IBD Meta-analysis

Taylor Reiter Luiz Irber . . . Phillip Brooks Alicia Gingrich
C. Titus Brown

July 9, 2020

Introduction

Metagenomics captures the functional potential of microbial communities through DNA sequencing of genes and organisms. Metagenomics has been used to profile many human microbial communities, including those that change in or contribute to disease. In particular, human gut microbiomes have been extensively characterized for their potential role in diseases such as obesity (Greenblum, Turnbaugh, and Borenstein 2012), type II diabetes (Qin et al. 2012), colorectal cancer (Wirbel et al. 2019), and inflammatory bowel disease (Lloyd-Price et al. 2019; Morgan et al. 2012; Hall et al. 2017; Franzosa et al. 2019). Inflammatory bowel disease (IBD) refers to a spectrum of diseases characterized by chronic inflammation of the intestines and is likely caused by host-mediated inflammatory responses at least in part elicited by microorganisms (Kostic, Xavier, and Gevers 2014). However, no causative or consistent microbial signature has been associated with IBD to date.

Statements about biology, determined once computation is all done

Although there is no consistent taxonomic or functional trend in the gut microbiome associated with IBD diagnosis, metagenomic studies conducted unto this point have left substantial portions of data unanalyzed. Reference-based pipelines commonly used to analyze metagenomic data from IBD cohorts such as HUMANn2 characterize on average 31%-60% of reads from the human gut microbiome metagenome, as many reads do not closely match sequences in reference databases (Franzosa et al. 2014; Lloyd-Price et al. 2019). To combat this issue, reference-free approaches like *de novo* assembly and binning are used to generate metagenome-assembled genome bins (MAGs) that represent species-level composites of closely related organisms in a sample. However, *de novo* approaches fail when there is low-coverage of or high strain variation in gut microbes, or with sequencing error (Olson et al. 2017). Even when performed on a massive scale, an average of 12.5% of reads fail to map to all *de novo* assembled organisms from human microbiomes (Pasolli et al. 2019).

Here we perform a meta-analysis of six studies of IBD gut metagenome cohorts comprising 260 CD, 132 UC and 213 healthy controls (see **Table 1**) (Lloyd-Price et al. 2019; Lewis et al. 2015; Hall et al. 2017; Franzosa et al. 2019; Gevers et al. 2014; Qin et al. 2010). First, we re-analyzed each study using a consistent k-mer-based, reference-free approach. We demonstrate that diagnosis accounts for a small but significant amount of variation between samples. Next, we used random forests to predict IBD diagnosis and to determine the k-mers that are predictive of UC and CD. Then, we use compact de Bruijn graph queries to reassociate k-mers with sequence context and perform taxonomic and functional characterization of these sequence neighborhoods. We find that strain variation is important (ADD MORE HERE AFTER CORNCOB). Our analysis pipeline is lightweight and is extensible to other association studies in large metagenome sequencing cohorts.

Results

Table 1: Six IBD cohorts used in this meta-analysis.

Cohort	**Cohort names**	**Country**	**Total**	**CD**	**UC**	**nonIBD**	**Reference**
iHMP	IBDMDB	USA	106	50	30	26	(Lloyd-Price et al. 2019)
PRJEB2054	MetaHIT	Denmark, Spain	124	4	21	99	(Qin et al. 2010)
SRP057027	NA	Canada, USA	112	87	0	25	(Lewis et al. 2015)
PRJNA385949	PRISM, STiNKi	USA	17	9	5	3	(Hall et al. 2017)
PRJNA400072	PRISM, LLDeep, and NLIBD	USA, Netherlands	218	87	76	55	(Franzosa et al. 2019)
PRJNA237362	RISK	North America	28	23	0	5	(Gevers et al. 2014)
Total			605	260	132	213	

Annotation-free approach for meta-analysis of IBD metagenomes.

Given that both reference-based and *de novo* methods suffer from substantial and biased loss of information in the analysis of metagenomes (Thomas and Segata 2019; Breitwieser, Lu, and Salzberg 2019), we sought a reference- and assembly-free pipeline to fully characterize each sample (**Figure 1**). K-mers, words of length k in nucleotide sequences, have previously been exploited for annotation-free characterization of sequencing data (reviewed by Rowe (2019)). K-mers are suitable for metagenome analysis because they do not need to be present in reference databases to be included in analysis and because they capture information from reads even when there is low coverage or high strain variation that preclude assembly. In particular, scaled MinHash sketching produces compressed representations of k-mers in a metagenome while retaining the sequence diversity in a sample (Pierce et al. 2019). Importantly, this approach creates a consistent set of hashes across samples by retaining the same hashes when the same k-mers are observed. This enables comparisons between metagenomes. Given these attributes, we use scaled MinHash sketches to perform metagenome-wide k-mer association with IBD-subtype. We refer to the scale MinHash sketches as *signature*, and to each subsampled k-mer in a signature as a *hash*.

We also implemented a consistent preprocessing pipeline to remove erroneous sequences that could falsely deflate similarity between samples. We removed adapters, human DNA, and erroneous k-mers, and filtered signatures to retain hashes that were present in multiple signatures. These preprocessing steps removed hashes that were likely to be errors while keeping hashes that were real but of low abundance in some signatures. 7,376,151 hashes remained after preprocessing and filtering.

K-mers capture variation due to disease subtype

In this study, we aimed to identify microbial signatures associated with IBD. However, given that biological and technical artifacts can differ greatly between metagenome studies (Wirbel et al. 2019), we first quantified these sources of variation. We calculated pairwise distance matrices using jaccard distance and cosine distance between filtered signatures, where jaccard distance captured sample richness and cosine distance captured sample diversity. We performed principle coordinate analysis and PERMANOVA with these distance matrices (**Figure 2**), using the variables study accession, diagnosis, library size, and number of hashes in a filtered signature (**Table 2**). Number

