Fall 2022

Interplay between Human Microbiota and Celiac Disease

John James Colgan

LOYOLA UNIVERSITY CHICAGO

INTERPLAY BETWEEN HUMAN GUT MICROBIOTA AND CELIAC DISEASE

A THESIS SUBMITTED TO
THE FACULTY OF THE GRADUATE SCHOOL IN
CANDIDACY FOR THE DEGREE OF
MASTER OF SCIENCE

PROGRAM IN BIOINFORMATICS

BY
JOHN COLGAN
CHICAGO, IL
AUGUST 2022

ACKNOWLEDGEMENTS

I would like to thank all of the people who made this thesis possible, starting with my thesis director Dr. Michael Burns, who proved to be an invaluable mentor throughout this entire process. Thank you Dr. Catherine Putonti who’s energy and enthusiasm helped draw me to bioinformatics, as well as all other bioinformatics faculty that I had the privilege of studying under. You are some of the most amazing and dedicated faculty I have ever encountered. In addition, I want to thank Dr. Mashkoor Choudhry for taking the time to serve on my thesis committee. You have been great fun to work with and I appreciate your interest in my research.

I would also like to thank the authors of each study I reanalyzed. Without their contributions to science this work would not have been possible.

Finally, I would like to thank my parents, Dr. John D. Colgan and Dr. Diana F. Colgan for supporting me throughout my time at Loyola and for encouraging me to always follow my dreams.

You all have helped make me the scientist I am today and I cannot thank you enough.

Dedicated to Jeannie.
So small, so sweet, too soon. You are loved and missed dearly.

If I have seen further, it is by standing on the shoulders of giants

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii

LIST OF TABLES ix

LIST OF FIGURES x

LIST OF ABBREVIATIONS xii

ABSTRACT xiii

INTRODUCTION 1

MATERIALS AND METHODS 6

RESULTS OF INDIVIDUAL ANALYSIS 10
- Bonder 10
- Garcia-Mazcarro 11
  - Duodenal Analysis 11
  - Fecal Analysis 14
  - Greengenes Vs Silva Taxonomy 16
  - Pre and Post Treatment Mexican CD Microbiome 17
- Bodkhe 18
  - Stool Analysis 18
  - LEFSe 20
  - Pathways 20

RESULTS OF POOLED DUODENUM ANALYSIS 23
- Pooled Raw Duodenum 23
  - Community Structure Analysis 23
  - LEFSe 25
  - Random Forest 26
- Pooled Filtered Duodenum 27
  - Community Structure Analysis 27
  - LEFSe 29
  - Random Forest 30
- Pooled Filtered and Scaled Duodenum 31
  - Community Structure Analysis 31
  - LEFSe 33
  - Random Forest 35
- Pooled Normalized Duodenum 36
  - Community Structure Analysis 36
  - LEFSe and Random Forest 38
- Duodenum PICRUSt2 Pathways 38

Clustering 38
Differentially Abundant Pathways 38
Random Forest 39
RESULTS OF POOLED FECES ANALYSIS 41
Raw Analysis 41
Community Structure Analysis 41
LEFSe 44
Random Forest 45
Filtered Analysis 45
Community Structure Analysis 46
LEFSe 48
Random Forest 50
Filtered and Scaled Analysis 50
Community Structure Analysis 50
LEFSe 52
Random Forest 53
Pooled Normalized Feces 54
Community Structure Analysis 54
LEFSe 56
Random forest 56
Pooled Pathways Feces Analysis 56
Clustering 56
Differentially abundant pathways 56
Random Forest 58

DISCUSSION 59
GFD's Impact on Healthy Individuals: Original Analysis Versus Reanalysis of Bonder *et Al.* 59
Alpha and Beta Diversity 59
Differentially Abundant Taxa and Functions 59
The Mexican CD Microbiome: Original Analysis Versus Reanalysis of Garcia-Mazcarro *et Al.* 61
Duodenal Microbiome Alpha and Beta Diversity 61
Differentially Abundant Duodenal Microbiota 61
Differential metabolic functions: Does the celiac microbiome cause vitamin deficiency 63
Stool Microbiota Alpha / Beta Diversity and Differentially Abundant Taxa 64
GFD and Mexican Patients 66
Greengenes versus SILVA Taxonomy 67
The Indian CD Microbiome: Original Analysis versus Reanalysis of Bodkhe *et Al.* 68
Fecal Alpha/Beta Diversity and Contaminating DNA 68
Differentially Abundant Taxa 70

Differential Metabolic Functions 72
Pooled Analysis 73
Indian CD FDRs as NIBD 73
Duodenal and Fecal Alpha and Beta Diversity Analysis 73
Duodenal Community Structure analysis 74
Differentially Abundant Duodenal Microbiota 74
Differentially abundant Duodenal Microbiota Function: Connections to Dysbiosis 76
PICRUSt2: An Improved Tool 78
Differentially Abundant Stool Microbiota 79
Random Forest 81
Previous Attempts to Unify CD Knowledge 82
Dealing with Noise and Batch Effects 83
Pooled Analysis and Dysbiosis 84

CONCLUSION 85
Reanalysis of Old Datasets Using New Tools 85
Regional Differences in CD 85
Metabolic Differences in CD 86
Small Intestinal Biopsies Versus Stool Samples 86
Summary 87

BIBLIOGRAPHY 88

VITA 99

LIST OF TABLES

Table 1. SILVA versus Greengenes taxonomy 17

LIST OF FIGURES

Figure 1. Alpha and beta diversity analysis of Bonder *et al*. 10

Figure 2. LEFSe results of Bonder *et al.* 11

Figure 3. Duodenal alpha and beta diversity analysis of Garcia-Mazcorro *et al.* 12

Figure 4. Duodenal LEFSe results of Garcia-Mazcorro *et al.* 13

Figure 5. Duodenal differentially abundant metabolic pathways of Garcia-Mazcorro *et al*. 14

Figure 6. Fecal alpha and beta diversity analysis of Garcia-Mazcorro *et al*. 15

Figure 7. Fecal LEFSe results of Garcia Mazcorro *et al*. 16

Figure 8. Alpha and beta diversity analysis of Bodkhe *et al.* 19

Figure 9. LEFSe results of Bodkhe *et al* 20

Figure 10. Differentially abundant metabolic pathways of Bodkhe *et al*. 22

Figure 11. Pooled raw duodenal alpha and beta diversity analysis 24

Figure 12. Pooled raw duodenal abundance plots 25

Figure 13. Pooled raw duodenal LEFSe results 26

Figure 14. Pooled filtered duodenal alpha and beta diversity analysis 28

Figure 15. Pooled filtered duodenal abundance plots 29

Figure 16. Pooled filtered duodenal LEFSe results 30

Figure 17. Pooled filtered and scaled duodenal alpha and beta diversity analysis 32

Figure 18. Pooled filtered and scaled duodenal abundance plots 34

Figure 19. Pooled filtered and scaled duodenal LEFSe results 35

Figure 20. Pooled normalized duodenal alpha and beta diversity 37

Figure 21. Normalized duodenal abundance plots 37

Figure 22. Pooled differentially abundant duodenal metabolic pathways 40

Figure 23. Pooled raw fecal alpha diversity 42

Figure 24. Pooled raw fecal abundance plots 43

Figure 25. Pooled raw fecal LEFSe results 44

Figure 26. Pooled filtered fecal alpha and beta diversity analysis 46

Figure 27. Pooled filtered fecal abundance plots 47

Figure 28. Pooled filtered fecal LEFSe results 48

Figure 29. Pooled filtered and scaled fecal alpha and beta diversity 50

Figure 30. Pooled filtered and scaled fecal abundance plots 51

Figure 31. Pooled filtered and scaled fecal LEFSe results 52

Figure 32. Normalized pooled fecal alpha and beta diversity 54

Figure 33. Normalized pooled fecal abundance plots 54

LIST OF ABBREVIATIONS

ASV Amplicon sequence variant, a denoised sequence produced by pipelines such as dada2.

CD Celiac disease, an inflammatory bowel disease caused by a response to gluten.

FDR First-degree relative, a direct relative i.e. a mother, father or sibling.

GCD Gluten containing diet, a normal diet.

GFD Gluten-free diet, a diet without gluten in it. The only approved treatment for CD.

HEPB Hepatitis B, a viral infection of the liver.

IBD Inflammatory bowel disease, diseases of the bowel characterized by an inflammatory response.

IBS Irritable bowel syndrome, diseases of the bowel absent of an inflammatory response.

NCGS Non-celiac gluten sensitivity, a sensitivity to gluten absent an immune response.

NIBD Non-inflammatory bowel disease, another term for IBS.

OTU Operational taxonomic unit, consensus sequence produced by the clustering of DNA sequences opposed to denoising.

RA Rheumatoid arthritis, an autoimmune disease in which the joints are the target of attack.

RCD Refractory celiac disease, CD which is non-responsive to the GFD.

ABSTRACT

Celiac disease (CD) is an autoimmune disorder of the small intestine in which gluten, an energy-storage protein in wheat and other cereals, elicits an immune response leading to villous atrophy. Despite a strong genetic component, the disease arises sporadically over the lifetime leading us to hypothesize the microbiome might be a trigger. Here, we re-examined 16S rRNA data from 3 prior studies examining celiac disease and the microbiome with newer computational tools: the dada2 and PICRUSt 2 pipelines. Our results both confirmed findings of previous studies and generated new data regarding the celiac microbiome of India and Mexico. The datasets were also pooled to determine whether any taxonomic or metabolic features remained consistent across the world using a variety of data transformations to control for batch effects. Our results showed the celiac microbiome displays dysbiosis without a discernable pattern, likely indicating that perturbations in the CD microbiome are a result of the disease rather than a cause of the disease. Data from PICRUSt 2 further confirms this, showing connections to the CD metabolome which are supported by previous research examining dysbiotic microbiomes.

INTRODUCTION

Celiac Disease (CD) is an inflammatory bowel disease (IBD) of the small intestine, in which the protein gluten, found in wheat and barley, causes an inflammatory response that degrades the lining of the small intestine, specifically the villi (Valitutti *et al.,* 2019) . If left untreated, patients acquire a host of health problems, including malnutrition, osteoporosis and, in rare cases, cancer (Valitutti *et al.,* 2019). CD is estimated to affect 1-2% of the world's population, with Western countries having the bulk of those affected; however, many cases go undiagnosed (Valitutti *et al.,* 2019). The disease can present at almost any age with intestinal and non-intestinal symptoms, such as diarrhea, bloating, pain, nausea, insomnia, and migraines (Valitutti *et al.,* 2019). Currently, there exists only one treatment for CD, being on a gluten-free diet (GFD), in which patients exclude gluten from their diet to prevent the inflammatory response (Valitutti *et al.,* 2019).

A genetic predisposition for CD is predicted among people with the HLA DQ2 and DQ8 haplotypes, but recent twin studies found the concordance of these genes is not 100%, suggesting environmental factors contribute to the disease (Greco *et al.,* 2002). Furthermore, recent research found CD subjects have a dysbiotic gut microbiome, although no clear pattern for the celiac microbiome has, as yet, been defined (Valitutti *et al.,* 2019). In addition, although several studies found evidence the celiac microbiome harbors an excess of the bacterial taxa that might cause inflammation, no specific bacterium or microbial community configuration has been linked to causing an inflammatory response to gluten (Valitutti *et al.,* 2019). The inflammation characteristic of celiac diseases may be caused by metabolic activity of the suspect microbes, whereby byproducts of microbial metabolism modulate the immune system. This type of effect was demonstrated for the short chain fatty-acid, butyrate, which is a by-product of bacterial digestion of fiber. In mice, butyrate was shown to ameliorate symptoms of rheumatoid arthritis, another autoimmune disorder, thus demonstrating the ability of symbiotes to modulate the immune system (Rosser *et al.,* 2020).

In addition, it is known that changes in diet can have drastic consequences on microbiota composition. Since a gluten-free diet is the treatment for CD, some researchers posited that the absence of dietary gluten skews the microbiome profile; in which case the celiac microbiome is an effect of diet and not a cause of disease. To address this question, one study examining how the microbiota of healthy patients adapts to a gluten-free diet was included (Bonder *et al.,* 2016). It found the changes to the microbiota were minimal, although the researchers used “dated” 16S rRNA analysis techniques.

Many other studies examined the role of the microbiome in CD and other IBDs, and used the best computational tools available at the time. Yet, bioinformatics and computational biology develop rapidly, with new tools constantly being developed that vastly out-perform their predecessors. For instance, previous 16S rRNA microbiome analysis was conducted using pipelines that generate operational taxonomic unit tables or OTUs. This is a process in which sequences are compared and binned based on a user-defined similarity threshold (usually 79% similarity), From this bin, a single sequence is pulled, with the entirety of the bin classified as this single sequence’s identification. In contrast, newer amplicon sequence variant producing pipelines, such as DADA2, use quality scores in conjunction with machine learning to distinguish sequences that differ as little as a single nucleotide, giving a far more granular and accurate picture as to what organisms are represented by a given sequence (Callahan *et al*., 2017). Previous investigation into data generated by different pipelines has already demonstrated that ASV and OTU generating pipelines result in different counts of ASVs and OTUs, and that different pipelines generate different data in regards to sequence taxonomic assignment (Allali *et al*., 2017). PICRUSt2 has a database which is 20 times larger than its predecessor, and uses Metacyc pathways as its high level output (Douglas *et al*., 2020, Babera *et al* 2019, Czech, 2020, Louca, Doebelia, 2018, Caspi *et al*., 2007). These pathways are determined by looking at community-wide aggregations of enzyme counts to predict differences in metabolic function. This is in opposition to PICRUSt1 (Langille *et al.,* 2013) which simply outputs differentially abundant genes, of which can be active in several metabolic pathways. Further evolution of the analysis of microbiome data has led to the development of algorithms such as LEFSe, to detect significant differences in community structure. LEFSe differs from methods taken to analyze RNA seq data in that it uses a less stringent p-value in association with a LEFSe LDA score (a measure of effect size) to determine which taxa are not only significantly different between cases, but with a large enough difference to produce an effect on the host, thus producing more biologically relevant results (Segata *et al.,* 2011). Several databases exist for taxonomic assignment, many previous analyses have utilized Greengenes, last updated in 2013 (DeSantis *et al*., 2006). This database has fallen out of favor and has largely been replaced by SILVA (Yilmaz *et al*., 2014). Previous work has shown that while many entries between the two databases do indeed overlap, considerable differences do exist (Balvočiūtė, Huson, 2017). This begs the question: are conclusions drawn from older analysis methods still relevant? Or should data from previous studies be reanalyzed with new tools to obtain the best results possible.

Here, we reanalyzed data from three previous studies, pooling their results to examine the associations between CD and the microbiome, and determine whether any taxa or metabolic pathway is associated with CD, and whether previous findings using older analysis techniques differ from analysis with new tools. Our analysis was conducted using the dada2 pipeline to generate taxonomic classification for each read. This data was then passed off to Phangorn to create a phylogenetic tree and PICRUSt 2 to obtain functional analysis of the microbes. These data were then analyzed using microbiome analyst, to obtain alpha and beta diversity metrics and identify differentially abundant taxa and metabolic pathways.

The first study that was reanalyzed, *The influence of a short-term gluten-free diet on the human gut microbiome,* by Bonder *et al*., examined stool samples from 21 participants to test for microbiome changes associated with the transition from a gluten-free (GF) to a gluten-containing diet (GD), using QIIME (Caparaso *et al.,* 2010), PICRUSt (Langille *et al.,* 2013) and the greengenes database (DeSantis *et al.,* 2006, Bonder et al., 2016). We also re-analyzed data from *First Insights into the Gut Microbiota of Mexican Patients with Celiac disease and Non-Celiac Gluten Sensitivity,* by Garcia-Mazcorro et al. (Garcia-Mazcorro *et al*., 2016). This study compared the microbiomes of 12 celiac patients,12 non-celiac, gluten sensitive patients, and 12 controls on GD and resampled 6 months after strict adherence to the GFD, using QIIME (Caparaso *et al.,* 2010), PICRUSt (Langille *et al.,* 2013), and Greengenes (DeSantis *et al.,* 2006) analysis of paired stool and duodenum samples. A third dataset we re-analyzed was *Comparison of Small Gut and Whole Gut Microbiota of First-Degree Relatives With Adult Celiac Disease Patients and Controls* by Bodkhe *et al*. (Bodkhe *et al.,* 2019). The original study included 23 untreated celiac patients, 15 first-degree relatives without celiac disease and 24 controls with hepatitis B or functional dyspepsia with paired stool and biopsy samples taken from each participant. To supplement this data, 19 healthy stool samples from Chaudhari *et al. (*Chaudhari *et al*., 2020) and 17 from Dubey *et al*. (Dubey *et al.,* 2018) were used in the analysis of the stool samples from Bodkhe *et al.*. It is known that diet (Pace, Crowe, 2016, Human Microbiome Consortium, 2012) and environment (Stearns *et al*., 2017) each play a significant role in shaping the gut microbiome. In this work the studies of interest each used patient cohorts from across the globe. As geographical location is known to correlate with distinct gut microbial structure (Arumaguam *et al*., 2011), we anticipate seeing some differences across studies among the control samples. In addition to standard to standard batch-effects that need to be accounted for when performing inter-study analyses, we incorporated additional control samples, where applicable, to increase our statistical power to determine CD-specific changes. We also included data taken from *The influence of a short-term gluten-free diet on the human gut microbiome* by Bonder *et al.* Which included 21 healthy patients on a GFD for 4 weeks, and GCD for 4 weeks, with participants sampled weekly.

Across the four studies we were able to incorporate 166 participants, with 31 celiac patients, 12 non-celiac gluten sensitive patients, 15 first-degree relatives, 24 patients with functional dyspepsia or hepatitis B, and 84 controls, making this work one of the largest meta-analysis examining CD across both the duodenal and stool microbiomes. These data were then pooled to examine what similarities exist in the celiac microbiome across geographical regions.

MATERIALS AND METHODS

Studies with available data were gathered using search queries “celiac microbiome”, “celiac disease and the microbiome”, “celiac disease and gut microbiota”, and “celiac disease and gut-microbiome”. Studies which were selected examined the v4 variable region of the 16s ribosomal subunit (rRNA).

Collected sequences were prepared for dada2 (Callahan *et al.,* 2016). This was done using Cutadapt (Martin, 2011) and the following command “cutadapt -g ADAPTERSEQUENCE1 -g ADAPTERSEQUENCE2 -o output input”. Adapter sequences were provided by the Materials and Methods sections of the respective parent study. This was done for all studies with the exception of Bodkhe et al., in which the adapter sequences were removed using the trimLeft = c(20,20) in dada2’s filterAndTrim step. Next, the sequences were passed to dada2. We used the same steps as the original paper for Bodkhe et al., since the parent study also used dada2. Both Garcia-Mazcarro et al. and Bonder et al. used single-end sequencing; adjustments were made to the pipeline in accordance with the developer’s advice on the dada2 FAQ page for running the pipeline with single-end data. Taxonomy was assigned in dada2 using the Silva nr99 v138 training set.

UPGMA phylogenetic trees for each dataset were then constructed using the R package Phangorn (Schliep, 2011, Schliep, 2017) with the following commands

```
sequences<-getSequences(seqtab.final)
alignment <- AlignSeqs(DNAStringSet(sequences), anchor=NA)
phang.align <- phyDat(as(alignment, "matrix"), type="DNA")
dm <- dist.ml(phang.align)
treeUPMA <- upgma(dm)
```

Next, the data generated by dada2 was prepared for PICRUSt2 by creating an .fasta file of the ASV sequences and .biom table using the following R commands:

```
biomTable<-make_biom(t(seqtab.nochim), sample_metadata = NULL,
observation_metadata = NULL   id = NULL,matrix_element_type = "int")
write_biom(biomTable, biom_file="table.biom")
asvTable<-seqtab.nochim
write.table(asvTable, file="ASVTableNewDataDADAbimera.txt", row.names=TRUE ,
sep="\t")
write.fasta(sequences = as.list(sequences) , names = as.list(sequences), nbchar = 80,
file.out = "ASV.fasta")
```

This was then passed to PICRUSt2 (Douglas *et al*., 2020, Babera *et al* 2019, Czech, 2020, Louca, Doebelia, 2018) and run using the default parameters. The resulting data was then passed to microbiomeAnalyst (Dhariwal *et al*., 2017, Chong *et al*., 2020) . Filtering in microbiomeAnalyst was done in accordance with each respective parent study's methods in mind, and no transformation or refraction was performed. For weighted unifrac, unweighted unifrac, Shannon diversity index, Simpson diversity index, Chao1 diversity index, RNA seq, metagenome seq, and random forest default parameters were used. For LEFSe, features with a P-value (unadjusted) less than 0.1 and LDA score with an absolute value of 2 or more were identified as significant. For the pooled analysis, ASV tables from each study were merged using dada2's mergesequencetables. This ASV table was then assigned a taxonomy and used for the downstream processing. The pooled datasets were too large to generate an alignment using phangorn. Clustal-o (Lee *et al*., 2022) was used locally, with the alignment being passed off to FastTree (Price *et al*., 2010) using the -gtr and -nt options. Studies with paired stool samples and biopsies were split into biopsy and stool sample sets, and analyzed separately. Before filtering, ASVs without taxonomic assignment below Kingdom level were excluded. This removed 23,336 sequences. The data was filtered, removing ASVs with a count less than 4 or prevalence in less than 10% of samples. Features with a variance of less than 10% based on the interquartile range were also removed. This removed 5065 ASVs leaving 1619 ASVs for the remaining analysis. Samples with a library size of less than 3000 ASVs were removed, the data was then analyzed without filtering, normalization or scaling, with filtering, with filtering and with total sum scaling, and with the procedure described in Gibbons *et al.* (Gibbons *et al*., 2018). Taxa without a phylum assignment were removed from the pooled analysis. Analysis of pathway data was conducted using default parameters for each parent study, and default parameters plus refraction for the pooled analysis. Analysis of pathway data was done using Bray-curtis, RNA seq, metagenomeSeq, LEFSe and random forest, all of which were run using default parameters.

A phyloseq (McMurdie, Holmes, 2013) object was created and used to merge ASVs with identical taxonomy using phyloseq's glom_taxa method. ASVs were collapsed at the genus level, leaving only ASVs with genus level assignments. This reduced the original unfiltered ASV table from 57,943 ASVs to 799. The resulting ASV table was then uploaded to microbiomeAnalyst using the same protocol as above.

Discrepancies were found with the original controls in Bodkhe et al. To remedy this, new controls were found by searching for studies with accessible data using search queries: "Indian microbiome" and "Indian gut-microbiota". Only studies examining the 16S v4 variable regions were used. One study from the Delhi area of India, the same as Bodkhe *et al*., and the other from populations of Indians living in both rural and urban areas. These sequences were prepared and analyzed using the protocol described above. All three datasets had a large number of ASVs with unassigned taxonomy. To determine the identity of these sequences, ASVs without taxonomic assignment and those with no taxonomic assignment below the kingdom level were tabulated using a proprietary python script. ASVs with no taxonomic assignment were removed and added to a fasta file. ASVs with kingdom level assignment (bacteria) were allowed to remain. 10% of the sequences were then pulled from the resulting fasta file and clustered in mega using 95% similarity threshold (Kumar *et al*., 2018). This threshold was chosen as it represents the variability of the v4 16s variable region. One sequence was then pulled from each cluster and assigned an identity using BLAST with default parameters(Altschul *et al*., 1990).

Garcia-Mazcarro et al. contained data on GFD and GCD. The data was first analyzed looking only at disease state (celiac, NCGS or control) then looking at differences due to disease state and diet (celiac GCD, celiac GFD, etc).

Comparison in ASV assignment between Greengenes and Silva was carried out by assigning taxonomy to the Garcia-Mazcorro ASV table using both Silva nr99 v138 and Greengenes v12 databases. The resulting taxonomy tables were then analyzed for differences using a proprietary python script.

RESULTS OF INDIVIDUAL ANALYSIS

Bonder *et al.*

Bonder *et al.* had a small but significant change in alpha diversity indexes for Shannon (p = 0.0448 ANOVA) and Simpson diversity indexes(p =0.0039 ANOVA), but not Chao1 (p = 0.2743. Figure 1A), with the GFD having a higher alpha diversity compared to GCD.

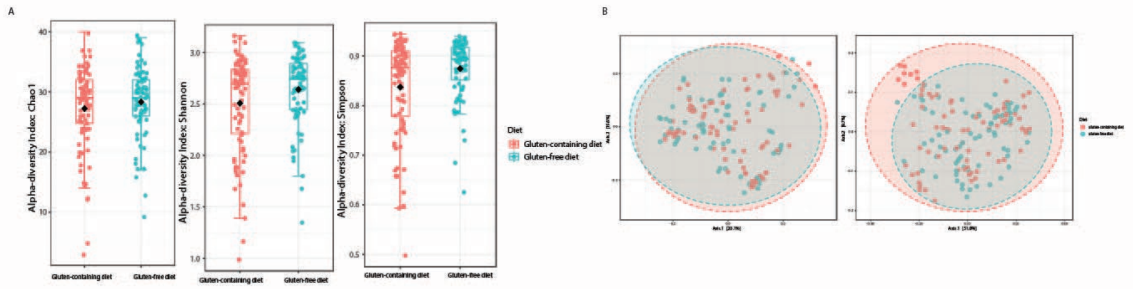

Figure 1 alpha and beta diversity analysis of Bonder et al.: A) The Chao1 alpha diversity (left) of the two study groups GFD is increased, but not significantly (p = 0.2743). Shannon diversity (middle) of the two study groups showed that GFD had a significant increase in alpha-diversity (p = 0.0448). Simpson alpha diversity index (right) showed GFD had a signifincant increase in diversity (p= 0.0039). B) Unweighted unifrac (left) and weighted unifrac (right) failed to produce clustering as a factor of diet.

No clustering was apparent for both weighted and unweighted unifrac values (unweighted unifrac p > 0.42, weighted unifrac p < 0.0508; PERMANOVA, Figure 1B). A single ASV was identified as being differentially abundant. This ASV corresponded to the genus *Faecalibacterium*, and was higher on GFD (Figure 2, FDR corrected p < 0.1, LEfSe LDA > 2.0). This ASV was noted as being significant in both metagenome-Seq and RNA-seq (DeSeq2, Love *et al* 2014) as well. Additionally, this ASV was noted as being the best predictor for diet status from the random forest analysis, however this model had a class error rate of 0.226 for GCD and 0.487 for GFD. No differentially abundant pathways, as predicted by PICRUSt2 were identified (Douglas *et al*., 2020). Random Forest class error rates for this model were 0.36 for GCD and 0.5 for GFD with an OOB error of 0.35.

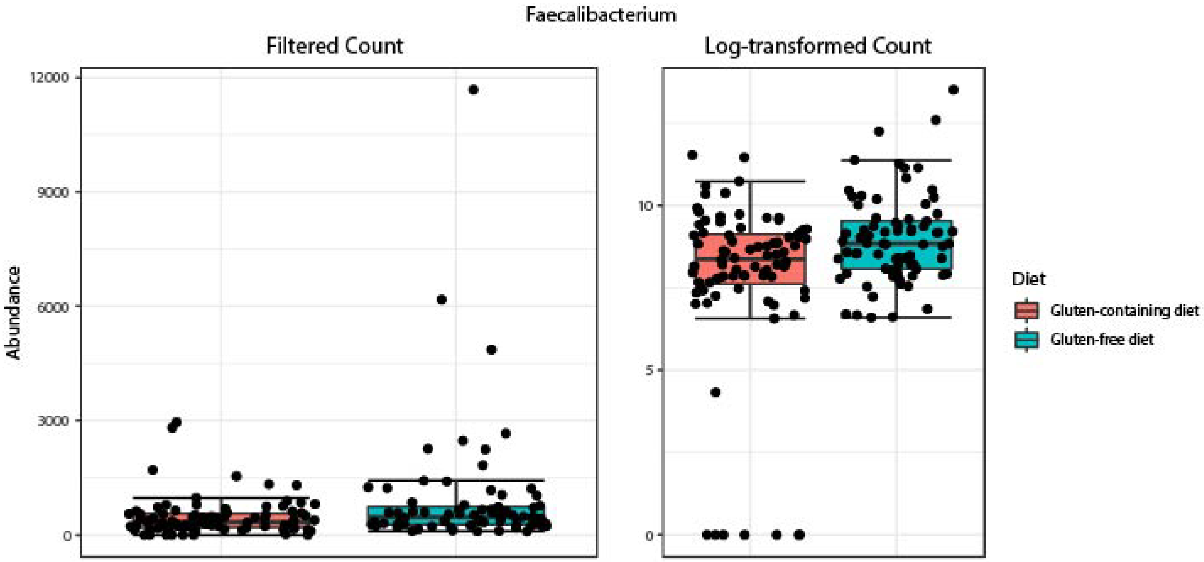

Figure 2 LEFSe results of Bonder et al: Boxplots showing the abundance of the genus Faecalibacterium between GFD and GCD. The left graph shows the raw group ASV abundance and the right log-transformed abundances. There was a signficant increase in this ASV in the gluten-free diet study group (FDR = 0.017, LDA = 2.21)

Garcia-Mazcorro *et al.*

Duodenal analysis

Duodenum biopsies from Garcia-Mazcorro *et al.* showed lowered alpha-diversity in CD patients compared to controls, with NIBD having the highest diversity, though these results were not significant for Chao1 (Figure 3A, Chao1 p = 0.1922, Shannon p = 0.046243, Simpson p = 0.09176 ANOVA). No clustering was evident using weighted and unweighted unifrac (Figure 3B, unweighted unifrac, weighted unifrac p < 0.034, PERMANOVA). The duodenum of celiac patients was characterized as having elevated ASVs belonging to *Phyllobacterium*, *Azospira*, and *Stenotrophomonas*, while the duodenum of NCGS patients had elevated ASV corresponding to the genera *Neisseria* and *Streptococcus*, with celiac and controls having similar average ASV counts for each. NCGS and control biopsies had similar levels of the genus *Fusobacterium*, with celiac patients having lowered ASVs corresponding to *Fusobacterium* (Figure 4).

