IBD Meta-analysis

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

13 September, 2020

1 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.

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. Using random forests, we determine that k-mers that are predictive of UC and CD originate from a core set of microbial genomes. We find that stochastic loss of diversity in this core set of microbial genomes is a hallmark of CD, and to a lesser extent, UC. While reduced diversity is responsible for the majority of disease signatures, multiple strains are enriched in disease. While we are more likely to find these strains in IBD metagenomes, these strains are present in low abundance in nonIBD as well. Our findings highlight the need for strain-level analysis of metagenomic data sets, and provide future avenues for research into IBD therapeutics.

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

Cohort	Name	Country	Total	CD	UC	nonIBD	Reference
iHMP	IBDMDB	USA	106	50	30	26	[@lloyd2019]
PRJEB2054	MetaHIT	Denmark, Spain	124	4	21	99	[@qin2010]
SRP057027	NA	Canada, USA	112	87	0	25	[@lewis2015]
PRJNA385949	PRISM, STiNKi	USA	17	9	5	3	[@hall2017]
PRJNA400072	PRISM, LLDeep, and NLIBD	USA, Netherlands	218	87	76	55	[@franzosa2019]
PRJNA237362	RISK	North America	28	23	0	5	[@gevers2014]
Total			605	260	132	213	

2 Results

2.1 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 gut metagenomes of IBD patients (**Figure 1**). K-mers, words of length k in nucleotide sequences, have previously been used 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 scaled MinHash sketches as *signatures*, and to each sub-sampled k-mer in a signature as a *hash*.

We also implemented a consistent preprocessing pipeline to remove sequencing errors 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 low abundance. In total, 7,376,151 hashes remained across all samples after preprocessing and filtering.

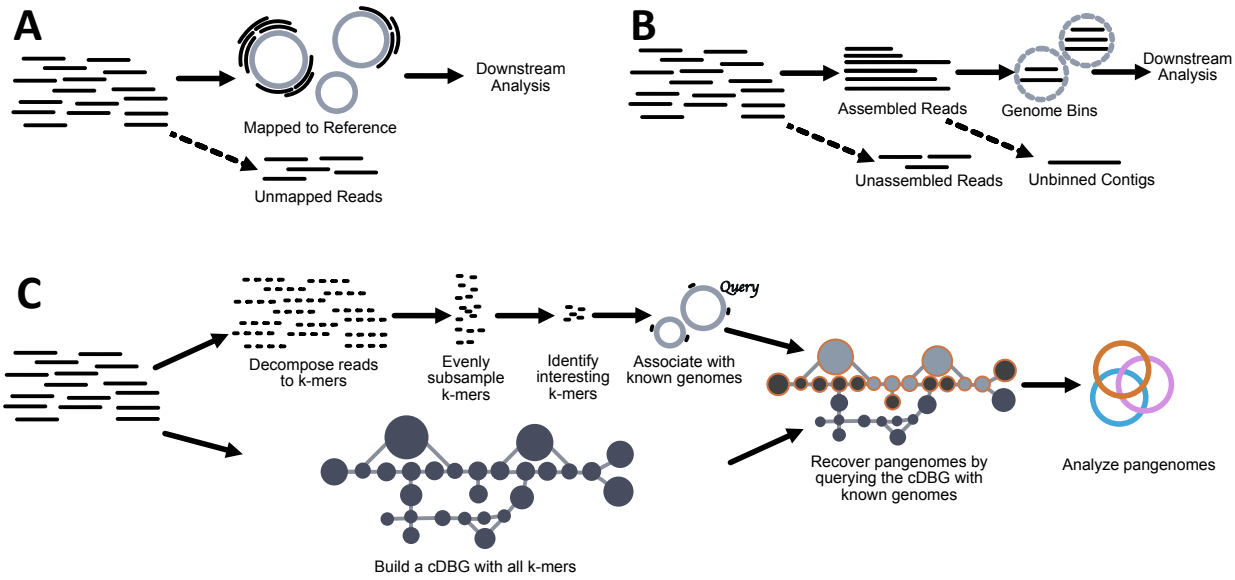

Figure 1: *Comparison of common metagenome analysis techniques with the method used in this paper. Short read metagenomes consist of 50-300 bp reads derived from sequencing DNA from environmental samples.* ***A*** *Reference-based metagenomic analysis. Reads are compared to genomes, genes, or proteins 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, approximately 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 metapangenome depicted in orange, blue, and pink.*

2.2 K-mers capture variation due to disease subtype

We first sought to understand whether variation due to IBD diagnosis is detectable in gut metagenomes. 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 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, indicating that there is a small but detectable signal of IBD subtype in stool metagenomes.

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. All test were significant at $p < .001$.*

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%

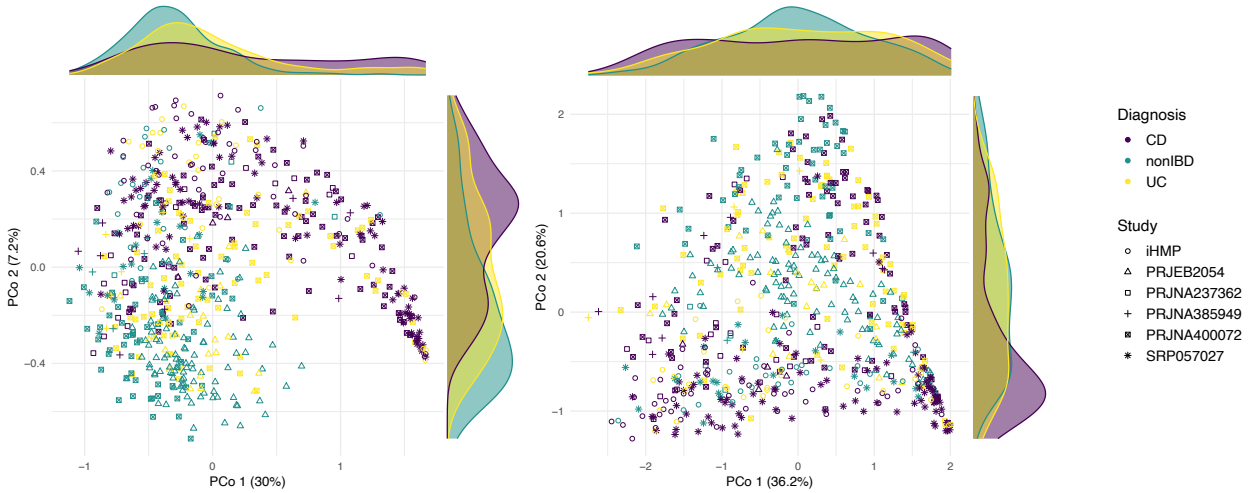

Figure 2: *Principle coordinate analysis of metagenomes from IBD cohorts performed on filtered signatures.* ***A*** *Jaccard distance.* ***B*** *Angular distance.*

2.3 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 nonIBD subtype. We used random forests because of the interpretability of feature importance via variable importance measurements. To assess whether disease signatures generalized across study populations, 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 data set (e.g. many more hashes than samples), we first used variable selection to narrow the set of predictive hashes in the training set (Janitza, Celik, and Boulesteix 2018; Degenhardt, Seifert, and Szymczak 2017). 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% (**Table 4**, **Figure S1**), outperforming a previously published model built on metagenomic data alone (Franzosa et al. 2019).

We next sought to understand whether there was a consistent biological signal captured among classifiers by evaluating the fraction of shared hashes between models. We intersected each set of hashes used to build each optimized classifier. Nine hundred thirty two hashes were shared between all classifiers, while 3,859 hashes were shared between at least five studies (**Figure S2**). 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 out sized proportion

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

Validation.study	Selected.hashes	Percent.of.total.hashes
PRJNA385949	41701	0.57%
PRJNA237362	40726	0.55%
iHMP	39628	0.54%
PRJEB2054	35343	0.48%
PRJNA400072	32578	0.44%
SRP057027	29264	0.40%

Table 4: *Accuracy of random forest classifiers built with different underlying representations of IBD metagenomes when applied to each validation set.*

Validation.Study	Hash.model	Marker.gene.model	Hash.model.of.marker.genes
SRP057027	75.9	86.4	71.7
PRJNA237362	71.4	75.0	64.3
PRJEB2054	69.4	19.1	15.5
PRJNA385949	52.9	52.9	41.2
PRJNA400072	50.9	48.1	47.4
iHMP	49.1	44.2	46.5

of variable importance in the optimized classifiers. After normalizing variable importance across classifiers, 40.2% of the total variable importance was held by shared hashes, 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.