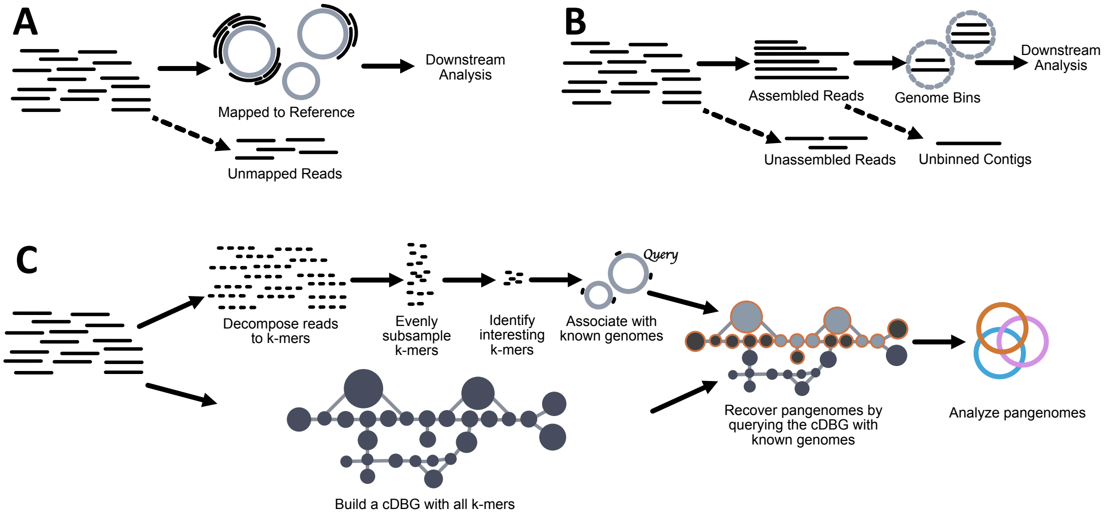

Figure 1: Comparison of common metagenome analysis techniques with the method used in this paper. Metagenomes consist of short (~50-300 bp) reads derived from sequencing DNA from environmental samples. **A** Reference-based metagenomic analysis. Reads are compared to genomes, genes, or proteins in reference databases to determine the presence and abundance of organisms and proteins in a sample. Unmapped reads are typically discarded from downstream analysis. **B** *De novo* metagenome analysis. Overlapping reads are assembled into longer contiguous seqeunces (~500bp-150kbp, (Vollmers, Wiegand, and Kaster 2017)) and binned into metagenome-assembled genome bins. Bins are analyzed for taxonomy, abundance, and gene content. Reads that fail to assemble and contigs that fail to bin are usually discarded from downstream analysis. **C** Annotation-free approach for meta-analysis of metagenomes. We decompose reads into k-mers and subsample these k-mers, selecting k-mers that evenly represent the sequence diversity within a sample. We then identify interesting k-mers using random forests, and associate these k-mers with genomes in reference databases. Meanwhile, we construct a compact de Bruijn graph (cDBG) that contains all k-mers from a metagenome. We query this graph with known genomes that contain our interesting k-mers to recover sequence diversity nearby our query sequences in the cDBG. In the colored cDBG, light grey nodes indicate nodes that contain at least one identical k-mer to the query, while nodes outlined in orange indicate the nearby sequences recovered via cDBG queries. The combination of all orange nodes produces a sample-specific pangenome that represents the strain variation of closely-related organisms within a single metagenome. We repeat this process for all metagenomes and generate a single pangenome depicted in orange, blue, and pink.

of hashes in a filtered signature accounts for the highest variation, possibly reflecting reduced diversity in stool metagenomes of CD and UC patients (reviewed in (Schirmer et al. 2019)). Study accounts for the second highest variation, emphasizing that technical artifacts can introduce biases with strong signals. Diagnosis accounts for a similar amount of variation as study, demonstrating that there is a small but detectable signal of IBD subtype in stool metagenomes.

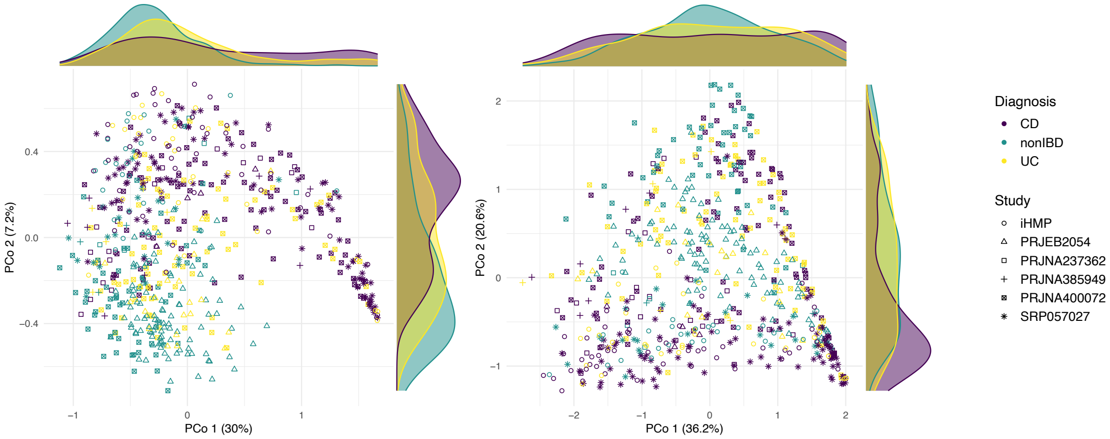

Figure 2: Principle coordinate analysis of metagenomes from IBD cohorts performed on filtered signatures. **A** Jaccard distance. **B** Angular distance.

Table 2: Results from PERMANOVA performed on Jaccard and Angular distance matrices. Number of hashes refers to the number of hashes in the filtered signature, while library size refers to the number of raw reads per sample. * denotes p < .001.

Variable	**Jaccard distance**	**Angular distance**
Number of hashes	9.9%*	6.2%*
Study accession	6.6%*	13.5%*
Diagnosis	6.2%*	3.3%*
Library size	0.009%*	0.01%*

Given that number of hashes in a filtered signature accounted for the highest source of variation, we sought to understand whether this reflected reduced diversity in stool metagenome of CD and UC patients. We created a diversity metric by dividing the number of hashes in a filtered signature by the total number of observed hashes across all samples. We observed that k-mer diversity in CD and UC is lower than in non-IBD, in concordance with similar findings from many sequencing studies (CITATIONS).

Hashes are weakly predictive of IBD subtype

To evaluate whether the variation captured by diagnosis is predictive of IBD disease subtype, we built random forests classifiers to predict CD, UC, or non-IBD. We selected random forests because of the interpretability of feature importance via variable importance measurments. We used a leave-one-study-out cross-validation approach where we built and optimized a classifier using five cohorts and validated on the sixth.