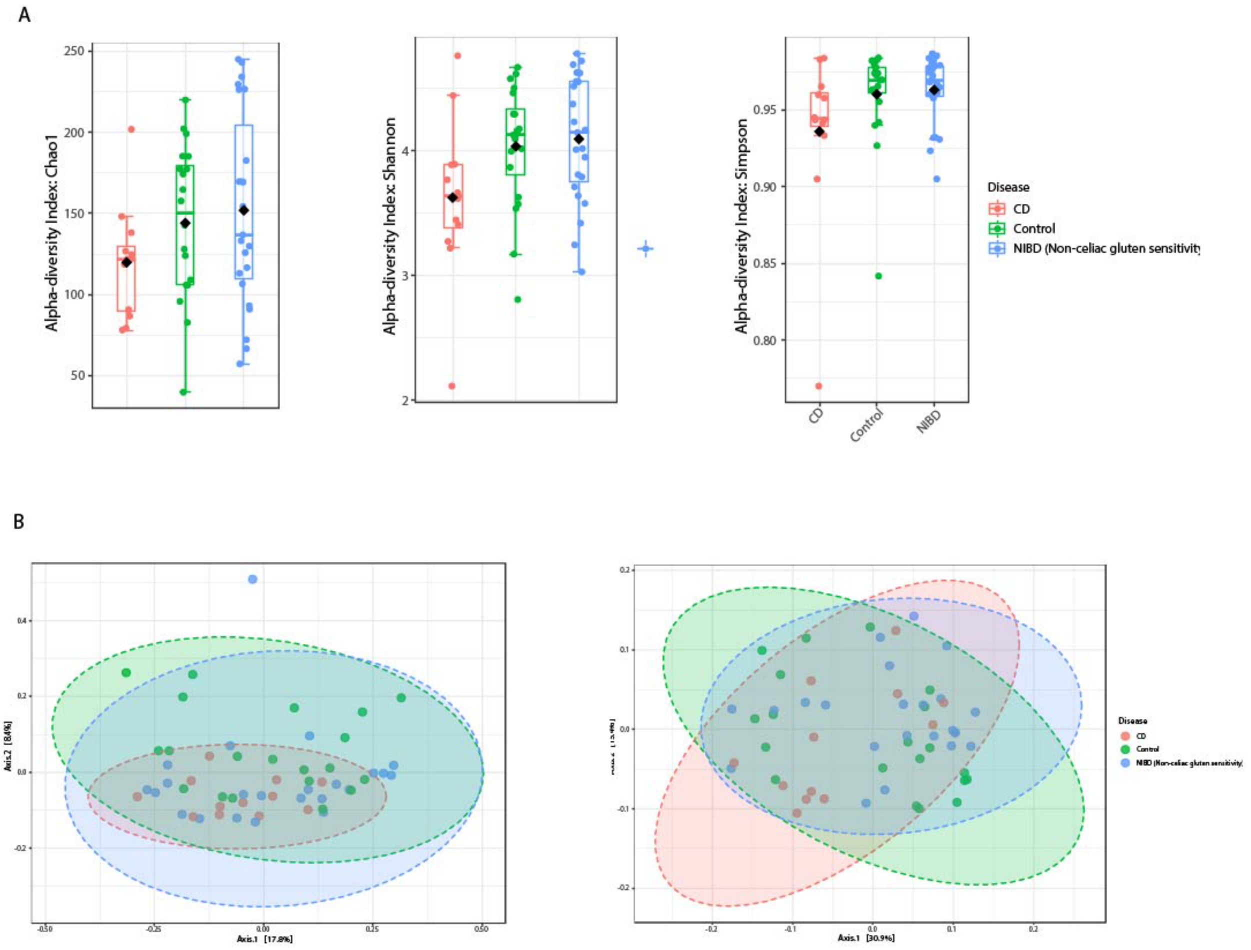

Figure 3 duodenal alpha and beta diversity analysis of Garcia-Mazcarorro et al. A) Alpha-diversity analysis of duodenum samples. No test was significant (P> .05). B) Unifrac analysis depicting clustering ofsamples based on community structure and relatedness of samples. Neither weighted unifrac (right) nor unweighted unifrac (left) wereable to produce accurate clusters for CD, NIBD, and controls.

RNA-seq (edgeR Robinson *et al*., 2010) analysis of pathways showed that the bacteria of celiac patients contained fewer taxa capable of menaquinone biosynthesis with pathways 1,4-dihydroxy-6-naphthoate biosynthesis II (PWY 7371) superpathway of demethyl menaquinone -6-biosynthesis II (PWY 7373), and superpathway of menaquinone-8 biosynthesis II(PWY 6263) being lowered in celiac patients. All features had LEfSE LDA scores greater than 2.0 and p-values below 0.1 for RNA-seq (EdgeR), but were not identified as significant using metagenome-Seq (Figure 5).These pathways had similar average counts for healthy and NCGS. Random forest analysis of pathways identified pwy 7371 and 6263 as being important predictors of disease state, with mean decrease in accuracy of 0.00024 and 0.0002 respectively. The random forest model had an OOB error of 0.662 and class errors of 0.583 for celiac, 0.789 for controls, and 0.478 for NCGS.

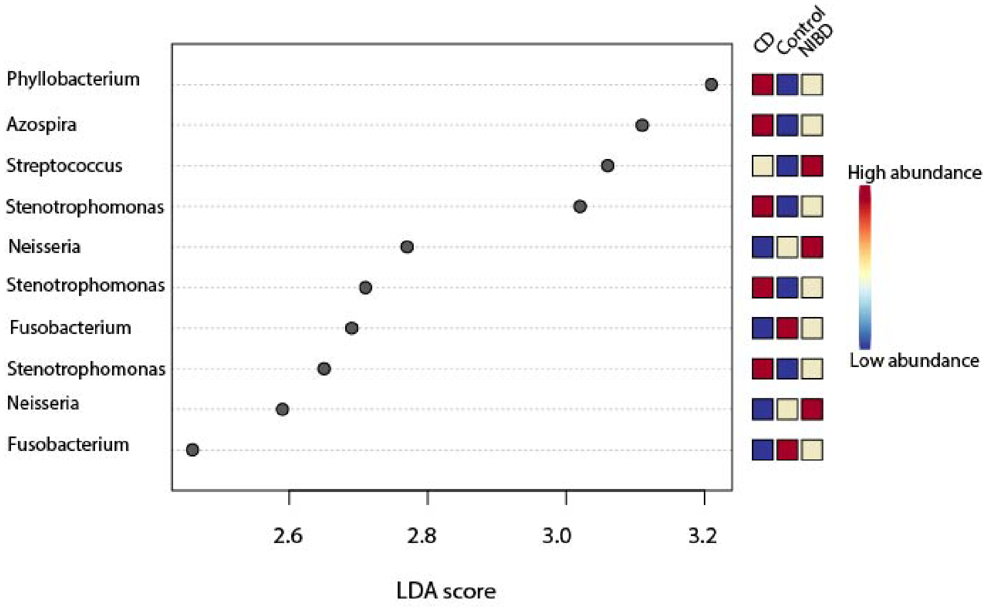

Figure 4 duodenal LEFSe results of Garcia-Mazcorro et al: LEFSe output of differentially abundant taxa from Duodenum biopsies of CD, NIBD (non-celiac gluten sensitivity), and controls. All taxa had P-values of 0.1or less and LDA scores of 2.0 or more. It should be noted that LDA scores are the main output from LEFSe and should be used to identify significant taxa opposed to traditional methods using adjusted or unadjusted P-values. Abundance is shown on the boxes to the right of the LDA scores with red colors indicating higher abundance relative to the other study groups and blue representing lower.

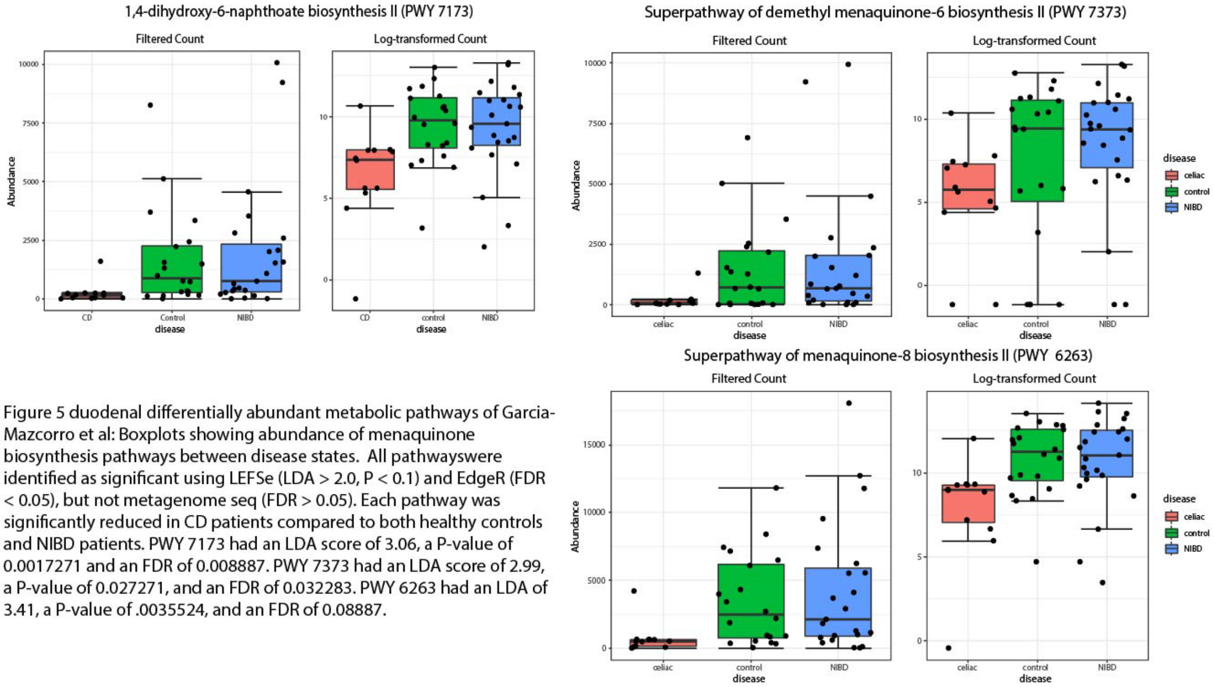

Figure 5 duodenal differentially abundant metabolic pathways of Garcia-Mazcorro et al: Boxplots showing abundance of menaquinone biosynthesis pathways between disease states. All pathwayswere identified as significant using LEFSe (LDA > 2.0, P < 0.1) and EdgeR (FDR < 0.05), but not metagenome seq (FDR > 0.05). Each pathway was significantly reduced in CD patients compared to both healthy controls and NIBD patients. PWY 7173 had an LDA score of 3.06, a P-value of 0.0017271 and an FDR of 0.008887. PWY 7373 had an LDA score of 2.99, a P-value of 0.027271, and an FDR of 0.032283. PWY 6263 had an LDA of 3.41, a P-value of .0035524, and an FDR of 0.08887.

Fecal analysis

Fecal samples from Garcia-Mazcorro *et al.* showed no significant differences in alpha diversity (Chao1 p = 0.79019, Shannon p = 0.61687, Simpson p = 0.58068 ANOVA, Figure 6A) or beta diversity, with no clusters forming for weighted and unweighted unifrac (p < 0.145 unweighted unifrac, p< 0.179 weighted unifrac PERMANOVA, Figure 6B). Stool samples of celiac patients had elevated levels of ASVs corresponding to *Pseudomonas* and *Novispirillum*, and lowered ASVs corresponding to *Haemophilus* while NCGS had elevated ASVs corresponding to *Clostridia* and *Collinsella*. Control samples had an abundance of Ruminococcus and *Bifidobacterium* (p < 0.1, LDA > 2.0, Figure 7), with NCGS and celiac having reduced ASV counts.

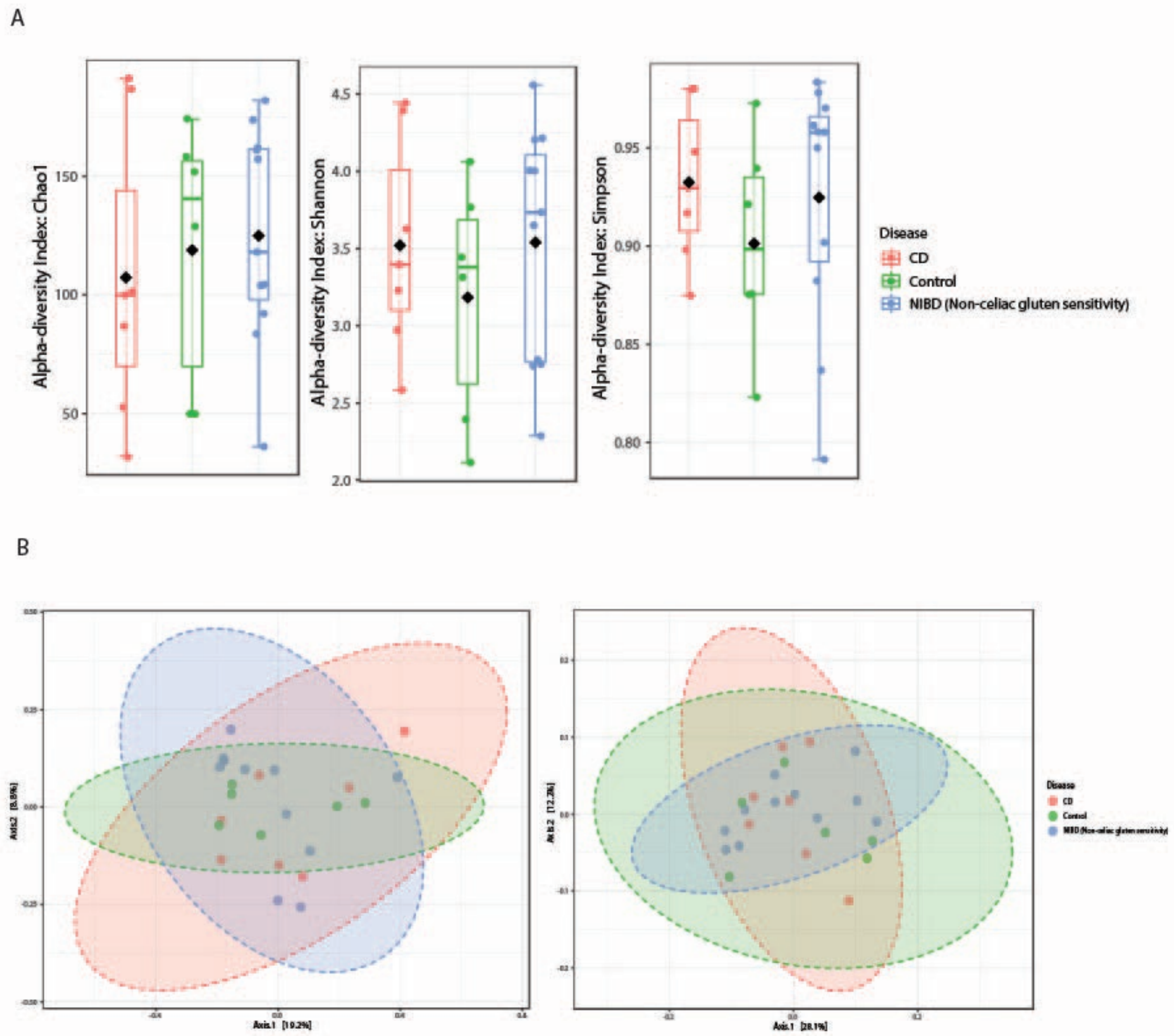

Figure 6 fecal alpha and beta diversity analysis of Garcia-Mazcorro et al: A) Alpha diversity analysis of stool samples. No significant difference was noted in Chao1, Shannon or Simpson diversity indexes (P > 0.05). B) Weighted unifrac (right) and unweighted unifrac (left) were unable to produce clusters based on disease.

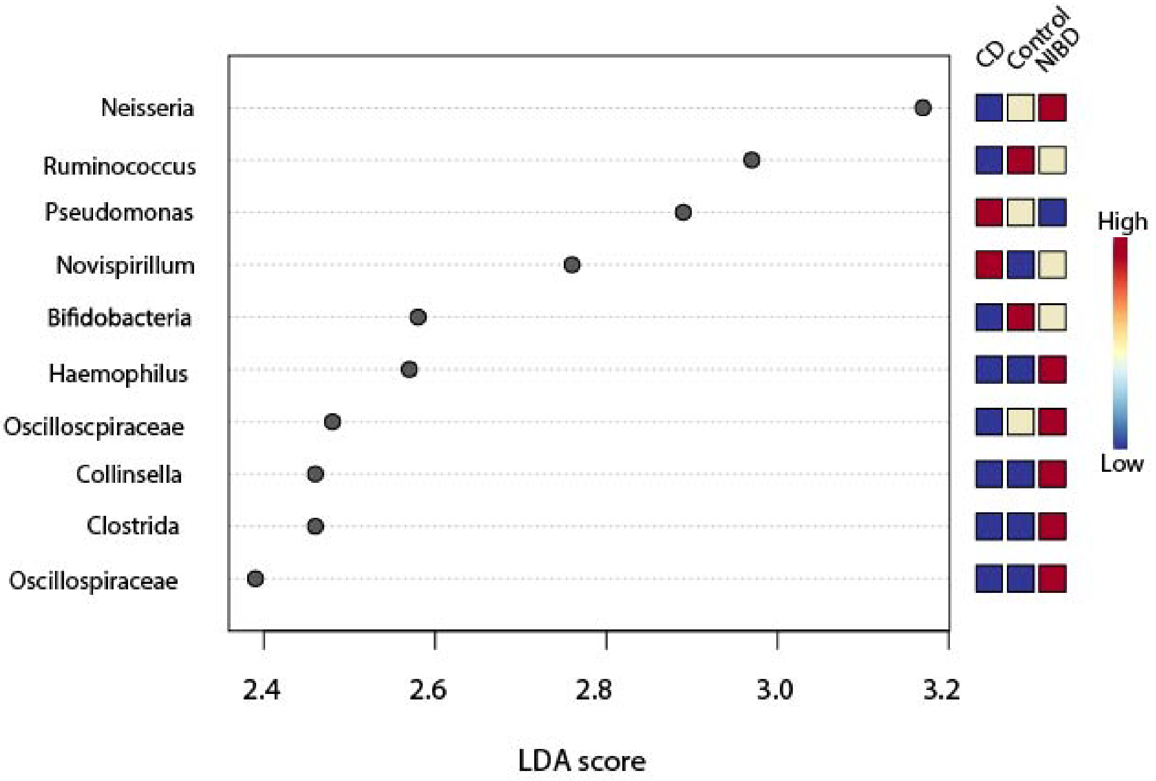

Figure 7 fecal LEFSe results of Garcia-Mazcorro et al: LEFSe analysis of stool microbiota of Mexican healthy, NIBD, and CD patients. All features had LDA scores greater than 2.0 and P-values less than 0.1. Blue colors indicate a lowered presence relative to other disease states while red colors indicate an elevated presence relative to other disease states.

114 differentially abundant pathways were identified with edgeR, none were shared between LEFSe, metagenomeSeq and RNA seq, however 14 were shared between LEFSe and EdgeR. 319 differentially abundant KEGG orthologs were identified as being differentially abundant between LEFSe and RNA seq, with none being identified as significant using metagenomeSeq (Ogata *et al*., 1999).

Greengenes vs SILVA taxonomy

This data was also analyzed using the Greengenes database version 12 to determine to what extent pipeline choice impacts the results. Of the top 10 largest effect sizes producing taxa in the duodenum, 2 were assigned different taxonomy between Greengenes v12 and Silva nr99 v138 (DeSantis *et al*., 2006, Yilmaz *et al., 2014*). ASV 14 was identified by Silva as *Stenotrophomonas*, while Greengenes assigned it as *Clostridiales* and ASV 20 was identified as *Neisseria* by Silva and *Actinobacillus* by Greengenes (Table 1A).

For stool samples, the only one ASV had a mismatching assignment with ASV 124 being identified as *Oscillospriaceae* and *Ruminococcaceae* by Silva and Greengenes respectively. Despite this seeming consensus among the largest effect size producing taxa, there was only a 6.3% overall similarity in ASV assignment (Table 1B). Both analyses had the same overall similarity as the ASV table used in each case was the same.

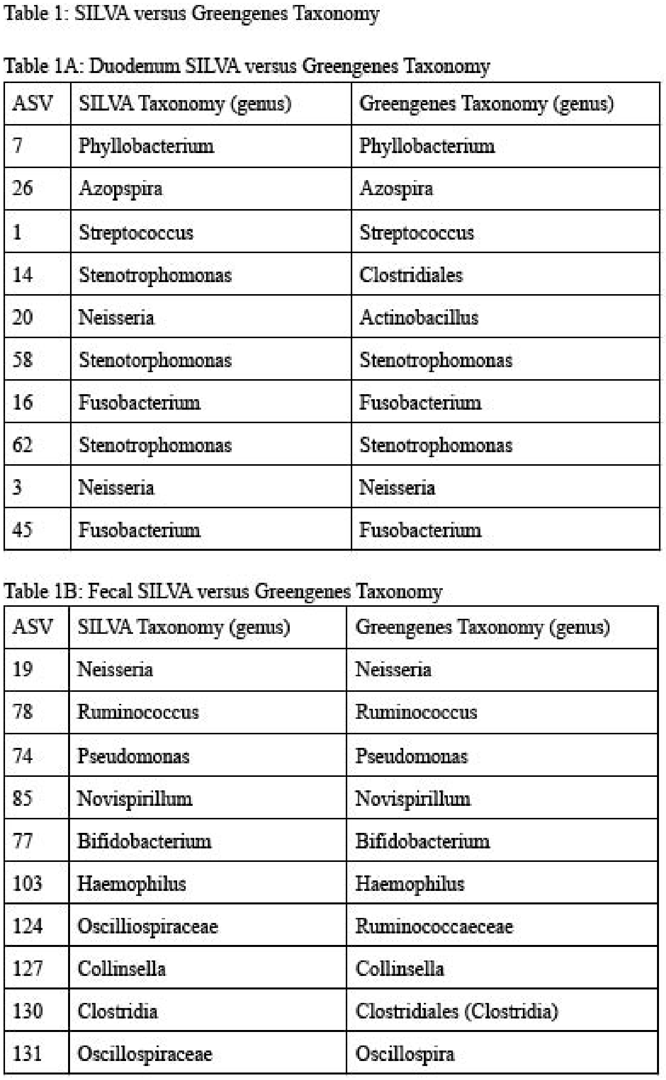

Table 1: SILVA versus Greengenes Taxonomy

Table 1A: Duodenum SILVA versus Greengenes Taxonomy

ASV	SILVA Taxonomy (genus)	Greengenes Taxonomy (genus)
7	Phyllobacterium	Phyllobacterium
26	Azopspira	Azospira
1	Streptococcus	Streptococcus
14	Stenotrophomonas	Clostridiales
20	Neisseria	Actinobacillus
58	Stenotorphomonas	Stenotrophomonas
16	Fusobacterium	Fusobacterium
62	Stenotrophomonas	Stenotrophomonas
3	Neisseria	Neisseria
45	Fusobacterium	Fusobacterium

Table 1B: Fecal SILVA versus Greengenes Taxonomy

ASV	SILVA Taxonomy (genus)	Greengenes Taxonomy (genus)
19	Neisseria	Neisseria
78	Ruminococcus	Ruminococcus
74	Pseudomonas	Pseudomonas
85	Novispirillum	Novispirillum
77	Bifidobacterium	Bifidobacterium
103	Haemophilus	Haemophilus
124	Oscilliospiraceae	Ruminococcaeceae
127	Collinsella	Collinsella
130	Clostridia	Clostridiales (Clostridia)
131	Oscillospiraceae	Oscillospira

Pre and post-treatment Mexican CD microbiome

Garcia-Mazcarro *et al.* was then analyzed looking at disease state and diet together, dividing each category into GFD and GCD time points, to look for differences in taxa which remained the same regardless of diet. There were no discernible differences in alpha or beta diversity for both stool samples and biopsies between disease states. ASVs corresponding to *Pseudomonas* and *Stenotrophomonas* (LEFSe LDA > 2.0) were elevated in celiac biopsies and stool regardless of dietary status. No significant pathways were identified as being differentially abundant between time points for both stool and biopsies. No KEGG orthologs identified biopsies or feces.

Bodkhe *et al.*

Stool analysis

For the individual study analysis of Bodkhe *et al.*, biopsies and non-celiac samples were removed, leaving just stool samples to be analyzed with controls taken from other datasets. The stool samples showed an elevated Shannon diversity index at the feature level in celiac patients, however at every taxonomic level (genus through phylum) celiacs were characterized by lowered alpha diversity (Shannon index feature p = 1.0071 $*10^{-12}$ ANOVA, Shannon genus p = 3.3627$*10^{-31}$ANOVA, Shannon family p = 6.7394 $*10^{-28}$ANOVA, Shannon order p = 3.0946$*10^{-29}$ANOVA, Shannon Class, p = 2.9985$*10^{-31}$ANOV, Shannon Phylum p = 9.7104$*10^{-31}$ANOVA, Figure 8A). Beta-diversity analysis showed clustering for controls on a basis of region, with celiac samples clustering distinctly from either control cluster (p< .05, Bray-Curtis index PERMANOVA, Figure 8B). Unifrac also displayed clustering at the feature level for both weighted and unweighted measures (p< 0.001 PERMANOVA, Figure 8C). Upon investigation of the abundance plots, it was noted that many sequences from the three studies had a large degree of unassigned reads (Figure 8D). Of the original 38,005 ASVs, 4710 had no taxonomic assignment and 11527 had only phylum level assignment. Those without taxonomic assignment were removed, leaving 33,295 ASVs (87%). 10% of the reads without taxonomic assignment (both unclassified and without assignment below phylum) were pulled for clustering in mega. This created 32 clusters, with one cluster comprising 96% of the data. 97% of the DNA in the clusters correspond to uncultured 16S bacterial DNA, 1% was determined to be contaminating human DNA and the remaining 1% was split between viral DNA, fungal DNA and gDNA from *Bacteroidetes*, *Akkermansia*, and *Bifidobacterium*. Reanalysis of the data with unassigned bacteria removed (only those without taxonomic assignment) showed a similar trend as earlier, with the CD alpha diversity being elevated when compared to controls at the feature-level, and lowered in higher taxonomic (Shannon feature p = 1.6529*10^{-5} ANOVA, Shannon genus p = 1.6491*10^{-24}ANOVA, Shannon Family p = 9.4673*10^{-21}ANOVA, Shannon order p = 4.4937*10^{-16}ANOVA, Shannon class p = 5.527*10^{-10} ANOVA, Shannon phylum p = 2.7197*10^{-13} ANOVA, p < .05, Figure 8D, Figure 8E). Indicating that the increase in diversity was due to the presence of unclassified bacteria. Weighted and unweighted unifrac displayed similar results as previously as well, however clustering was no longer noted at taxonomic levels genus and higher (p< 0.001, PERMANOVA, Figure 8F).]

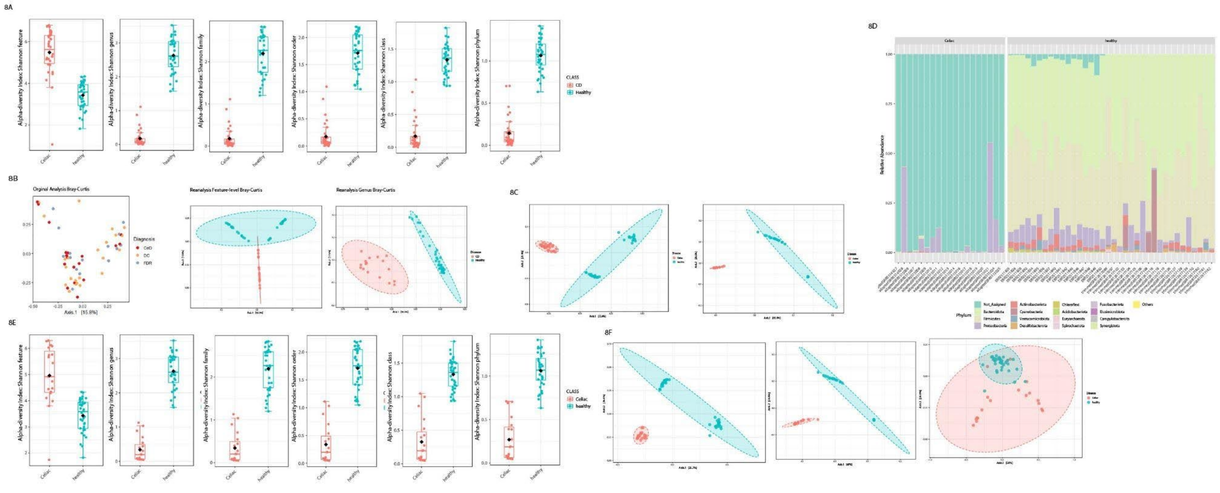

Figure 8 alpha and beta diversity of Bodkhe et al A) Shannon diversity index the from genus to phylum level. CD patients had enriched stool diversity at the feature level, while controls had increased diversity at the genus through phylum levels (Shannon index feature P = 1.0071 *10-12, Shannon genus P = 3.3627*10-31, Shannon family P = 6.7394 *10-28, Shannon order P = 3.0946*10-29, Shannon Class, P = 2.9985*10-31, Shannon Phylum P = 9.7104*10-31). B) Original analysis of Bray-Curtis index versus reanalysis at the feature and genus level. Reanalysis showed distinct clustering at feature and genus levels (P< 0.01). C) Unifrac unweighted (left) and weighted (right) analysis of CD and non-CD stool samples showed distinct clustering (P<0.001). D) Stacked area abundance bar plots at the phylum level. CD samples were characterized by a large degree of unassigned taxa. E) Shannon diversity of stool samples with unassigned taxa (Kingdom level removed). CD samples had increased alpha diversity at the feature level but loweralpha diversity at all other taxonomic levels (Shannon feature P = 1.6529*10-5, Shannon genus P = 1.6491*10-24, Shannon Family P = 9.4673*10-21, Shannon order P = 4.4937*10-16, Shannon class P = 5.527*10-10. E) Shannon phylum (P = 2.7197*10-13). F) Unweighted unifrac (right) and weighted unifrac (feature-level middle, genus right) analysis. Weighted unifrac only produced clusters at the feature level while unweighted did for all taxonomic levels (P< 0.001).