Many hashes were identifiable when compared 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). 77.7% of hashes from all classifiers were identifiable and anchored to 1,161 genomes (**Figure 3 A**). In contrast, 69.4% of shared hashes anchored to 41 genomes (**Figure 3 B**). These shared 41 genomes held an additional 10.3% of variable importance over the shared hashes because some genomes contain additional hashes not shared across all models. Using GTDB taxonomy, we find 38 species represented among the 41 genomes (**Figure 3 C**). These genomes represent a microbial core important for IBD subtype classification.

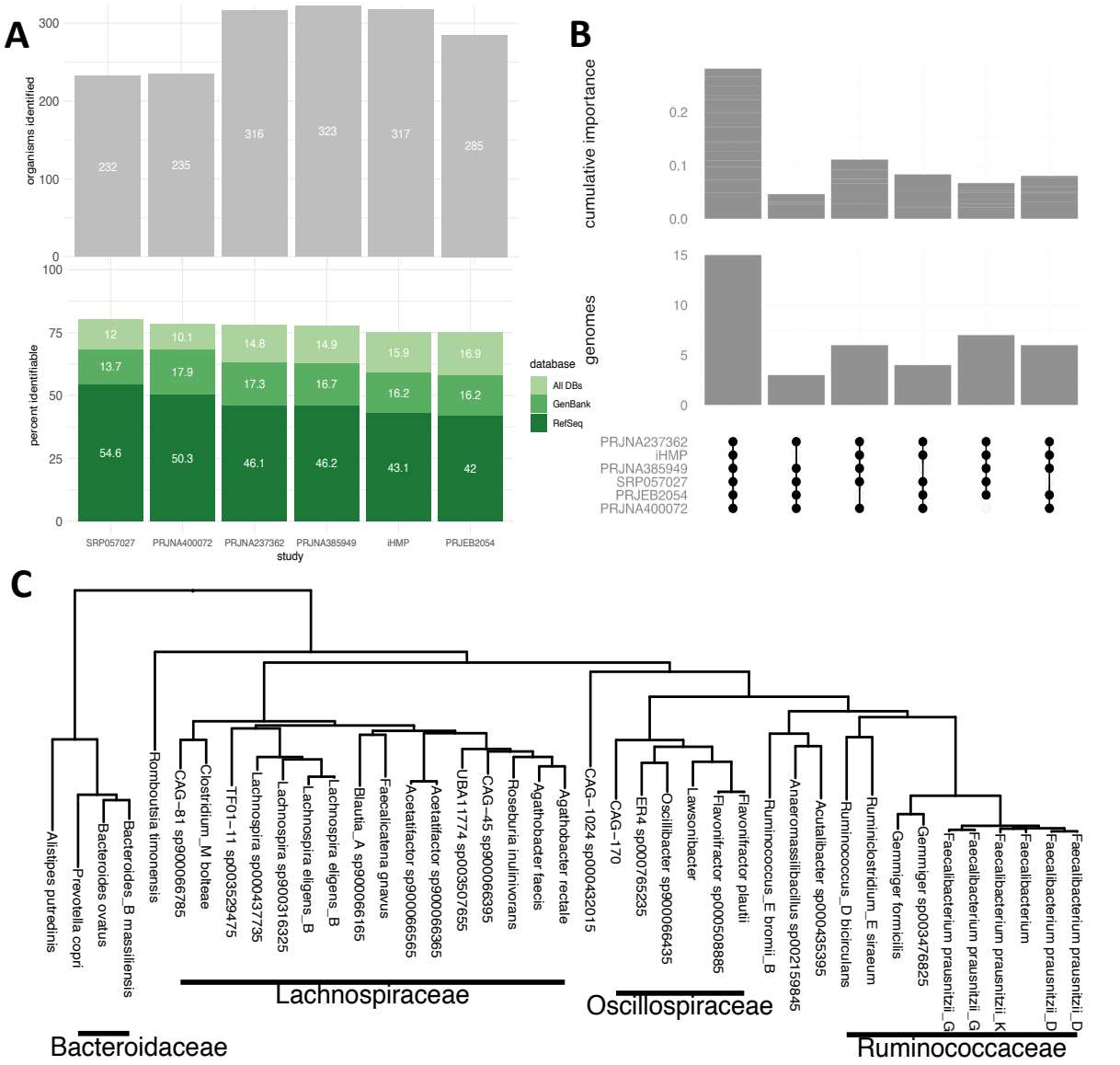

Figure 3: ***A*** *Hashes in random forest classifiers anchor to 1,161 genomes across three databases, accounting for 75.1-80.3% of all predictive hashes in each model. Models are labelled by validation study.* ***B*** *The 3,859 hashes that are shared between the majority of models anchor to 41 genomes. These genomes account for 50.5% of cumulative normalized variable importance for hashes across all models.* ***C*** *The 41 shared genomes are annoated as 38 species in the GTDB taxonomy.*

2.4 Marker genes dominate signatures of IBD

While k-mer based signatures allow us to use all sequencing data in a metagenome and quickly compare against all known genomes, hashes lack sequence context and do not represent function. Given this, we next sought to uncover the functional potential of the shared hashes. To annotate each hash, we reasoned that a hash with predictive importance would be nearby the genomic feature driving the predictive signal in both genomic DNA and the DNA assembly graph. Therefore, we performed assembly graph queries on each metagenome using each gene in the 41 shared genomes. We then identified all gene query neighborhoods in which each hash occurred, and transferred those annotations to the hash.

We noticed that many hashes were annotated as marker genes: 440 hashes accounting for 7.5% of variable

importance across all models annotated to bacterial single copy marker genes (Parks et al. 2015; Na et al. 2018), as well as 16S and 23S ribosomal RNA (**Figure 4 A**). Given the substantial fraction of variable importance attributable to these genetic elements, we were curious how well models built from marker genes alone would perform in IBD subtype classification. However, we wanted to only look at marker gene abundances from the microbial core. We first performed assembly graph queries using the shared 41 genomes to retrieve all reads in the neighborhoods of those genomes (**See Supplement**). We then built random forest classifiers using the same approach as with hashes, but using read abundances of 14 ribosomal marker genes and 16S rRNA (Woodcroft 2018). Classification accuracy across all models was similar to the k-mer based model (**Table 4**), however marker gene models performed marginally better at CD classification and marginally worse at UC classification (**Figure S1**). Both the hash model and the marker gene model ranked a 16S rRNA sequence from the genus *Acetatifactor* as having the highest variable importance across studies, demonstrating that while based on different data features, both model types extract similar information.

The marker gene model performed similarly as the k-mer model for all studies except PRJEB2054. This study was sequenced with 36 base pair reads. It performed poorly due to decreased prediction of marker genes from reads. While the hash model performs well even with very short reads, this has limited technological application given ever-increasing sequencing read lengths.

While hash and marker gene models performed similarly, we were curious whether they captured the same information or whether they captured overlapping but distinct characteristics about IBD subtype. Therefore, we used reads used to build the marker gene model to build a hash model from marker genes alone. These hash models of marker genes performed worse than both the full hash model and marker gene model (**Table 4**}, indicating that each model captures separate, additional information important for IBD classification. The full marker gene model likely out-performs the hash model of marker genes given both subsampling and shorter sequence lengths used by k-mers, both of which obscure taxonomic relationships between short sequences.

To test whether marker gene model accuracies were uniquely driven by marker gene abundance in the 41 shared genomes, we next built a model from all marker genes in the metagenome. XXXXX.

We also calculated the difference between marker gene sequences identified in the 41 shared genome neighborhoods and those identified in the full metagenomes. XXXXX. (Preliminarily, half are in the neighborhoods – so no matter if full abundtrim models are better or worse, argues that the 41 genomes are a useful framework to learn more about IBD.)

