392. Retrieved from //www.sciencedirect.com/science/article/pii/S1931312814000638

Gevers, D., Kugathasan, S., Knights, D., Kostic, A. D., Knight, R., & Xavier, R. J. (2017). A microbiome foundation for the study of crohn's disease. *Cell Host & Microbe*, *21*(3), 301–304.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., ... Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science (New York, N.Y.)*, *312*(5778), 1355–1359. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3027896/

Glenn, T. C. (2011). Field guide to next-generation dna sequencers. *Molecular Ecology Resources*, *11*(5), 759–769.

Gray, M. W., Sankoff, D., & Cedergren, R. J. (1984). On the evolutionary descent of organisms and organelles: A global phylogeny based on a highly conserved structural core in small subunit ribosomal rna. *Nucleic Acids Research*, *12*(14), 5837–5852.

Hajian-Tilaki, K. (2013). Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627.

Hamady, M., & Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Research*, *19*(7), 1141–1152. Retrieved from http://genome.cshlp.org/content/19/7/1141.abstract N2

Hsu, C.-W., Chang, C.-C., Lin, C.-J., & others. (2003). A practical guide to support vector classification.

Hyams, J., Markowitz, J., Otley, A., Rosh, J., Mack, D., Bousvaros, A., ... others.

(2005). Evaluation of the pediatric crohn disease activity index: A prospective multicenter experience. *Journal of Pediatric Gastroenterology and Nutrition*, *41*(4), 416–421.

Jalanka, J., Salonen, A., Salojärvi, J., Ritari, J., Immonen, O., Marciani, L., ... others. (2014). Effects of bowel cleansing on the intestinal microbiota. *Gut*, gutjnl–2014.

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., ... others. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology*, *7*.

Khosravi, A., Yáñez, A., Price, J. G., Chow, A., Merad, M., Goodridge, H. S., & Mazmanian, S. K. (2014). Gut microbiota promote hematopoiesis to control bacterial infection. *Cell Host & Microbe*, *15*(3), 374–381. http://doi.org/http://dx.doi.org/10.1016/j.chom.2014.02.006

Kinross, J. M., Darzi, A. W., & Nicholson, J. K. (2011). Gut microbiome-host interactions in health and disease. *Genome Medicine*, *3*(3), 14. http://doi.org/10.1186/gm228

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., ... Ley, R. E. (2012). A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets. *PLoS Computational Biology*, *9*(1), e1002863. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3542080/

Kumar, R., & Indrayan, A. (2011). Receiver operating characteristic (roc) curve for medical researchers. *Indian Pediatrics*, *48*(4), 277–287.

Laass, M. W., Roggenbuck, D., & Conrad, K. (2014). Diagnosis and classification of crohn's disease. *Autoimmunity Reviews*, *13*(4–5), 467–471. http://doi.org/http://dx.doi.org/10.1016/j.autrev.2014.01.029

Lee, J. C., & Gutell, R. R. (2012). A comparison of the crystal structures of eukaryotic and bacterial ssu ribosomal rnas reveals common structural features in the hypervariable regions. *PloS One*, *7*(5), e38203.

Leffler, D. A., & Lamont, J. T. (2015). Clostridium difficile infection. *New England Journal of Medicine*, *372*(16), 1539–1548.

Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, *124*(4), 837–848. http://doi.org/http://dx.doi.org/10.1016/j.cell.2006.02.017

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from http://CRAN.R-project.org/doc/Rnews/

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews:*

Data Mining and Knowledge Discovery, *1*(1), 14–23.

Manichanh, C., Borruel, N., Casellas, F., & Guarner, F. (2012). The gut microbiota in ibd. *Nat Rev Gastroenterol Hepatol*, *9*(10), 599–608. Retrieved from http://dx.doi.org/10.1038/nrgastro.2012.152

Mazmanian, S. K., Round, J. L., & Kasper, D. L. (2008). A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature*, *453*(7195), 620–625. Retrieved from http://dx.doi.org/10.1038/nature07008

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., ... Hugenholtz, P. (2012). An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, *6*(3), 610–618.

McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLOS Computational Biology*, *10*(4), e1003531. http://doi.org/10.1371/journal.pcbi.1003531

Morgan, H., Xochitl C. (2012). Chapter 12: Human microbiome analysis. *PLOS Computational Biology*, *8*(12), 1–14. http://doi.org/10.1371/journal.pcbi.1002808

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., ... Huttenhower, C. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, *13*(9), R79. http://doi.org/10.1186/gb-2012-13-9-r79

Noecker, C., McNally, C. P., Eng, A., & Borenstein, E. (2017). High-resolution characterization of the human microbiome. *Translational Research*, *179*, 7–23. Retrieved from //www.sciencedirect.com/science/article/pii/S193152441630127X

Nsubuga, A. M., Robbins, M. M., Roeder, A. D., Morin, P. A., Boesch, C., & Vigilant, L. (2004). Factors affecting the amount of genomic dna extracted from ape faeces and the identification of an improved sample storage method. *Molecular Ecology*, *13*(7), 2089–2094. http://doi.org/10.1111/j.1365-294X.2004.02207.x

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., ... Guyer, M. (2009). The nih human microbiome project. *Genome Research*, *19*(12), 2317–2323.