LEFSe

LEFSe analysis had mixed results with ASVs corresponding to *Prevotella 9* and *Bifidobacterium* elevated or reduced levels depending upon ASV. Healthy samples had elevated

Pseudobutyrivibrio and *Acinetobacter*, while celiac patients had an elevated ASV identified as *Bacteroidales* (Figure 9). All taxa had LDA scores larger than 2.0 and P-values below or equal to 0.1.

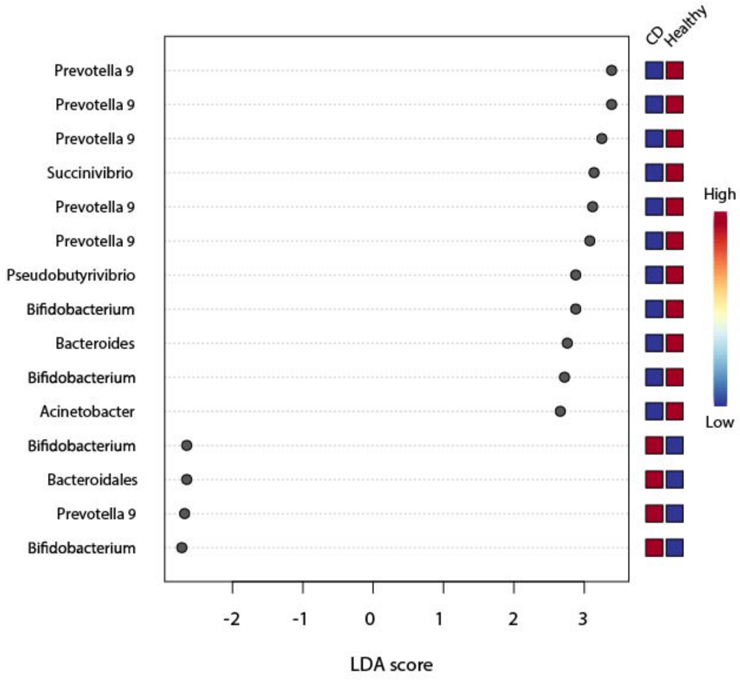

Figure 9 LEFSe results of Bodkhe et al: LEFSe analysis of differentially abundant Indian stool taxa from both health and CD patients. Blue indicates a lower relative abundance, while red indicates a higher relative abundance. All taxa had LDA scores with an absolute value of 2.0 or greater and P-values less than or equal to 0.1 .

Pathways

Analysis of pathways showed that samples clustered in accordance with region, with a mixed control/celiac cluster forming (both sets of data taken from the Delhi region of India) and control cluster. No clustering occurred on a basis of disease state. 69 pathways were identified as significant between EdgeR, LEFSe and metagenomeSeq. The top 10 pathways with the largest effect size included anaerobic gondoate biosynthesis (PWY 7663), incomplete reductive TCA cycle (P42 PWY), L-lysine biosynthesis II (PWY 2941), cis-vaccenate biosynthesis (PWY 5973), super pathway of adenosylcobalain salvage from cobinamide II (PWY 6269) adenosylcobalamin biosynthesis from adenosylcobinamide -GDP I (PWY 5509)(Figure 10), lipid IVA biosynthesis (*E. coli*)(NAGLIPASYN), superpathway of adenosylcobalaimin salvage from coinamide I(COBALSYN PWY)(Figure 12), preQ biosynthesis (PWY 6703) and Kdo transfer of lipid IVA (*Chlamydia*)(PWY 6467) with all pathways but PWY 2941 being lowered in CD compared to controls. All pathways had an LDA score greater than or equal to 4.14 and P-values less than 0.1 (Figure 10).   Random forest analysis of pathways produced an OOB error of 0.0755 with class errors of 0.235 and 0 for celiac and healthy respectively. None of the best predictors matched what was identified as significant in the combined LEFSe-rna- seq(EdgeR)-metagenome-Seq analysis.

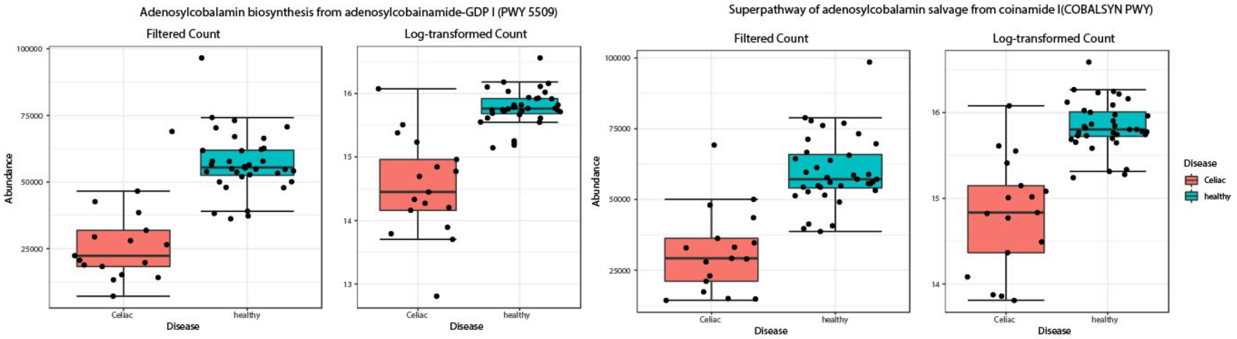

Figure 10. differentially abundant metabolic pathways of Bodkhe et al: Box plots showing the abundance of vitamin B-producing pathways in Indian fecal samples. Plots on the left side shows the abundance and log-transformed abundance of adenosylcobalamin biosynthesis from adenosylcobinamide-GDP I (PWY 5509) in both CD (red) and healthy (blue) Indian stool samples. This pathway was significantly lowered in CD patients (P = 7.41817*10-6 P = 1.3328 * 10-9 , LDA = 4.17). Similarly, the plots on the right side show the abundance of the superpathway of adenosylcobalamin salvage from coinamide I(COBALSYN PWY) in both healthy and CD stool samples of Indian patients. This pathway was also significantly lowered in CD samples (P = 3.9879 * 10-6, P = 3.1067 * 10 -9, LDA = 4.15).

RESULTS OF POOLED DUODENUM ANALYSIS

Pooled Raw Duodenum

Community Structure Analysis

6684 ASVs were included in the raw analysis, after the removal of unassigned taxa. Chao1 alpha diversity showed that CD and NIBD had similar alpha-diversities with healthy samples having a lower and less varied distribution. Shannon showed that CD had a similar average alpha -diversity compared to the other study groups and Simpson showed CD having a lowered average alpha-diversity, however this was not significant. (Figure 11A). When examining CD versus non-CD a similar trend was observed, with Chao1 showing an increased diversity for CD versus non-CD, similar levels for Shannon, and lowered alpha-diversity for CD. Once again these results were not significant. (Figure 11B). When looking at samples as a factor of geographic region it was observed that Indian samples had a higher alpha diversity across metrics, though this was only significant for Chao1 ($p = 0.0014783$ ANOVA , Figure 11C).

Clustering was only achieved as a basis of region rather than disease or CD status using unweighted unifrac ($p< 0.001$ PERMANOVA). A similar trend was observed for weighted Unifrac, with segregation of samples only occurring on a basis of geographic region($p< 0.001$ PERMANOVA, Figure 11D-F).

Stacked area bar plots of percentage abundance showed significant differences in community structure as a factor of both disease and region of isolation. Indian CD patients had more *Fusobacteria* compared to NIBD counterparts with Mexican CD patients having less compared to NIBD and controls.

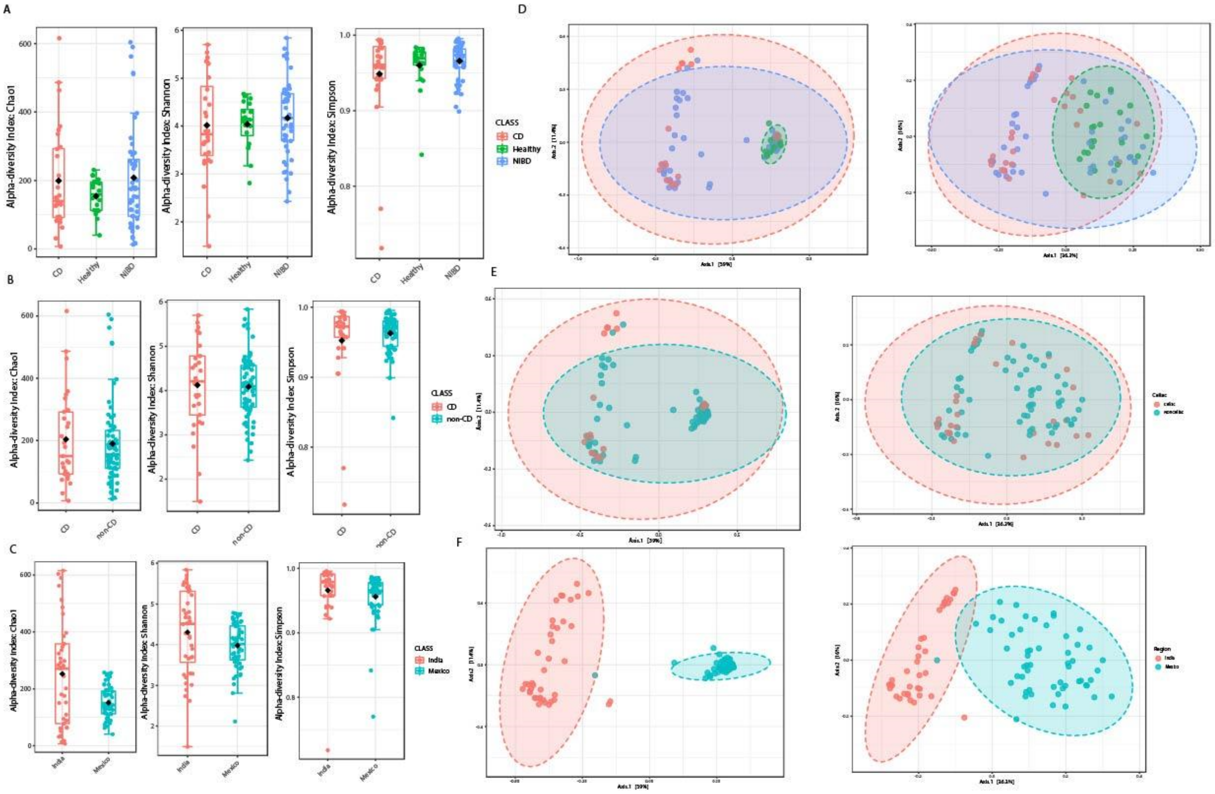

Figure 11. Alpha and beta diversity of raw duodenum data. A) Alpha-diversity as a factor of disease, with red representing CD, green healthy and blue NIBD. No significant difference was noted for Chao1, Shannon or Simpson (P >0.05). B) Alpha-diversity as a factor of CD status with red representing CD and blue non-CD. No significant difference was noted (P >0.05). C) Alpha-diversity as a factor of region with red representing Indian samples and blue representing Mexicansamples. Indian samples were significantly enriched for Chao1 (P= 0.0014783) but not Shannon or Simpson (P > 0.05). D) Unweighted (left) and weighted (right) unifrac analysis of CD (red), NIBD(blue), and healthy(green). E) Unweighted (left) and weighted (right) unifrac analysis of CD (red) versus non-CD (blue). F) Unweighted (left) and weighted (right) unifrac analysis, with Indian samples in red and Mexican samples in blue. All unfrac analysis had P-values less than 0.001.

Mexican CD patients tend to have less *Bacteroidota* compared to non-CD counterparts, with the opposite trend being observed in Indian samples. In most cases, samples tended to look more like those taken from the same region as opposed to those from the same disease (Figure 12).

LEFSe

When looking at a factor of disease, LEFSe identified 25 differentially abundant taxa. 4 ASVs corresponding to *Acinetobacter* were identified, with all being enriched only in CD. 3 ASVs of *Fusobacterium* were found to be reduced in CD and found in similar abundance in controls and NIBD patients as well as 2 ASVs of *Haemophilus* following a similar trend.

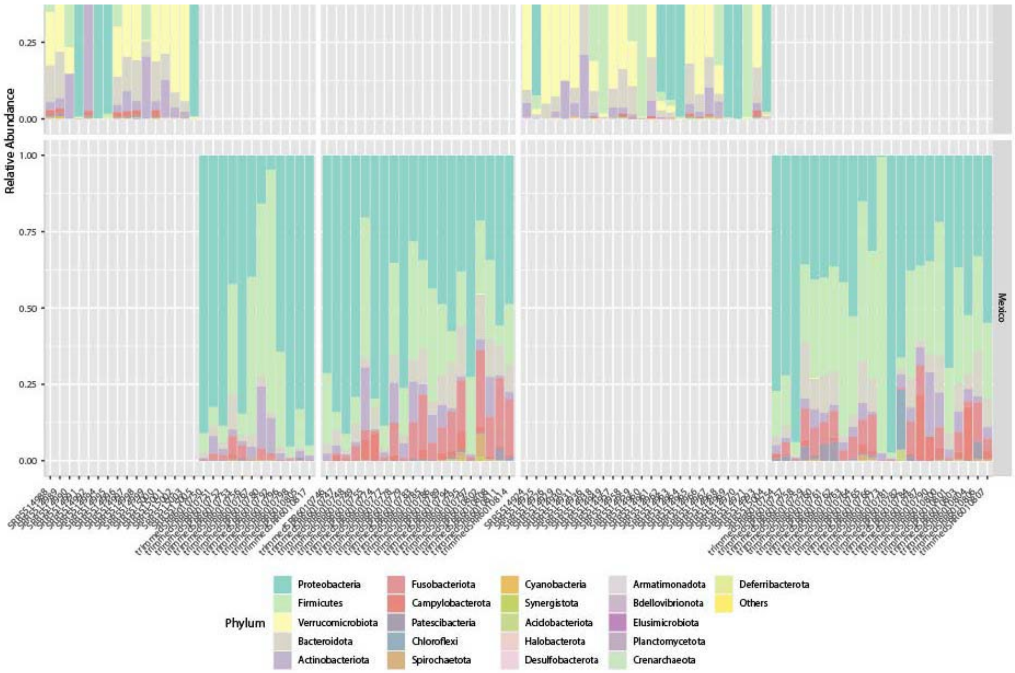

Figure 12. pooled raw duodenal abundance plots: Raw duodenum stacked area plots showing percentage abundance at the phylum level. Differences in community structure were noted as a factor of both disease and region of isolation

One ASV of *Rothia* was found to be abundant in both CD and controls with a lowered abundance in NIBD. 2 ASVs of *Pseudomonas* were also identified with both being elevated in CD and controls compared to NIBD (Figure 13A). When examining CD versus non-CD, 9 taxa were identified, all of which were enriched in CD compared to non-CD. 8 of the 9 ASVs were identified as *Acinetobacter* with the remaining being identified as *Moraxellaceae* (Figure 13B). 149 taxa were identified as significant when looking at the region, with only two genera matching those from the diseased (CD vs. NIBD vs. healthy) . With *Pseudomonas* being elevated in Indian samples and *Haemophilus* in Mexican (Figure 13C).

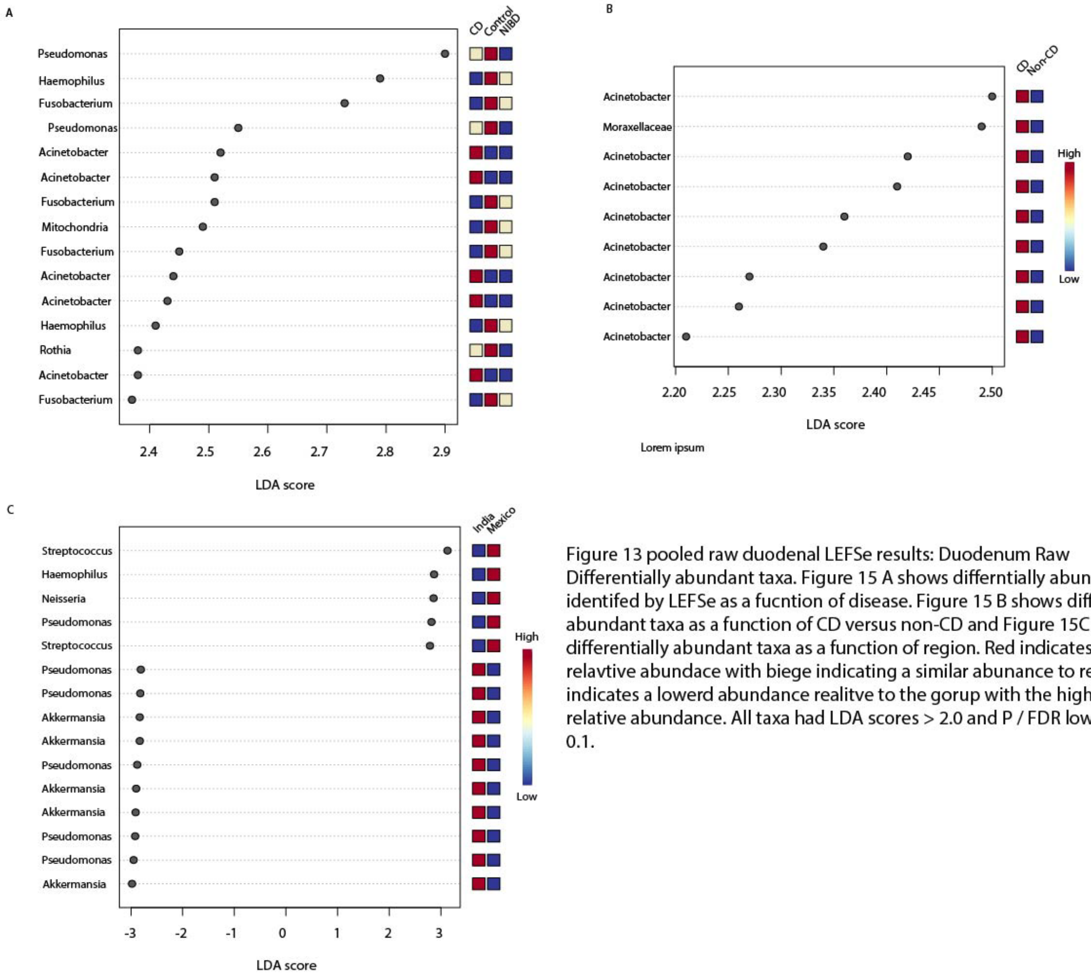

Figure 13 pooled raw duodenal LEFSe results: Duodenum Raw Differentially abundant taxa. Figure 15 A shows differntially abundant taxa identifed by LEFSe as a fucntion of disease. Figure 15 B shows diffentailly abundant taxa as a function of CD versus non-CD and Figure 15C shows differentially abundant taxa as a function of region. Red indicates a higher relavtive abundace with biege indicating a similar abunance to red. Blue indicates a lowerd abundance realitve to the gorup with the highest relative abundance. All taxa had LDA scores > 2.0 and P / FDR lowerd than 0.1.

Random Forest

Random forest as a factor of disease had an OOB of 0.5 with class erros of 0.964, 1.0 and 0.0208 for CD, NIBD and healthy respectively. When looking at a factor of CD versus non-CD the model had an OOB of 0.302 with class errors of 0.966 and 0.0149 for CD and non-CD respectively. When looking at a factor of geographic region, the model had an OOB of 0.0833 with class errors of 0.195 for the Indian data and 0.0 for Mexican data.

Pooled Filtered Duodenum

Filtering removed 4681 low abundance features and 201 low variance features leaving 1802 features for analysis.

Community Structure Analysis

Chao1 showed that CD and NIBD had higher alpha-diversity than healthy, though this was not significant (p = 0.37608 ANOVA). Shannon index showed that NIBD had the highest diversity followed by healthy then CD. This once again was not significant (p = 0.48161 ANOVA). Simpson index showed that CD had the lowest alpha diversity with NIBD having the highest, although this also was not significant (p = 0.19143 ANOVA, Figure 14A).
Chao1 and Shannon indices showed that CD and non-CD had similar alpha-diversities (p = 0.19143, p = 0.91247, ANOVA). Simpson showed that CD had a lowered average alpha-diversity, however this was not deemed significant (p =0.30967 ANOVA, Figure 14B). Chao1 (p = 0.031825 ANOVA), Shannon (p = 0.24744 ANOVA) and Simpson (p = 0.664 ANOVA) indices all showed that Indian samples had a greater alpha-diversity, though this was only significant for Chao1 (Figure 14C).

Clustering was only achieved as a basis of geographic region rather than disease or CD status using unweighted unifrac (p< 0.001 PERMANOVA) A similar trend was observed for weighted unifrac, with segregation of samples only occurring as a basis of region (p< 0.001 PERMANOVA, Figure 14D-F). Stacked area bar plots of percentage abundance once again showed differences in community structure as both a factor of disease and region, with Indian samples having large proportions of *Verrucomicrobitoa*. Indian CD samples were characterized by an abundance of *Bacteriodata* with Mexican CD samples having the opposite trend. Mexican samples Contained *Fusobacteriota*, which appeared to be largely absent from Indian samples, with Mexican CD samples having lowered abundance compared to NIBD and controls.

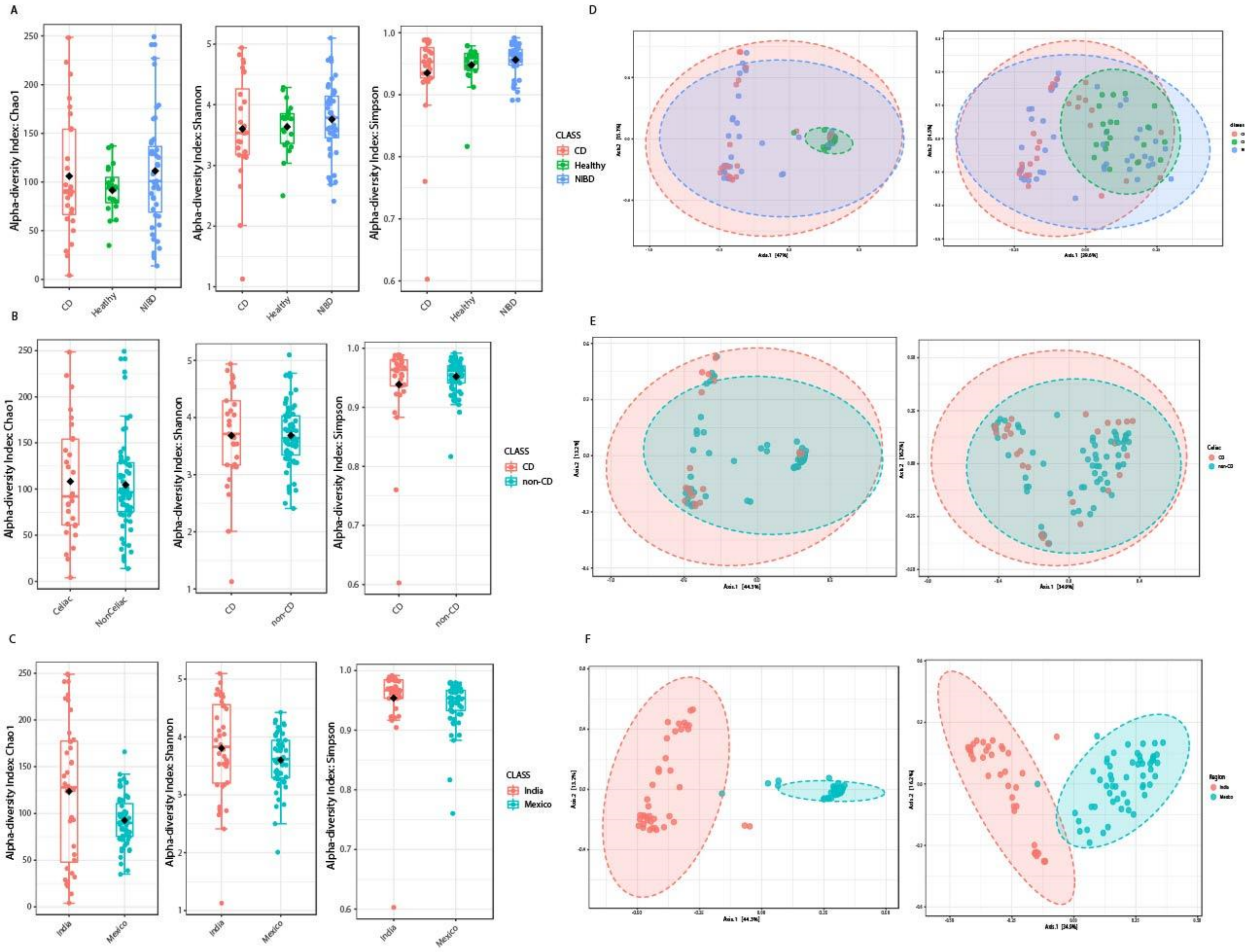

Figure 14. pooled filtered alpha and beta diversity analysis: A) Alpha-diversity as a factor of disease, with red repsenting CD, green healthy, and blue NIBD. No significant difference was noted for Chao1, Shannon or Simpson (P >0.05).B) Alpha-diversity as a factor of CD status with red representing CD and blue non-CD. No significant difference was noted (P >0.05). C) Alpha-diversity as a factor of region, with red representing Indian samples and blue representing Mexican samples. Indian samples were significantly enriched for Chao1 (P = 0.031825) but not Shannon or Simpson (P > 0.05). D) Unweighted (left) and weighted (right) unifrac analysis of CD (red), NIBD (blue,) and healthy (green). E) Unweighted (left) and weighted (right) unifrac analysis of CD (red) versus non-CD (blue). F) Unweighted (left) and weighted (right) unifrac analysis of Indian samples in red and Mexican samples in blue. All unifrac analysis yieldedP-values less than 0.001.

Once again, samples from the same region tended to have a similar community structure compared to those of a similar disease state (Figure 15).

LEFSe

47 differentially abundant taxa were identified using LEFSe. 3 ASVs of *Akkermansia* were identified as being elevated in both CD and NIBD. 3 ASVs of *Fusobacterium* were identified as being lowered only in CD. Two genera of *Pseudomonas* were found to be elevated in both controls and CD with a lowered relative abundance in NIBD.

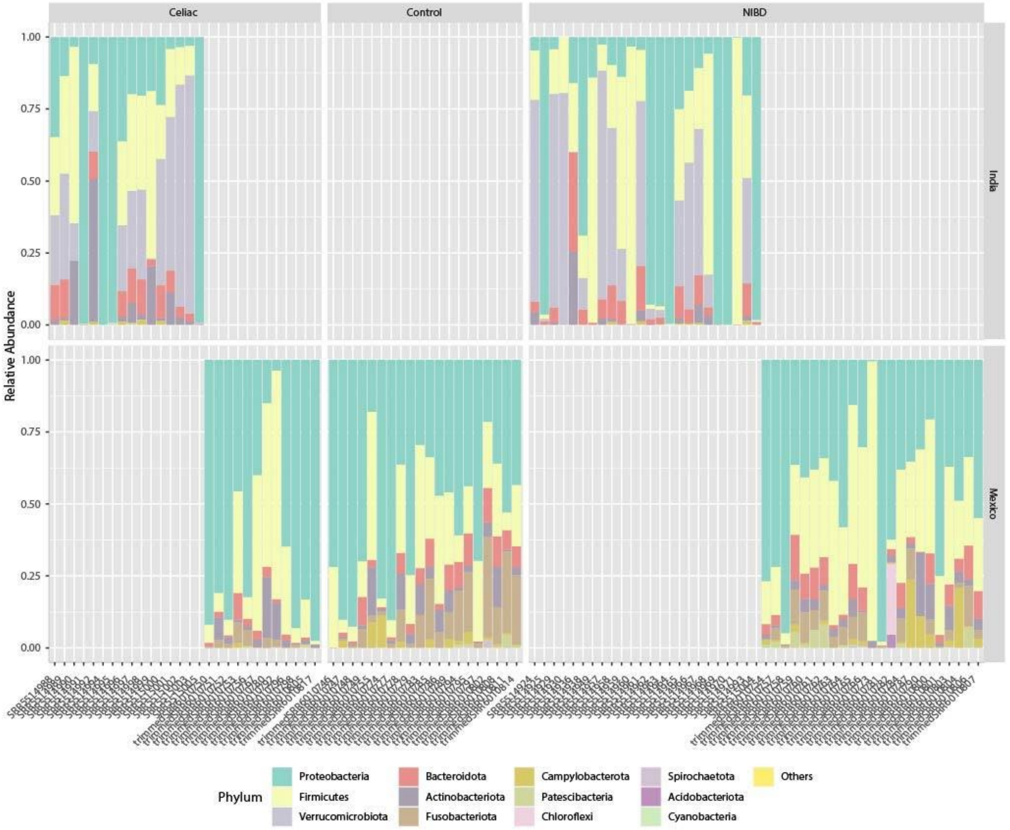

Figure 15 pooled filtered duodenal abundance plots: Taxonomic distribution of filtered duodenal biospies at the phylum level. Shown are percent abundance stacked area barplots of duodenum biospies from Indian CD and NIBD patientsand Mexican CD, healthy, and NIBD patients at the phylum level. Differences in community structure were noted as both a factor of disease and region of isolation.

Haemophilus was found to be elevated in both NIBD and CD and lowered in CD. 1 ASV of *Moraxellaceae* and two *Acinetobacter* (a member of the family Moraxellaceae) were identified as being elevated in CD only (Figure 16A). CD versus non-CD showed 16 differentially abundant ASVs, all of which were elevated in CD. 14 of the ASVs corresponded to *Acinetobacter* with a single ASV belonging to the family *Moraxellaceae* (Figure 16B). Regional analysis identified 155 differentially abundant taxa with 3 overlapping from the previous analysis. *Neisseria* was associated with Mexican samples while *Pseudomonas* and *Akkermansia* were associated with Indian samples (Figure 16C) All identified taxa had LEFSe LDA scores of 2.0 or greater and FDR/p-values less than or equal to 0.1.