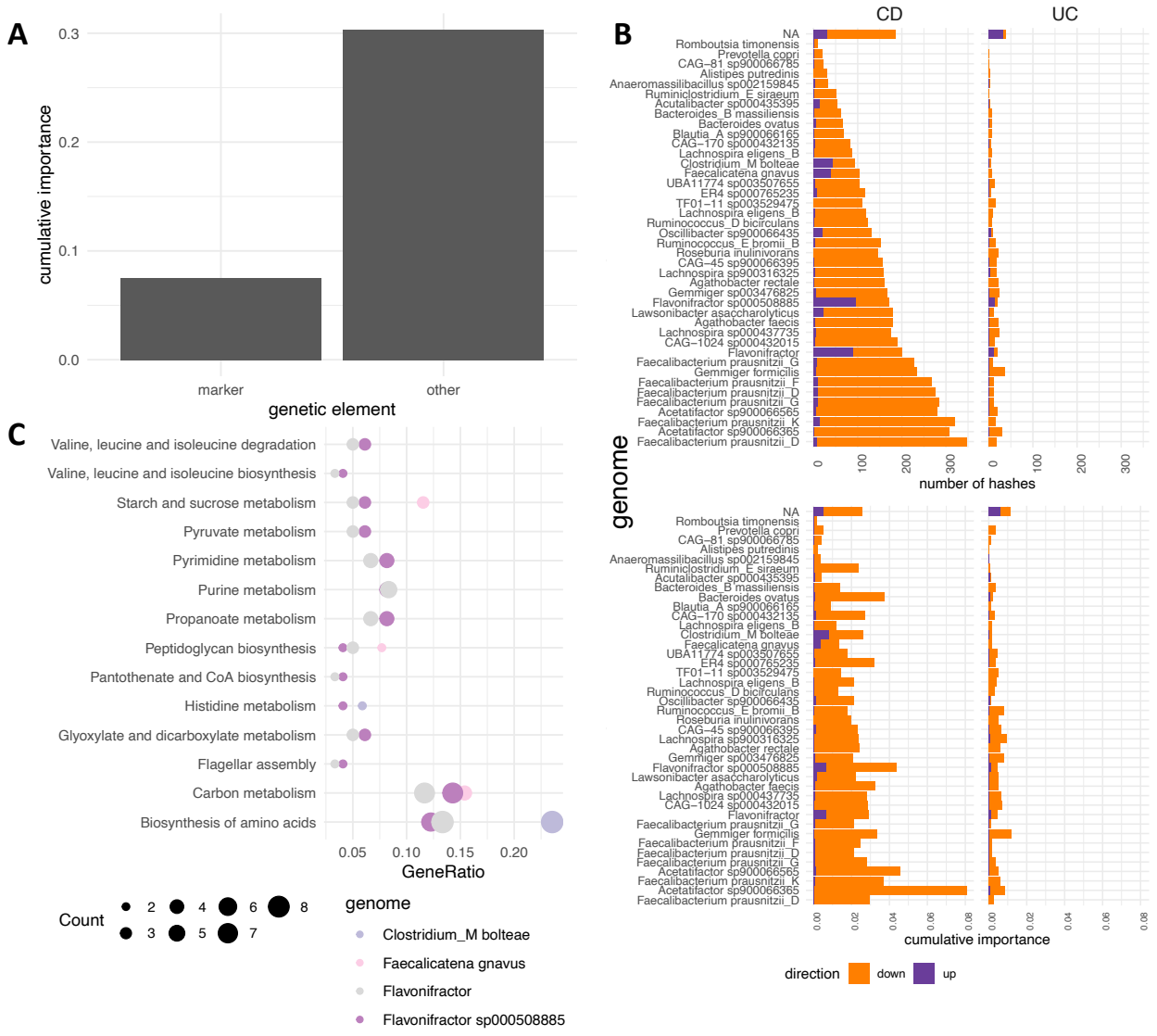

Figure 4: *Decrease in marker genes drives differences between CD and nonIBD.* ***A*** *Many shared hashes annotate as marker genes or ribosomal RNAs. 11.4% of the 3859 shared hashes annotate as marker genes, accounting for 7.5% of variable importance across all models.* ***B*** *The majority of hashes are less abundant in IBD than in nonIBD. However, hashes that anchor to four genomes in CD and two genomes in UC are more abundant.* ***C*** *More abundant hashes in CD are enriched in metabolic pathways and contain few marker genes. Only pathways that are significantly enriched in two of the four genomes are depicted.*

2.5 Decreased diversity punctuated by strain enrichment explains IBD

To better understand microbial signatures of IBD captured by the hash model, we performed differential abundance analysis on the shared hashes (Martin et al. 2020). Many more hashes were differentially abundant in CD than UC (1815 hashes versus 166 hashes, respectively), reflecting the heterogeneity of the UC gut microbiome (Franzosa et al. 2019). Differentially abundant marker genes segregated to the fraction of hashes that were decreased in abundance in CD. This, combined with results from results from marker gene models, indicates that decrease in diversity captured by marker genes is a hallmark of the CD gut microbiome. The majority of non-marker genes with decreased abundance in CD appear to be driven by general decrease in diversity of the gut microbiome, not systematic loss of specific functional potential.

Two strains of *Faecalibacterium prausnitzii* have the largest number of hashes with decreased abundance compared to nonIBD (**Figure 4 B**). *F. prausnitzii* is an obligate anaerobe and a key butyrate producer in the gut, and plays a crucial role in reducing intestinal inflammation (Lopez-Siles et al. 2017). *F. prausnitzii* is extremely sensitive to oxygen, though may be able to withstand oxygen exposure for up to 24 hours depending on the availability of metabolites for extracellular electron transfer (Lopez-Siles et al. 2017). *Acetatifactor* (GTDB species *Acetatifactor sp900066365*) has hashes with the largest variable importance with decreased abundance compared to nonIBD (**Figure 4 B**).
Acetatifactor is a bile-acid producing bacteria associated with a healthy gut, but limited evidence has associated it with decreased abundance in IBD (Pathak et al. 2018). In UC, *Gemmiger formicilis* has both the largest variable importance and number of hashes with decreased abundance compared to nonIBD (**Figure 4 B**). *G. formicilis* is a strictly anaerobic bacteria that produces both formic acid butyric acid (GOSSLING and Moore 1975).

While most of the microbial core decreases in abundance, a substantial portion of hashes from four genomes in CD and two in UC are more abundant in disease (**Figure 4 B**). These four belong to *Faecalicatena gnavus* (referred to as *[Ruminococcus] gnavus* in NCBI taxonomy and IBD literature) and *Clostridium bolteae* in CD, and two genomes in the *Flavonifractor* in CD and UC. KEGG enrichment analysis on more and less abundant hashes demonstrated enrichment of pathways dominated by marker genes (e.g. Ribosomes, tRNA biosynthesis) in the less abundant hashes from CD in all four genomes, indicating a general decrease in abundance for these organisms. However, metabolic pathways such as starch and sucrose metabolism, propanoate metabolism, and peptidoglycan synthesis are enriched in the more abundant hashes in CD (**Figure 4 C**). This shift is indicative of strain-specific enrichment in CD.

2.6 Enriched strains are more abundant in but not exclusive to IBD

- kmer accumulation curves
- strain comparisons for function/determine if there is a "clade" that is more abundant in disease (as is supported by hashes)
- Include tie in with *R gnavus* disease isolate that is in ~10 "healthy" metagenomes but that hasn't been observed in nonIBD metagenomes before (written up on metapangenome ORFs in the supplement right now).

3 Discussion

IBD is a heterogeneous disease characterized by periods of activity and dormancy. While the underlying etiology is poorly understood, IBD arises from a complex interaction between host genetics, environment, and the gut microbiome. Here we present a new method to examine microbial associations of disease, and using this method uncover signatures of IBD subtype. These signatures demonstrate consistent loss of diversity of specific microorganisms, particularly in CD. Meanwhile, four strains are enriched in CD and two in UC, potentially indicating niche partitioning in response to IBD-associated perturbations. While our classifiers are not accurate enough to be clinically relevant, the conserved signatures we detect warrant further research and may yield new therapeutics for IBD treatment.

