IBD Meta-analysis

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

April 29, 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 reads 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 (Franzosa et al. 2014; Lloyd-Price et al. 2019). Reads fail to map when there is no closely related organism or sequence in the reference database. Reads that do not map to references are typically ignored in downstream analysis. To combat these issues, reference-free approaches like *de novo* assembly and binning are used to generate metagenome-assembled genome bins (MAGs). MAGs represent species-level composites of closely related organisms in a sample, and thus often more closely recapitulate genomes found 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), meaning some sequences are not assembled or binned. As with reference-based approaches, these reads are typically left unanalyzed.

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. Our analysis pipeline is lightweight and relies on well-documented and maintained software, making it extensible to other large cohorts of metagenomic sequencing data.

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)

Cohort	Cohort names	Country	Total	CD	UC	nonIBD	Reference
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, 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 (CITATION). K-mers are superior to alignment and assembly in metagenome analysis because: 1) k-mers enable exact matching, which is fast and requires little computational resources; 2) k-mers do not need to be present in reference databases to be included in analysis; and 3) k-mers capture information from reads even when there is low coverage or high strain variation, both of which preclude assembly. However, k-mers are complex sets given that there are approximately nˆk k-mers in a nucleotide sequence (e.g. 4.7 Mbp genome such as *Escherichia coli* K-12 (substrain MGI655) contains approximately 4.6 million k-mers).

However, only a fraction of these are needed to recapitulate similarity measurements using other metrics (Pierce et al. 2019). Thus we used scaled MinHash sketching as implemented in sourmash to produce a compressed representation of the k-mers contained in each sample (Pierce et al. 2019). 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 refer to the subsampled representative set of k-mers as a *signature*, and to each subsampled k-mer in a signature as a *hash*. 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.

Using this method, adapter sequences, human DNA, and sequencing errors can falsely inflate or deflate similarity measurements. As such, we also used a consistent preprocessing pipeline to adapter trim, remove human DNA, and k-mer trim (e.g. remove erroneous k-mers). Because k-mer trimming retains some erroneous k-mers, we further filtered signatures to retain hashes that were present in multiple signatures. This removed hashes that were likely to be errors while keeping hashes that were real but of low abundance in some signatures. There were 46,267,678 distinct hashes across all samples, of which 7,376,151 remained after filtering.

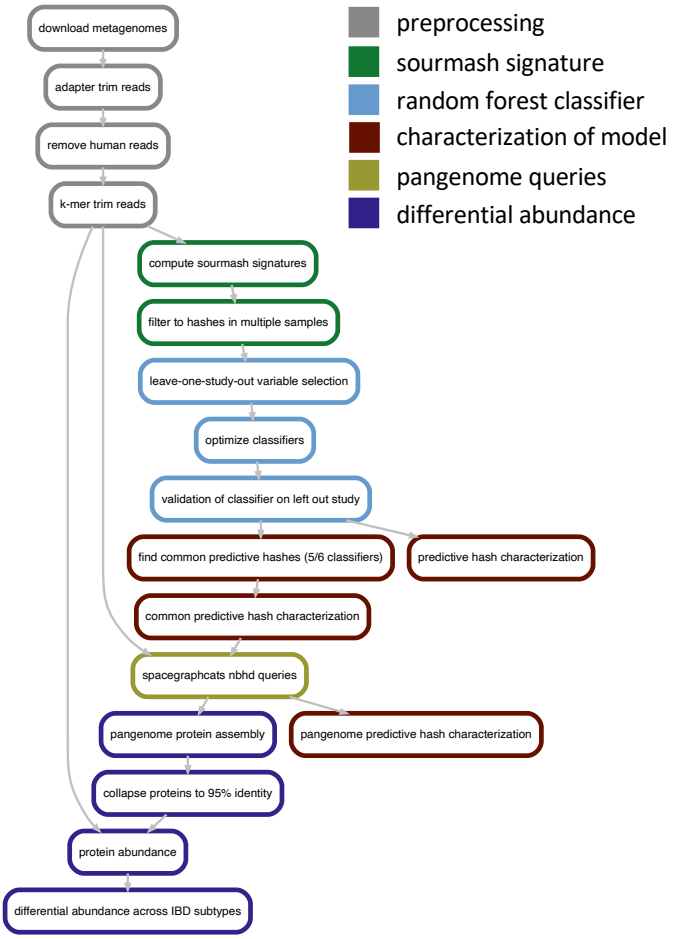

Figure 1: Overview of pipeline used in this paper.