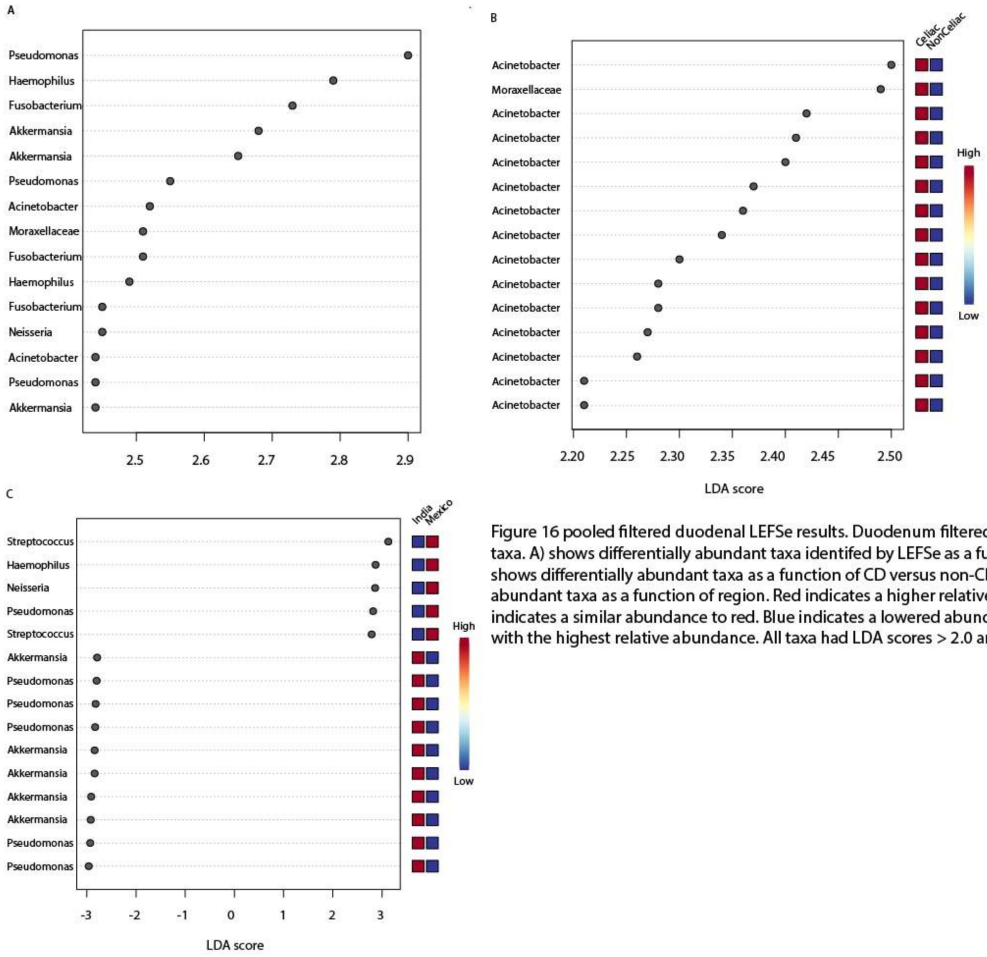

Figure 16 pooled filtered duodenal LEFSe results. Duodenum filtered differentially abundant taxa. A) shows differentially abundant taxa identifed by LEFSe as a function of disease. B) shows differentially abundant taxa as a function of CD versus non-CD. C) shows differentially abundant taxa as a function of region. Red indicates a higher relative abundance, while biege indicates a similar abundance to red. Blue indicates a lowered abundance relative to the group with the highest relative abundance. All taxa had LDA scores > 2.0 and P / FDR lower than 0.1.

Random Forest

When looking at a factor of disease, random forest analysis had an OOB of 0.453 with class errors of 0.821 for CD, 0.85 for control, and 0.0638 for NIBD. When looking as a factor of cd versus non-CD, the model had an OOB of 0.253 with class errors of 0.793 for CD and 0.0152 for non-CD. When looking at a factor of region the model had an OOB of 0 as well as class errors of 0 for both Mexican and Indian samples.

Pooled Filtered and Scaled Duodenum

The data was next analyzed using the same filtering parameters as earlier with total sum scaling applied.

Community Structure analysis

When looking at a factor of disease, Chao1 showed the NIBD had the highest alpha diversity and healthy samples having a lowered diversity, with CD in between. This however, was not significant (p = 0.27608 ANOVA) . Shannon showed that CD and healthy samples had similar alpha diversities with NIBD having the highest, however this was not significant (p = 0.48161 ANOVA). Simpson diversity index showed that CD had the lowest average alpha diversity with healthy and NIBD having elevated averages, once again this was not significant (p =0.19143 ANOVA, Figure 17A). Chao1 and Shannon indices showed that CD and non-CD had similar alpha-diversities (p = 0.83982, p = 0.91247 ANOVA) with Simpson showing CD with a lowered average alpha diversity, however this was not significant (p = 0.30967 ANOVA, Figure 17B). Chao1, Shannon and Simpson showed that Indian samples had greater alpha-diversity, though this was only significant for Chao1 (p = 0.027265 ANOVA, Figure 17C).

Clustering was only achieved as a basis of geographic region rather than disease or CD status using unweighted unifrac (p< 0.001 PERMANOVA). A similar trend was observed for weighted Unifrac, with segregation of samples only occurring as a basis of region (p< 0.001 PERMANOVA, Figure 17D-F). No clustering was apparent as a factor of disease or CD status

(unweighted unifrac disease p< 0.001, weighted unifrac disease p < 0.001, unweighted unifrac CD status p < 0.133, weighted unifrac p < 0.124 PERMANVOA)

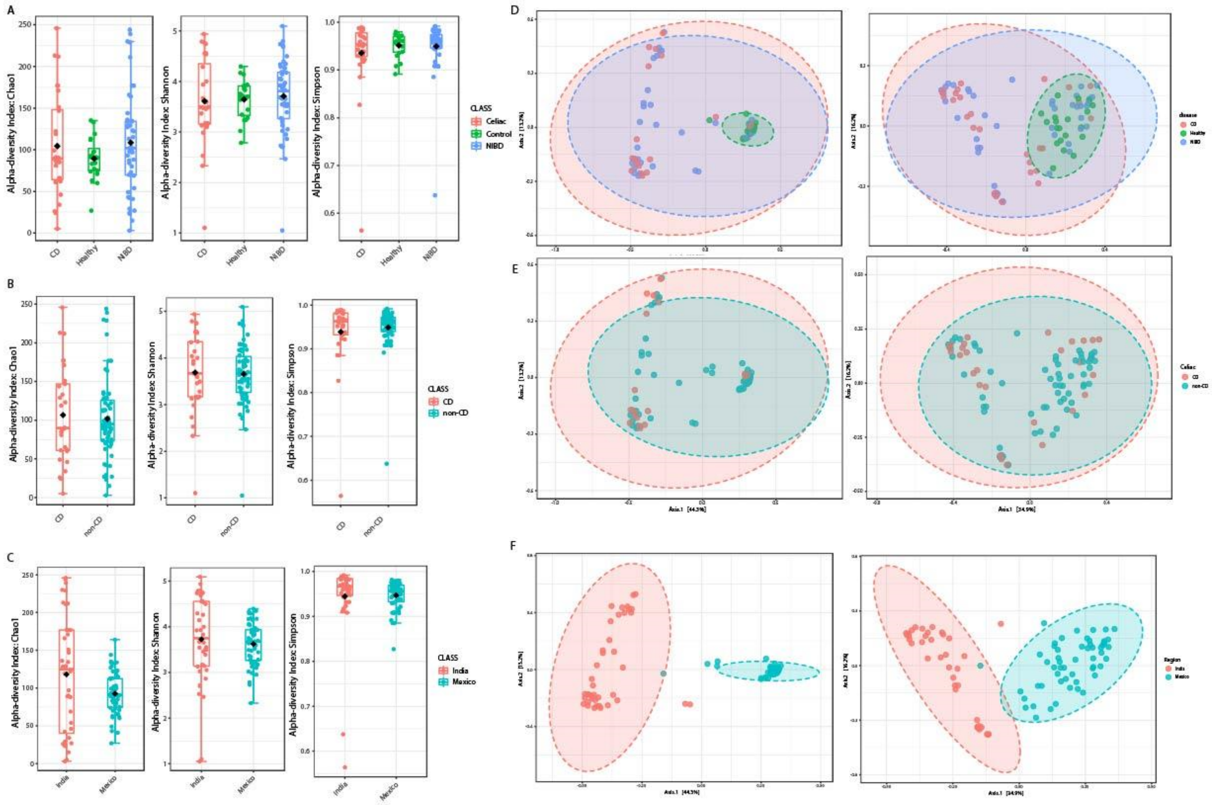

Figure 17 pooled filtered and scaled duodenal alpha and beta diversity analysis: A) Alpha-diversity as a factor of disease, with red representing CD, green healthy and blue SNIBD. No significant difference was noted for Chao1, Shannon or Simpson (P >0.05).B) Alpha-diversity as a factor of CD status with red representing CD and blue non-CD. No significant difference was noted (P >0.05). C) Alpha-diversity as a factor of region, with red representing Indian samples and blue representing Mexican samples. Indian samples were significantly enriched for Chao1 (P = 0.027265) but not Shannon or Simpson (P > 0.05). D) Unweighted (left) and weighted (right) unifrac analysis of CD (red) NIBD(blue) and Healthy(green). E) Unweighted (left) and weighted (right) unifrac analysis of CD(red) versus non-CD (blue). F) Unweighted (left) and weighted (right) unifrac analysis, with Indian samples in red and Mexican in blue. All unifrac analysis yielded P-values less than 0.001.

Stacked area percentage abundance box plots showed similar trends as earlier with samples differing as both a factor of disease and region of isolation. Similar to previous results of the pooled duodenum community structure analysis, samples isolated from the same geographic region were more similar to each other compared to samples isolated from the same disease (Figure 18).

LEFSe

When looking at a factor of disease, LEFSe identified 107 differentially abundant taxa. 4 ASVs of *Pseudomonas* were identified as being elevated in CD and controls but lowered in NIBD. Three ASVs of *Neisseria* were identified as being lowered in CD and elevated in controls and NIBD. Two ASVs of *Haemophilus* were identified as being elevated in controls, and for NIBD two ASVs of *Fusobacterium* were lowered in abundance for CD and two ASVs of *Akkermansia* were elevated in CD and NIBD, (Figure 19A). 25 ASVs were identified as differentially abundant when looking at CD versus non-CD. Of the 15 LDA highest score taxa, 13 ASVs were identified as *Acinetobacter*, and two as belonging to the family *Moraxellaceae* (Figure 19B). When looking at a function of the region of isolation, 500 significant ASVs were identified, with overlap of previously identified genre occurring with *Haemophilus* and *Neisseria*, both of which were associated with the Mexican cohort microbiome and *Akkermansia* which was associated with the Indian cohort microbiome (Figure 19C). All ASVs had LDA scores of 2.0 or greater and FDR-adjusted p-values less than or equal to 0.1.

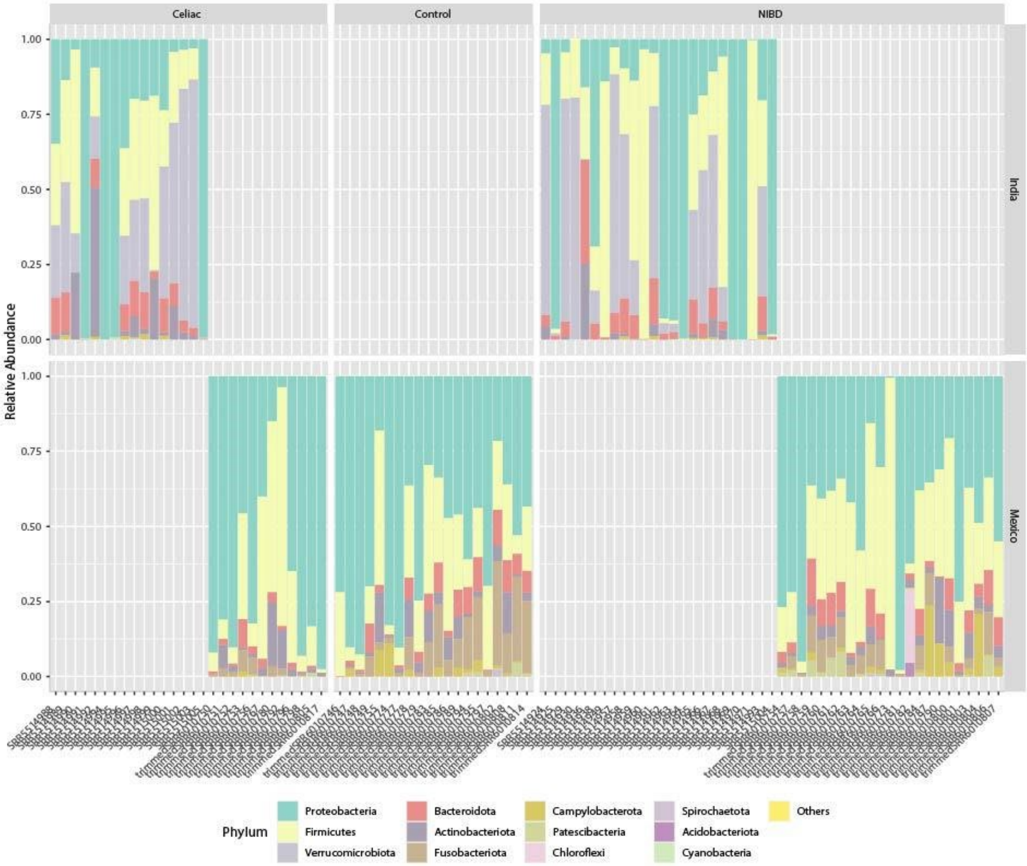

Figure 18 pooled filtered and scaled duodenal abundance plots: Taxonomic distribution of filtered and scaled duodenal biopsies at the phylum level. Shown are percent abundance stacked area barplots of duodenum biopsies fromIndian CD and NIBD patients and Mexican CD, healthy, and NIBD patients at the phylum level. Differences in community structure were noted as both a factor of disease and region of isolation.

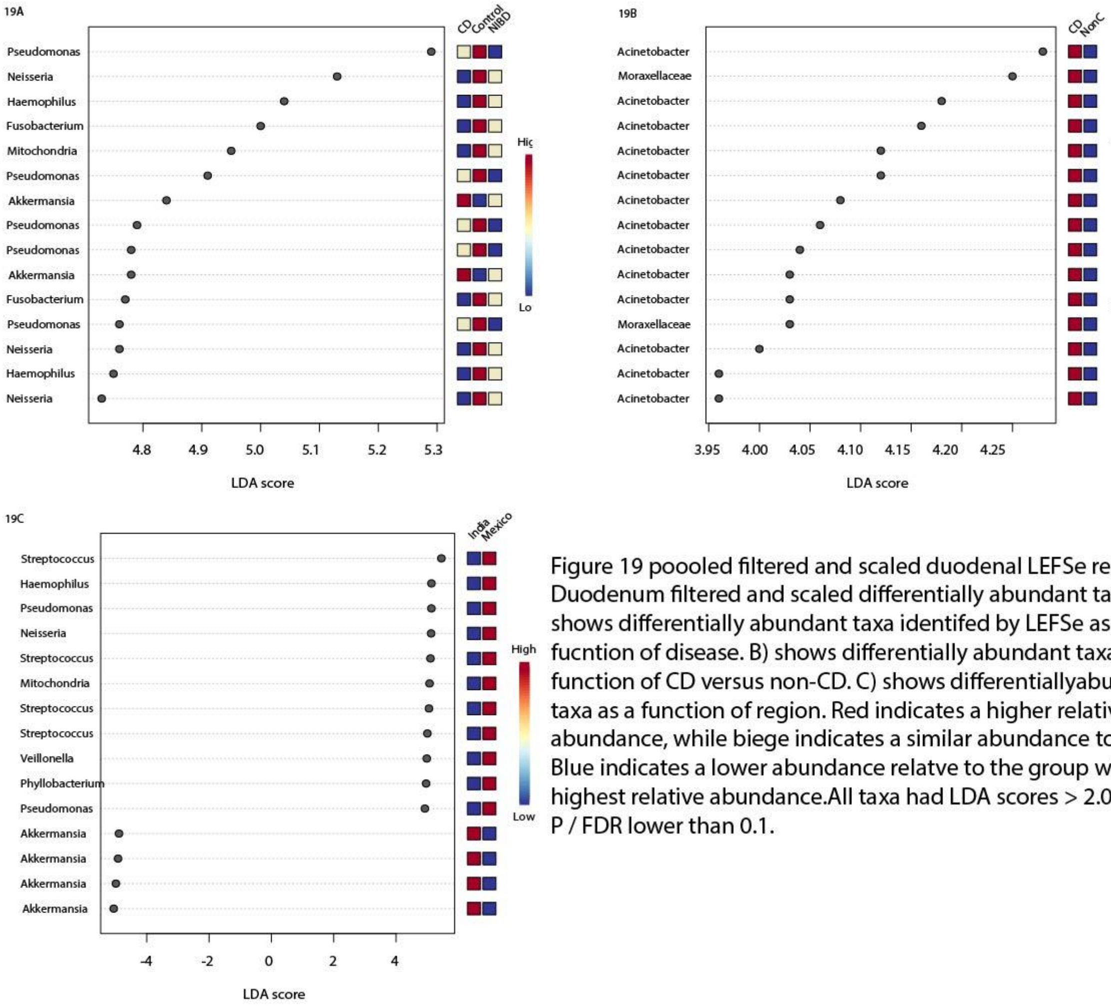

Figure 19 poooled filtered and scaled duodenal LEFSe results: Duodenum filtered and scaled differentially abundant taxa. A) shows differentially abundant taxa identifed by LEFSe as a fucntion of disease. B) shows differentially abundant taxa as a function of CD versus non-CD. C) shows differentiallyabundant taxa as a function of region. Red indicates a higher relative abundance, while biege indicates a similar abundance to red. Blue indicates a lower abundance relatve to the group with the highest relative abundance.All taxa had LDA scores > 2.0 and P / FDR lower than 0.1.

Random Forest

When looking at a factor of disease the model had an OOB 0.463, with class errors for CD of 0.786, 0.85 for healthy and 0.106 for NIBD. When looking at a factor of CD status the model had an error of 0.242 with class errors of 0.759 for CD and 0.0152 for non-CD. When looking at a factor of region, the model had an OOB of 0.0105 with class errors of 0 for India and 0.0182 for Mexico.

Pooled Normalized Duodenum

Community structure analysis

The normalization procedure from Gibbons et al. left 799 ASVs out of a total of 6684 ASVs . After filtering 719, taxa remained with 80 low variance features removed.

Both the control normalized and non-CD normalized duodenum samples had significantly lowered alpha-diversity for both CD and NIBD patients (Shannon control normalized p = 7.8156*10^{-8}, Simpson control normalized p = 7.8156*10^{-5} , Shannon non-CD normalized p = 3.5834*10^{-6}, Simpson non-CD normalized p = 0.0001319, Shannon non-CD normalized p = 0.034427 ANOVA, Figure 20A-B). No significant clustering was achieved using weighted unifrac for either normalization group (p > 0.05 PERMANOVA). Interestingly, the results from unweighted unifrac showed many samples plotted on the same point in the plane plotting the first two principal coordinates.This was true for both the control normalized set and non-CD normalized set and most dramatic for controls; where all samples were collapsed as a single point. This resulted in disease samples being plotted around the controls in a far less organized structure. To determine whether this represented clustering, other analysis methods were used (Bray-Curtis, Shannon-Jenson, Jaccard). All of these clustering methods did not generate significant clusters. Furthermore, none the clusters were significant (p> 0.05 PERMANOVA, Figure 20A,C)

No differences were noted in the stacked percent abundance bar plots for the control duodenum normalized data set nor the non-CD normalized dataset (Figure 21A, B).

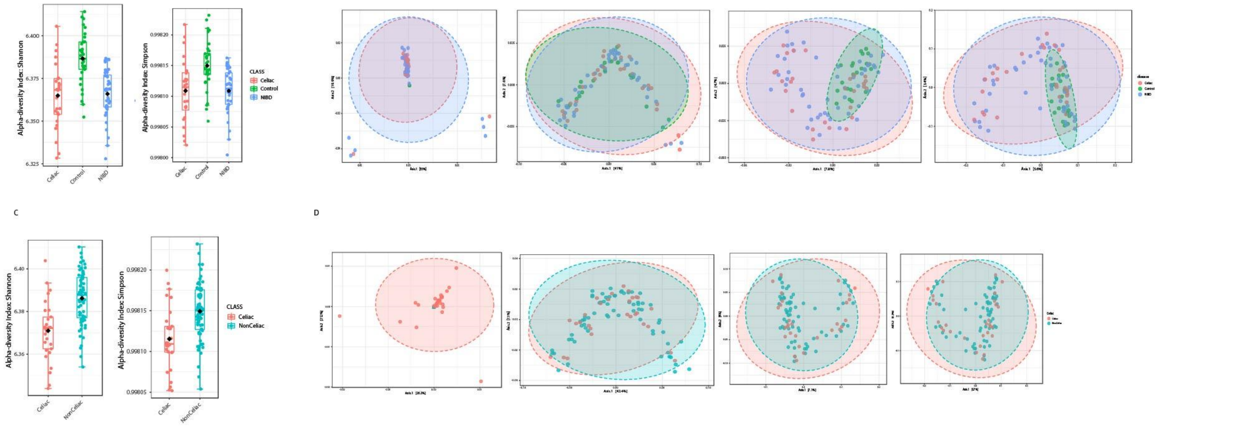

Figure 20. Normalized duodenum alpha and beta diversity. A) Control normalized alpha diversity. The left plot shows group Shannon diversity for the control normalized dataset with disease samples having lower alpha-diversity compared to healthy samples. This was observed in both Shannon (right) and Simpson (left) indices (hannon control normalized P = 7.8156*10-8, Simpson control normalized P = 7.8156*10-5). Red shows CD patients, green healthy and blue NIBD. B) shows clustering methods, with the left most being unweighted unifrac, followed by weighted unifrac, Bray-Curtis, and Jaccard index. No significant clustering was achieved on the basis of disease for any of the methods (P> 0.05). C) shows non-CD normalized alpha diversity with Shannon index on the left and Simpson on the right, with red showing CD patients and blue non-CD patients. CD patients had significantly lower alpha-diversity compared to non-CD patients for both tests (Simpson non-CD normalized P= 0.0001319, Shannon non-CD normalized P = 0.034427). D) Beta-diversity of non-CD normalized data, with unweighted unifrac on the far left, followed by weighted unifrac, Bray-Curtis and Jaccard inidices. No significant clustering was achieved on the basis of CD status for any method (P >0.05).

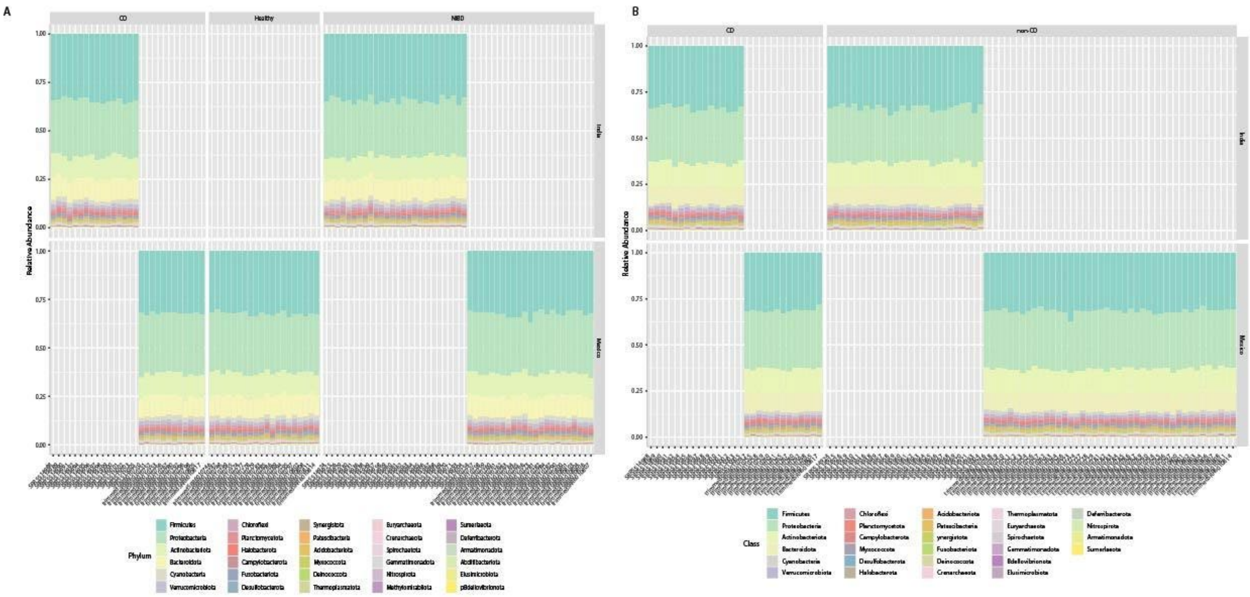

Figure 21. Normalized abundance plots. A) Control normalised abundance plots of Indian and Mexican duodenal biopsies. B) Non-celiac normalized abundance plots of Indian and Mexican biopsies.

LEFSe and Random Forest

After normalization no significant taxa were identified using LEFSe for either normalization group. Random forest for control normalization had an OOB error of 0.66 with class errors of 0.929 for CD, 0.879 for control and 0.25 for NIBD. Random forest for non-CD normalized had an OOB of 0.309 with class error of 1 for CD and 0.0147 for non-CD. The control normalized set had an OOB of 0.0722 for geographic region with class errors of 0.143 for Indian cohort samples and 0.0182 for Mexican cohort samples. Geographic region analysis of the non-CD normalized data had an OOB of 0.0309 with class errors of 0.0476 and 0.0182 for Indian and Mexican samples respectively.

Duodenum PICRUSt 2 Pathways

The PICRUSt 2 data was analyzed using microbiomeAnalyst's default parameters. This dataset was much smaller and far less noisy than the ASVs data generated by DADA2; thus repeated analysis using different transformations was not necessary. No samples had to be excluded due to low library size. 12 low abundance features were removed and 41 low variance features were removed using default settings, leaving 368 features.

Clustering

Clustering by Bray-Curtis using MetaCyc pathway data was unable to generate accurate clusters for disease state, CD status or geographic region (Bray-Curtis disease $p < 0.034$, Bray-Curtis CD status $p < 0.097$, Bray-Curtis region $p < 0.001$ PERMANOVA).

Differentially abundant pathways

MetaCyc pathway analysis identified 16 shared pathways between RNA-seq (EdgeR), metagenomeSeq and LEFSe with pathways for norspermidine biosynthesis (PWY 6562), L-lysine fermentation to acetate and butanoate (P163 PWY) , methylaspartate cycle (PWY 6728), L-glutamate degradation V (P162 PWY), UDP-2, 3-acetamido-2, 3-dideoxy-a-D-mannuronate biosynthesis (PWY 7090) and glycogen degradation III (PWY 5767) lowered in CD. Pathways

L-ornithine biosynthesis I (GLUTORN), L-histidine biosynthesis (HISTSYN), methanogenesis from acetate (METH-ACETATE PWY), dTDP-N-acetylhomosamine biosynthesis (PWY 7315), L-tryptophan biosynthesis (TRPSYN), guanosine deoxyribonucleotides de novo biosynthesis II (PWY 7222) were elevated in CD/ NIBD and lowered in controls. All features were identified as significant between LEFSe, RNA seq (EdgeR), and metagenomeSeq (RNA seq/ metagenomeSeq p < 0.1, LEFSe LDA > 2.0, Figure 22A)

When looking at CD vs non-CD 9 features , UMP biosynthesis (PWY 5686), L-glutamate degradation V (via hydroxyglutarate)(P162 PWY, guanosine deoxyribonucleotides de novo biosynthesis (PWY 6125), and superpathway of purine nucleotides de novo biosynthesis I (PWY 841) were higher in CD, while pathways superpathway of arginine and polyamine biosynthesis (ARG+polyamineSYN), tetrapyrrole biosynthesis I from glutamate (PWY 5188), L-lysine fermentation to acetate and butanoate (P163 PWY), reductive acetyl coenzyme A pathways I (homoacetogenic bacteria)(CODH PWY), and tetrapyrrole biosynthesis from glycine (PWY 5189) were elevated in non-CD. All pathways had LDA scores with an absolute value of 2.5 or greater and RNA(EdgeR)/metagenomeSeq p < 0.1 (Figure 22).

Random forest

Random forest analysis of duodenum MetaCycMetaCyc pathways had an OOB error rate of 0.469 with class error rates of 0.464 for CD, 0.6 for control, and 0.394 for NIBD. Random forest using CD vs non-Cd had an OOB error rate of 0.272 with class error rates of 0.552 for CD and 0.115 for Non-CD. When looking at the region the sample was taken from, the model had an OOB error of 0.037 with class errors of 0.0769 for Indian samples and 0.0182 for Mexican samples.

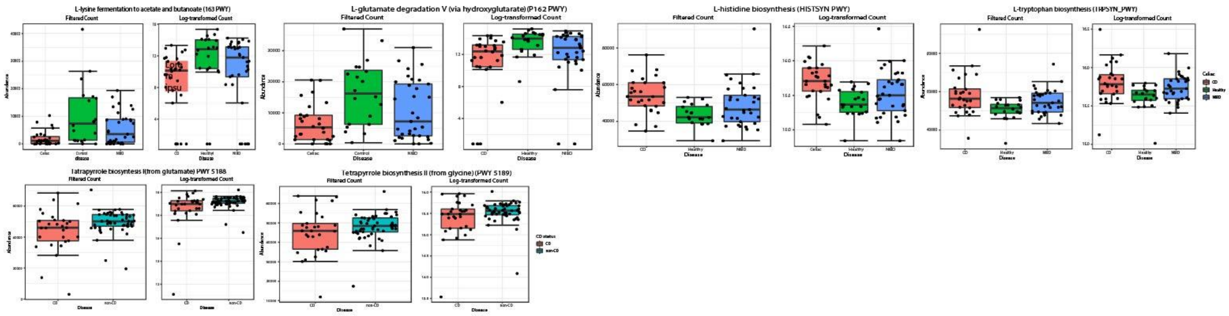

Figure 22 pooled differentially abudant duodenal metabolic pathways: Figure 22A differentially abudndant pathways as a factor of disease with CD shown in red, NIBD blue and controls green. All pathways were identified as significant across LESFSe (LDA > 2.0) as well as metagenome and RNA seq (P > .1). Figure 22 B differentially abdundant pathways as a factor of CD status with CD in red and non-CD in blue. All pathways were identifed as signifcant using LEFSe (LDA >2.0) RNA and metagenome seq(P >0.1).