While we find conserved signatures in IBD subtype, we find no evidence for disease-specific microbiomes, or pangenomes of the organisms that comprise them. Instead, almost all k-mers (or genes) are observed in CD, UC, and nonIBD, suggesting stochastic loss of pangenome diversity during disease. Similarly, while a few strains are enriched in IBD microbiomes, these strains are all detected in nonIBD at low frequency. This includes (lab-proven *F. gnavus*). HOW DOES THIS FIT WITH THE WORKING MODEL OF IBD BIOLOGY?

These patterns in part explain the inconsistent results generated in IBD subtype characterization, where no consistent microbiological signal has emerged in human gut microbiomes other than loos of diversity (CITATIONS). However, the results presented herein demonstrate the need for reference- and assembly-free

analysis of metagenomes. Strain-level resolution was essential for the detection of enriched organisms, but this resolution is precluded by assembly and reference-based methods. Recent large-scale assembly efforts have dramatically improved our catalog of diversity for human microbiomes (Pasolli et al. 2019; Nayfach et al. 2019; Almeida et al. 2019), however many sequences that are signatures of IBD are not in these databases. K-mer-based analysis combined with assembly graph queries provides a necessary window into strain-level dynamics in metagenomes.

Our models consistently performed the most poorly on the iHMP cohort. The iHMP tracked the emergence and diagnosis of IBD through time series profiling of emergent cases (CITE). We selected the first sample in each time series for this analysis. Given that our model performed poorly on these samples, this may suggest that disease onset is a distinct biological process. One avenue of future research is analysis of these time series samples for emergence of disease signatures.

While k-mer-based analysis revealed signatures of IBD subtype, 9.1% of shared hashes were uncharacterized by reference databases or assembly graph queries. These hashes may represent strain variants of the microbial core we detected, or may be novel sequences from other organisms, plasmids, or viruses. Targeted graph-based queries may reveal the identity of these elements and their relationship to IBD.

While we apply our pipeline to IBD classification, it is extensible to other large meta cohorts of metagenomic sequencing data. This method may be particularly suitable for disease such as colorectal cancer, where a recent meta analysis using a marker gene approach was successful in classifying colorectal samples from health controls (Wirbel et al. 2019). Our method may bring strain-level resolution and generate hypothesis for further research.

Taken together, XXXXX.

4 Methods

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

4.1 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.

4.2 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`.

4.3 Random forest classifiers

We built random forests classifier to predict CD, UC, and non-IBD status using filtered signatures (hash models), marker genes in the shared 41 genomes (marker gene models), signatures from reads that were detected as marker genes (hash models of marker genes), and marker genes in the full metagenome (full marker gene models).

For models from signatures, 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 importance (Janitza, Celik, and Boulesteix 2018). It is based on permutation 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). 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.

For the marker gene models, we generated marker gene abundances for 14 ribosomal marker genes and 16S rRNA using singleM (Woodcroft 2018). We then followed the same model building procedure as the hash models.

4.4 Anchoring predictive hashes to genomes

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.

4.5 Compact de Bruijn graph queries for predictive genes and genomes

To annotate hashes with functional potential, we first extracted open reading frames (ORFs) from the shared 41 genomes using prokka, and annotated ORFs with EggNog (Seemann 2014; Huerta-Cepas et al. 2019). When then used spacegraphcats `multifasta_query` to create a hash:gene map. Spacegraphcats retrieves k-mers in the compact de Bruijn graph neighborhood of a query gene, and hashing these k-mers via sourmash generates a hash:gene map (Brown et al. 2020; Brown and Irber 2016). Because genomes with shared 31-mers may annotate the same hash, we allowed hashes to be annotated multiple times. This was particularly appropriate for hashes from highly conserved regions, e.g. 16S ribosomal RNA.

We used spacegraphcats `search` to retrieve k-mers in the compact de Bruijn graph neighborhood of the shared genomes (Brown et al. 2020). We then used spacegraphcats `extract_reads` to retrieve the reads and `extract_contigs` to retrieve unitigs in the compact de Bruijn graph that contained those k-mers, respectively. These reads were used to generate marker gene abundances for the 41 shared genomes for the marker gene random forest models.

4.6 Differential hash abundance analysis

To determine whether shared hashes were differentially abundant from nonIBD in UC or CD, we used corncob (Martin et al. 2020). We used all hash abundances from sourmash signatures to determine hash library size, and then compared hash abundances between disease groups 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. We performed enrichment analysis using the R package clusterProfiler (Yu et al. 2012).

5 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.

Brown, C Titus, Dominik Moritz, Michael P O'Brien, Felix Reidl, Taylor Reiter, and Blair D Sullivan. 2020. "Exploring Neighborhoods in Large Metagenome Assembly Graphs Using Spacegraphcats Reveals Hidden Sequence Diversity." *Genome Biology* 21 (1): 1–16.

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.

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.

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.

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.

GOSSLING, JENNIFER, and WEC Moore. 1975. "Gemmiger Formicilis, N. Gen., N. Sp., an Anaerobic Budding Bacterium from Intestines." *International Journal of Systematic and Evolutionary Microbiology* 25 (2): 202–7.

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.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, et al. 2019. "EggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses." *Nucleic Acids Research* 47 (D1): D309–D314.

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.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. "MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph." *Bioinformatics* 31 (10): 1674–6.

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.

Lopez-Siles, Mireia, Sylvia H Duncan, L Jesús Garcia-Gil, and Margarita Martinez-Medina. 2017. "Faecalibacterium Prausnitzii: From Microbiology to Diagnostics and Prognostics." *The ISME Journal* 11 (4): 841–52.

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.

Na, Seong-In, Yeong Ouk Kim, Seok-Hwan Yoon, Sung-min Ha, Inwoo Baek, and Jongsik Chun. 2018. "UBCG: Up-to-Date Bacterial Core Gene Set and Pipeline for Phylogenomic Tree Reconstruction." *Journal of Microbiology* 56 (4): 280–85.

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*.

Parks, Donovan H, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. 2015. "CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes." *Genome Research* 25 (7): 1043–55.

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.

Pathak, Preeti, Cen Xie, Robert G Nichols, Jessica M Ferrell, Shannon Boehme, Kristopher W Krausz, Andrew D Patterson, Frank J Gonzalez, and John YL Chiang. 2018. "Intestine Farnesoid X Receptor Agonist and the Gut Microbiota Activate G-Protein Bile Acid Receptor-1 Signaling to Improve Metabolism." *Hepatology* 68 (4): 1574–88.

Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19.

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.

Seemann, Torsten. 2014. "Prokka: Rapid Prokaryotic Genome Annotation." *Bioinformatics* 30 (14): 2068–9.

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.

Woodcroft, B. 2018. "Singlem."

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.

Yuan, Cheng, Jikai Lei, James Cole, and Yanni Sun. 2015. "Reconstructing 16S rRNA Genes in Metagenomic Data." *Bioinformatics* 31 (12): i35–i43.