Given the high-dimensional structure of this dataset (e.g. many more hashes than samples), we first used the vita method to select predictive hashes in the training set (Janitza, Celik, and Boulesteix 2018; Degenhardt, Seifert, and Szymczak 2017). Vita variable selection is based on permuation of variable importance, where p-values for variable importance are calculated against a null distribution that is built from variables that are estimated as non-important (Janitza, Celik, and Boulesteix 2018). This approach retains important variables that are correlated (Janitza, Celik, and Boulesteix 2018; Seifert, Gundlach, and Szymczak 2019), which is desirable in omics-settings where correlated features are often involved in a coordinated biological response, e.g. part of the same operon, pathways, or genome (Stuart et al. 2003; Sabatti et al. 2002). Variable selection reduced the number of hashes used in each model to 29,264-41,701 (**Table 3**). Using this reduced set of hashes, we then optimized each random forests classifier on the training set, producing six optimized models. We validated each model on the left-out study. The accuracy on the validation studies ranged from 49.1%-75.9% (**Figure 3**), outperforming a previously published model built on metagenomic data alone (Franzosa et al. 2019).

Table 3: Number of hashes retained after Vita variable selection for each of 6 classifiers. Classifiers are labelled by the validation study that was held out from training.

Validation study	**Selected hashes**
iHMP	39628
PRJEB2054	35343
PRJNA237362	40726
PRJNA385949	41701
PRJNA400072	32578
SRP057027	29264

We next sought to understand whether there was a consistent biological signal captured among classifiers by evaluating the fraction of shared hashes selected by variable selection between models. We intersected each set of hashes used to build each optimized classifier (**Figure 3**). Nine hundred thrity two hashes were shared between all classifiers, while 3,859 hashes were shared between at least five studies. The presence of shared hashes between classifiers indicates that there is a weak but consistent biological signal for IBD subtype between cohorts.

Shared hashes accounted for 2.8% of all hashes used to build the optimized classifiers. If shared hashes are predictive of IBD subtype, we would expect that these hashes would account for an outsized proportion of variable importance in the optimized classifiers. To calculate the relative variable importance contributed by each hash, we first normalized the variable importance values within each classifier by dividing by the total variable importance (e.g. sum to 1 within each classifier). We then normalized the variable importance across all classifiers by dividing by the total number of classifiers (e.g. divided by six so the total variable importance of all hashes across all classifiers summed to 1). 40.2% of the total variable importance was held by the 3,859 hashes shared between at least five classifiers, with 21.5% attributable to the 932 hashes shared between all six classifiers. This indicates that shared hashes contribute a large fraction of predictive power for classification of IBD subtype.

Some predictive hashes anchor to known genomes

We next evaluated the identity of the predictive hashes in each classifier. We first compared the predictive hashes against sequences in reference databases. We used sourmash `gather` to anchor predictive hashes to known genomes (Pierce et al. 2019). We compared our predictive hashes against all microbial genomes in GenBank, as well as metagenome-assembled genomes from three recent *de novo* assembly efforts from human microbiome metagenomes (Pasolli et al. 2019; Nayfach et al. 2019; Almeida et al. 2019). Between 75.1-80.3% of hashes anchored to 1,161 genomes (**Figure 4**). This indicates that 19.7-24.9% of hashes that are predictive of IBD subtype represent sequences not in reference databases.

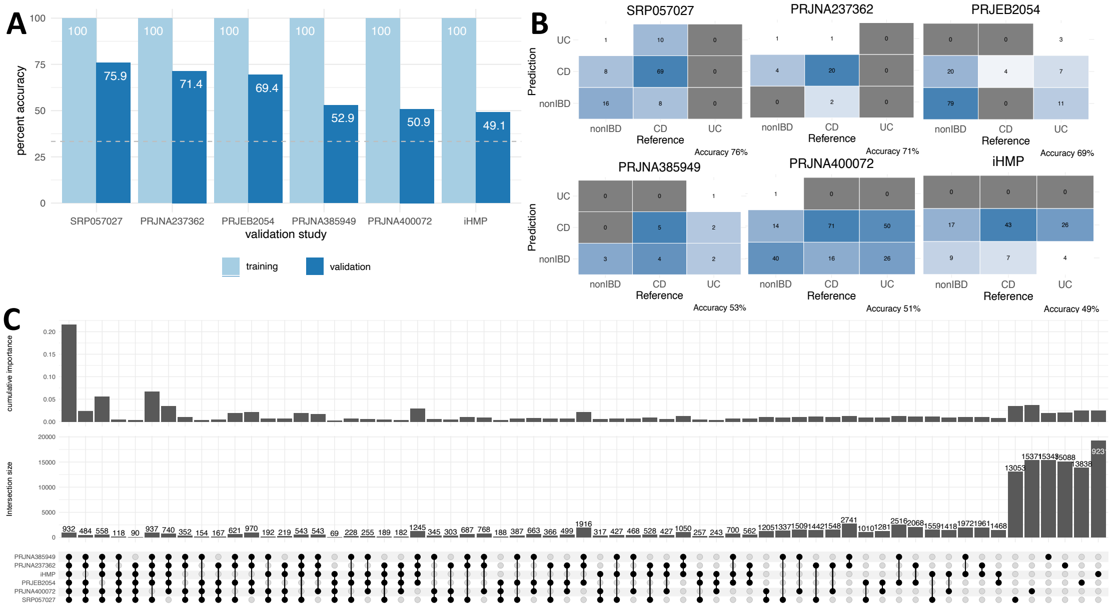

Figure 3: Random forest classifiers weakly predict IBD subtype. **A** Accuracy of leave-one-study-out random forest classifiers on training and validation sets. The validation study is on the x axis. **B** Confusion matrices depicting performance of each leave-one-study-out random forest classifier on the validation set. **C** Upset plot depicting intersections of sets of hashes as well as the cumulative normalized variable importance of those hashes in the optimized random forest classifiers. Each classifier is labelled by the left-out validation study.

The 3,859 hashes shared between at least five classifiers anchored to only 41 genomes (**Figure 4**). Futher, these 41 genomes accounted for 50.5% of the total variable importance, a 10.3% increase over the hashes alone. This means that tehse genomes contain additional predictive hashes not shared between at least five classifiers.

In contrast to all hashes, only 69.4% of these hashes were identifiable among the 3,859 shared hashes, a decrease of 5.7-10.9%. This indicates that hashes that are more likely to be important for IBD subtype classification are less likely to be anchored to genomes in reference databases.

Using sourmash lca classify to assign GTDB taxonomy, we find 38 species represented among the 41 genomes. The genome that anchors the most variable importance is **Acetatifactor sp900066565**. (Add %phyla/etc? Is it even worth analyzing these that much when everything changes after spacegraphcats?) However, we observe that while most genomes assign to one species, 19 assign to an additional one or more distantly related genomes that likely represent contamination from the assembly and binning process. When we take the Jaccard index of these 41 genomes, we observe little similarity despite contamination (**Figure 4**). Therefore, we proceeded with analysis with the idea that each of the 41 genomes is a self-contained entity that captures distinct biology.