RESULTS OF POOLED FECES ANALYSIS

Raw Analysis

Community Structure Analysis

6684 ASVs were included in the raw analysis. Chao1, Shannon and Simpson indices showed that CD and NIBD had a higher alpha diversity than controls ($p = 1.7268^{-27}$,$p = 5.2366*10^{-15}$, $p = 2.7*10^{-7}$ANOVA , Figure 23A). Chao1, Shannon, and Simpson all showed that CD had higher alpha diversity than non-CD ($p =0.0066874$, $p = 9.0994*10^{-5}$, $p = 2.2775*10^{-10}$ANOVA , Figure 23B). When filtering as a function of geographic region, it was found that Indian cohort samples had the highest alpha diversity across metrics, with Mexican and American cohorts following respectively ($p =3.0216*10^{-30}$, $p = 8.2875*10^{-16}$, $8.2946*10^{-7}$ ANOVA , Figure 23C). Both weighted and unweighted unifrac analysis failed when using the raw pooled datasets, likely due to dataset size.

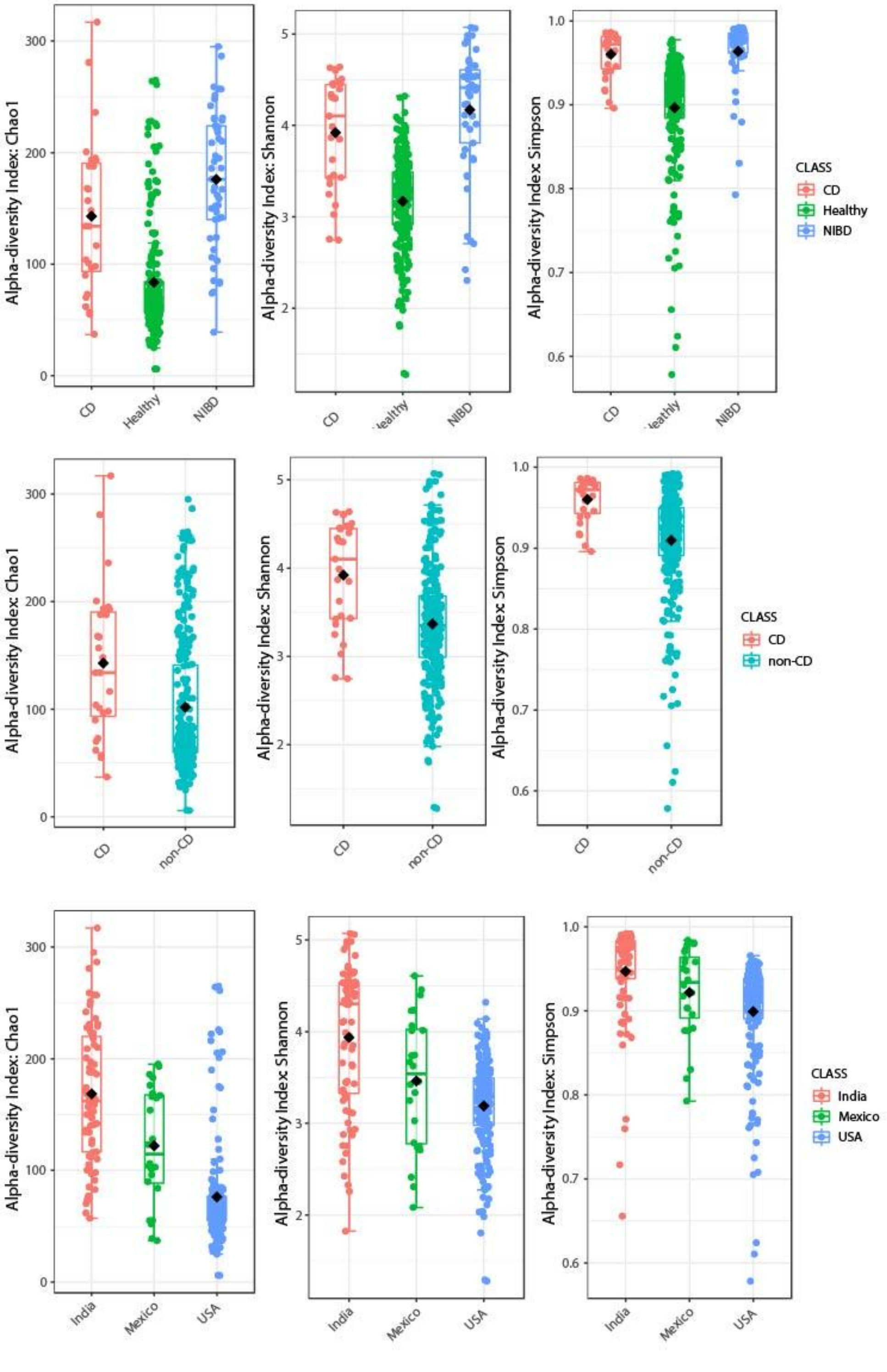

Figure 23. pooled raw fecal alpha and beta-diveristy analysis: A) Chao1, Shannon, and Simpson diversity all showed that NIBD and CD had higher alpha-diversities than healthy, with CD shown in red, healthy in green, and NIBD in blue (P = 1.7268-27,P = 5.2366*10-15, P = 2.7*10-7). B) CD versus non-CD alpha-diversity, with non-CD shown in red and CD shown in blue. Chao1, Shannon and Simpson all showed that non-CD had reduced alpha-diversity (P=0.0066874, P = 9.0994*10-5, P = 2.2775*10-10). C) Regional alpha diversity with Indian samples in red, Mexican in green, and American in blue. All metrics showed that Indian samples were significantly enriched with American samples having the lowest alpha diversity (P =3.0216*10-30, P = 8.2875*10-16, 8.2946*10-7).

Stacked area bar plots of percentage abundance showed differences in community structure in accordance with both region and disease, with Healthy American stool having a far less varied community than either Mexican or Indian cohort cohort stool. Mexican cohort cohort CD samples appeared to have a more uniform constructruction compared to Indian samples with an abundance of *Proteobacteria* compared to their NIBD and healthy counterparts This was similar in Indian cohort cohort samples. Similar to the duodenum analysis, samples were more similar to those from the same region as opposed to those of the same condition (Figure 24).

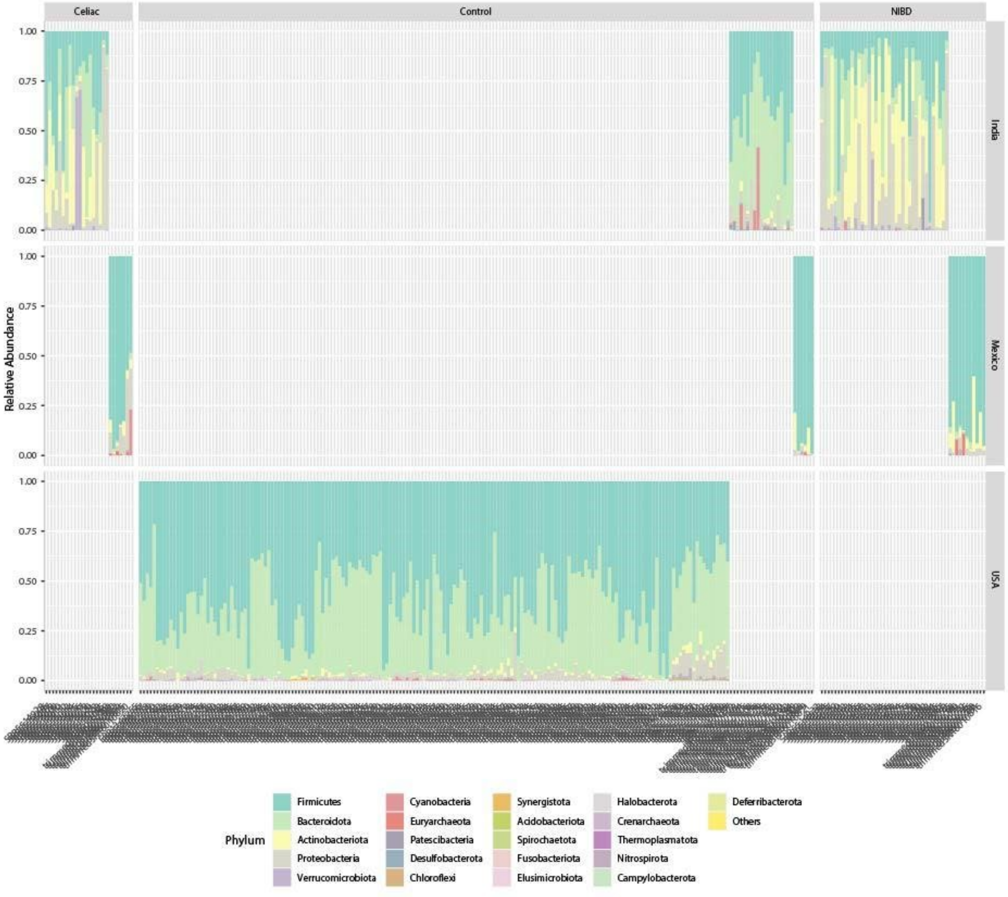

Figure 24 pooled raw fecal abundance plots. Stacked area plots of percent abundance of raw stool samples as a factor of region and disease. Variation in community structure as a factor of disease and region of isolation was noted.

LEFSe

LEFSe identified 138 differentially abundant ASV. Of the top 15 ASVs reported by LEFSe, 8 were *Bifidobacterium*, 2 were *Akkermansia,* 2 were *Enterobactericeae*, 1 was *Prevotella*, and 1 *Klebsiella,* which were all reduced in controls compared to other disease states (Figure 25A). 72 ASVs were identified as a factor of CD. Of the top 15 reported, all were enriched in CD compared to non-CD with 5 corresponding to *Prevotella*, 3 to *Lactobacillus*, 2 to *Proteobacteria*, 2 to *Escherichia-Shigella,* 1 to *Bifidobacterium*, and 1 to *Oscillospiraceae* UCG-002 (Figure 25B). 189 ASVs were identified as a factor of geographic region. None were enriched in American controls. Of those enriched on the basis of regions that overlapped with the diseased results, *Prevotella*, *Bifidobacterium* were also enriched in Indian samples (Figure 25C). All features had LEFSe LDA scores of 2.0 or greater and FDR corrected p-values below 0.1.

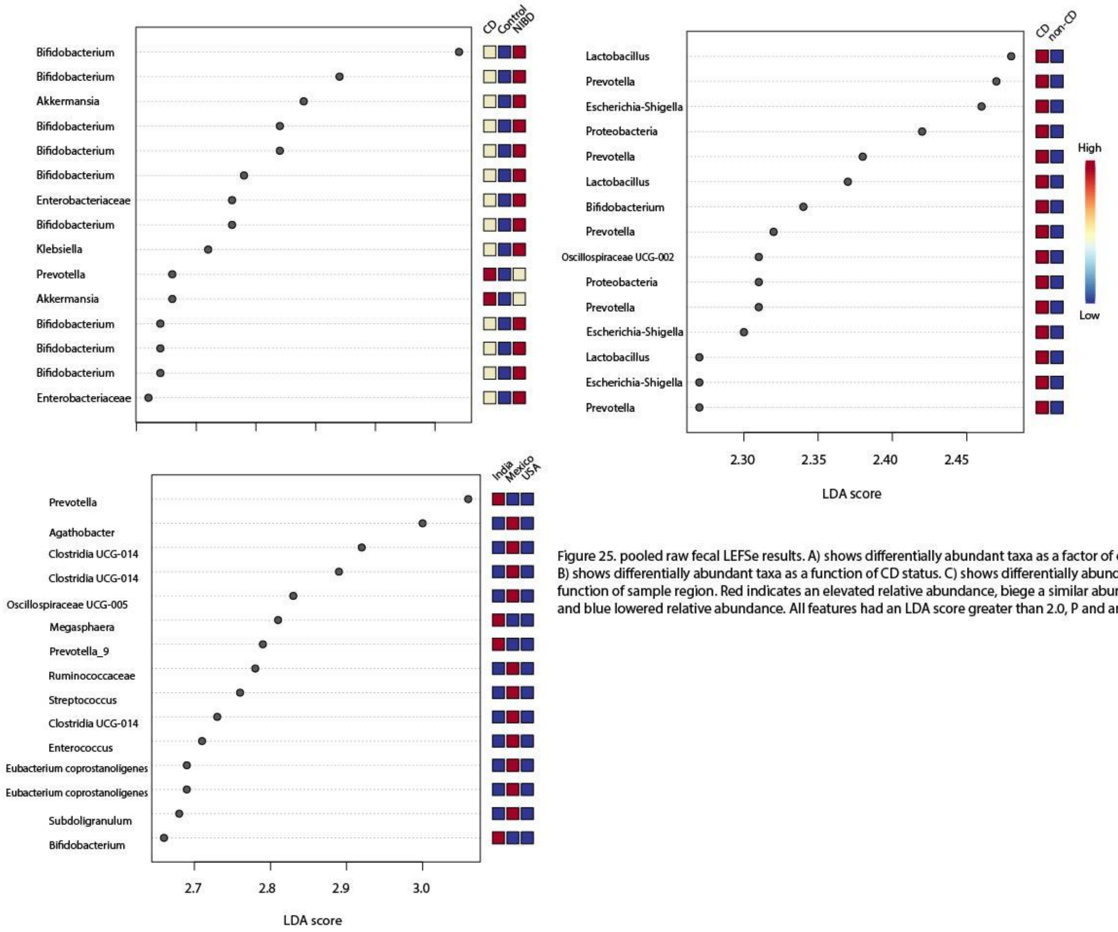

Figure 25. pooled raw fecal LEFSe results. A) shows differentially abundant taxa as a factor of disease. B) shows differentially abundant taxa as a function of CD status. C) shows differentially abundant taxa as a function of sample region. Red indicates an elevated relative abundance, biege a similar abundance to red, and blue lowered relative abundance. All features had an LDA score greater than 2.0, P and an FDR less than 0.1.

Random forest

When looking at a factor of disease, the model had an OOB of 0.156 with class errors of 0.963, 0.001 and 0.306 for CD, healthy and NIBD respectively. When using CD status, the model had an OOB of 0.0942 with class errors of 0.963 for CD and 0 for non-CD.

Filtered analysis

Filtering removed 4689 low abundance features and 200 low variance features leaving 1795 ASVs.

Community structure analysis

Chao1 showed that NIBD had the highest alpha-diversity, followed by CD and controls ($p = 7.1026 *10^{-28}$) Shannon and Simpson indices had a similar trend ($p = 8.006*10^{-23}$, $p = 3.789*10^{-13}$, Figure 26A) Chao1 showed that CD samples had higher alpha diversity compared to non-CD (p =0.0090813). A similar trend was observed for both Shannon (p = 0.0015761) and Simpson ($3.3357*10^{-7}$,figure 28). Chao1 showed that Indian samples had the highest alpha-diversity, followed by Mexican and American samples ($p = 3.1866*10^{-21}$). This trend was mirrored by Shannon ($1.3147*10^{-11}$) and Simpson ($p = 8.9321*10^{-7}$, Figure 26C). Unweighted and weighted unifrac showed no separation as a factor of disease (Figure 28D). Unweighted and weighted unifrac failed to produce clustering on the basis of CD status (Figure 26E) Unweighted unifrac showed clear segregation of samples on the basis of geographic region, while weighted unifrac showed that Mexican and Indian cohort samples overlapped with clear segregation from American cohort cohort stool samples (p< 0.001 PERMANOVA, Figure 26F).

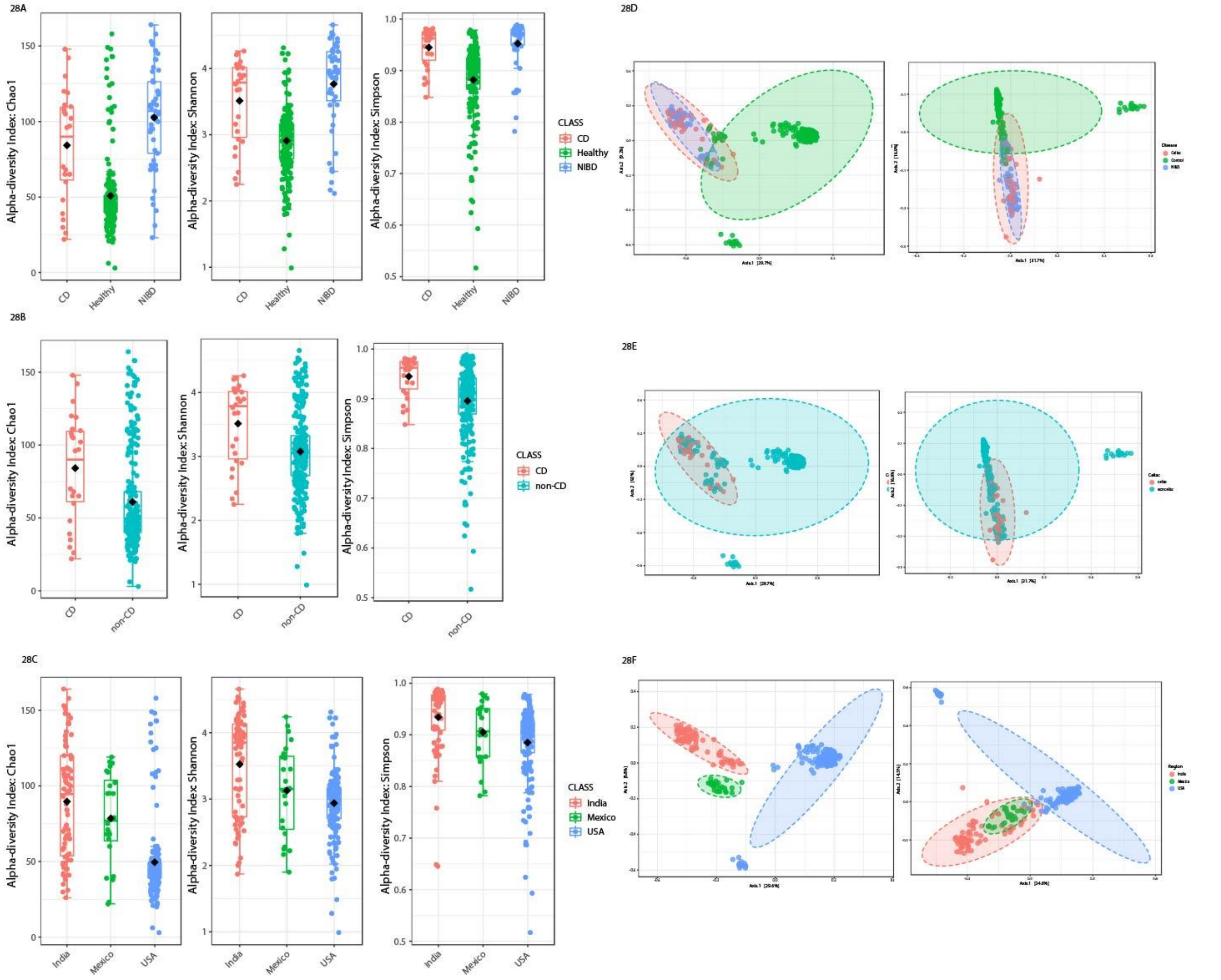

Figure 26 pooled filtered fecal Alpha and beta-diversity anlysis. A) Chao1, Shannon, and Simpson diversity all showed that NIBD and CD had higher alpha-diversities than healthy, with CD shown in red, healthy in green, and NIBD in blue (P = 7.1026 *10-28, P = 8.006*10-23, P = 3.789*10-13). B) shows CD versus non-CD alpha-diversity, with non-CD shown in red and CD shown in blue. Chao1, Shannon and Simpson indices all showed that non-CD had reduced alpha-diversity (P=0.0090813, P = 0.0015761, 3.3357*10-7). C) Regional alpha diversity with Indian samples in red, Mexican in green and American in blue. All metrics showed that Indian samples were significantly enriched with American samples having the lowest alpha diversity (P = 3.1866*10-21, 1.3147*10-11, P = 8.9321*10-7). D) Weighted (right) and unweighted (left) unifrac analysis showing clustering on the basis of disease, with CD shown in red, healthy in green, and NIBD in blue. No clustering was achieved on the basis of disease. E) Unweighted (left) and weighted (right) unifrac of CD (red) versus non-CD (blue). No clustering was achieved as a result of the analysis. F) Unweighted (left) and weighted (right) unifrac analysis of American, Indian and Mexican stool samples, with Indian in red, Mexican in green, and American in blue. Clustering was achieved on the basis of region with clear separation for unweighted unifrac (P < 0.001) and separation of American from Indian and Mexican samples for weighted unifrac (P <0.001).

Stacked area plots of percent abundance showed that Indian samples had the most varied community structure with healthy samples having larger proportions of *Bacteriodota* with NIBD having a larger proportion of *Actinobacteria* and CD having *Proteobacteria*. Mexican CD patients also had elevated *Proteobacteria* compared to other samples from the same region, however not to the same extent as Indian cohort cohort samples. Healthy samples from the

United States had a rather homogenous community structure characterized by a dominance of both *Bacteroidetes* and *Fimicutes* with small populations of *Proteobacteria* (Figure 27).

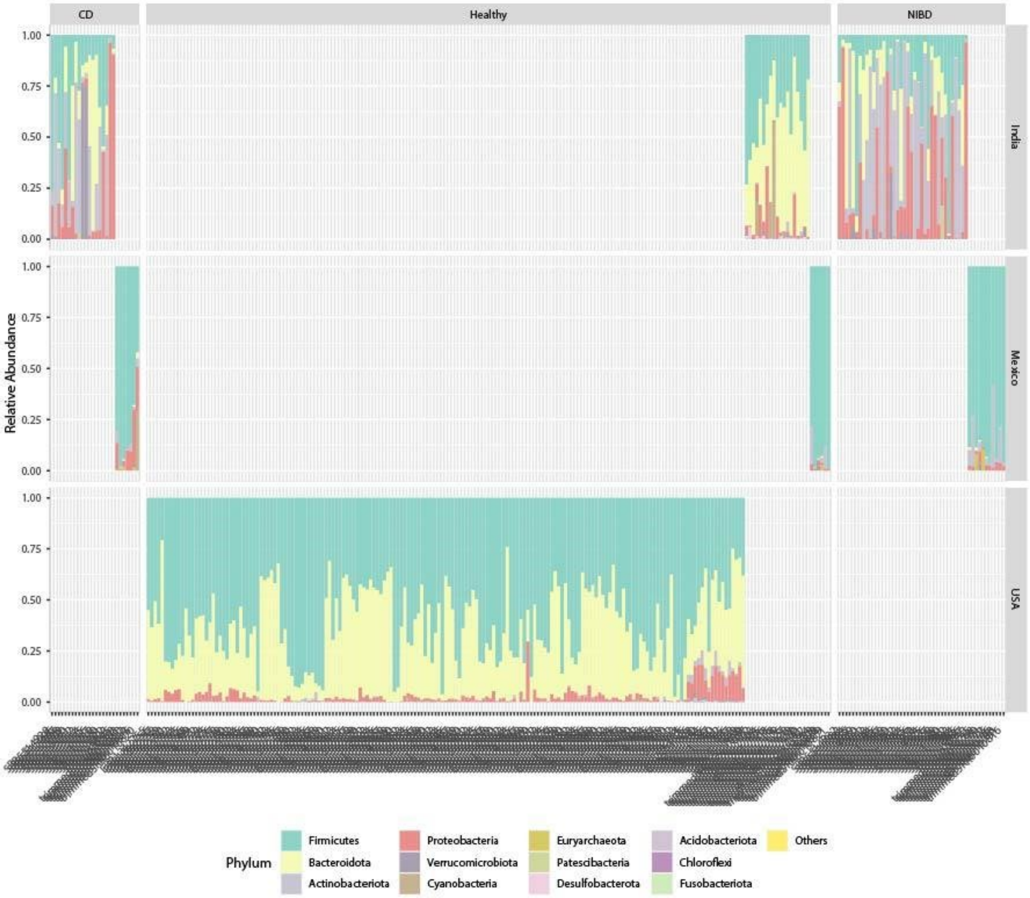

Figure 27 Pooled filtered fecal abundance plots: Stacked area plots of percent abundance of filtered stool samples as a factor of region and disease. Variation in community structure as a factor of disease and region of isolation was noted.

LEFSe

LESe identified 132 taxa associated with disease. Of the top 15 LDA score producing taxa, all of them were reduced in healthy samples and increased in both NIBD and CD. 9 ASVs correspond to *Bifidobacterium*, 2 to *Akkermansia*, and 1 ASV each of *Escherichia-Shigella*, *Eubacterium coprostanoligenes* and *Prevotella*, (Figure 30A). 76 ASVs were identified as a function of CD versus non-CD. Of the top 15 effect size producing taxa, all were reduced in non-

CD samples with *Escherichia-Shigella* made up 3 of the 15 ASVs, *Prevotella* making up 3, *Lactobacillus* 2, and *Oscillospiraceae* UCG-002, *Proteobacteria*, *Akkermansia*, *Eubacterium copropstaoligenes*, *Enterobacteriaceae, Bifidobacterium* and *Catenibacterium* making up 1 ASV each (Figure 28B). 183 as a function of the region of sample origin.Of these, *Bifidobacterium*,and *Prevotella* overlapped with diseased results (Figure 28C). All taxa had LDA scores greater than 2.0 and P/FDR values less than 0.1.

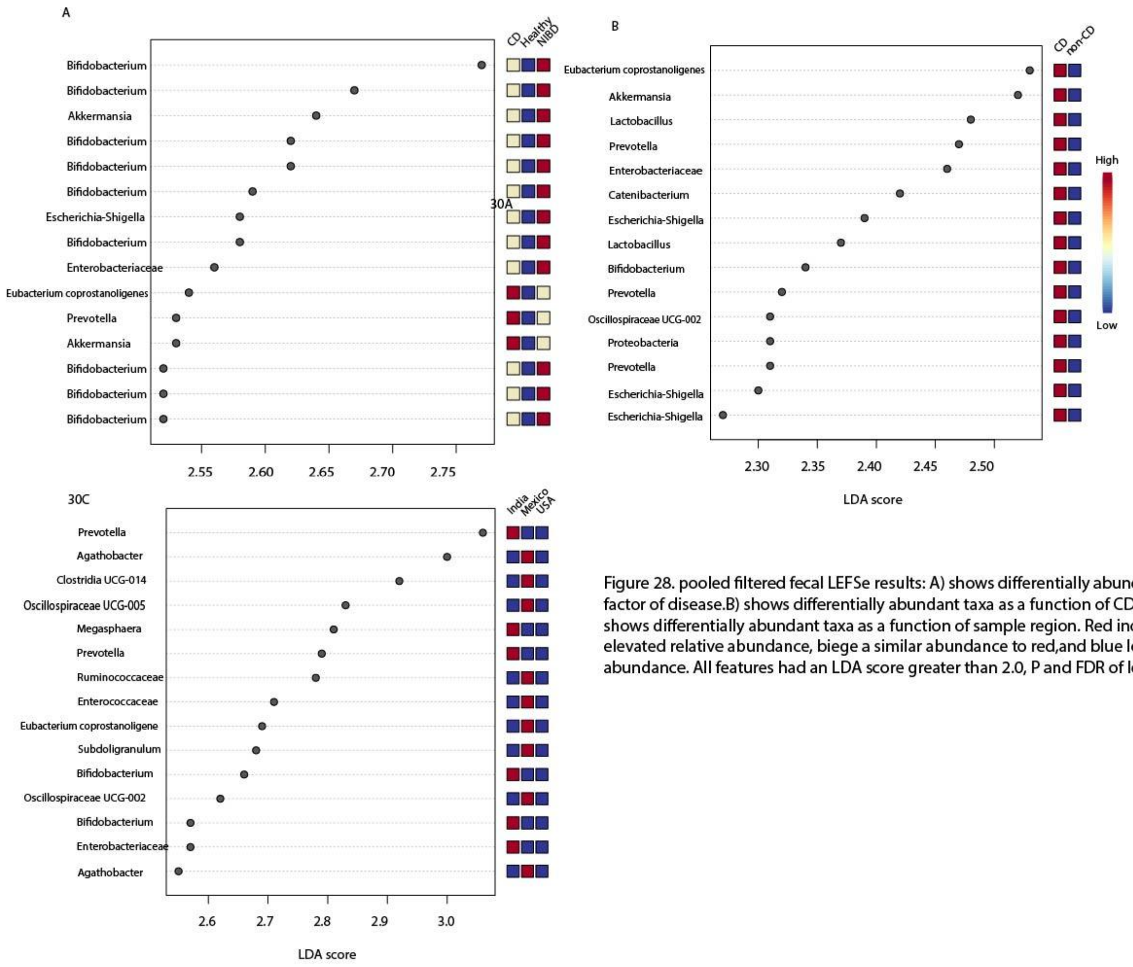

Figure 28. pooled filtered fecal LEFSe results: A) shows differentially abundant taxa as a factor of disease.B) shows differentially abundant taxa as a function of CD status. C) shows differentially abundant taxa as a function of sample region. Red indicates an elevated relative abundance, biege a similar abundance to red,and blue lowered relative abundance. All features had an LDA score greater than 2.0, P and FDR of less than 0.1.

Random forest

Random forest had an OOB of 0.108 when using CD, NIBD and healthy with class errors of 0.704, 0.143 and 0.0197 respectively. When using CD versus non-CD, the model had an OOB of 0.0789 with class errors of 0.815 for CD and 0 for non-CD respectively. When looking at a factor of geographic region, the model had an OOB of 0.0251 with class errors of 0.0519 for CD 0.125, for Mexico, and 0 for the United States.

Filtered and scaled analysis

Community structure analysis

Chao1, Shannon, and Simpson showed NIBD having an elevated alpha diversity compared to both CD and healthy, followed by CD and healthy respectively (p = 7.1026*10^{-28},p = 8.0086*10^{-23}, p = 3.3789*10^{-13}, ANOVA, Figure 29A). CD had higher alpha diversity compared to non-CD for Chao1, Shannon, and Simpson (p =0.0090813,p = 0.0015761,p = 3.357*10^{-7} ANOVA, Figure 29B). Chao1 found that Indian samples had the highest diversity followed Mexican and American samples respectively. This was mirrored by Shannon and Simpson (p = 3.1866*10^{-21}p = 1.31478*10^{-11}, p = 8.9231*10^{-7} ANOVA, Figure 29C). Unifrac analysis only produced clustering for geographic region using unweighted unifrac, indicating that low abundance taxa were mainly responsible for the separation seen (p< 0.001, Figure 29D-F).