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

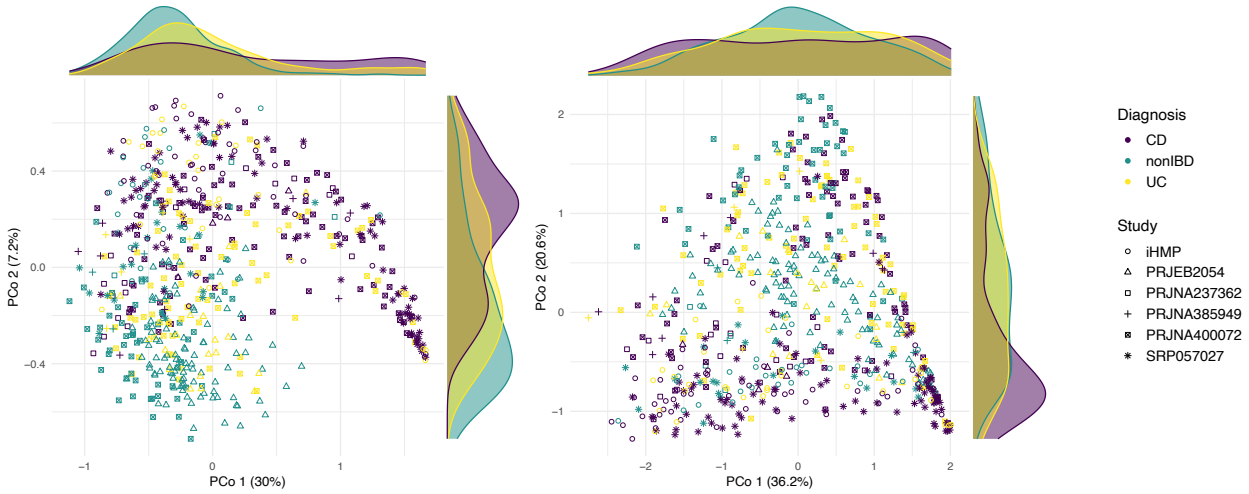

Figure 2: Principle coordinate analysis of metagenomes from IBD cohorts performed on filtered signatures. **A** Jaccard distance. **B** Angular distance.

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%*	
Study accession	6.6%*	

Variable	Jaccard distance	Angular distance
Diagnosis	6.2%*	
Library size	0.009%*	

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 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 (CITATIONS). 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 previously published models built on metagenomic data alone (CITATIONS).

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

Validation study	Selected hashes
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 reassembly efforts from human microbiome metagenomes (Pasolli et al. 2019; Nayfach et al. 2019; Almeida et al. 2019). Between 75.1-80.3% of of hashes anchored to 1,161 genomes (**Figure 4**). However, 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. In contrast to all hashes, only 69.4% of these hashes were identifiable, 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.