Poynard, T., Halfon, P., Castera, L., Munteanu, M., Imbert-Bismut, F., Ratziu, V., ... others. (2007). Standardization of roc curve areas for diagnostic evaluation of liver fibrosis markers based on prevalences of fibrosis stages. *Clinical Chemistry*, *53*(9), 1615–1622.

Pushkarev, D., Neff, N. F., & Quake, S. R. (2009). Single-molecule sequencing of an

individual human genome. *Nature Biotechnology*, *27*(9), 847–850.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F. O. (2012). The silva ribosomal rna gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, gks1219.

Ravel, J., & Wommack, K. E. (2014). All hail reproducibility in microbiome research. *Microbiome*, *2*(1), 8. Retrieved from `http://dx.doi.org/10.1186/2049-2618-2-8`

Robbins-Pianka, A. (2015). *Advanced computational tools for analyzing microbial communities for energy production environments* (PhD thesis). *ProQuest Dissertations and Theses*. Retrieved from `https://search.proquest.com/docview/1728061050?accountid=14657`

Russo, I., Luciani, A., De Cicco, P., Troncone, E., & Ciacci, C. (2012). Butyrate attenuates lipopolysaccharide-induced inflammation in intestinal cells and crohn's mucosa through modulation of antioxidant defense machinery. *PLoS One*, *7*(3), e32841.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., ... others. (2011). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *39*(suppl 1), D38–D51.

Schnabl, B., & Brenner, D. A. (2014). Interactions between the intestinal microbiome and liver diseases. *Gastroenterology*, *146*(6), 1513–1524. Retrieved from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3996054/`

Shaw, K. A., Bertha, M., Hofmekler, T., Chopra, P., Vatanen, T., Srivatsa, A., ... others. (2016). Dysbiosis, inflammation, and response to treatment: A longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome Medicine*, *8*(1), 75.

Shergill, A. K., Lightdale, J. R., Bruining, D. H., Acosta, R. D., Chandrasekhara, V., Chathadi, K. V., ... DeWitt, J. M. (2015). The role of endoscopy in inflammatory bowel disease. *Gastrointestinal Endoscopy*, *81*(5), 1101–1121.e13. `http://doi.org/http://dx.doi.org/10.1016/j.gie.2014.10.030`

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in r. *Bioinformatics*, *21*(20), 3940–3941.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25.

Takahashi, K., Nishida, A., Fujimoto, T., Fujii, M., Shioya, M., Imaeda, H., ... Sugimoto, M. (2016). Reduced abundance of butyrate-producing bacteria species in the fecal microbial community in crohn's disease. *Digestion*, *93*(1), 59–65.

Terheggen, G., Lanyi, B., Schanz, S., Hoffmann, R. M., Böhm, S. K., Leifeld, L.,

... Kruis, W. (2008). Safety, feasibility, and tolerability of ileocolonoscopy in inflammatory bowel disease. *Endoscopy*, *40*(08), 656–663. http://doi.org/10.1055/s-2008-1077445

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & Hijum, S. A. F. T. van. (2013). Data mining in the life sciences with random forest: A walk in the park or lost in the jungle? *Briefings in Bioinformatics*, *14*(3), 315. http://doi.org/10.1093/bib/bbs034

Travis, A. C., Pievsky, D., & Saltzman, J. R. (2012). Endoscopy in the elderly. *The American Journal of Gastroenterology*, *107*(10), 1495–1501.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., ... Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480–484. Retrieved from http://dx.doi.org/10.1038/nature07540

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., & Gordon, J. I. (2007). The human microbiome project: Exploring the microbial part of ourselves in a changing world. *Nature*, *449*(7164), 804–810. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3709439/

Van de Peer, Y., Chapelle, S., & De Wachter, R. (1996). A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Research*, *24*(17), 3381–3391.

Vandeputte, D., Falony, G., Vieira-Silva, S., Tito, R. Y., Joossens, M., & Raes, J. (2015). Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut*. Retrieved from http://gut.bmj.com/content/early/2015/06/11/gutjnl-2015-309618.abstract N2

Větrovsky, T., & Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PloS One*, *8*(2), e57923.

Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal dna amplification for phylogenetic study. *Journal of Bacteriology*, *173*(2), 697–703.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., ... Knight, R. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J*, *10*(7), 1669–1681. Retrieved from http://dx.doi.org/10.1038/ismej.2015.235

Yang, B., Wang, Y., & Qian, P.-Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics*, *17*(1),

135.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., ... others. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*(7402), 222–227.

Youngster, I., Sauk, J., Pindar, C., Wilson, R. G., Kaplan, J. L., Smith, M. B., ... Hohmann, E. L. (2014). Fecal microbiota transplant for relapsing clostridium difficile infection using a frozen inoculum from unrelated donors: A randomized, open-label, controlled pilot study. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *58*(11), 1515–1522. Retrieved from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4017893/`