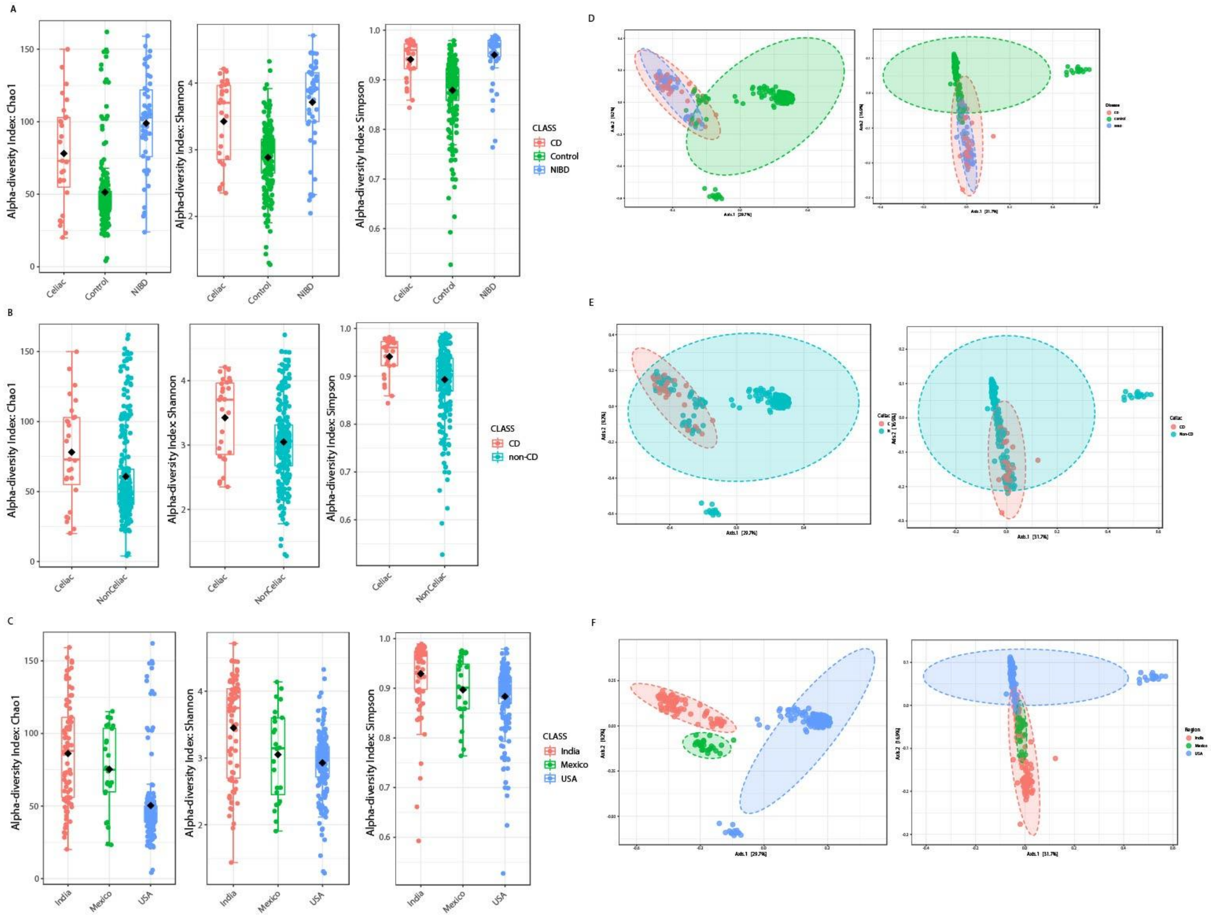

Figure 29. Filtered and scaled fecal alpha and beta diversity analysis. A) Alpha diversity of CD (red), healthy (green) and NIBD (Blue) stool samples. NIBD and CD were found to have higher alpha-diversity based on the Chao1(left), Shannon (middle) and Simpson (right) indices (P = 7.1026*10-28, P = 8.0086*10-23, P = 3.3789*10-13). B) Alpha diveristy of CD versus non-CD showed that CD (red) had higher alpha diversity than non-CD (blue) across the Chao1 (right), Shannon (middle) and Simpson (right) indices (P =0.0090813, P = 0.0015761, P = 3.357*10-7). C) Alpha diversity as a function of region showed that American samples (blue) had the lowest alpha diversity compared to Mexican (green)and Indian(blue) samples (P = 3.1866*10-21, P= 1.31478*10-11, P= 8.9231*10-7). D) shows unweighted (left) and weighted (right) unifracanalysis of CD (red), healthy (green) and NIBD (blue). No clustering as a function of disease was observed. E) Unweighted (left) and weighted (right) unifrac analysis of of CD (red) versus non-CD (blue). No significant clustering was achieved a result of CD. E) unweighted (left) and weighted (right) unifrac analysis of Indian (red), American (blue), and Mexican (green) stool samples. Clustering was achieved as a result of region for unweighted unifrac (P <0.001), but not weighted unifrac analysis.

Stacked area bar plots of abundance showed a similar trend as what was noted in the filtered analysis (Figure 30).

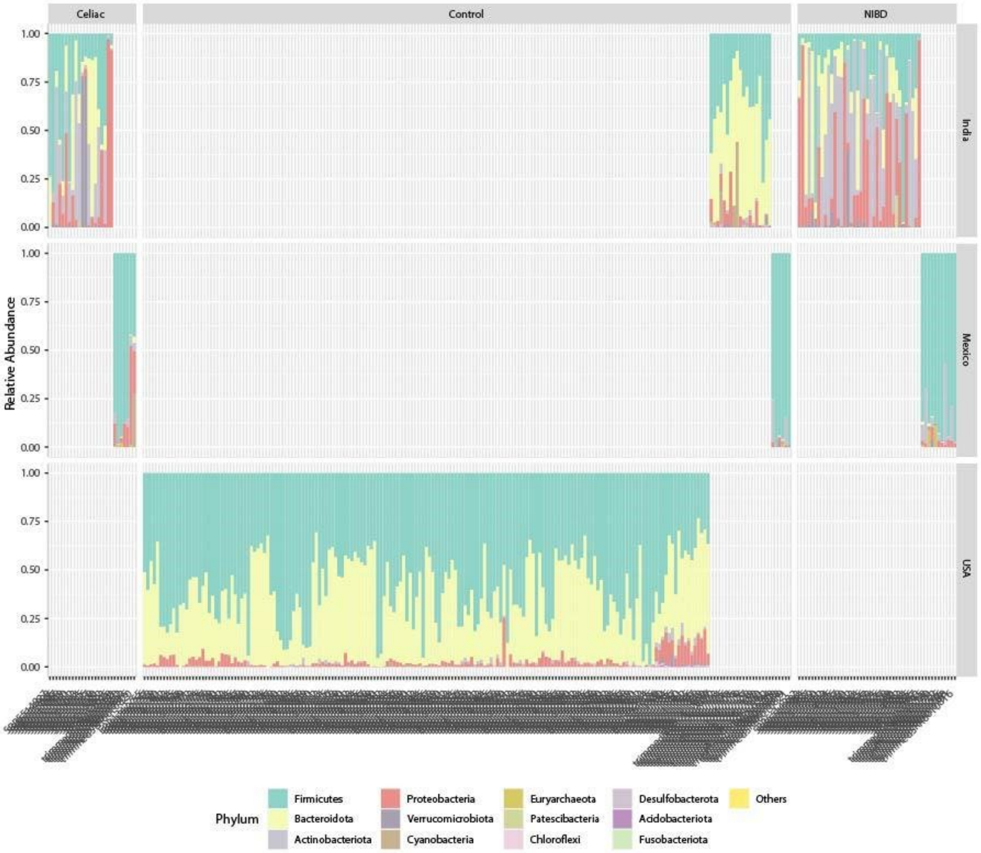

Figure 30 filtered and scaled fecal abundance plots: Stacked percent abundance bar plots of filtered and scaled data showing community structure at the phylum level. Differences were noted due to both region and disease state.

LEFSe

LEFSe identified 500 significant features when examining CD, NIBD and healthy samples. Healthy samples had an abundance of *Bacteroides* (2 ASVs), *Prevotella*, *Ruminococcaceae*, *Oscillospiraceae* UCG 002, *Alistipes*, *Lachnospiraceae* (2 ASVs), *Subdoligranulum*(2 ASVs), *Dialister*, and *Eubacterium eligens* (Figure 31A). When looking at CD versus non-CD, 415 taxa were identified with *Escherichia-Shigella* and *Eubacterium coprostanoligenes* being elevated in CD and *Lachnospiraceae* (3 ASVs), *Subdoligranulum* (2 ASVs) *Eubacterium eligens*, *Dialister*, *Alistipes*, *Oscillospiraceae* UCG-002, *Ruminococcaceae*, *Bacteroides* and *Prevotella* being elevated in non-CD (Figure 31B). 498 ASVs were identified when looking at the region. It was found that *Prevotella* was elevated in Indian samples, *Oscillospiraceae* UCG-002 in Mexican, and *Ruminococcaceae, Alistipes* and *Bacteroides* in American (Figure 31C).

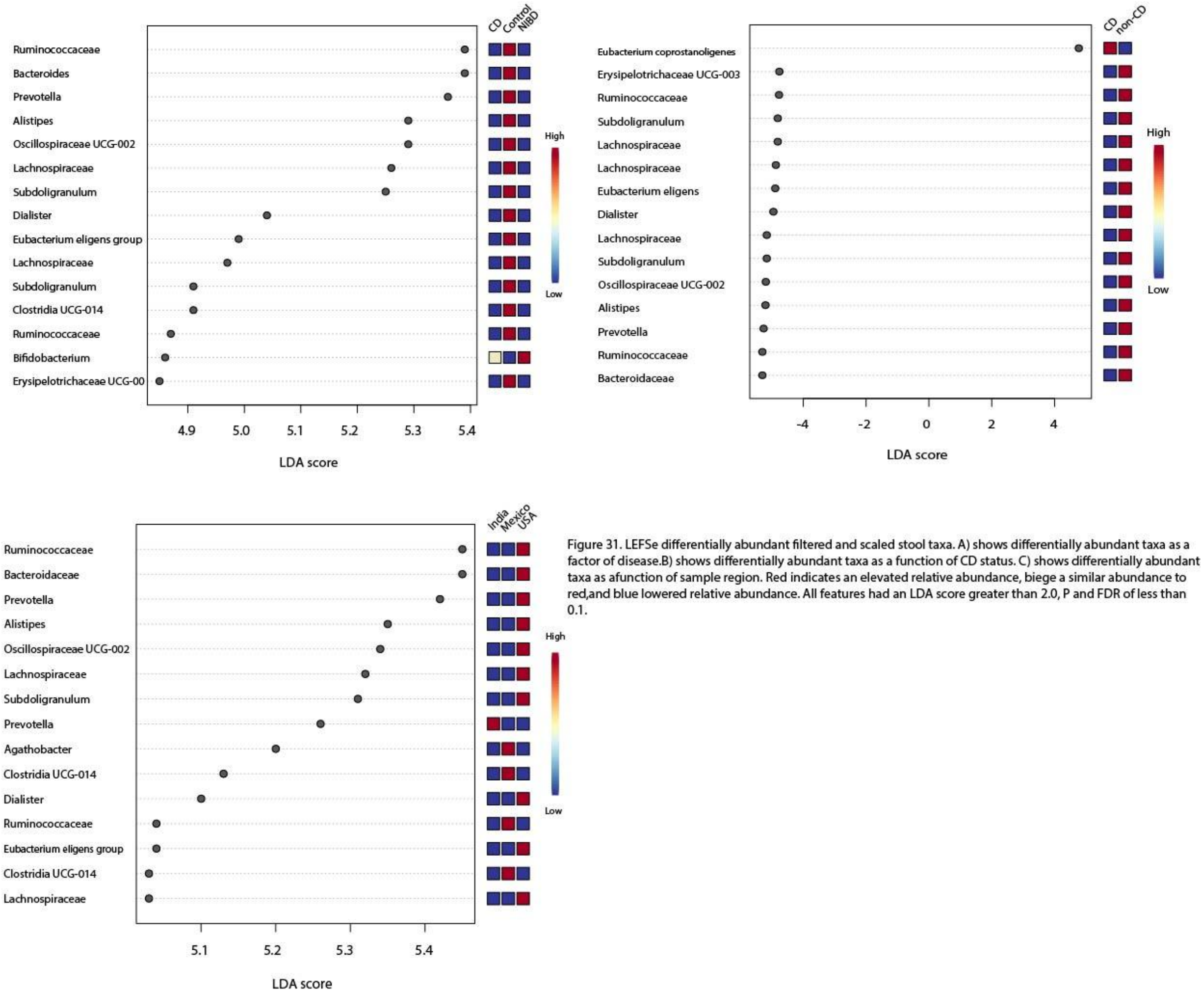

Figure 31. LEFSe differentially abundant filtered and scaled stool taxa. A) shows differentially abundant taxa as a factor of disease.B) shows differentially abundant taxa as a function of CD status. C) shows differentially abundant taxa as afunction of sample region. Red indicates an elevated relative abundance, biege a similar abundance to red,and blue lowered relative abundance. All features had an LDA score greater than 2.0, P and FDR of less than 0.1.

Random Forest

Random forest had an OOB of 0.111 with class errors of 0.741 for CD, 0.0197 for healthy, and 0.143 for NIBD. For CD, versus non-CD the model had an OOB of 0.0789 with class errors of 0.815 for CD and 0 for non-CD. Random forest had an OOB of 0.208 for control normalized with class errors of 1 for CD, 0 for controls, and 0.939 NIBD. Random forest analysis of non-CD had an OOB error of 0.0968 with class error rates of 1 for CD and 0 for non-CD. When examining the region, the model had an OOB of 0.0251 with class errors of 0.0519 for Indian, 0.125 for Mexican, and 0 for American samples.

Pooled Normalized Feces

Normalization of the pooled feces samples yielded 799 ASVs compared to non-normalised data which yielded 6684 ASVs. .

Community structure analysis

Both control normalized and non-CD normalized fecal samples had lowered alpha diversity for NIBD and CD for both Shannon and Simpson diversity indices (Shannon control normalized p = 0.0031876, Simpson control normalized p = $1.2784*10^{-6}$, Simpson non-CD normalized, Shannon non-CD normalize p = 0.020527, ANOVA Figure 32A,32B). The Simpson index of non-CD normalized data showed lowered fecal alpha diversity for both CD and NIBD; however these results were not significant (p = 0.12809 ANOVA).

No clustering was achieved for weighted unifrac for either normalization group. Unweighted unifrac analysis plotted all samples directly on top of each other, and was similarly unable to accurately cluster. To determine whether these were indicative of biologically relevant clusters, non-phylogenetic methods were utilized, with both Bray-Curtis and Jaccard indices failing to produce clusters for either normalization group (Figure 32C,32D).

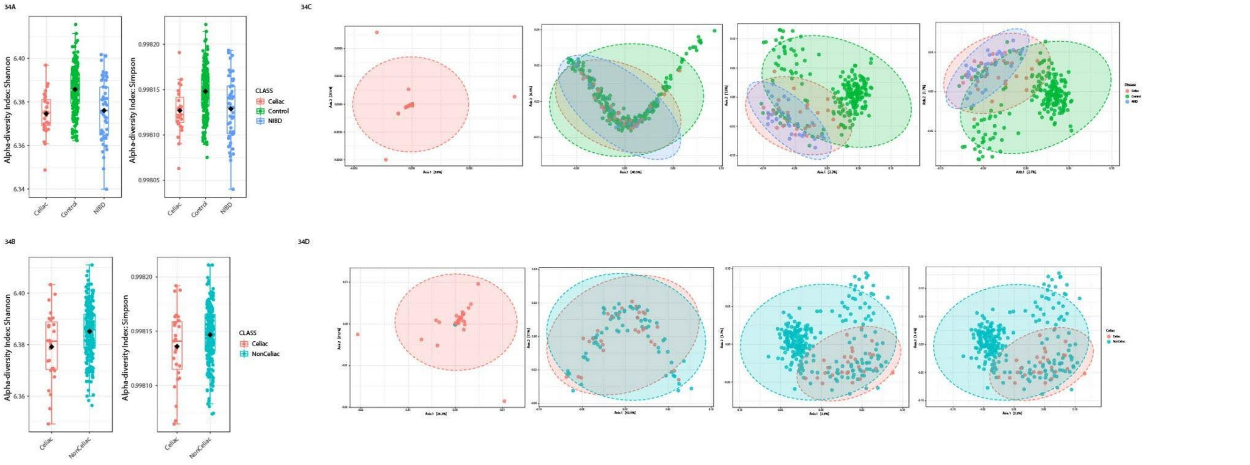

Figure 32. Normalized fecal alpha and beta diversity: A) Disease alpha diversity of control normalized samples. Shannon (left) and Simpson (right) alpha diversity of CD (red), healthy (green) and NIBD (blue). Healthy stool samples had higher alpha diversity comapared to diseased samples (Shannon control normalized p = 0.0031876, Simpson control normalized P = 1.2784*10-6). B) CD versus non-CD alpha diversity of non-CD normalized samples. Shannon (left) and Simpson (right) alpha-diversity of CD (red) versus non-CD samples (blue). CD samples had lower alpha-diversity compared to non-CD, though this was only significant for the Shannon index (non-CD normalized P = 0.020527) C) Beta-diversity analysis of control normalized samples with unweighted unifrac (far left) followed by weighted unifrac, Bray-curtis and Jaccard indices. No significant clustering as a function of disease was identified. D) Beta-diversity analysis of non-CD normalised samples with unweighted unifrac (far left), weighted unifrac, Bray-Curtis and Jaccard indices. No significant clustering was achieved with any of the analysis methods.

Stacked area percent abundance bar plots showed no difference in community structure as both a function of disease and geographic region (control normalized, Figure 33A). This was similar for the non-CD normalized data (Figure 33B).

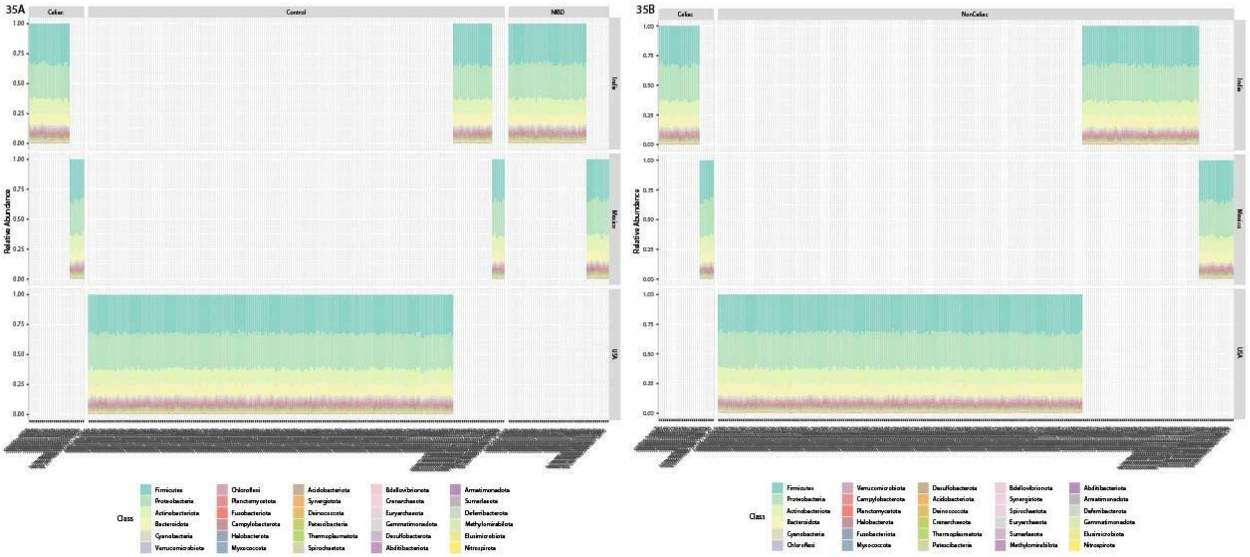

Figure 33. Stacked area percent abundance of normalized data. A) Control normalized percent abundance plots of control normalized data at the phylum level. B) Stacked area percent normalized abundance plots of non-CD normalized data at the phylum level. No differences in community structure were noted for either normalization technique.

LEFSe

LEFSe failed to identify any differentially abundant taxa between the study groups for both control normalized and non-CD normalized.

Random forest

Random forest analysis of control normalized data had an OOB 0.168 with class errors of 1 for CD, 0.0197 for healthy and 0.327 for NIBD. Random forest analysis of non-CD normalized data had an OOB 0.0968 with class errors of 1 for CD and 0 for non-CD.

Pooled Pathways Feces Analysis

Clustering

Analysis of stool sample pathway data showed that controls tended to cluster (Bray-Curtis) distinctly from NIBD and CD with the two diseased states forming a much lower mixed cluster. Clustering based on CD status (Bray-Curtis) created a looser non-CD cluster and CD cluster with CD and non-CD found in each respectively. Clustering (Bray-Curtis) based upon the region created a tight American cluster and a single mixed Indian/Mexican cluster. Clustering on regions had significant overlap with clustering based upon disease state, however, many healthy Indian samples clustered within the American cluster, which represented only healthy stool samples.

Differentially abundant pathways

There were 198 different pathways shared between LEFSe, RNA seq (DeSeq2), and metagenomeSeq with the top 10 largest effect size producing features being aerobic respiration I (cytochrome c) (PWY 3781), acetylene degradation (anaerobic) (P161 PWY), incomplete reductive TCA cycle (P42 PWY), sucrose degradation III (sucrose invertase) (PWY 621), peptidoglycan maturation (meso-diaminopimelate containing) (PWY0 1586), pyruvate fermentation to propionate I (P108 PWY), TCA cycle V (2-oxoglutarate synthase) (PWY 6969), bifidobacterium (P124 PWY), fatty acid elongation (FASYN_ELONG_PWY), and superpathway of L-alanine biosynthesis (PWY0 1061). With PWY 3781, P42 PWY, PWY 6969, and FASYN_ELONG_PWY being lowered in CD and NIBD compared to controls, and lowest in NIBD compared to CD. Pathways P161 PWY, PWY 621 PWY0 1586, P124 PWY and PWY0 1061 are elevated in CD and NIBD compared to controls. PWY0 1061 was elevated in CD compared to NIBD and P161 PWY, PWY0 1586, and P124 PWY were elevated in NIBD compared to CD. All pathways had LDA scores with an absolute value of 2.5 or greater.

269 pathways were shared between RNA seq, metagenomeSeq and LEFSe when examining CD vs. non-CD, with all pathways having LDA scores with an absolute value greater than or equal to 4.1 and were found to be reduced in the stool of CD patients. These pathways were incomplete reductive TCA cycle (P42 PWY), superpathway of pyrimidine ribonucleosides salvage (PWY 7196), superpathway of L-alanine biosynthesis (PWY0 1061), sucrose degradation III (sucrose invertase) (PWY 621), 6-hydroxymethyl-dihydropterin diphosphate biosynthesis I (PWY 6147), 6-hydroxymethyl-dihydropterin diphosphate biosynthesis III (Chlamydia) (PWY 7539), phosphopantothenate biosynthesis I (PANTO PWY), dTDP-B-L-rhamnose biosynthesis (DTDPRHAMSYN PWY), L-isoleucine biosynthesis IV (PWY 5104), gondoate biosynthesis (anaerobic) (PWY 7663).

Random Forest

Random forest analysis of MetaCyc pathways from fecal samples had an OOB error of 0.193 with class errors for CD being 0.963, NIBD 0.654 and control 0.0415. When looking at CD vs non-CD, the model had an OOB error 0.1 with class errors for CD being 1.0 and 0 for non-CD. When examining the geographic region, the model had an OOB error of 0.0259 with class errors of 0.0441 for Indian samples, 0.167 for Mexican samples, and 0 for American samples.

DISCUSSION

GDF's impact on healthy individuals: original analysis versus reanalysis of Bonder *et al.*

Alpha and beta diversity

Since celiac patients are only offered one treatment–a GFD–it is important to separate pathologic microbiome changes that might be causative for celiac disease from benign microbiome changes that happen when someone switches to a GFD. To control for these GFD-associated microbiome changes, this work took data from a previous study that tracked microbiome changes in healthy patients eating a GFD (Bonder *et al.*). That study used QIIME, PICRUSt, and the Greengenes database and found that the transition from GCD to GFD had little to no impact on beta-diversity of samples, concluding that the transition did not alter bacterial diversity. In contrast, our analysis detected a small but significant difference in the alpha diversity between diets, with GFD samples having a higher average alpha diversity than GCD (Chao1, Shannon, Simpson, Figure 1A).

Differentially abundant taxa and functions

Originally a small but significant change in beta diversity during the transition from GCD to GFD was reported (Wilcoxon p-value = 0.024, using weighted and unweighted unifrac values, Bonder *et al.*, 2016). PCoA analysis also showed samples tended to cluster on the basis of individual of isolation regardless of diet, than diet. In contrast, our analysis detected no differences in beta-diversity or unifrac (weighted or unweighted, Figure 1B). Furthermore, no clustering as a factor of diet was apparent.

The original report noted that the species *Ruminococcus bromii* and *Roseburia faecis,* and the *Veillonellaceae* family had lowered abundance in GFD, while the families *Victivallaceae, Clostridiaceae,* and *Coriobacteriaceae*, the order *ML615J-28,* and the genus *Slackia* all increased in abundance in GFD (Bonder *et al*., 2016). Our analysis found the only one ASV that was differentially abundant corresponded to the genus *Faecalibacterium,* LEFSe LDA >= 2.0, with a higher abundance in GFD samples (Figure 2). The previous analysis noted no significantly differentially abundant pathways in the transition period. Likewise, we noted no differentially abundant MetaCyc pathways during the transition.

It has been noted that members of *Faecalibacterium*, specifically *F. prausnitzii*, are less abundant in both the treated and untreated celiac microbiome, than to both the healthy and untreated celiac microbiome (Herrán *et al*., 2017, De Palma *et al*., 2010). *F. prausnitzii* are known for producing butyrate, a short-chain fatty acid known to exert an anti-inflammatory effect by driving the differentiation of T-cells into anti-inflammatory T-regulatory cells (Zhou *et al*., 2018). Our work suggests the reduction in *F. prausnitzii* is unlikely from a GDF, because this result is replicated in healthy controls. Furthermore, other perturbations to the microbiome seen in treated celiac disease (such as lowered alpha and beta diversity) were not noted in healthy patients on a GFD, once again demonstrating that these changes are likely due to the disease rather than treatment. Overall, our findings found little impact on microbial community composition and metabolic pathways when healthy patients are placed on a GFD.

The Mexican CD microbiome: original analysis versus reanalysis of Garcia-Mazcorro *et al.*

Duodenal microbiome alpha and beta diversity

Another celiac microbiome study, by Garcia-Mazcorro *et al.*, was also re-examined. That study used QIIME, PICRUSt, and the GreenGenes database and included patients with NCGS, a condition where gluten triggers symptoms similar to celiac disease, however there is identifiable inflammatory reaction of villous degradation unlike in celiac disease. Although duodenal biopsies of celiac patients had a lowered alpha-diversity (Shannon diversity index), they showed no differences in clustering for weighted or unweighted unifrac. Our analysis mirrored this result, with lowered alpha-diversity in celiac samples across metrics, with the only significant change being in the Shannon diversity index (ANOVA test, p = 0.046243, Figure 3A). Similar to the original study, there were no differences in beta diversity or clustering for weighted and unweighted unifrac values (Figure 3B).

Differentially abundant duodenal microbiota

The original study used LEFSe to identify differentially abundant taxa, and found the duodenum of celiac patients was characterized by a lowered abundance of OTUs corresponding to *Bacteroidetes* and *Fusobacteria* and an elevated abundance of OTUs belonging to *Novisprillium*. The microbiome of NCGS patients was characterized by elevated OTUs belonging to *Actinobacillus* and *Ruminococcaceae*. Controls had an elevated abundance of OTUs belonging to *Sphingobacterium* compared to both CD and NCGS. Our analysis found that the biopsies of celiac patients had elevated ASVs belonging to *Azospira*, *Phyllobacterium* and *Stenotrophomonas*. Also, both *Streptococcus* and *Neisseria* were elevated in NCGS, with both taxa having similar average group aduncance in CD and controls. *Fusobacterium* was found to be lowered in CD, with similar average group abundances in NCGS and controls (Figure 4).

Both *Azospira* and *Phyllobacterium* are nitrogen-fixing bacteria commonly found in the roots of plants (Jha *et al*., 2020, Jiao *et al*., 2015). Finding these species is likely a misidentification, as these samples are unlikely to be found in the human gut. *Stenotrophomonas*, on the other hand, has already been linked to inflammatory bowel disease, found in elevated abundance in both Crohn's disease and ulcerative colitis (Knösel *et al*., 2009, Walujkar *et al*., 2018). Furthermore, this bacteria dominates the microenvironment near the small intestinal gut epithelium in dysbiotic mice (Bertolini *et al*., 2019). It is possible this bacteria is also found in close association with the gut-epithelium of the duodenum in humans, perhaps producing similar effects as those found in mice with ulcerative colitis. Previous studies noted elevated *Fusobacterium* in CD (Di Biase *et al*., 2021); however, our analysis and the original analysis both showed that this genus is lowered in CD but not NCGS, illustrating a distinction between the two conditions. *Fusobacterium* is considered to be a "bad" bacteria, as it has been found to be overly abundant in colorectal cancer, where, among other signaling mechanisms, it can act on T-cells to inhibit the immune response, thus worsening the cancer (Kelly *et al*., 2018). As CD is ultimately mediated by CD8+ and CD4+ T-cells (Han *et al*., 2013), it is counterintuitive to expect elevated abundance of *Fusobacterium* to worsen the disease. More likely, the lowered prevalence of *Fusobacterium* upregulates T-cell mediated immune responses, as in both the previous analysis and our analysis. Another explanation for this discrepancy is that it is possible that the *Fusobacterium* detected as overabundant in CD and colorectal cancer are actually two different strains, each leading to its own disease state. An alternative possibility is that the observed changes reflect regional differences in the celiac microbiome of Mexican CD patients as diet and other factors corresponding to region can greatly affect microbial composition.