Unknown but predictive hashes represent novel pangenomic elements

Given that 30.6% of hashes shared between at least five classifiers did not anchor to genomes in databases, we next sought to characterize these hashes. We reasoned that many unknown but predictive hashes likely originate from closely related strain variants of identified genomes and sought to recover these variants. We performed compact de Bruijn graph queries into each metagenome sample with the 41 genomes that contained predictive hashes (CITATION: SPACEGRAPHCATS). This produced pangenome neighborhoods for each of the 41 genomes.

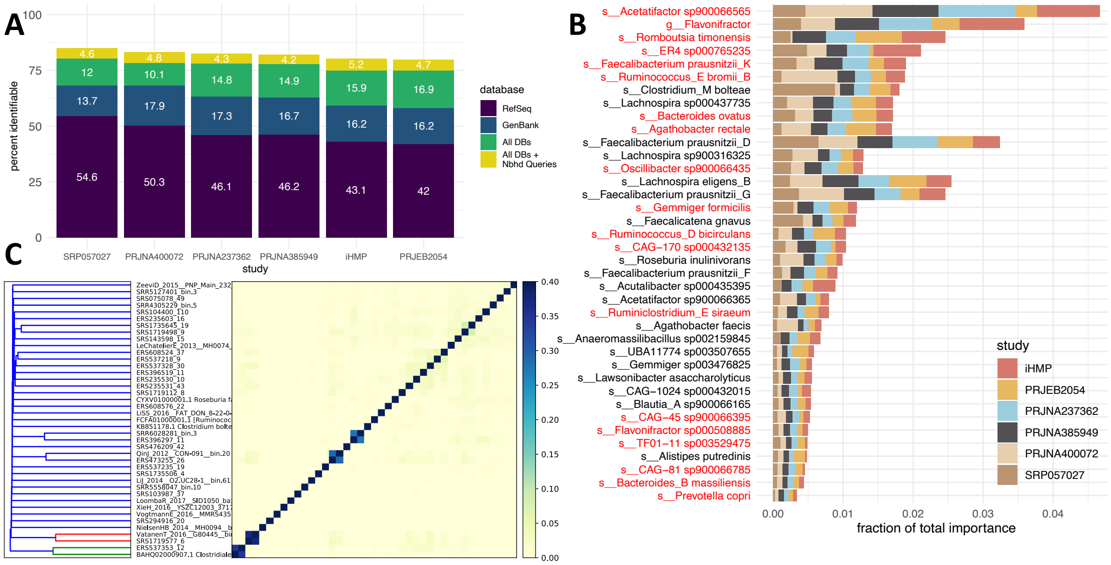

Figure 4: Some predictive hashes from random forest classifiers anchor to known genomes. **A** 75.1-80.3% of all hashes used to train classifiers anchor to known genomes in RefSeq, GenBank, or human microbiome metagenome-assembled genome databases. A further 4.2-5.6% of hashes anchor to pangenomes of a subset of these genomes. **B** The 3,859 hashes shared between at least five classifiers anchor to 41 genomes. Genomes account for different amounts of variable importance in each model. Genomes are labelled by 38 GTDB taxonomy assignments. Genomes labelled in red were classified as multiple distantly related species, likely indicating contamination. **C** Jaccard similarity between 41 genomes. The highest similarity between genomes is 0.37 and is shared by genomes of the same species, while most genomes have no similarity. This indicates that each genome represents distinct nucleotide sequence.

86.1% of hashes shared between at least five classifiers were in the pangenomes of the 41 genomes, a 16.7% increase over the 41 genomes alone. This suggests that at least 16.7% of shared hashes originate from novel strain-variable or accessory elements in pangenomes.

Further, these pangenomes captured an additional 4.2-5.2% of all predictive hashes from each classifier, indicating that pangenomes contain novel sequences not captured in any database (**Figure 4**). The pangenomes also captured 74.5% of all variable importance, a 24% increase over the 41 genomes alone. This indicates that pangenomic variation contributes substantial predictive power toward IBD subtype classification.

Pangenomic neighborhood queries disproportionately impact the variable importance attributable to specific genomes (**Figure 5**). While most genomes maintained a similar proportion of importance with or without pangenome queries, three pangenomes shifted dramatically. While an *Acetatifactor* species anchored the most importance prior to pangenome construction, the specific species of *Acetatifactor* switched from *sp900066565*, to *sp900066365*. Conversely, *Faecalibacterium prausnitzii_ D* increased from anchoring ~2.9% to ~10.5% of the total variable importance. These results indicate that strain variation is more important and less characterizable for prediction of IBD subtype in some species that in others.

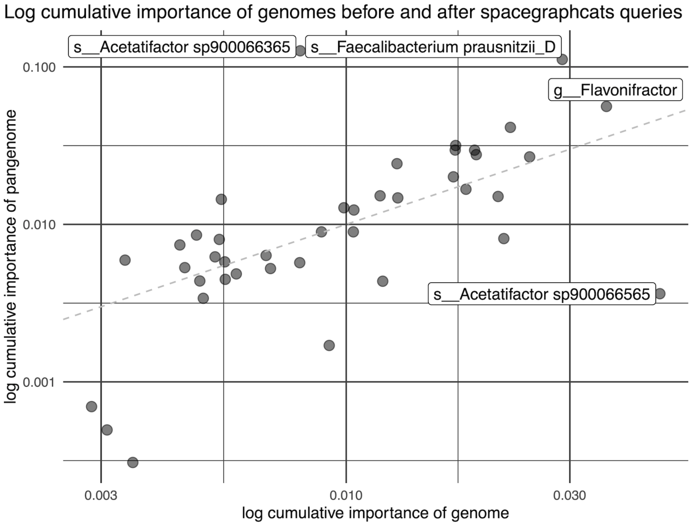

Figure 5: Pangenome neighborhoods generated with cDBG queries recover strain variation that is important for predicting IBD subtype. While the variable importance attributable to some genomes does not change with cDBG queries, other genomes increase by more than 7%.

Per sample diversity is reduced in IBD

We next sought to understand the functional potential of the recovered pangenomes. To build a composite pangenome for each of our 41 query genomes, we assembled each query neighborhood individually, extracted open reading frames using prokka, and then clustered genes and gene fragements at 90% identity. Pangenomes ranged in size from 4,661-29,571 (mean = 15,897, sd = 6,991) representative gene sequences. The smallest pangenome belonged to *Romboutsia timonensis*, for which isolates from the same genus have 2,852-3,662 coding domain sequences (Gerritsen et al. 2019). The largest pangenome was *Faecalibacterium prausnitzii*, a ubiquitous member of the human gut microbiome with high genome plasticity (Fitzgerald et al. 2018).