Supplementary material

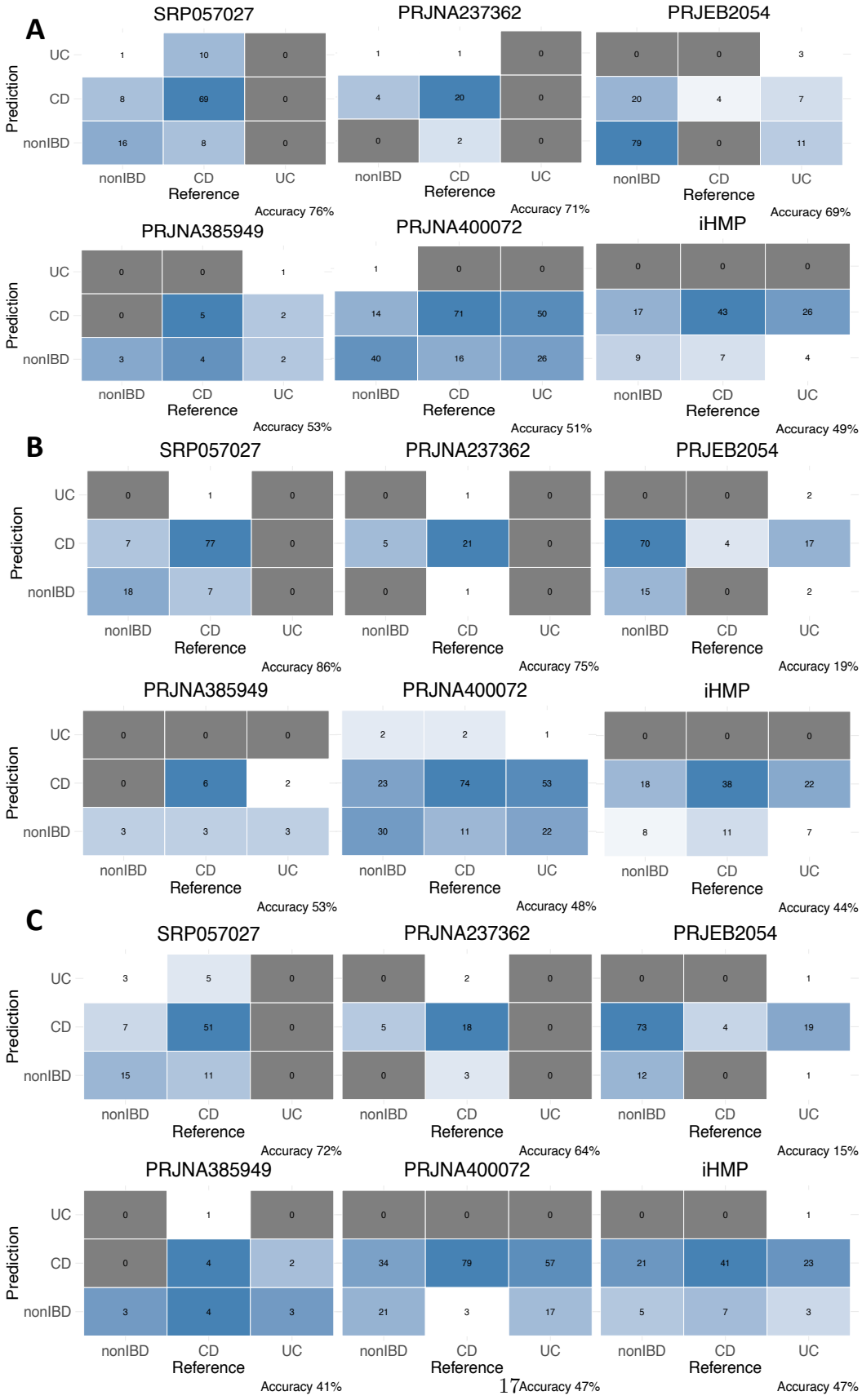

Figure S1: *Confusion matrices for leave-one-study-out random forest models evaluated on the validation set. **A** hash model. **B** marker gene model. **C** hash model of marker genes.*

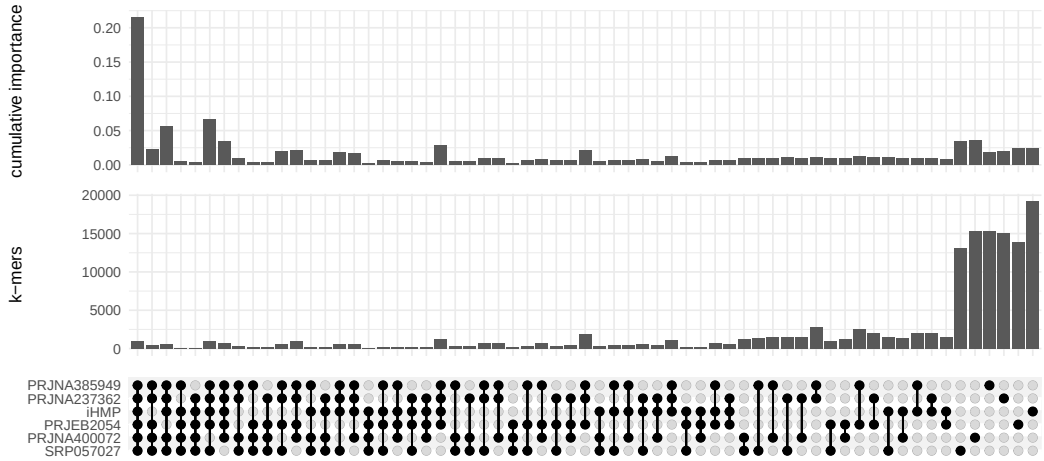

Figure S2: *Hash models share a large fraction of predictive hashes. Upset plot depicting intersections of sets of hashes as well as the cumulative normalized variable importance of those hashes in the optimized random forest classifiers. Each classifier is labelled by the left-out validation study.*

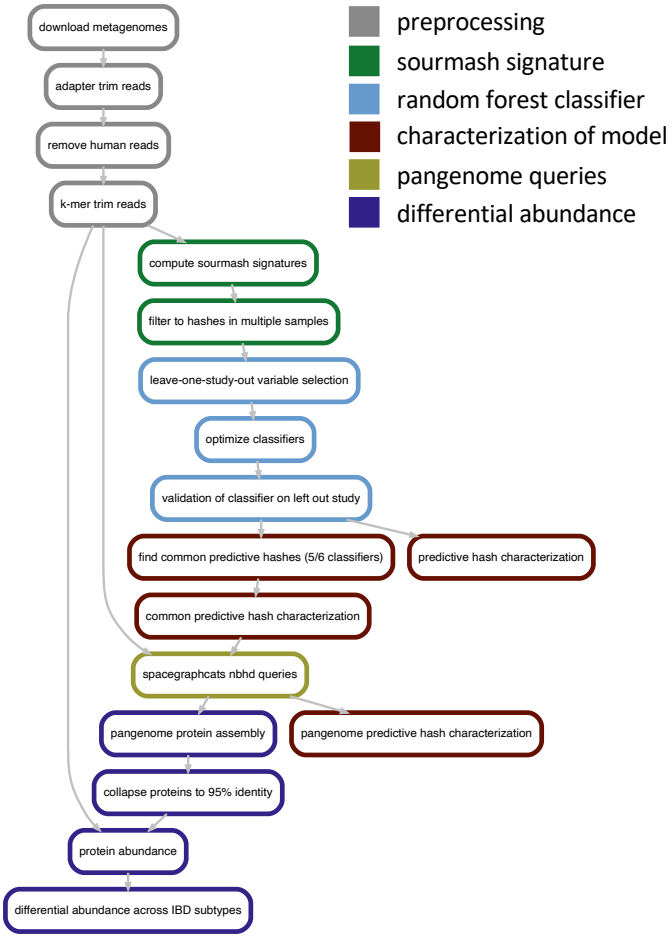

Figure S3: *Simplified directed acyclic graph of the steps used in our pipeline, color coded by the section of the pipeline each step corresponds to. The steps in blue were performed six times, each time with a different validation study.*

5.1 Description of IBD metagenome study cohorts

Below we present a description of each of the six cohorts used in this meta analysis. Each description is presented as was found in the original publication of each cohort.

iHMP (Lloyd-Price et al. 2019):

> Five medical centres participated in the IBDMDB: Cincinnati Children's Hospital, Emory University Hospital, Massachusetts General Hospital, Massachusetts General Hospital for Children, and Cedars-Sinai Medical Center. Patients were approached for potential recruitment upon presentation for routine age-related colorectal cancer screening, work up of other gastrointestinal (GI) symptoms, or suspected IBD, either with positive imaging (for example, colonic wall thickening or ileal inflammation) or symptoms of chronic diarrhoea or rectal bleeding. Participants could not have had a prior screening or diagnostic colonoscopy. Potential participants were excluded if they were unable to or did not consent to provide tissue, blood, or stool, were pregnant, had a known bleeding disorder or an acute gastrointestinal infection, were actively being treated for a malignancy with chemotherapy, were diagnosed with indeterminate colitis, or had undergone a prior, major gastrointestinal surgery such as an ileal/colonic diversion or j-pouch. Upon enrollment, an initial colonoscopy was performed to determine study strata. Subjects not diagnosed with IBD based on endoscopic and histopathologic findings were classified as 'non-IBD' controls, including the aforementioned healthy individuals presenting for routine screening, and those with more benign or non-specific symptoms. This creates a control group that, while not completely 'healthy', differs from the IBD cohorts specifically by clinical IBD status. Differences observed between these groups are therefore more likely to constitute differences specific to IBD, and not differences attributable to general GI distress.