Both *Streptococcus* and *Neisseria* were elevated in controls, with CD and celiac having similar group averages, highlighting a possible distinction between the NGCS microbiome and celiac microbiome. *Streptococcus* was previously noted as elevated in patients with NCGS (Garcia-Mazcorro *et al*., 2018) as well as patients with functional dyspepsia (Bodkhe *et al*., 2019). However, in the past it was noted that *Neisseria* was elevated in the duodenum of Italian CD patients (D'Argenio *et al*., 2016). Here we demonstrated that *Neisseria* is found in similar levels in CD and control participants. This may indicate that the microbiome of Italian and Mexican patients differ, or previous work may have identified a different species of *Neisseria*. Together these results highlight the differences of microbiome structure between CD and NCGS.

Differential metabolic functions: Does the celiac microbiome cause vitamin K deficiency?

Previous work detected no differentially abundant pathways in the duodenum of CD, NCGS and controls. Our re-analysis of the PICRUSt2 pathways output, however, detected three pathways that were differentially abundant (Figure 5). These three pathways were identified as significant using both LESFe and EdgeR; and all three pathways, PWY 7371, 7373, and 6263, are involved in the synthesis of menaquinones. These results were validated by random forest analysis, with PWY 7371 and 6263 being ranked among the top-10 predictors of disease state. Menaquinones include nutrients such as vitamin K2, which are produced almost exclusively by gut microbes in mammals (Conly *et al*., 1992). Several case studies have noted that CD patients have vitamin K deficiencies (Gonzalez *et al.,* 2019, Hussaini *et al* 1999), however previously no explanation had been established. It is possible that these deficiencies may be due to a lowered abundance of vitamin K-producing bacteria, as indicated by the reduced abundance of pathways 7371, 7373, and 6263. This result may also explain some of the lower alpha-diversity in the celiac microbiome, as menaquinones commonly serve as microbial growth factors. Furthermore, menaquinones have been shown to be growth factors of *Faecalibacterium*, perhaps explaining the previously noted deficiency of the genus (Fenn *et al*., 2017). A reductionion in menaquinone producing taxa would result in a microbiome which is either lower in population and or diversity, either of which would result in a sample with reduced diversity.

Together these results indicate that the celiac and NCGS microbiomes are distinct in composition, with the duodenum of celiac patients being characterized by an abundance of *Stenotrophomonas* and deficiency of *Fusobacterium,* while the duodenum of NCGS patients is characterized by an abundance of *Neisseria and Streptococcus*. Furthermore, the celiac microbiome is functionally distinct from that of controls or the NGCS microbiomes, as is evident by the reduced presence of 3-menaquinone-producing pathways. Deficiencies in these three pathways may explain the known deficiencies in vitamin K among CD patients, as well as the reduced alpha-diversity seen in the duodenum of celiac patients. Furthermore, menaquinones are known growth factors for several "good" bacteria; and loss of menaquinone producing taxa is known to promote dysbiosis and the loss of beneficial genera of bacteria observed in CD patients (Conly, Stein, 1992).

Stool microbiota alpha / beta diversity and differentially abundant taxa

Garcia-Mazcorro *et al.* also obtained stool samples from patients before and after GFD treatment. However, that original study was unable to get full participation from their sample with only 5 of 12 NCGS, 6 of 12 controls and 3 of 6 CD submitting stool samples for both a

GFD and a GCD. Because of this, the original authors did not focus on analyzing stool samples. Nevertheless, they did unexpectedly detect a shift in the proportion of *Bacteroidetes* and *Firmicutes* for all samples, regardless of disease state, with the abundance of *Bacteroidetes* being lowered across samples. No pathways were identified as being differentially abundant.

Our analysis noted no differences in alpha diversity between study groups, as well as no differences in beta diversity nor clustering using both weighted and unweighted unifrac values (Figure 6 A/B). LEFSe identified *Pseudomonas* and *Novispirillum* as being elevated in celiac stool samples, *Ruminococcus and Bifidobacterium* being elevated in controls, and *Haemophilus, Oscillospiraceae, Collinsella, Clostridia,* and *Oscillospiraceae* as being elevated in NCGS (Figure 7). *Novisprillum* were previously noted as associated with celiac disease; *Pseudomonas* is known to be elevated in CD patients (Vittasalo *et al*., 2014), and to elicit an inflammatory response in the gut (Lin, Kazmierczk, 2017). *Novisiprilum* was noted as being more abundant in the duodenum in the previous analysis (Garcia-Mazcorro *et al*., 2018). *Ruminococcus* and *Bifidobacterium* were previously shown to be less abundant in fecal samples of CD (Bibbò *et al*., 2020). *Collinsella* was previously noted as being a pro-inflammatory bacteria (Astbury *et al*., 2020); however this genus was reduced in CD but elevated in NCGS. This is interesting as NCGS is a condition in which an inflammatory response is unlikely the cause of symptoms, with NCGS patients having normal biomarkers for inflammation (Dale *et al*., 2021), unlike untreated CD patients. *Haemophilus* was noted as being reduced in CD patients only. *Haemophilus* is known to be underrepresented in the microbiome of patients with rheumatoid arthritis (RA). Both conditions are considered autoimmune diseases, illustrating a similarity between celiac and other autoimmune diseases (Di Sante *et al*., 2021).

GFD and Mexican patients

We next examined samples based upon dietary status. The study included initial data taken from untreated NCGS and CD (GCD) and data taken 6 months post-treatment (GFD). All patients, regardless of disease status were sampled after adhering to a strict GFD for 6 months. It was noted that some symptoms of CD persist in many patients after the transition to GFD despite strict adherence to a GFD, in a condition dubbed refractory celiac disease (RCD). To determine whether microbes are contributing to RCD, dietary status was considered in the search for taxa and metabolic features that remain the same pre- and post-treatment. LEFSe analysis of duodenum biopsies showed several ASVs of *Pseudomonas* that were elevated either pre-treatment (ASVs 4 and 22), post-treatment (ASVs 52 and 29), or regardless of treatment (ASV 11). *Stentotrophomonas* and *Pseudomonas* were identified as being elevated in CD regardless of treatment status. Both taxa are commonly found in association with the mucosa in IBD (Knösel *et al*., 2009, Walujkar *et al*., 2018, Vittasalo *et al*., 2014), with *Pseudomonas* specifically known for producing an inflammatory response in the bowel (Lin, Kazmierczk, 2017). It may be that the persistent association of these bacteria with the mucosa, despite dietary intervention, causes a persistent inflammatory response in RCD, despite the removal of gluten from the diet. It was also found that *Anaerostipes* was elevated in the pre-treatment celiac biopsies. *Anaerostipes* is known to produce butyrate (Rivière *et al*., 2016) and is also known to be elevated in FDRs of CD patients (Bodhke *et al*., 2019). This may indicate that butyrate-producing bacteria are more abundant in the pre-treatment state. Notably, however, this finding is at odds with claims by Bonder *et al*., which found an increase in the butyrate-producing genus, *Faecalibacterium*, due to the GFD. This likely indicates that these differences are due to the disease itself rather than diet.

Our analysis of fecal samples, pre-treatment vs. post-treatment, again detected elevated *Pseudomonas* in CD, regardless of treatment status. *Haemophilus*, *Methanobrevibacter*, and *Collinsella* were all elevated in NCGS, regardless of disease state. Previously, *Methanobrevibacter* was noted as increased in IBS (Takakura, Pimetel, 2020), once again illustrating differences between the microbial communities of CD and NCGS, and potential similarities between NCGS and IBS. Interestingly, *Neisseria* was elevated in CD pre-treatment, perhaps indicating that *Neisseria* associates with CD only at pre-treatment; however the low sample size due to a lack of patient participation greatly reduces the power of these findings.

Greengenes versus SILVA taxonomy

There was a difference in the identity of the significant taxa detected in the previous study and our analysis. To understand whether this difference is due to pipeline (QIIME *vs.* dada2) or database (GreenGenes v12 *vs.* SILVA nr 99 v138), we assigned taxonomy to our ASV table, using both GreenGenes and SILVA databases. If our findings replicated the original study using GreenGenes rather than SILVA, it would indicate that OTU-generating pipelines produce similar results as ASV pipelines, meaning the old analysis is still valid. If the taxonomy with SILVA and GreenGenes are similar however, then this would indicate the need to reanalyze older data with new tools to obtain more accurate results. We found that the most abundant taxa between greengenes and SILVA remained 80% similar for the duodenum and 90% similar for feces (Table 1). However the resulting taxonomy table from Greengenes and SILVA only had an overall similarity of 6.3%. This indicates that the most abundant taxa share identity between

GreenGenes and SILVA, thus illustrating that this database is most likely the cause of the discrepancies between the results of our study and the original.

This difference in results illustrates the need of current researchers to reanalyze old datasets. Although scientists conduct analyses using best practices in their time, computational biology and bioinformatics are rapidly evolving fields, with new tools and analysis techniques being produced constantly. Databases like SRA and ENA make data freely available and easily accessible, so it is simple to perform analyses like ours on older data and extract new and relevant results.

The Indian CD microbiome: original analysis *vs.* reanalysis of Bodhke *et al.*

Fecal alpha/beta diversity and contaminating DNA

Bodkhe *et al*. originally included paired biopsies and stool samples taken from 23 untreated CD patients, 24 First-degree relatives (FDRs), and 23 patients with functional dyspepsia or hepatitis B (HEPB). Their study treated FDRs as CD patients in the pre-diseased state, and patients with functional dyspepsia/ HEPB as controls. Our analysis aimed to compare the microbiome of Indian patients compared to healthy controls, for that reason both the FDRs and HEPB patients were left out of the analysis of the individual study, but included in the pooled analysis with the disease state: Non-inflammatory bowel disease. New controls were pulled from Chaudhari *et al*. and Dubey *et al*. Dubey *et al*. was conducted in the same region of India (Delhi) and thus serves as the best control since the diets on the Indian subcontinent are regionally specific giving different parts of the country wildly different microbial compositions in their intestinal flora (e.g. data from Chaudhari *et al*. was collected from a rural region of India.) Together, 19 negative stool controls were pulled from Chaudhari *et al.* and 17 from

Dubey *et al*. No publicly available datasets containing healthy Indian duodenum biopsies were found, however the biopsies were included in the pooled analysis.

The original report found no significant differences in alpha diversity and no clustering using Bray-Curtis. Our analysis found higher alpha diversity (Shannon index) in CD at the feature level, with alpha diversity being lower compared to controls at all taxonomic levels (Figure 8).

Beta-diversity using Bray-Crustis produced 3 clusters, 2 control clusters and a diseased CD cluster. The two control clusters likely reflected microbiome differences due to regional diet, as both control sets were taken from different regions of India. Interestingly, CD clustered separately from both, rather than with the healthy samples from the same region, perhaps indicating differences in community structure (Figure 8B). Unweighted Unifrac showed clustering for controls and CD for all taxonomic levels, once again with the 3 sample groups clustering distinctly from each other. Weighted unifrac showed clustering of controls and celiacs distinctly for feature level, however this pattern disappeared for genus- phylum levels (Figure 8C). The results with Shannon diversity and Weighted unifrac were puzzling as the communities should be closer in diversity for higher taxonomic levels, as opposed to the feature level. Indicating that the differences in both diversity and community structure are derived from unassigned bacterial ASVs. These ASVs are indeed bacterial in nature, as they were classified as such, indicating that better classification of non-Western microbiomes is desperately needed to better understand the contribution of the gut-microbiome in these understudied regions.

Examination of the composition of communities uncovered a high proportion of unassigned reads in the diseased state: Of the original 38,005 ASVs, 4710 had no taxonomic assignment, and 11,527 bacteria were not classified below the kingdom level, making 43% of the reads from this set uniformative (Figure 8D). These sequences were removed, and 10% clustered by 95% similarity in mega. 95% similarity was chosen as it represents the sequence similarity of the v4 variable region and thus one cluster should encapsulate most of the potential 16S v4 sequences. Clustering created 32 clusters, with 1 cluster representing 96% of the sequences. One sequence from each cluster was removed and passed to BLAST for assignment. BLAST analysis showed that 96% of the unclassified DNA had greatest sequence similarity to uncultured 16S rRNA records, with the remaining 3 percent split between contaminianting human, viral, fungal, and bacterial gDNA. The uncultured bacterial DNA may represent DNA chimeras; however they were not identified as chimeric using dada2's remove-chimera denovo method nor vsearch's reference-based removal method. Thus, it is likely these taxa represent organisms that have not yet been classified but are common to the Indian microbiome. Nevertheless, as these taxa exerted more of an effect on unweighted unifrac (which does not take taxon abundance into account, compared to weighted unifrac), they are likely to only be present in small numbers within the Indian microbiome. Removing these taxa generated more robust results, with the alpha and beta diversity plots remaining more or less static, through the taxonomic levels, and with CD having lowered alpha diversity at the feature, genus, and family levels (Figure 8C, 8D).

Differentially abundant taxa

The original report found both FDRs and CD had fewer ASVs belonging to *Dorea* and *Akermansia*. It was noted that CD had a lowered abundance of *Prevotella*. FDRs and CD had an increase in ASVs corresponding to *Pediococcus, Intestinibacter, Blautia* and *Dorea* (Bodhke *et al*., 2019). Our analysis showed an elevation of *Prevotella-9* in controls with ASVs 4, 5, 6, 8, and

10. Interestingly, one ASV 206 of *Prevotella-9* was elevated in CD. Healthy stool samples also had elevated ASVs of *Pseudobutyrivibrio, Acinetobacter,* and *Bacteroides*. *Bifidobacterium* was elevated in both CD and controls with ASV 38 being elevated in CD and ASV 50 elevated in controls. An ASV belonging to *Bacteroidales* was also noted as being elevated in CD (Figure 9).

Prevotella has been identified as a potentially inflammatory bacteria (Scher *et al*., 2013), however it has also been noted as being elevated in non-western populations, with the highest enrichment being those from the Indian subcontinent (Prasoodanan *et al*., 2021). It was found that strains of *Prevotella* taken from Western and non-Western populations tend to cluster separately from one and other, with the Western populations tending to have pro-inflammatory *Prevotella* and non-Western populations having strains of carbohydrate-degrading *Prevotella* (Prasoodanan *et al*., 2021, Wu *et al*., 2013). Non-Western populations tend to have a diet that is more rich in plant matter compared to Western populations, likely indicating that these enriched genera utilize carbohydrates from the diet (Wu *et al*., 2013), and that diet may explain the differences in bacterial function. Furthermore, it was found that *Prevotella* is indeed enriched in the gut of Western IBD patients, however these bacteria tend to be closely related to pro-inflammatory oral strains of *Prevotella* (Prasoodanan *et al* 2021, Olbjørn *et al*., 2019). Previous studies looking at the Italian pediatric CD microbiome found that stool of CD children had deficiencies in *Prevotella* compared to non-CD children, once again suggesting that *Prevotella* may play a beneficial role in the gut (Di Biase *et al*,. 2021). Together these results likely show that the healthy Indian fecal microbiome is enriched in species/strains of *Prevotella* that degrade dietary carbohydrates, while the diseased Indian CD microbiome is enriched in potentially pro-inflammatory strains of *Prevotella*.

Pseudobutyrivibrio is a butyrate producing bacteria that encoded for many genes for plant-derived polysaccharide utilization, with butyrate being one of the end products (Kopečný *et al*., 2003). *Acinetobacter* was previously noted as being lower in stool samples of Indian patients, via the original results of Bodkhe *et al*. Elevated abundance of *Bacteroides* was previously noted in stool of CD patients and children at risk for the development of celiac disease (Di Biase *et al*., 2021). Previous studies noted a reduction of *Bifidobacterium* in the stool of CD patients (Olivares *et al*., 2015), but we detected just one ASV of *Bifidobacterium,* perhaps reflecting species or strain differences in the *Bifidobacterium* associated with CD and controls. *Bacteroidales* were also found to be significantly reduced in patients with IBD, specifically Crohn's disease (Gevers *et al*., 2014).

Together these results show significant overlap between the fecal microbiomes of Indian CD patients and CD patients from around the world.

Differential metabolic functions

Several pathways for the salvage/production of adenosylcobalamin were identified as being reduced in the stool of Indian CD patients compared to healthy controls (COBALSYN PWY, PWY 5509, Figure 10). Adenosylcobalamin, or vitamin B12, is another critical growth factor found in microbial communities. Previous work has demonstrated that B vitamins, including vitamin B12 are widely shared in the gut-microbiome with many species lacking genes critical for the production of B vitamins (Magnúsdóttir *et al*., 2015). Oral vitamin B2 supplements were shown to increase the diversity of species and ameliorate signatures of dysbiosis in fecal samples of patients with Crohn's disease (Pham *et al*., 2021). Another study found deficiencies in vitamin B led to a proinflammatory state, illustrating another connection between vitamin B and its potential contributions to IBD. Furthermore, symptoms of IBD were alemorated when paired with vitamin B supplementation (Gominak, 2016). It appears that these vitamins promote a diverse gut microbiome and the absence of B vitamin producing bacteria and B vitamins seems to positively correlate with worsening of IBD symptoms.

Pooled analysis

Indian CD FDRs as NIBD

FDRs of CD patients are noted to have microbiomes that are atypical: distinct from the healthy microbiome and similar to the CD microbiome (for this reason both duodenal biopsies and fecal samples from Bodkhe et al. were included as NIBD rather than healthy, as such perturbations may confuse the analysis). In any case, it was noted that dysbiotic samples (NIBD and CD) from both India and Mexico more closely resembled each other rather than healthy samples. This may be indicative of dysbiosis as a result of disease, rather than disease as a result of dysbiosis.

Duodenal and fecal alpha and beta diversity analysis

Both duodenum and stool samples from the pool analysis clustered based upon geographic region rather than disease or CD status, indicating that factors such as diet and geographic region are more influential than disease in terms of microbial community structure. This comes as no surprise, as both factors heavily influence microbiome composition. Surprisingly, little differences in alpha-diversity were found between disease states (Figure 11, 14, 17, 23, 26, 29)

Previously, it was noted that CD patients have stool and duodenum microbiome with lowered alpha diversity compared to healthy controls; however our analysis across 376 samples did not replicate this finding. This, together with the findings from our beta-diversity/clustering analysis, may show that the CD and healthy microbiomes do not differ as much as previously thought in terms of community structure.

Duodenal Community Structure Analysis

When looking at the stacked abundance bar plots, it is obvious that CD and NIBD samples, while distinct from controls, were more like controls from the same region than samples taken from other parts of the world with the same condition (Figure 12, 15, 18, 24, 27, 30).

Differentially abundant duodenal microbiota

Fusobacterium was consistently found to be reduced in CD patients from raw and scaled data (Figure 13, 16, 19) but not from normalized data. This genus of bacteria was also reduced in CD patients in our reanalysis of the Mexican data alone. In some regards, it makes sense for the genus to be reduced in CD patients, as it has been demonstrated to inhibit the response of T-cells, with CD being a T-cell mediated disease. Both Indian and Mexican samples appeared to have somewhat similar levels of *Fusobacteria*, the phylum this genus belongs to. However this phylum was absent from non-CD samples from India. This perhaps indicates that the trend observed in Indian and Mexican CD samples is once again a function of disease-induced dysbiosis rather than dysbiosis inducing the disease. If *Fusobacterium* deficiencies indeed play a role in the progression of CD, then one would expect similar levels of *Fusobacterium* to be present in both Indian and Mexican non-CD patients.

Haemophilus was also found as abundant in healthy samples compared to CD samples, a confirmation of our finding in the re-analysis of the Mexican samples (Figure 13, 16, 19). As stated previously, this bacteria was found to be enriched in the microbiomes of patients with RA, which is notable because RA and CD are similar in regards to biomarkers, as well as a positive response to the removal of gluten from the diet (Lerner, Matthias, 2015). The depletion of this bacteria in both conditions perhaps represents another connection between the two diseases. It should be noted that this bacteria was also found to be enriched in the Mexican samples, meaning that this genus correlated with geographic region and was thus lower in Indian samples. This result, while robust across raw, filtered, and scaled data, may simply be noise due to batch effects. In a previous profiling of the Indian microbiome, *Haemophilus* was not found to constitute the Indian gut microbiome in significant numbers, further supporting that this result is simply from batch effects (Chaudhari *et al*., 2020).

Similarly, *Akkermansia* was found to associate both with CD and region of origin, with this taxa being elevated in both CD biopsies and Indian biopsies (Figure 13,16, 19). This genus of bacteria was previously identified as less abundant in Italian pediatric CD patients, and was noted as beneficial and found in association with the gut lining (Xu *et al.,* 2020, Dao *et al.,* 2016). This likely indicates that this genre is beneficial and not contributing to the disease. With these points considered this is likely a false-positive as a result of a lack of healthy Indian duodenal biopsies

Acinetobacter gave the strongest signal from CD-associated microbiota (Figure 13B, 16B, 19B). This genus, as previously noted, was elevated in Brazilian CD patients and was also linked to Crohn's disease in non-Western populations (El Mouzan *et al.*, 2018). This result perhaps found an association between CD and the microbiome that is not detected by studies conducted in the United States and Europe. Previous research on the microbiome of post-menopausal women found that the species *Acinetobacter radioresistens* was positively correlated, but weakly associated with high levels of C-reactive protein. C-reactive protein is an inflammatory blood marker which has been observed as elevated in IBD patients and is used to determine whether a patient is suffering from IBD or non-inflammatory IBS, with IBD patients having elevated levels (Menees *et al*., 2015). C-reactive protein has also been found to be elevated in CD patients (Tetzlaff *et al.,* 2017), perhaps illustrating a connection between *Acinetobacter* and the systemic inflammation observed in CD. More research is needed to understand whether this association has anything to do with the progression of the disease, or if it is simply due to regional effects such as diet.

No definitive microbial signature was identified in the duodenum of CD patients from India and Mexico, but that does not mean one does not exist. The duodenum is just one of several chambers of the small intestine that are impacted by the disease. The duodenum also has the lowest concentration of bacteria in the small intestine (Brown, Esterházy, 2021). Concentrations of bacteria increase over the length of the large intestine, meaning that the concentration of bacteria within the duodenum may be too small to exert an effect on the host (Brown, Esterházy, 2021). Other chambers of the small intestine should also be evaluated to see whether their communities mirror the dysbiosis of the duodenum.

Differentially abundant duodenal microbiota function: connections to dysbiosis

One pathway (P163 PWY, Figure 22) which generates both acetate and butyrate, was found to be reduced in CD patients compared to controls and NIB. A decrease in acetate production may in turn lower levels of acetate in CD patients. Acetate is used by 95% of butyergenenic taxa; and acetate concentrations directly correlate with butyrate concentrations(Barcenilla *et al*., 2000). Thus, this imbalance in pathways for the production and consumption of acetate may reduce butyrate production, potentially worsening inflammation and dysbiosis. While no butyergenenic taxa were identified in the pooled analysis, several were identified in each individual re-analysis, with this result perhaps explaining why. This is due to the fact that this pathway was found to be universally lowered in CD, despite geographic region of isolation, perhaps indicating that it is the metabolome of CD which should be the focus of our study rather than the microbiome.

Differences in amino acid synthesis and degradation were observed with CD (Figure 22). Affected patients had elevated pathways for the production of ornithine, histidine and tryptophan, while non-CD upregulated other pathways, including ones that degrade glutamate or used glutamate as a substrate. Tryptophan, glutamate, and ornithine were previously linked to CD in previous studies; however, these results were generated using peripheral sera (Naluai *et al*., 2018). There is little in the literature regarding whether metabolic data taken from blood sera reflects activity by the gut metabolome; however, previous work demonstrated that tryptophan and histidine are elevated in stool of CD patients (Di Cagno *et al*., 2011). Furthermore, DSS-induced colitis in murine models also showed an increase in several amino acids, including glutamate and tryptophan, further supporting the PICRUSt2 results (Xie *et al*., 2021). Data generated using PICRUSt2 serve as a proxy as to what is occurring in the microbiome and as to what metabolites are present. This is because PICRUSt2 aligns ASVs to reference genomes, annotates the genomes, and makes predictions on the abundance of a given pathway based on enzyme count. This gives information as to differences in the potential to perform a given pathway, but does not actually provide information as to whether those genes are being expressed and the metabolite being produced or substrate consumed. That being said, other *in* *silico* studies also detected elevated amino acid synthesis in IBD-mediated dysbiosis (Heinken *et al*., 2021) and the results of our study are in line with those findings.

As noted previously, stool samples from Indian CD patients were characterized by a deficiency in pathways for the production of vitamin B12. In the pooled duodenum analysis, CD patients were deficient in a pathway generating the precursor to all of the B12 vitamins (Figure 22). This pathway (PWY 5188) begins with glutamate and ends with tetrapyrrole. Tetrapyrroles serve as the precursors of many metal-binding compounds, such as B-vitamins (colbamine) and hemes. As previously noted, the presence of B vitamins in a microbial community is able to induce a more diverse environment (Lovley *et al*., 1996). Cobalamins and hemes are both used as electron acceptors in environments poor in oxygen. Cooperative electron transport among microbes has been identified across several environments (Lovley *et al*., 1996, Hederstedt *et al.,* 2020) and exists within the gut (Light *et al*., 2018). Lowered abundance of pathways producing vitamin B and K, vitamin B precursors, and flavodoxin may indicate a breakdown in the shared electron transport chain in the CD microbiome. Whether the breakdown is a symptom or cause of dysbiosis is impossible to know without metabolic data, so investigation into the CD metabolome should be done to verify these results.

PICRUSt2: An improved tool

Overall, the results generated using PICRUSt2 corroborated the limited information found in the literature, showing that new computational tools are able to generate valid results from older data. More results regarding the CD metagenome and metabolome were generated using PICRUSt2 compared to PICRUSt1, which was used by many of the previous studies. This result is of no surprise given that PICRUSt2 is able to incorporate more user generated data due to its use of ASVs rather than just reference GreenGenes OTUs as is the case in PICRUSt1. Furthermore, PICRUSt2 has a reference database that is 20x larger than that of PICRUSt1, with functional information being reported as both pathways and individual KEGG modules. The pathway data generated by PICRUSt2 is regarded as the highest level output, with its predictions being made using enzyme counts (ECs). This is much more useful compared to just KEGG modules, which are the only output of PICRUSt1, as many enzymes can belong to more than just a single pathway and be implicated in the generation of several metabolites. By looking at the aggregate enzyme counts, PICRUSt2 is able to better predict which pathways have the potential to be elevated in a given microbial community. Similar to re-analysing old data with ASV producing pipelines can gleen better more clear results of community differences, PICRUSt2 can be used to reanalyze older data to get more accurate metabolic and metagenomic information regarding old data and can be used to search for functional trends in a given disease using publically available data.

Differentially abundant stool microbiota

Many taxa identified as less abundant in CD were also identified in American studies of CD stool samples (Figure 25, 28, 31). *Alistipes* and *Ruminococcaceae* were both repeatedly identified as less abundant in CD and NIBD stool and associated with healthy samples. Another bacteria, *Escherichia-Shigella,* was identified in CD stool samples across all analysis methods. This bacteria was previously isolated from stool of American CD patients and shown to induce immune responses (De Palma *et al.,* 2012). Our results mirror that analysis, with this bacteria elevated in CD stool samples.

Lachnospiraceae was also found enriched in CD stool (Figure 25, 28,31). This bacteria was previously found overabundant in stool samples of children at-risk for CD (Leonard *et al*., 2021). Our results suggest this increase in abundance persists once the disease is active in adults. This is somewhat confusing, as this family of bacteria is commonly identified as a “good” bacteria. One study, seeking to profile members of this family found that while, by and large, most members are beneficial to the host, some members are associated with IBD (Vacca *et al*., 2020). This reflects a limitation of 16S-based analysis: it cannot identify some ASVs with high specificity. It is possible that there are species and strains of *Lachnospiraceae* that are associated with CD that are impossible to identify with our chosen analysis methods.

Stool, however, should not be the focus of studies examining celiac disease, since it does not necessarily reflect the small intestinal microbiome’s composition. As previously mentioned, CD is active in the upper chambers of the small intestine, namely the duodenum and jejunum. Biopsies from these tissues are more difficult to obtain compared to stool samples, but should likely be the gold-standard when studying CD microbiome, as these samples represent the chambers of the gut where the disease is active.

A previous meta-analysis (Sze, Schloss, 2018) of sequencing data from colorectal cancer stool and tissue samples also found that features associated with disease were not uniform across samples, but rather had a patchy distribution, with some studies having a strong signal indicating that a taxa was highly associated with the disease, while in other studies the signal was reduced or absent. To these researchers, this indicated that these taxa may be associated with the disease, and while some may worsen patient outcomes, are not a required component of the mechanism of pathogenesis in colorectal cancer. While this study was conducted on an entirely separate disease, the same may be true of CD, with some “bad” taxa associating with the disease and worsening symptoms and recovery but not actually playing a casual role in the prognosis of a patient from inactive to active CD.

Random forest

Random forest is a supervised method of machine learning for the classification of disease state based upon microbial abundance data. Machine learning algorithms such as random forest have been of interest; they represent a non-invasive method of diagnosis in CD. If these algorithms can accurately identify disease state from stool abundance data, with a high sensitivity, then stool sampling may represent a better diagnostic than duodenal biopsies which require patients to undergo a surgical procedure and are the gold standard for CD diagnosis. Random forest has been demonstrated to accurately predict the origin of fecal samples. For instance, in Roguet *et al*., a random forest algorithm predicted whether a given stool sample came from a cat, dog, pig, deer, or human using microbial abundance data (Roguet *et al*., 2018) . It should be noted that these communities should be expected to deviate significantly from each other as they are from entirely different host taxa. Microbiomes are highly adapted to their host, and it should be of no surprise that these communities are easy to tell apart. Our implementation of random forest analysis was able to accurately predict the country of origin of both the duodenal and fecal pooled analysis, however across the studies and pooled analysis, it was unable to accurately predict disease status. It is known that geographic region can have drastic impacts on the composition of microbial communities within the host due to a variety of factors including diet (Human Microbiome Consortium, 2012, Mancabelli *et al.*, 2017). This may illustrate that despite having different disease states, CD and NIBD patients have a similar microbiome to healthy individuals of the same population as opposed to CD or NIBD patients from other continents. This is further supported by unifrac analysis of the individual studies and Bray-Curtis analysis of the pooled studies which illustrated that samples tended to cluster together on the basis of country of origin rather than disease. This does not however mean that random forest classification is not applicable to IBD and CD, with Chehoud *et al.* being able to accurately predict IBD status using both bacterial and fungal data, perhaps demonstrating the importance of fungi within the gut microbiome. If anything this result, reinforces the idea that the differences observed in CD community structure are not uniform and likely do not play a contributing role in the progression of the disease.