We next investigated the number of ORFs observed in each pangenome within each sample. The mean number of ORFs observed in the pangenome from a single sample was lower than nonIBD for thirty-nine of 41 pangenomes for CD and 37 of 41 pangenomes for UC (ANOVA $p < .05$, Tukey's HSD $p < .05$).

Only the pangenome of *Clostridium bolteae* had a higher mean number of observed ORFs per sample in CD than nonIBD. *C. bolteae* is a virulent and opportunistic bacteria detected in the human gut microbiome that is more abundant in diseased than healthy guts (Finegold et al. 2005; Lozupone et al. 2012). *C. bolteae* is associated with disturbance succession in which the stable gut consortia is compromised (Lozupone et al. 2012), and has increased gene expression during gut dysbiosis (Lloyd-Price et al. 2019).

Given these associations, we performed differential abundance analysis on the *C. bolteae* pangenome between CD and nonIBD. We compared our results against study of virulence-causing gene in *C. bolteae* (Lozupone et al. 2012), and find that 24 of 41 previously identified orthologs are significantly induced in CD. Seven of these orthologs are associated with response to oxidative stress. (OXIDATIVE STRESS IBD BIO TIE IN).

We then performed gene enrichment analysis on the differentially abundant genes with KEGG ortholog annotations in *C. bolteae.* While many KEGG pathways are significant, flagellar assembly had the second lowest p value (17 genes). Bacterial flagellin is a dominant antigen in Crohn's disease but not ulcerative colitis (Lodes et al. 2004; Duck et al. 2007).

Only *Faecalicatena gnavus* (*Ruminococcus gnavus* in NCBI taxonomy) showed no difference in the mean number of genes per sample between CD and nonIBD and UC and nonIBD. *F. gnavus* is an aerotolerant anaerobe, one clade of which has only been found in the guts of IBD patients (Hall et al. 2017). *F. gnavus* also produces an inflammatory polysaccharide that induces TNFa secretion in a response mediated by toll-like receptor 4 (Henke et al. 2019). We performed differential abundance analysis between CD and nonIBD as well as UC and non IBD to understand whether the pangenome, but not pangenome diversity, varied between disease states. While 5,984 genes were differentially abundant in CD, only 197 were less abundant in UC. This suggests that *F. gnavus* is different from nonIBD in CD alone.

We next investigated whether the gene cluster thought to be involved in biosynthesis of the inflammatory polysaccharide was significantly induced in CD. We identified 19 of 23 ORFs in the *F. gnavus* pangenome that matched the putative genes in the cluster, all of which were more abundant in CD. Further, two subsets, one containing 5 ORFs and one contain 7, were co-located on two contiguous sequences, indicating these genes do form a cluster. We then investigated whether this gene cluster was present in non-IBD samples, and found an average of more than 100 reads that mapped per gene in the cluster in 10 of 213 nonIBD metagenomes. This indicates that while more abundant in CD, it is also identifiable within healthy human gut microbiomes.

We also genes involved in oxidative stress resistance that are more abundant in CD. This includes one super oxide dismutase and five NADH oxidases.

While this evidence supports the idea that *F. gnavus* is harmful in CD, we see some genes that are more abundant in CD that are beneficial for gut health. For example, we find 10 a-L-fucosidases. Tryptophan metabolism. ?

In three pangenomes, we see a higher mean number of genes observed per sample for UC than CD or nonIBD. These include *R. timonensis*, *Anaeromassilibacillus*, and *Actulibacter*. ?

No evidence of disease-specific pangenome

Given that the mean number of proteins observed for most pangenomes in each sample was overwhelming lower for CD and UC, we wanted to know whther there was a disease-specific pangenome for each organism. Using count matrices detailing the number of reads that mapped to each gene from each sample, we generated gene accumulation curves. We find for most genomes, the majority of genes are observed in CD, UC, and nonIBD, suggesting there is no disease-specific pangenome. This in part explains heterogenous study findings in IBD gut microbiome investigations (CITATIONS) and underscores that IBD is a spectrum of diseases characterized by intermittent health and dysbiosis.

There are notable excpetions to this trend. One of 41 pangenome accumulation curves did not saturate for UC, while 10 did not saturate for CD. *C. bolteae* does not saturate in UC. One hundred seventy-one of 16,822 genes were not observed in UC, many of which had no annotated function.

Ten of 41 pangenome accumulation curves did not saturate for CD samples. On average, 366 genes were unobserved in CD. The largest number of unobserved genes was 2,089 in CAG-1024 pangenome.

Accessory elements, species abundance, and different strains contribute to disease-specific microbiome

The lower number of genes observed in individual samples could be driven by lower abundance of the organism in the sample, by fewer strains present in the sample, or fewer accessory elements in the pangenomes of strains that are present. Given this, we next sought to understand the source of differences in teh gene content of the gut microbiomes in CD and UC. We performed differential abundance analysis between CD and nonIBD and UC and nonIBD for all pangenomes. We then searched for the presence of marker genes. We reasoned that if we identified no marker genes among differentially abundant genes, accessory elements were responsible for disease-specific signatures. Conversely, if we identified marker genes among significantly different genes, then the abudnance of an organism likely differed. Lastly, if marker genes were both more and less abundant, different strains were likely present in CD or UC.

We find almost no marker genes in any pangenome among genes that are more abundant in UC. However, we see the presence of marker genes in less abundant genes for three organisms, including *Gemminger formicilis*. This indicates that while some organisms are less abundant in UC, differences in non-marker genes, e.g. accessory elements, are a greater source of differentiation.

Conversely, two pangeomes contained marker genes that were more abundant in CD: *C. bolteae* and *F. gnavus*. For *C. bolteae*, 95% of marker genes were detected among more abundant genes, while 23% of marker genes were detected among less abundant genes. This suggests that *C. bolteae* is more abundant in CD. For *F. gnavus*, 60% of marker genes were detected among more abundant genes, while 68% were detected in less abundant genes. This suggests that a different strain of *F. gnavus* is more abundant in CD, which matches previous findings from gut microbiome metagenome investigations (Hall et al. 2017).