PRJEB2054 (Qin et al. 2010):

> As part of the MetaHIT (Metagenomics of the Human Intestinal Tract) project, we collected faecal specimens from 124 healthy, overweight and obese individual human adults, as well as inflammatory bowel disease (IBD) patients, from Denmark and Spain.

SRP057027 (Lewis et al. 2015):

> Children and young adults less than 22 years of age were enrolled at the time of initiation of EN or anti-TNF therapy for treatment of active CD (defined as the Pediatric Crohn's Disease Activity Index [PCDAI] >10) at The Hospital for Sick Children in Toronto, ON, Canada; IWK Health Centre, Halifax, NS, Canada; and the Children's Hospital of Philadelphia, Pennsylvania. Participants in this observational cohort study were prescreened for eligibility and recruited from clinic or during inpatient hospitalization. Exclusion criteria included presence of an ostomy, treatment with probiotics within 2 weeks of initiating EN, treatment with anti-TNF therapy within 8 weeks of starting EN, or treatment with EN within 1 week of initiating anti-TNF therapy. The study protocol was approved by the institutional review boards at all participating institutions. Informed consent was obtained from all young adults and the parents/guardians of children less than 18 years of age.

PRJNA385949 (Hall et al. 2017):

> Samples from the PRISM study, collected at Massachusetts General Hospital: A subset of the PRISM cohort was selected for longitudinal analysis. A total of 15 IBD cases (nine CD, five UC, one indeterminate colitis) were enrolled in the longitudinal stool study (LSS). Three participants with gastrointestinal symptoms that tested negative for IBD were included as a control population. Enrollment in the study did not affect treatment. Stool samples were collected monthly, for up to 12 months. The first stool sample was taken after treatment had begun. Comprehensive clinical data for each of the participants was collected at each visit. At each collection, a subset of participants were interviewed to determine their disease activity index, the Harvey-Bradshaw index for CD participants and the simple clinical colitis activity index (SCCAI) for UC participants. Samples collected at Emory University: To increase the number of participants in our analysis, a

subset of the pediatric cohort STiNKi was selected for whole metagenome sequencing including five individuals with UC and nine healthy controls. All selected UC cases were categorized as non-responders to treatment. Stool samples were collected approximately monthly for up to 10 months. The first sample from participants in the STiNKi cohort is before treatment started, and subsequent samples are after treatment started. Stool collection and DNA extraction methods are detailed in Shaw et al.

PRJNA400072 (Franzosa et al. 2019):

PRISM cohort description and sample handling: PRISM is a referral centre-based, prospective cohort of IBD patients; 161 adult patients (>18 years old) enrolled in PRISM and diagnosed with CD, UC, and non-IBD (control) were selected for this study, with diagnoses based on standard endoscopic, radographical and histological criteria. The PRISM research protocols were reviewed and approved by the Partners Human Research Committee (re. 2004-P-001067), and all experiments adhered to the regulations of this review board. PRISM patient stool samples were collected at the MGH gastroenterolgy clinic and stored at -80C before DNA was extracted.

Validation cohort description and sample handling: The validation cohort consisted of 65 patients enrolled in two distinct studies form the Netherlands; 22 controls were enrolled in the LifeLines DEEP general population study and 43 patients with IBD were enrolled in a study at the Department of Gastroenterology and Hematology at the University Medical Center Groningen. Patients enrolled in both studies collected stool using the same protocol: a single stool sample was collected at home and then frozen within 15 min in a conventional freezer. A research nurse visited all participants at home to collect home-frozen stool samples, which were then transported and stored at -80C. The stool samples were kept frozen before DNA was extracted.

PRJNA237362 (Gevers et al. 2014):

A total of 447 children and adolescents (<17 years) with newly diagnosed CD and a control population composed of 221 subjects with noninflammatory conditions of the gastrointestinal tract were enrolled to the RISK study in 28 participating pediatric gastroenterology centers in North America between November 2008 and January 2012.

5.2 Construction of human microbiome metagenome assembled genome databases

While GenBank contains hundreds of thousands of isolate and metagenome-assembled genomes, we augmented the number of genomes by creating sourmash databases for all medium- and high-quality metagenome-assembled genomes from three recent human microbiome metagenome *de novo* assembly efforts (Pasolli et al. 2019; Nayfach et al. 2019; Almeida et al. 2019). The databases are available at in the OSF repository, "Comprehensive Human Microbiome Sourmash Databases" at the URL https://osf.io/hza89/. While we are aware that contamination in both GenBank and from these studies could introduce contamination into our analysis, we reasoned that the increase we observed in identifiable hashes when we did not restrict ourselves to RefSeq was worth the trade.

To generate the databases, we downloaded the medium- and high-quality metagenome-assembled genomes and used sourmash `compute` with parameters `k 21,31,51`, `--track-abundance`, and `--scaled 2000`. We then used sourmash `index` to generate databases for `k = 31`. Below we detail the contents of each database.

- Pasolli et al. (2019): contains 70,178 high- and 84,545 medium-quality MAGs assembled from 9,428 human microbiome samples. Samples originate from stool (7,783), oral cavity (783), skin (503), vagina (88), and maternal milk (9). Original Data Download: http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html
- Almeida et al. (2019): contains 40,029 high- and 65,671 medium-quality MAGs assembled from 11,850 human microbiome samples. All samples originate from stool. Original Data Download: ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/umgs_analyses/mags-gut_qs50.tar.gz

- Nayfach et al. (2019): contains 24,345 high- and 36,319 medium-quality MAGs assembled from 3,810 human gut microbiome samples. Original Data Download: https://github.com/snayfach/IGGdb

5.3 41 genome accessions and taxonomy

Genomes are available for download at https://osf.io/ungza/

5.4 Contamination in 41 shared genomes

We identified 41 genomes that were important for IBD subtype classification across six models. We used assigned GTDB taxonomy to each genome. 38 species represented among the 41 genomes. 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 S4**). Therefore, we proceeded with analysis with the idea that each of the 41 genomes is a self-contained entity that captures distinct biology.

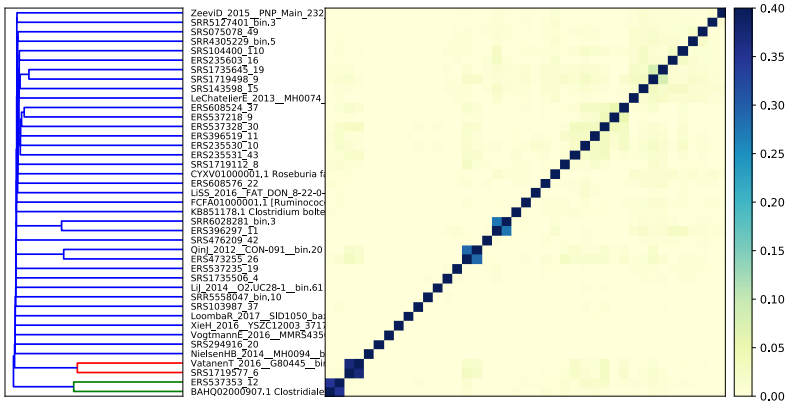

Figure S4: *Jaccard similarity between 41 genomes. The highest similarity between genomes is 0.37 and is shared by genomes of the same species, while most genomes have no similarity. This indicates that each genome represents distinct nucleotide sequence.*

5.5 Characterization of unknown but predictive hashes through assembly graph queries

Given that 30.6% of shared hashes did not anchor to genomes in databases, we sought to characterize these hashes. We reasoned that many unknown but predictive hashes likely originate from closely related strain variants of identified genomes, or from closely-related sequences not assembled or binned during the original genome analysis. We sought to recover these variants. We performed assembly graph queries into each metagenome sample with the 41 genomes that contained shared hashes, producing a pangenome for each query genome within each metagenome sample. Combining pangenomes from all metagenomes, we generated a metapangenome for each of the 41 original query genomes. 90.9% of shared hashes were in the 41 metapangenomes, a 21.5% increase over the genomes alone. This suggests that at least 21.5% of shared hashes originate from strain-variable or accessory elements in pangenomes.