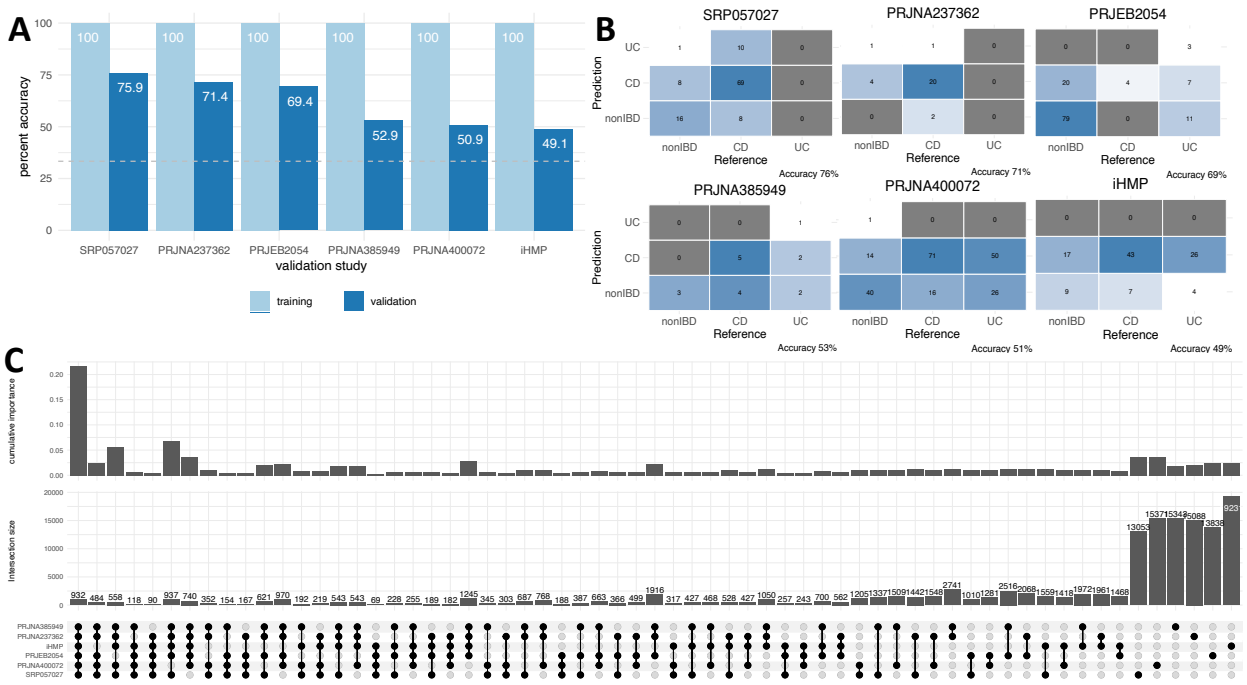

Figure 3: Random forest classifiers weakly predict IBD subtype. **A** Accuracy of leave-one-study-out random forest classifiers on training and validation sets. The validation study is on the x axis. **B** Confusion matrices depicting performance of each leave-one-study-out random forest classifier on the validation set. **C** Upset plot depicting intersections of sets of hashes with variable importance in each random forests classifier.

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 one or more distantly related genomes. 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. 86.1% of unknown hashes shared between at least five classifiers were in the pangenomes of