For thirty-one of 41 pangenomes, the majority of marker genes were significantly less abundant in CD, indicating that these organisms are less abundant in CD. However, for 10 pangenomes, we detect few marker genes in significantly less abundant genes. This includes *Prevotella copri*, *Bacteroidees massilensis*, *Bacteroides ovatus*, and two organisms from the genus *Flavonifractor*. Differences in these pangenomes are likely attributable to accessory elements. **ARE THERE CONFLICTING REPORTS ON THE GOOD/BAD OF THESE ORGS? COULD MAKE SENSE IF DRIVEN BY STRAIN VARIATION**

Operons in differentially abundant genes (tmp title)

Given that all genes detected from the *F. gnavus* inflammatory polysaccharide biosynthetic gene cluster were significantly induced in CD, and that subsets of these sequences were colocated on single contiguous sequences, we reasoned that other biologically meaningful genes were likely to occur in clusters. Using results from differential abundance analysis, we searched for gene clusters of five or more genes. We selected five as a signal:noise compromise, as five was the smallest consecutive stretch detected in the *F. gnavus* cluster.

We find no evidence of gene clusters that are more abundant in UC. Conversely, we find many gene clusters in XX pangenomes that are more abundant in CD. XXX

Random forests on genes (tmp title)

While our k-mer based random forest models were weakly predictive of diagnosis, the contents of the pangenomes of the 41 most predictive organisms uncovered interesting associations that were not apparent from k-mers alone. While k-mers allow us to look at all of the data and compare against tall known genome, they are brittle to evolutionary distance and don't recapitulate function. Given this, we built new random forest classifiers using the same leave-one-study-out design and using gene counts from the 41 pangenomes as predictive variables.

Discussion

We present XXX.

In this investigation, we find that gut microbiomes from both UC and CD suffer from stochastic loss of diversity.

While *C. bolteae* and *R. gnavus* emerge as bad actors in the pathophysiology of CD, no similar signal is detected for UC. This suggests that while both diseases are associated with lower diversity, CD is uniquely exacerbated by microbes that become more abundant during disease.

Methods

All code associated with our analyses is available at www.github.com/dib-lab/2020-ibd/

IBD metagenome data acquisition and processing

We searched the NCBI Sequence Read Archive and BioProject databases for shotgun metagenome studies that sequenced fecal samples from humans with Crohn's disease, ulcerative colitis, and healthy controls. We included studies sequenced on Illumina platforms with paired-end chemistries and with sample libraries that contained greater than one million reads. For time series intervention cohorts, we selected the first time point to ensure all metagenomes came from treatment-naive subjects.

We downloaded metagenomic fastq files from the European Nucleotide Archive using the "fastq_ftp" link and concatenated fastq files annotated as the same library into single files. We also downloaded iHMP samples from idbmdb.org. We used Trimmomatic (version 0.39) to adapter trim reads using all default Trimmomatic paired-end adapter sequences (`ILLUMINACLIP:{inputs/adapters.fa}:2:0:15`) and lightly quality-trimmed the reads (`MINLEN:31 LEADING:2 TRAILING:2 SLIDINGWINDOW:4:2`) (Bolger, Lohse, and Usadel 2014). We then removed human DNA using BBMap and a masked version of hg19 (Bushnell 2014). Next, we trimmed low-abundance k-mers from sequences with high coverage using khmer's `trim-low-abund.py` (Crusoe et al. 2015).

Using these trimmed reads, we generated scaled MinHash signatures for each library using sourmash (k-size 31, scaled 2000, abundance tracking on) (Brown and Irber 2016). At a scaled value of 2000, an average of one k-mer will be detected in each 2000 base pair window, and 99.8% of 10,000 base pair windows will have at least one k-mer representative. We selected a k-mer size of 31 because of its species-level specificity (Koslicki and Falush 2016). A signature is composed of hashes, where each hash represents a k-mer contained in the original sequence. We retained all hashes that were present in multiple samples, and refer to these as filtered signatures.

Principle Coordinates Analysis

We used jaccard distance and cosine distance implemented in `sourmash compare` to pairwise compare filtered signatures. We then used the `dist()` function in base R to compute distance matrices. We used the `cmdscale()` function to perform principle coordinate analysis (Gower 1966). We used ggplot2 and ggMarginal to visualize the principle coordinate analysis (Wickham et al. 2019). To test for sources of variation in these distance matrices, we performed PERMANOVA using the `adonis` function in the R vegan package (Oksanen et al. 2010). The PERMANOVA was modeled as `~ diagnosis + study accession + library size + number of hashes`.

Random forest classification

We built a random forest classifier to predict CD, UC, and non-IBD status using filtered signatures. First, we transformed sourmash signatures into a hash abundance table where each metagenome was a sample, each hash was a feature, and abundances were recorded for each hash for each sample. We normalized abundances by dividing by the total number of hashes in each filtered signature. We then used a leave-one-study-out validation approach where we trained six models, each of which was trained on five studies and validated on the sixth. To build each model, we first performed vita variable selection on the training set as implemented in the Pomona and ranger packages (Degenhardt, Seifert, and Szymczak 2017; Wright and Ziegler 2015). Vita variable selection reduces the number of variables (e.g. hashes) to a smaller set of predictive variables through selection of variables with high cross-validated permutation variable importace (Janitza, Celik, and Boulesteix 2018). Using this smaller set of hashes, we then built an optimized random forest model using tuneRanger (Probst, Wright, and Boulesteix 2019). We evaluated each validation set using the optimal model, and extracted variable importance measures for each hash for subsequent analysis. To make variable importance measures comparable across models, we normalized importance to 1 by dividing variable importance by the total number of hashes in a model and the total number of models.

Characterization of predictive k-mers

We used sourmash `gather` with parameters `k 31` and `--scaled 2000` to anchor predictive hashes to known genomes (Brown and Irber 2016). Sourmash `gather` searches a database of known k-mers for matches with a query (Pierce et al. 2019). We used the sourmash GenBank database (2018.03.29, https://osf.io/snphy/), and built three additional databases from medium- and high-quality metagenome-assembled genomes from three human microbiome metagenome reanalysis efforts (https://osf.io/hza89/) (Pasolli et al. 2019; Nayfach et al. 2019; Almeida et al. 2019). In total, approximately 420,000 microbial genomes and metagenome-assembled genomes were represented by these four databases. We used the sourmash `lca` commands against the GTDB taxonomy database to taxonomically classify the genomes that contained predictive hashes. To calculate the cumulative variable importance attributable to a single genome, we used an iterative winner-takes-all approach. The genome with the largest fraction of predictive k-mers won the variable importance for all hashes contained within its genome. These hashes were then removed, and we repeated the process for the genome with the next largest fraction of predictive k-mers.