Further, these metapangenomes captured an additional 4.2-5.2% of all predictive hashes from each classifier, indicating that metapangenomes contain novel sequences not captured in any database (**Figure S5**). The metapangenomes also captured 74.5% of all variable importance, a 24% increase over the 41 genomes alone.

Table S1: *Identifiers, GTDB and NCBI taxonomy for the 41 shared genomes.*

genome	GTDB	NCBI
ERS235530_10.fna	s__CAG-1024 sp000432015	Clostridium sp. CAG:1024
ERS235531_43.fna	s__Faecalibacterium prausnitzii_F	NA
ERS235603_16.fna	s__Agathobacter rectale	[Eubacterium] rectale
ERS396297_11.fna	s__Lachnospira eligens_B	[Eubacterium] eligens
ERS396519_11.fna	s__Lawsonibacter asaccharolyticus	Clostridium phoceensis
ERS473255_26.fna	s__Faecalibacterium prausnitzii_G	NA
ERS537218_9.fna	s__Gemmiger sp003476825	Faecalibacterium sp. UBA2
ERS537235_19.fna	s__Bacteroides_B massiliensis	NA
ERS537328_30.fna	s__Faecalibacterium prausnitzii_K	NA
ERS537353_12.fna	g__Flavonifractor	NA
ERS608524_37.fna	s__Gemmiger formicilis	NA
ERS608576_22.fna	s__Ruminococcus_E bromii_B	NA
GCF_000371685.1_Clos_bolt_90B3_V1_genomic.fna	s__Clostridium_M bolteae	Clostridium bolteae 90B3
GCF_000508885.1_ASM50888v1_genomic.fna	s__Flavonifractor sp000508885	Clostridiales bacterium VE2
GCF_001405615.1_13414_6_47_genomic.fna	s__Agathobacter faecis	Roseburia faecis strain 2789
GCF_900036035.1_RGNV35913_genomic.fna	s__Faecalicatena gnavus	[Ruminococcus] gnavus
LeChatelierE_2013__MH0074__bin.19.fa	s__CAG-45 sp900066395	NA
LiJ_2014__O2.UC28-1__bin.61.fa	s__Ruminiclostridium_E siraeum	[Eubacterium] siraeum
LiSS_2016__FAT_DON_8-22-0-0__bin.28.fa	s__CAG-170 sp000432135	Firmicutes bacterium CAG:
LoombaR_2017__SID1050_bax__bin.11.fa	s__Anaeromassilibacillus sp002159845	Anaeromassilibacillus sp. A
NielsenHB_2014__MH0094__bin.44.fa	s__Prevotella copri	NA
QinJ_2012__CON-091__bin.20.fa	s__Faecalibacterium prausnitzii_G	NA
SRR4305229_bin.5.fa	s__Roseburia inulinivorans	NA
SRR5127401_bin.3.fa	s__UBA11774 sp003507655	NA
SRR5558047_bin.10.fa	s__Alistipes putredinis	NA
SRR6028281_bin.3.fa	s__Lachnospira eligens_B	[Eubacterium] eligens
SRS075078_49.fna	s__TF01-11 sp003529475	Clostridium sp. CAG:75; Cl
SRS103987_37.fna	s__ER4 sp000765235	Oscillibacter sp. ER4
SRS104400_110.fna	s__Lachnospira sp900316325	NA
SRS143598_15.fna	s__Lachnospira sp000437735	NA
SRS1719112_8.fna	s__Oscillibacter sp900066435	NA
SRS1719498_9.fna	s__Acetatifactor sp900066565	Clostridium
SRS1719577_6.fna	s__Faecalibacterium prausnitzii_D	NA
SRS1735506_4.fna	s__Bacteroides ovatus	NA
SRS1735645_19.fna	s__Acetatifactor sp900066365	Firmicutes bacterium CAG:
SRS294916_20.fna	s__Romboutsia timonensis	NA
SRS476209_42.fna	s__Ruminococcus_D bicirculans	NA
VatanenT_2016__G80445__bin.9.fa	s__Faecalibacterium prausnitzii_D	NA
VogtmannE_2016__MMRS43563715ST-27-0-0__bin.70.fa	s__CAG-81 sp900066785	uncultured Clostridium sp.
XieH_2016__YSZC12003_37172__bin.63.fa	s__Acutalibacter sp000435395	Firmicutes bacterium CAG:
ZeeviD_2015__PNP_Main_232__bin.27.fa	s__Blautia_A sp900066165	uncultured Blautia sp.; Rum

This indicates that uncharacterizable sequences contribute substantial predictive power toward IBD subtype classification.

Recovery of metapangenomic variation disproportionately impacts the variable importance attributable to specific genomes (**Figure S5**). While most genomes maintained a similar proportion of importance with or without expansion by neighborhood queries, three metapangenomes shifted dramatically. While an *Acetatifactor* species anchored the most importance prior to pangenome queries, 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. This is likely in part driven by re-association of marker genes with genomes given that marker genes are difficult to assemble and bin in metagenomes. Strain-variable regions are also likely recovered (Brown et al. 2020).

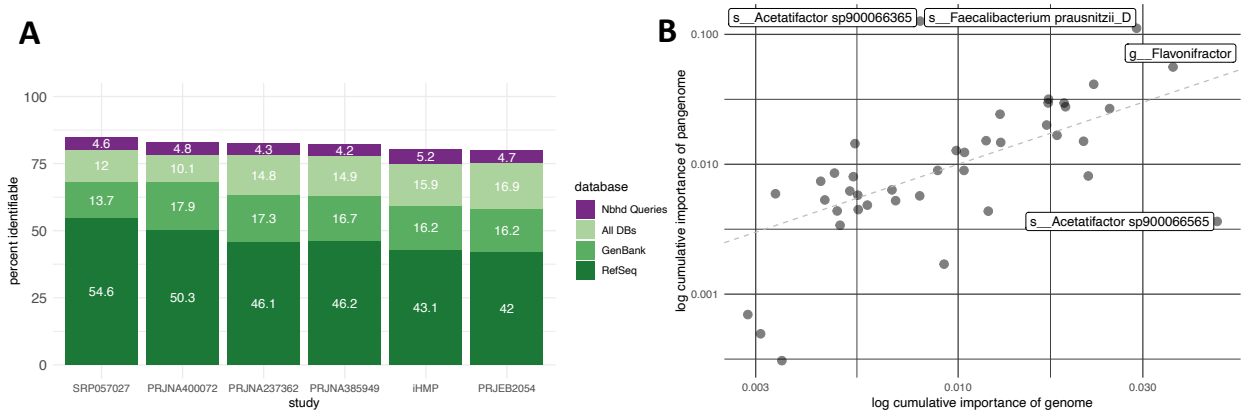

Figure S5: ***A*** *Some hashes anchor to known genomes in RefSeq, GenBank, or human microbiome metagenome-assembled genome databases. An additional approximately 5% of hashes anchor to metapangenome of the 41 shared genomes.* ***B*** *Metapangenome neighborhoods generated with compact de Bruijn graph queries recover strain variation that is important for predicting IBD subtype. While the variable importance attributable to some genomes does not change with assembly graph queries, other genomes increase by more than 7%.*

5.6 Comparing IBD metagenome analysis by assembly

While gene-based queries successfully annotated our shared hashes, we were curious how well an assembly-based approach could characterize pangenome graph neighborhoods. To build a gene catalog for each metapangenome, we assembled each pangenome individually and extracted open reading frames (ORFs). We then clustered ORFs and ORF fragments from pangenomes in the metapangenome at 90% identity.