Previous attempts unify CD knowledge

There are several sources of information describing gut-dysbiosis associated with celiac disease, however these resources are limited either in quantity of information, or to a single region of the world. One such study (Leonard *et al., 2015*) examined 500 infants at risk for the development of celiac disease over the course of five years to identify changes in the gut-microbiota and metabolome associated with the development of celiac disease. While this study did include a large sample size, it focused on developed Western countries, and is mainly representative of populations following a Western diet. This does not give an accurate representation of all the potential bacteria which could be associated with celiac disease as diet, antibiotic use and other factors associated with Western countries can have a large impact on microbiome composition. Furthermore, this study utilized mostly stool and blood samples, which serve as proxies for studying celiac disease, which is mainly active in the duodenum. There are also existing databases which serve to unify data from other studies under one archive, gutMDisorder is one such resource. GutMDisorder is a good start, but is lacking in information on CD with only eleven entries (under categories “celiac disease”- 7 entries, “coeliac disease”- 1 entry and “gluten-free diet”-3 entries) as relating to the CD microbiome, of which only three had information regarding the species level (Cheng *et al*., 2020). This is problematic as associations above the genus level are rather uninformative as the function of bacteria can vary wildly even at levels as low as strain; thus information about genus and higher classification, while useful in exploration of ecosystem structure, inform little regarding mechanisms of disease development or targets of intervention. Thus more work is needed to unify information taken from previous studies under a single resource to better understand the relationships between CD and the microbiome.

Dealing with noise and batch effects

One of the challenges of working with microbiome data is its sheer size. With ASV tables often consisting of thousands of taxa, this produces considerable noise for analysis and makes drawing conclusions from such data difficult. Thus normalization techniques have been introduced to correct for such errors. There is debate in the field as to whether such techniques are biologically relevant with some scientists going as far to say that results generated using any sort of transformation are not meaningful (McMurdie, Holmes, 2014). In our analysis, total sum-scaling, filtering and raw analysis produced robust results, with significantly differentially abundant taxa, diversity metrics and community structure all remaining similar despite transformations being applied to the data. The normalization technique taken from Gibbons *et al.* produced distinct results, showing a decrease in alpha-diversity amongst CD patients and no significantly differentially abundant taxa.

Pooled analysis and dysbiosis

Each study individually produced different results than the pooled analysis. This could be due to either noise or batch effects, both of which were corrected for and did not produce results showing that any specific disease causing taxa was enriched across studies, but instead showed that bacteria associated with dysbiosis, or an unbalanced community structure were associated with CD. This was further amplified by the results of the PICRUSt2 analysis, which again demonstrated connections to dysbiosis as a whole, but no real evidence that the CD microbiome plays a causative role in the progression of the disease. It would instead appear that the microbial community found in CD patients is a result of the disease. The procedure described in Gibbons *et al.* produced results which were distinct from these. As the results of this method are not robust, *i.e.* they are not mirrored by the other techniques, this procedure may be too stringent in its attempts to correct for batch effects. On the other hand, the results of the other studies do not seem to indicate that there is a CD microbial signature. Perhaps indicating that while the results of the normalization described in Gibbons *et al.* are distinct from the others, they are ultimately showing the same thing, that the CD microbiome is not causative of the disease and its differences are regionalized. Investigation and consensus is desperately needed to deduce what normalization techniques produce the most biologically relevant data.

CONCLUSION

Reanalysis of old datasets using new tools

Older OTU generating pipelines such as QIIME and Mothur have been used to conduct metagenomic studies of the gut for years. Such tools rely on a binning approach based on a user defined similarity threshold to denoise samples, with taxonomy assigned using a single sequence from each bin. Newer pipelines such as dada2 instead opt to use machine learning and quality scores associated with bases to denoise sequencing files and can differentiate ASVs with as little as a single base-pair of difference, thus giving a far more granular picture of the microbiome.

Regional differences in CD

Our analysis confirmed results of these older OTU based tools, as well as generating new data of CD-associated taxa in samples from both India and Mexic,. with deficiencies in *Fusobacterium* being found in both the original analysis and our reanalysis. Our reanalysis also found that *Stenotrophomonas*, a bacteria associated with IBD, was enriched in Mexican CD datasets, which was not found previously. Reanalysis of Bodhke *et al.* derived results that were more dissimilar to the original report, with CD samples characterized by a reduction in butyrate-producing taxa and an increase in *Prevotella-9*.

The datasets were then pooled to determine whether a global CD pattern exists regarding the CD microbiome using a variety of filtering and normalization techniques. Our results show that, while the CD microbiome is indeed distinct from that of the healthy microbiome, diseased samples are more similar to those taken from the same region rather than those of the same disease, likely indicating that dysbiosis seen in CD patients is a result of the disease rather than a contributing factor to the disease. Thus, one difficulty in interpreting our data was that people from different parts of the globe had location-specific microbial profiles making comparing CD cohorts difficult.

Metabolic differences in CD

Many of these studies also used PICRUSt1. In our analysis PICRUSt2 was used instead, with PICRUSt2 having a 20-fold larger database, theoretically giving it the power to derive more accurate results regarding community function. Our analysis found deficiencies in pathways for the production of electron accepting products in diseased and CD samples throughout both individual studies and pooled analysis. This trend was detected previously in dysbiotic microbiomes. We also found perturbations in amino-acid synthesis in CD patients, which was also confirmed by other studies investigating the CD microbiome. These results confirm that PICRUSt2 is indeed a valid predictor of microbial function using 16S data.

Small intestinal biopsies versus stool samples

Our studies focused on biopsies taken from the duodenum and stool samples. CD is active in the small intestine, meaning that stool samples are likely not reflective of the chambers of the gut where the disease is actually active. The duodenum represents a more logical choice of study for the CD microbiome, however it is not the only chamber of the small intestine affected by the disease. While we found no microbial signature associated with CD duodenal samples in our pooled analysis, that does not mean that the small intestinal microbiome is not involved. Investigation into the jejunal and ileal microbiome should also be pursued, where there is a higher concentration of bacteria and thus a greater possibility of these taxa exerting an effect on the host.

Summary

Overall, our findings indicate that the dysbiosis observed in CD is likely a result of the disease rather than a contributing factor. Analysis of data from any geographic region individually produces results showing potentially relevant differentially abundant taxa, however these results do not hold up across pooled analysis indicating that they likely are not contributing to the disease, or perhaps, contributing to the disease within this specific cohort in a manner that is not generalizable to the global population. This is further supported by the PICRUSt2 functional data, with connections to other traits observed in dysbiotic communities. Our results also show that reanalysis of old data is both needed and relevant as newer, more accurate results can be generated using previously analyzed sequencing data.

BIBLIOGRAPHY

Allali, I., Arnold, J.W., Roach, J., Cadenas, M.B., Butz, N., Hassan, H.M., Koci, M., Ballou, A., Mendoza, M., Ali, R., et al. (2017). A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. BMC Microbiol. *17*, 194..

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410..

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

Astbury, S., Atallah, E., Vijay, A., Aithal, G.P., Grove, J.I., and Valdes, A.M. (2020). Lower gut microbiome diversity and higher abundance of proinflammatory genus Collinsella are associated with biopsy-proven nonalcoholic steatohepatitis. Gut Microbes *11*, 569–580..

Balvočiūtė, M., and Huson, D.H. (2017). SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? BMC Genomics *18*, 114..

Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., and Stamatakis, A. (2019). EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. Syst. Biol. *68*, 365–369..

Barcenilla, A., Pryde, S.E., Martin, J.C., Duncan, S.H., Stewart, C.S., Henderson, C., and Flint, H.J. (2000). Phylogenetic relationships of butyrate-producing bacteria from the human gut. Appl. Environ. Microbiol. *66*, 1654–1661..

Bertolini, M., Ranjan, A., Thompson, A., Diaz, P.I., Sobue, T., Maas, K., and Dongari-Bagtzoglou, A. (2019). Candida albicans induces mucosal bacterial dysbiosis that promotes invasive infection. PLoS Pathog. *15*, e1007717..

Bibbò, S., Abbondio, M., Sau, R., Tanca, A., Pira, G., Errigo, A., Manetti, R., Pes, G.M., Dore, M.P., and Uzzau, S. (2020). Fecal Microbiota Signatures in Celiac Disease Patients With Poly-Autoimmunity. Frontiers in Cellular and Infection Microbiology *10*. https://doi.org/10.3389/fcimb.2020.00349.

Bodkhe, R., Shetty, S.A., Dhotre, D.P., Verma, A.K., Bhatia, K., Mishra, A., Kaur, G., Pande, P., Bangarusamy, D.K., Santosh, B.P., et al. (2019). Comparison of Small Gut and Whole Gut Microbiota of First-Degree Relatives With Adult Celiac Disease Patients and Controls. Front. Microbiol. *10*, 164..

Bonder, M.J., Tigchelaar, E.F., Cai, X., Trynka, G., Cenit, M.C., Hrdlickova, B., Zhong, H., Vatanen, T., Gevers, D., Wijmenga, C., et al. (2016). The influence of a short-term gluten-free diet on the human gut microbiome. Genome Med. *8*, 45..

Brown, H., and Esterházy, D. (2021). Intestinal immune compartmentalization: implications of tissue specific determinants in health and disease. Mucosal Immunol. *14*, 1259–1270..

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods *13*, 581–583..

Callahan, B.J., McMurdie, P.J., and Holmes, S.P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. *11*, 2639–2643..

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

Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C., et al. (2007). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Research *36*, D623–D631. https://doi.org/10.1093/nar/gkm900.

Chaudhari, D.S., Dhotre, D.P., Agarwal, D.M., Gaike, A.H., Bhalerao, D., Jadhav, P., Mongad, D., Lubree, H., Sinkar, V.P., Patil, U.K., et al. (2020). Gut, oral and skin microbiome of Indian patrilineal families reveal perceptible association with age. Sci. Rep. *10*, 5685..

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. Nucleic Acids Res. *48*, 7603..

Chong, J., Liu, P., Zhou, G., and Xia, J. (2020). Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. Nat. Protoc. *15*, 799–821..

Conly, J.M., and Stein, K. (1992). The production of menaquinones (vitamin K2) by intestinal bacteria and their role in maintaining coagulation homeostasis. Prog. Food Nutr. Sci. *16*, 307–343..

Czech, L. gappa: A toolkit for analyzing and visualizing phylogenetic (placement) data (Github).

Dale, H.F., Johannessen, J.C.H., Brønstad, I., and Lied, G.A. (2021). Assessment of Markers of Gut Integrity and Inflammation in Non-Celiac Gluten Sensitivity After a Gluten Free-Diet. Int. J. Gen. Med. *14*, 9459–9470..

De Palma, G., Nadal, I., Medina, M., Donat, E., Ribes-Koninckx, C., Calabuig, M., and Sanz, Y. (2010). Intestinal dysbiosis and reduced immunoglobulin-coated bacteria associated with coeliac disease in children. BMC Microbiol. *10*, 63..

De Palma, G., Kamanova, J., Cinova, J., Olivares, M., Drasarova, H., Tuckova, L., and Sanz, Y. (2012). Modulation of phenotypic and functional maturation of dendritic cells by intestinal bacteria and gliadin: relevance for celiac disease. J. Leukoc. Biol. *92*, 1043–1054..

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

Dhariwal, A., Chong, J., Habib, S., King, I.L., Agellon, L.B., and Xia, J. (2017). MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res. *45*, W180–W188..

Di Biase, A.R., Marasco, G., Ravaioli, F., Dajti, E., Colecchia, L., Righi, B., D'Amico, V., Festi, D., Iughetti, L., and Colecchia, A. (2021). Gut microbiota signatures and clinical manifestations in celiac disease children at onset: a pilot study. J. Gastroenterol. Hepatol. *36*, 446–454..

Di Cagno, R., De Angelis, M., De Pasquale, I., Ndagijimana, M., Vernocchi, P., Ricciuti, P., Gagliardi, F., Laghi, L., Crecchio, C., Guerzoni, M.E., et al. (2011). Duodenal and faecal microbiota of celiac children: molecular, phenotype and metabolome characterization. BMC Microbiol. *11*, 219..

Di Sante, G., Gremese, E., Tolusso, B., Cattani, P., Di Mario, C., Marchetti, S., Alivernini, S., Tredicine, M., Petricca, L., Palucci, I., et al. (2021). Haemophilus parasuis (Glaesserella parasuis) as a Potential Driver of Molecular Mimicry and Inflammation in Rheumatoid Arthritis. Front. Med. *8*, 671018..

Douglas, G.M., Maffei, V.J., Zaneveld, J.R., Yurgel, S.N., Brown, J.R., Taylor, C.M., Huttenhower, C., and Langille, M.G.I. (2020). PICRUSt2 for prediction of metagenome functions. Nat. Biotechnol. *38*, 685–688..

Dubey, A.K., Uppadhyaya, N., Nilawe, P., Chauhan, N., Kumar, S., Gupta, U.A., and Bhaduri, A. (2018). LogMPIE, pan-India profiling of the human gut microbiome using 16S rRNA sequencing. Sci Data *5*, 180232..

El Mouzan, M.I., Winter, H.S., Assiri, A.A., Korolev, K.S., Al Sarkhy, A.A., Dowd, S.E., Al Mofarreh, M.A., and Menon, R. (2018). Microbiota profile in new-onset pediatric Crohn's disease: data from a non-Western population. Gut Pathog. *10*, 49..

Fenn, K., Strandwitz, P., Stewart, E.J., Dimise, E., Rubin, S., Gurubacharya, S., Clardy, J., and Lewis, K. (2017). Quinones are growth factors for the human gut microbiota. Microbiome *5*, 161..

Garcia-Mazcorro, J.F., Rivera-Gutierrez, X., Cobos-Quevedo, O.D.J., Grube-Pagola, P., Meixueiro-Daza, A., Hernandez-Flores, K., Cabrera-Jorge, F.J., Vivanco-Cid, H., Dowd, S.E., and Remes-Troche, J.M. (2018). First Insights into the Gut Microbiota of Mexican Patients with Celiac Disease and Non-Celiac Gluten Sensitivity. Nutrients *10*. https://doi.org/10.3390/nu10111641.

Gevers, D., Kugathasan, S., Denson, L.A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S.J., Yassour, M., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe *15*, 382–392..

Gibbons, S.M., Duvallet, C., and Alm, E.J. (2018). Correcting for batch effects in case-control microbiome studies. PLoS Comput. Biol. *14*, e1006102..

Gillett, P.M. (2014). Early Microbial Markers of Celiac Disease. Journal of Clinical Gastroenterology *48*, 579–581. https://doi.org/10.1097/mcg.0000000000000159.

Gominak, S.C. (2016). Vitamin D deficiency changes the intestinal microbiome reducing B vitamin production in the gut. The resulting lack of pantothenic acid adversely affects the immune system, producing a "pro-inflammatory" state associated with atherosclerosis and autoimmunity. Medical Hypotheses *94*, 103–107. https://doi.org/10.1016/j.mehy.2016.07.007.

Gonzalez, J.J., Elgamal, M., Mishra, S., and Adekolujo, O.S. (2019). Severe Coagulopathy as a Rare Feature of Celiac Crisis in a Patient Previously Diagnosed with Celiac Disease. Am. J. Case Rep. *20*, 290–293..

Greco, L., Romino, R., Coto, I., Di Cosmo, N., Percopo, S., Maglio, M., Paparo, F., Gasperi, V., Limongelli, M.G., Cotichini, R., et al. (2002). The first large population based twin study of coeliac disease. Gut *50*, 624–628..

Han, A., Newell, E.W., Glanville, J., Fernandez-Becker, N., Khosla, C., Chien, Y.-H., and Davis, M.M. (2013). Dietary gluten triggers concomitant activation of CD4+ and CD8+ αβ T cells and γδ T cells in celiac disease. Proc. Natl. Acad. Sci. U. S. A. *110*, 13073–13078..

Hederstedt, L., Gorton, L., and Pankratova, G. (2020). Two Routes for Extracellular Electron Transfer in Enterococcus faecalis. J. Bacteriol. *202*. https://doi.org/10.1128/JB.00725-19.

Heinken, A., Hertel, J., and Thiele, I. (2021). Metabolic modelling reveals broad changes in gut microbial metabolism in inflammatory bowel disease patients with dysbiosis. NPJ Syst Biol Appl *7*, 19..

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214..

Hussaini, S.H., Ahmed, S., and Heatley, R.V. (1999). Celiac disease and hypoprothrombinemia. Nutrition *15*, 389–391..

Jha, P.N., Gomaa, A.-B., Yanni, Y.G., El-Saadany, A.-E.Y., Stedtfeld, T.M., Stedtfeld, R.D., Gantner, S., Chai, B., Cole, J., Hashsham, S.A., et al. (2020). Alterations in the Endophyte-Enriched Root-Associated Microbiome of Rice Receiving Growth-Promoting Treatments of Urea Fertilizer and Rhizobium Biofertilizer. Microb. Ecol. *79*, 367–382..

Jiao, Y.S., Yan, H., Ji, Z.J., Liu, Y.H., Sui, X.H., Zhang, X.X., Wang, E.T., Chen, W.X., and Chen, W.F. (2015). Phyllobacterium sophorae sp. nov., a symbiotic bacterium isolated from root nodules of Sophora flavescens. Int. J. Syst. Evol. Microbiol. *65*, 399–406..

Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. Protein Sci. *28*, 1947–1951..

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. Nucleic Acids Research *49*, D545–D551. https://doi.org/10.1093/nar/gkaa970.

Kelly, D., Yang, L., and Pei, Z. (2018). Gut Microbiota, Fusobacteria, and Colorectal Cancer. Diseases *6*. https://doi.org/10.3390/diseases6040109.

Knösel, T., Schewe, C., Petersen, N., Dietel, M., and Petersen, I. (2009). Prevalence of infectious pathogens in Crohn's disease. Pathol. Res. Pract. *205*, 223–230..

Kopečný, J., Zorec, M., Mrázek, J., Kobayashi, Y., and Marinšek-Logar, R. (2003). Butyrivibrio hungatei sp. nov. and Pseudobutyrivibrio xylanivorans sp. nov., butyrate-producing bacteria from the rumen. Int. J. Syst. Evol. Microbiol. *53*, 201–209..

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol. Biol. Evol. *35*, 1547–1549..

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. *31*, 814–821..

Leonard, M.M., Camhi, S., Huedo-Medina, T.B., and Fasano, A. (2015). Celiac Disease Genomic, Environmental, Microbiome, and Metabolomic (CDGEMM) Study Design: Approach to the Future of Personalized Prevention of Celiac Disease. Nutrients *7*, 9325–9336..

Leonard, M.M., Karathia, H., Pujolassos, M., Troisi, J., Valitutti, F., Subramanian, P., Camhi, S., Kenyon, V., Colucci, A., Serena, G., et al. (2020). Multi-omics analysis reveals the influence of genetic and environmental risk factors on developing gut microbiota in infants at risk of celiac disease. Microbiome *8*, 130..

Leonard, M.M., Valitutti, F., Karathia, H., Pujolassos, M., Kenyon, V., Fanelli, B., Troisi, J., Subramanian, P., Camhi, S., Colucci, A., et al. (2021). Microbiome signatures of progression toward celiac disease onset in at-risk children in a longitudinal prospective cohort study. Proc. Natl. Acad. Sci. U. S. A. *118*. https://doi.org/10.1073/pnas.2020322118.

Lerner, A., and Matthias, T. (2015). Rheumatoid arthritis–celiac disease relationship: Joints get that gut feeling. Autoimmun. Rev. *14*, 1038–1047..

Light, S.H., Su, L., Rivera-Lugo, R., Cornejo, J.A., Louie, A., Iavarone, A.T., Ajo-Franklin, C.M., and Portnoy, D.A. (2018). A flavin-based extracellular electron transfer mechanism in diverse Gram-positive bacteria. Nature *562*, 140–144..

Lin, C.K., and Kazmierczak, B.I. (2017). Inflammation: A Double-Edged Sword in the Response to Pseudomonas aeruginosa Infection. Journal of Innate Immunity *9*, 250–261. https://doi.org/10.1159/000455857.

Louca, S., and Doebeli, M. (2018). Efficient comparative phylogenetics on large trees. Bioinformatics *34*, 1053–1055..

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550..

Lovley, D.R., Coates, J.D., Blunt-Harris, E.L., Phillips, E.J.P., and Woodward, J.C. (1996). Humic substances as electron acceptors for microbial respiration. Nature *382*, 445–448..

Madeira, F., Pearce, M., Tivey, A.R.N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. https://doi.org/10.1093/nar/gkac240.

Magnúsdóttir, S., Ravcheev, D., de Crécy-Lagard, V., and Thiele, I. (2015). Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes. Front. Genet. *6*, 148..

Mancabelli, L., Milani, C., Lugli, G.A., Turroni, F., Ferrario, C., van Sinderen, D., and Ventura, M. (2017). Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. Environmental Microbiology *19*, 1379–1390. https://doi.org/10.1111/1462-2920.13692.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12..

McMurdie, P.J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One *8*, e61217..

McMurdie, P.J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput. Biol. *10*, e1003531..

Menees, S.B., Powell, C., Kurlander, J., Goel, A., and Chey, W.D. (2015). A Meta-Analysis of the Utility of C-Reactive Protein, Erythrocyte Sedimentation Rate, Fecal Calprotectin, and Fecal Lactoferrin to Exclude Inflammatory Bowel Disease in Adults With IBS. American Journal of Gastroenterology *110*, 444–454. https://doi.org/10.1038/ajg.2015.6.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *27*, 29–34..

Olbjørn, C., Småstuen, M.C., Thiis-Evensen, E., Nakstad, B., Vatn, M.H., Jahnsen, J., Ricanek, P., Vatn, S., Moen, A.E.F., Tannæs, T.M., et al. (2019). Fecal microbiota profiles in treatment-naïve pediatric inflammatory bowel disease – associations with disease phenotype, treatment, and outcome. Clinical and Experimental Gastroenterology *12*, 37–49. https://doi.org/10.2147/ceg.s186235.

Olivares, M., Albrecht, S., De Palma, G., Ferrer, M.D., Castillejo, G., Schols, H.A., and Sanz, Y. (2015). Human milk composition differs in healthy mothers and mothers with celiac disease. Eur. J. Nutr. *54*, 119–128..

Pace, L.A., and Crowe, S.E. (2016). Complex Relationships Between Food, Diet, and the Microbiome. Gastroenterol. Clin. North Am. *45*, 253–265..

Pandit, L., Cox, L.M., Malli, C., D'Cunha, A., Rooney, T., Lokhande, H., Willocq, V., Saxena, S., and Chitnis, T. (2021). Clostridium bolteae is elevated in neuromyelitis optica spectrum disorder in India and shares sequence similarity with AQP4. Neurol Neuroimmunol Neuroinflamm *8*. https://doi.org/10.1212/NXI.0000000000000907.

Pham, V.T., Fehlbaum, S., Seifert, N., Richard, N., Bruins, M.J., Sybesma, W., Rehman, A., and Steinert, R.E. (2021). Effects of colon-targeted vitamins on the composition and metabolic activity of the human gut microbiome- a pilot study. Gut Microbes *13*, 1–20..

Prasoodanan P K, V., Sharma, A.K., Mahajan, S., Dhakan, D.B., Maji, A., Scaria, J., and Sharma, V.K. (2021). Western and non-western gut microbiomes reveal new roles of Prevotella in carbohydrate metabolism and mouth-gut axis. NPJ Biofilms Microbiomes *7*, 77..

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

Rivière, A., Selak, M., Lantin, D., Leroy, F., and De Vuyst, L. (2016). Bifidobacteria and Butyrate-Producing Colon Bacteria: Importance and Strategies for Their Stimulation in the Human Gut. Front. Microbiol. *7*, 979..

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140..

Roguet, A., Eren, A.M., Newton, R.J., and McLellan, S.L. (2018). Fecal source identification using random forest. Microbiome *6*, 185..

Rosser, E.C., Piper, C.J.M., Matei, D.E., Blair, P.A., Rendeiro, A.F., Orford, M., Alber, D.G., Krausgruber, T., Catalan, D., Klein, N., et al. (2020). Microbiota-Derived Metabolites Suppress Arthritis by Amplifying Aryl-Hydrocarbon Receptor Activation in Regulatory B Cells. Cell Metab. *31*, 837–851.e10..

Scher, J.U., Sczesnak, A., Longman, R.S., Segata, N., Ubeda, C., Bielski, C., Rostron, T., Cerundolo, V., Pamer, E.G., Abramson, S.B., et al. (2013). Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. Elife *2*, e01202..

Schippa, S., Iebba, V., Barbato, M., Di Nardo, G., Totino, V., Checchi, M., Longhi, C., Maiella, G., Cucchiara, S., and Conte, M. (2010). A distinctive “microbial signature” in celiac pediatric patients. BMC Microbiology *10*, 175. https://doi.org/10.1186/1471-2180-10-175.

Schliep, K.P. (2011). phangorn: phylogenetic analysis in R. Bioinformatics *27*, 592–593..

Schliep, K., Potts, A.J., Morrison, D.A., and Grimm, G.W. (2017). Intertwining phylogenetic trees and networks. Methods Ecol. Evol. *8*, 1212–1220..

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. Genome Biol. *12*, R60..

Stearns, J.C., for the NutriGen Alliance, Zulyniak, M.A., de Souza, R.J., Campbell, N.C., Fontes, M., Shaikh, M., Sears, M.R., Becker, A.B., Mandhane, P.J., et al. (2017). Ethnic and diet-related differences in the healthy infant microbiome. Genome Medicine *9*. https://doi.org/10.1186/s13073-017-0421-5.

Sze, M.A., and Schloss, P.D. (2018). Leveraging Existing 16S rRNA Gene Surveys To Identify Reproducible Biomarkers in Individuals with Colorectal Tumors. MBio *9*. https://doi.org/10.1128/mBio.00630-18.

Takakura, W., and Pimentel, M. (2020). Small Intestinal Bacterial Overgrowth and Irritable Bowel Syndrome - An Update. Front. Psychiatry *11*, 664..

Tetzlaff, W.F., Meroño, T., Menafra, M., Martin, M., Botta, E., Matoso, M.D., Sorroche, P., De Paula, J.A., Boero, L.E., and Brites, F. (2017). Markers of inflammation and cardiovascular disease in recently diagnosed celiac disease patients. World J. Cardiol. *9*, 448–456..

Vacca, M., Celano, G., Calabrese, F.M., Portincasa, P., Gobbetti, M., and De Angelis, M. (2020). The Controversial Role of Human Gut Lachnospiraceae. Microorganisms *8*. https://doi.org/10.3390/microorganisms8040573.

Valitutti, F., Cucchiara, S., and Fasano, A. (2019). Celiac Disease and the Microbiome. Nutrients *11*. https://doi.org/10.3390/nu11102403.

Viitasalo, L., Kurppa, K., Ashorn, M., Saavalainen, P., Huhtala, H., Ashorn, S., Mäki, M., Ilus, T., Kaukinen, K., and Iltanen, S. (2018). Microbial Biomarkers in Patients with Nonresponsive Celiac Disease. Dig. Dis. Sci. *63*, 3434–3441..

Walujkar, S.A., Kumbhare, S.V., Marathe, N.P., Patangia, D.V., Lawate, P.S., Bharadwaj, R.S., and Shouche, Y.S. (2018). Molecular profiling of mucosal tissue associated microbiota in patients manifesting acute exacerbations and remission stage of ulcerative colitis. World J. Microbiol. Biotechnol. *34*, 76..

Wu, G.D., Bushmanc, F.D., and Lewis, J.D. (2013). Diet, the human gut microbiota, and IBD. Anaerobe *24*, 117–120..

Xie, D., Li, F., Pang, D., Zhao, S., Zhang, M., Ren, Z., Geng, C., Wang, C., Wei, N., and Jiang, P. (2021). Systematic Metabolic Profiling of Mice with Dextran Sulfate Sodium-Induced Colitis. J. Inflamm. Res. *14*, 2941–2953..

Ye, Y., and Doak, T.G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Comput. Biol. *5*, e1000465..

Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., and Glöckner, F.O. (2014). The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. Nucleic Acids Res. *42*, D643–D648..

Zhou, L., Zhang, M., Wang, Y., Dorfman, R.G., Liu, H., Yu, T., Chen, X., Tang, D., Xu, L., Yin, Y., et al. (2018). Faecalibacterium prausnitzii Produces Butyrate to Maintain Th17/Treg Balance and to Ameliorate Colorectal Colitis by Inhibiting Histone Deacetylase 1. Inflamm. Bowel Dis. *24*, 1926–1940..

VITA

John Colgan began his academic career at Loyola University Chicago in 2017, starting a bachelor's degree in Biology. In the Summer of 2018 to the Summer of 2019 he worked in the Milanovich Lab as an undergraduate research assistant under Dr. Joseph Milanovich PhD. During his tenure in the Milanovich lab, Mr. Colgan contributed to several projects including a PGA funded project examining the health and biodiversity of ponds in the Chicagoland area as well as taking part in field research at the Great Smoky Mountains National Park, examining the recovery of Salamander populations following the 2016 wildfires. Mr. Colgan also contributed to "Draft Genome Sequences of Six *Lactobacillus gasseri* and Three *Lactobacillus paragasseri* Strains Isolated from Female Bladder" (2019) as part of an engaged learning course taught by Dr. Catherine Putonti PhD. In the Fall of 2019 he added bioinformatics as a second major. In the summer of 2020, he was accepted into the Loyola University Chicago bioinformatics master's program, and began his work on his thesis "Interplay Between Human Gut Microbiota and Celiac Disease " under the supervision of Dr. Michael Burns PhD. Mr. Colgan graduated Magna Cum Laude with his bachelors of arts and sciences Biology. He also was an honors graduate in Bioinformatics, as well as a 7 time Dean's list laureate. Mr. Colgan will receive his Masters of Science in Bioinformatics pending successful defense of his thesis. Following his defense, Mr. Colgan will begin his PhD studying at the University of Chicago in the Committee for Molecular Metabolism and Nutrition.