To identify hashes that were predictive in at least five of six models, we took the union of predictive hashes from all combinations of five models, as well as from the union of all six models. We refer to these hashes as shared predictive hashes. We anchored variable importance of these shared predictive hashes to known genomes using sourmash `gather` as above.

Compact de Bruijn graph queries for predictive genomes

We used spacegraphcats `search` to retreive k-mers in the compact de Bruijn graph neighborhood of the genomes that matched predictive k-mers (CITATION). We then used spacegraphcats `extract_reads` to retreive the reads and `extract_contigs` to retreive unitigs in the compact de Bruijn graph that contained those k-mers, respectively.

Characterization of graph pangenomes

Pangenome signatures To evaluate the k-mers recovered by pangenome neighborhood queries, we generated sourmash signatures from the unitigs in each query neighborhood. We merged signatures from the same query genome, producing 41 pangenome signatures. We indexed these signatures to create a sourmash gather database. To estimate how query neighborhoods increased the identifiable fraction of predictive hashes, we ran sourmash `gather` with the pangenome database, as well as the GenBank and human microbiome metagenome databases. To estimate how query neighborhoods increased the identifiable fraction of shared predictive hashes, we ran sourmash `gather` with the pangenome database alone. We anchored variable importance of the shared predictive hashes to known genomes using sourmash `gather` results as above.

Differential abundance We used differential abundance analysis to determine which protein sequences in each pangenome were differentially abundant in IBD subtype. We used diginorm on each spacegraphcats query neighborhood implemented in khmer as `normalize-by-median.py` with parameters `-k 20 -C 20` (Crusoe et al. 2015). We then assembled each neighborhood from a single query with `megahit` using default parameters [CITATION:megahit], and annotated each assembly using prokka [CITATION: prokka]. We used CD-HIT to cluster nucleotide sequences within a pangenome at 90% identity and retained the representative sequence (Fu et al. 2012). We used Salmon to quantify the number of reads aligned to each representative gene sequence [CITATION:salmon]. Using these abundances, we used the R package corncob to perform differential abundance analysis between IBD subtype, using the likelihood ratio test with the formula `study_accession + diagnosis` and the null formula `study_accession` (Martin et al. 2020). We considered genes with p values $< .05$ after bonferonni correction as statistically significant.

Annotation of differentially abundant proteins We used EggNog to annotate the representative sequences in each pangenome [CITATION:eggnog]. We performed enrichment analysis using the R package clusterProfiler (Yu et al. 2012).

References

Almeida, Alexandre, Alex L Mitchell, Miguel Boland, Samuel C Forster, Gregory B Gloor, Aleksandra Tarkowska, Trevor D Lawley, and Robert D Finn. 2019. "A New Genomic Blueprint of the Human Gut Microbiota." *Nature* 568 (7753): 499.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20.

Breitwieser, Florian P, Jennifer Lu, and Steven L Salzberg. 2019. "A Review of Methods and Databases for Metagenomic Classification and Assembly." *Briefings in Bioinformatics* 20 (4): 1125–36.

Brown, C Titus, and Luiz Irber. 2016. "Sourmash: A Library for Minhash Sketching of Dna." *J. Open Source Software* 1 (5): 27.

Bushnell, Brian. 2014. "BBMap: A Fast, Accurate, Splice-Aware Aligner." Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Crusoe, Michael R, Hussien F Alameldin, Sherine Awad, Elmar Boucher, Adam Caldwell, Reed Cartwright, Amanda Charbonneau, et al. 2015. "The Khmer Software Package: Enabling Efficient Nucleotide Sequence Analysis." *F1000Research* 4.

Degenhardt, Frauke, Stephan Seifert, and Silke Szymczak. 2017. "Evaluation of Variable Selection Methods for Random Forests and Omics Data Sets." *Briefings in Bioinformatics* 20 (2): 492–503.

Duck, Wayne L, Mark R Walter, Jan Novak, Denise Kelly, Maurizio Tomasi, Yingzi Cong, and Charles O Elson. 2007. "Isolation of Flagellated Bacteria Implicated in Crohn's Disease." *Inflammatory Bowel Diseases* 13 (10): 1191–1201.

Finegold, SM, Y Song, C Liu, DW Hecht, P Summanen, E Könönen, and SD Allen. 2005. "Clostridium Clostridioforme: A Mixture of Three Clinically Important Species." *European Journal of Clinical Microbiology and Infectious Diseases* 24 (5): 319–24.

Fitzgerald, Cormac Brian, Andrey N Shkoporov, Thomas DS Sutton, Andrei V Chaplin, Vimalkumar Velayudhan, R Paul Ross, and Colin Hill. 2018. "Comparative Analysis of Faecalibacterium Prausnitzii Genomes Shows a High Level of Genome Plasticity and Warrants Separation into New Species-Level Taxa." *BMC Genomics* 19 (1): 931.

Franzosa, Eric A, Xochitl C Morgan, Nicola Segata, Levi Waldron, Joshua Reyes, Ashlee M Earl, Georgia Giannoukos, et al. 2014. "Relating the Metatranscriptome and Metagenome of the Human Gut." *Proceedings of the National Academy of Sciences* 111 (22): E2329–E2338.

Franzosa, Eric A, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J Haiser, Stefan Reinker, Tommi Vatanen, et al. 2019. "Gut Microbiome Structure and Metabolic Activity in Inflammatory Bowel Disease." *Nature Microbiology* 4 (2): 293.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. "CD-Hit: Accelerated for Clustering the Next-Generation Sequencing Data." *Bioinformatics* 28 (23): 3150–2.

Gerritsen, Jacoline, Bastian Hornung, Jarmo Ritari, Lars Paulin, Ger T Rijkers, Peter J Schaap, Willem M De Vos, and Hauke Smidt. 2019. "A Comparative and Functional Genomics Analysis of the Genus Romboutsia Provides Insight into Adaptation to an Intestinal Lifestyle." *BioRxiv*, 845511.

Gevers, Dirk, Subra Kugathasan, Lee A Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, et al. 2014. "The Treatment-Naive Microbiome in New-Onset Crohn's Disease." *Cell Host & Microbe* 15 (3): 382–92.

Gower, John C. 1966. "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis." *Biometrika* 53 (3-4): 325–38.

Greenblum, Sharon, Peter J Turnbaugh, and Elhanan Borenstein. 2012. "Metagenomic Systems Biology of the Human Gut Microbiome Reveals Topological Shifts Associated with Obesity and Inflammatory Bowel Disease." *Proceedings of the National Academy of Sciences* 109 (2): 594–99.