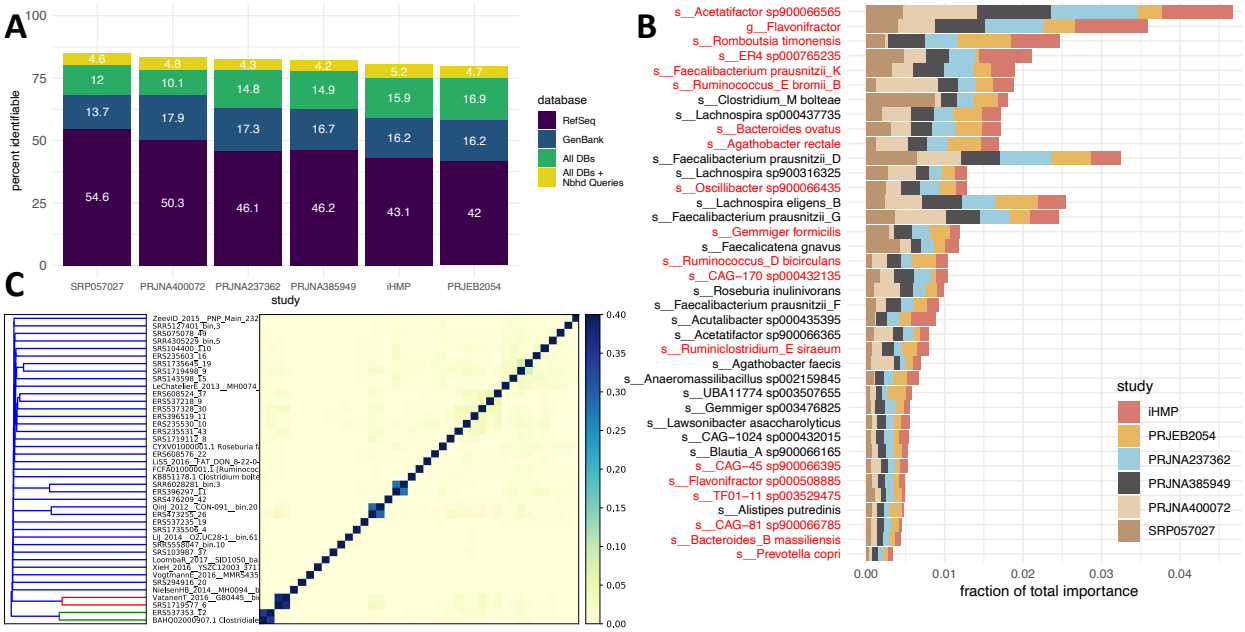

Figure 4: Some predictive hashes from random forest classifiers anchor to known genomes. **A** 75.1-80.3% of all hashes used to train classifiers anchor to known genomes in RefSeq, GenBank, or human microbiome metagenome-assembled genome databases. A further 4.2-5.6% of hashes anchor to pangenomes of 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 .37, while most genomes have no similarity. This indicates that each genome represents distinct nucleotide sequence.

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. These components are not recoverable by reference-based or *de novo* approaches, but are important for disease classification. 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 anchored by 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*. This suggests that pangenome queries might give a more complete picture of the strains involved in IBD (DOES THIS SUGGEST SOMETHING DIFFERENT/BETTER?).

Conversely, *Faecalibacterium prausnitzii_D* increased from anchoring ~2.9% to ~10.5% of the total variable importance, indicating that substantial pangenomic elements were hidden for this organism in particular. Or something.

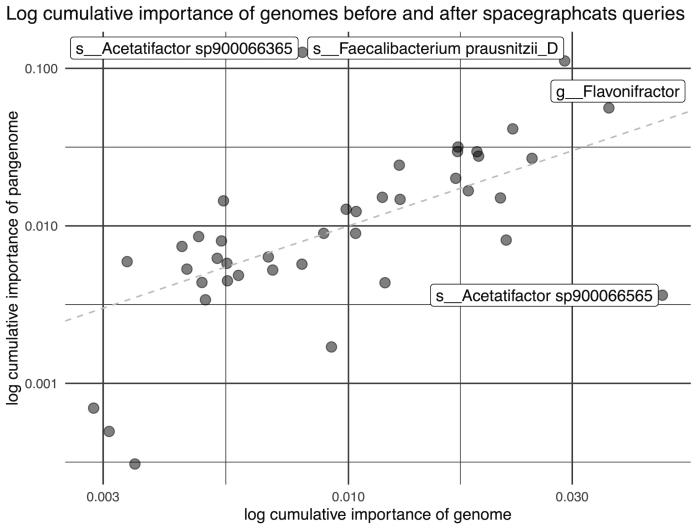

Figure 5: Pangenome neighborhoods reassign variable importance for some genomes.

Differential Abundance of Pangenomes

TBD

Discussion

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 idbudb.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 that have 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). 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; hashing k-mers to integers reduces storage space and improves computational run times when performing comparisons.

Although we adapter, quality, and k-mer trimmed our reads, some erroneous k-mers were likely included in the MinHash sequences, especially for low-coverage sequences. Therefore, we retained all hashes that were present in multiple samples. We 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) 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.

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 re-anchored these 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

Differential abundance

Compact de Bruijn graph queries for unknown predictive k-mers

We used the spacegraphcats `query_by_hashval` with a radius of 5 to retreive the compact de Bruijn graph neighborhood of unknown predictive k-mers.

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.

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.

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.

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.

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.

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.

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.

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*.