While the reads from all metapangenomes contain 90.9% of shared hashes, the metapangenome gene catalogs only contain 59.4% of shared hashes. While this loss is in part explained by ORF extraction and clustering, only 63.1% of shared hashes are in the assemblies themselves, demonstrating that assembly accounts for the largest loss of predictive hashes. Further, when we build random forest models of gene counts using the leave-one-study out approach, we observe a substantial decrease in prediction accuracy (**Table S2**). This indicates that some sequences that are important for IBD classification do not assemble.

Unassembled hashes occur in 40 of the 41 metapangenomes. Hashes that are unassembled are not more likely to hold higher variable importance than hashes that do not assemble (Welch Two Sample t-test p = .07; mean assembled = 0.00057, mean unassembled = 0.00072).

We next determined which shared hashes were not captured by assembly. Using gene neighborhood queries from the 41 shared genomes as described in the main text, many unassembled hashes were annotated as 16s and 23s ribosomal RNA, as well as genes encoding 30s and 50s ribosomal proteins. These sequences are difficult to assemble given their repetitive content, but are useful markers of taxonomy given their universal presence in bacterial genomes (Yuan et al. 2015; Parks et al. 2015; Woodcroft 2018).

Table S2: *Accuracy of model on each validation set.*

Validation.Study	Hash.model	Marker.gene.model	Gene.model
SRP057027	75.9	86.4	44.0
PRJNA237362	71.4	75.0	NA
PRJEB2054	69.4	19.1	NA
PRJNA385949	52.9	52.9	35.3
PRJNA400072	50.9	48.1	50.0
iHMP	49.1	44.2	44.3

While many hashes that are predictive of IBD subtype do not assemble, approximately 60% do. We next investigated how metapangenomes differed in CD, UC, and nonIBD based on these assembled fractions alone.

Given that reduced diversity of species in the gut microbiome is a hallmark of IBD (CITATIONS), we first investigated whether the diversity of metapangenome ORFs within a metagenome differed between CD and nonIBD and UC and nonIBD. For each metagenome, we counted the number of ORFs within each metapangenome against which any reads mapped. For 39 of 41 metapangenomes for CD and 37 of 41 metapangenomes for UC, the mean number of ORFs observed per metagenome was lower than nonIBD (ANOVA $p < 0.05$, Tukey's HSD $p < 0.05$). This indicates that the majority of metapangenomes in IBD microbiomes have lower diversity in observed ORFs than nonIBD microbiomes.

Only the metapangenome 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).

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*.

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* produces an inflammatory polysaccharide that induces TNFa secretion in a response mediated by toll-like receptor 4 (Henke et al. 2019).

While there is lower diversity of ORFs in IBD metapangenomes, we find limited evidence of disease-specific metapangenomes. We generated accumulation curves from ORF presence/absence across CD, UC, and nonIBD using metapangenome gene catalogs. While our assemblies were incomplete, we reasoned that by investigating the same set of genes for all samples, we could compare across groups. For most metapangenomes, the majority of genes are observed in CD, UC, and nonIBD. This in part explains heterogeneous study findings in IBD gut microbiome investigations (CITATIONS) and underscores that IBD is a spectrum of diseases characterized by intermittent health and dysbiosis.

ADD A GENE ACCUMULATION CURVE PANEL

Of all metapangenome accumulation curves, only *C. bolteae* does not saturate for UC, with 171 of 16,822 genes unobserved. Ten of 41 do not saturate for CD, with an average of 366 genes unobserved.

Given that both *C. bolteae* and *F. gnavus* demonstrated evidence for strain-specific enrichment in CD, we further performed differential abundance analysis using the metapangenomes. When we compared *F. gnavus* gene abundances in IBD against nonIBD, 5,984 genes were differentially abundant in CD while 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 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).

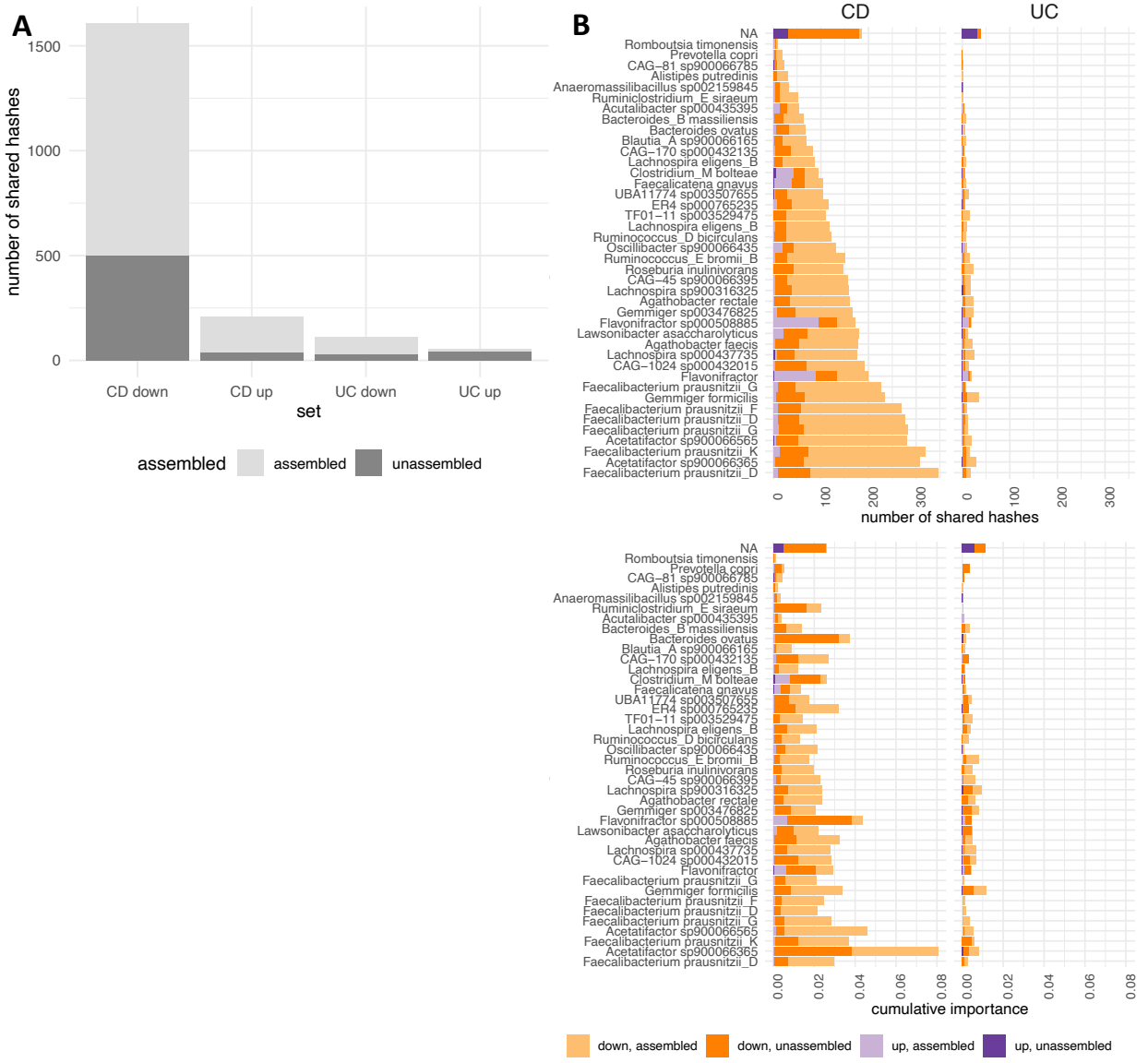

Figure S6: ***A*** *A large fraction of shared hashes do not assemble. The largest fraction segregates to those that are less abundant in CD than nonIBD.* ***B*** *Unassembled shared hashes are distributed across the 41 shared genomes.*

6 Supplementary Methods

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.

Pangenome assembly 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 (Li et al. 2015), and annotated each assembly using prokka (Seemann 2014). 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 (Patro et al. 2017), and BWA to quantify the number of mapped and unmapped reads (CITE: BWA MEM).

7 Supplementary References