Hall, Andrew Brantley, Moran Yassour, Jenny Sauk, Ashley Garner, Xiaofang Jiang, Timothy Arthur, Georgia K Lagoudas, et al. 2017. "A Novel Ruminococcus Gnavus Clade Enriched in Inflammatory Bowel Disease Patients." *Genome Medicine* 9 (1): 103.

Henke, Matthew T, Douglas J Kenny, Chelsi D Cassilly, Hera Vlamakis, Ramnik J Xavier, and Jon Clardy. 2019. "Ruminococcus Gnavus, a Member of the Human Gut Microbiome Associated with Crohn's Disease, Produces an Inflammatory Polysaccharide." *Proceedings of the National Academy of Sciences* 116 (26): 12672–7.

Janitza, Silke, Ender Celik, and Anne-Laure Boulesteix. 2018. "A Computationally Fast Variable Importance Test for Random Forests for High-Dimensional Data." *Advances in Data Analysis and Classification* 12 (4): 885–915.

Koslicki, David, and Daniel Falush. 2016. "MetaPalette: A K-Mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation." *MSystems* 1 (3): e00020–16.

Kostic, Aleksandar D, Ramnik J Xavier, and Dirk Gevers. 2014. "The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead." *Gastroenterology* 146 (6): 1489–99.

Lewis, James D, Eric Z Chen, Robert N Baldassano, Anthony R Otley, Anne M Griffiths, Dale Lee, Kyle Bittinger, et al. 2015. "Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease." *Cell Host & Microbe* 18 (4): 489–500.

Lloyd-Price, Jason, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, et al. 2019. "Multi-Omics of the Gut Microbial Ecosystem in Inflammatory Bowel Diseases." *Nature* 569 (7758): 655.

Lodes, Michael J, Yingzi Cong, Charles O Elson, Raodoh Mohamath, Carol J Landers, Stephan R Targan, Madeline Fort, Robert M Hershberg, and others. 2004. "Bacterial Flagellin Is a Dominant Antigen in Crohn Disease." *The Journal of Clinical Investigation* 113 (9): 1296–1306.

Lozupone, Catherine, Karoline Faust, Jeroen Raes, Jeremiah J Faith, Daniel N Frank, Jesse Zaneveld, Jeffrey I Gordon, and Rob Knight. 2012. "Identifying Genomic and Metabolic Features That Can Underlie Early Successional and Opportunistic Lifestyles of Human Gut Symbionts." *Genome Research* 22 (10): 1974–84.

Martin, Bryan D, Daniela Witten, Amy D Willis, and others. 2020. "Modeling Microbial Abundances and Dysbiosis with Beta-Binomial Regression." *Annals of Applied Statistics* 14 (1): 94–115.

Morgan, Xochitl C, Timothy L Tickle, Harry Sokol, Dirk Gevers, Kathryn L Devaney, Doyle V Ward, Joshua A Reyes, et al. 2012. "Dysfunction of the Intestinal Microbiome in Inflammatory Bowel Disease and Treatment." *Genome Biology* 13 (9): R79.

Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S Pollard, and Nikos C Kyrpides. 2019. "New Insights from Uncultivated Genomes of the Global Human Gut Microbiome." *Nature* 568 (7753): 505.

Oksanen, Jari, F Guillaume Blanchet, Roeland Kindt, Pierre Legendre, RB O'hara, Gavin L Simpson, Peter Solymos, M Henry H Stevens, and Helene Wagner. 2010. "Vegan: Community Ecology Package. R Package Version 1.17-4." *Http://Cran. R-Project. Org>. Acesso Em* 23: 2010.

Olson, Nathan D, Todd J Treangen, Christopher M Hill, Victoria Cepeda-Espinoza, Jay Ghurye, Sergey Koren, and Mihai Pop. 2017. "Metagenomic Assembly Through the Lens of Validation: Recent Advances in Assessing and Improving the Quality of Genomes Assembled from Metagenomes." *Briefings in Bioinformatics.*

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. "Extensive Unexplored Human Microbiome Diversity Revealed by over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle." *Cell* 176 (3): 649–62.

Pierce, N Tessa, Luiz Irber, Taylor Reiter, Phillip Brooks, and C Titus Brown. 2019. "Large-Scale Sequence Comparisons with Sourmash." *F1000Research* 8.

Probst, Philipp, Marvin N Wright, and Anne-Laure Boulesteix. 2019. "Hyperparameters and Tuning Strategies for Random Forest." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9 (3): e1301.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010. "A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing." *Nature* 464 (7285): 59.

Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, et al. 2012. "A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes." *Nature* 490 (7418): 55.

Rowe, Will PM. 2019. "When the Levee Breaks: A Practical Guide to Sketching Algorithms for Processing the Flood of Genomic Data." *Genome Biology* 20 (1): 199.

Sabatti, Chiara, Lars Rohlin, Min-Kyu Oh, and James C Liao. 2002. "Co-Expression Pattern from Dna Microarray Experiments as a Tool for Operon Prediction." *Nucleic Acids Research* 30 (13): 2886–93.

Schirmer, Melanie, Ashley Garner, Hera Vlamakis, and Ramnik J Xavier. 2019. "Microbial Genes and Pathways in Inflammatory Bowel Disease." *Nature Reviews Microbiology* 17 (8): 497–511.

Seifert, Stephan, Sven Gundlach, and Silke Szymczak. 2019. "Surrogate Minimal Depth as an Importance Measure for Variables in Random Forests." *Bioinformatics* 35 (19): 3663–71.

Stuart, Joshua M, Eran Segal, Daphne Koller, and Stuart K Kim. 2003. "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules." *Science* 302 (5643): 249–55.

Thomas, Andrew Maltez, and Nicola Segata. 2019. "Multiple Levels of the Unknown in Microbiome Research." *BMC Biology* 17 (1): 48.

Vollmers, John, Sandra Wiegand, and Anne-Kristin Kaster. 2017. "Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective-Not Only Size Matters!" *PloS One* 12 (1): e0169662.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686.

Wirbel, Jakob, Paul Theodor Pyl, Ece Kartal, Konrad Zych, Alireza Kashani, Alessio Milanese, Jonas S Fleck, et al. 2019. "Meta-Analysis of Fecal Metagenomes Reveals Global Microbial Signatures That Are Specific for Colorectal Cancer." *Nature Medicine* 25 (4): 679.

Wright, Marvin N, and Andreas Ziegler. 2015. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *arXiv Preprint arXiv:1508.04409*.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. "ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters." *Omics: A Journal of Integrative Biology* 16 (5): 284–87.