Supplementary Information

September 13, 2020

Description of IBD metagenome study cohorts

Below we present a description of each of the six cohorts used in this metaanalysis. 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 enrolment, 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 pipulation study and 43 patients with IBD were enrolled in a study at the Department of Gastroenterology and Hepatology 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.

Overview of pipeline

Accuracy of random forests classifiers

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 respository, “Comprehensive Human Microbiome Sourmash Databases” at the URL https://odf.io/hza89/. While we are

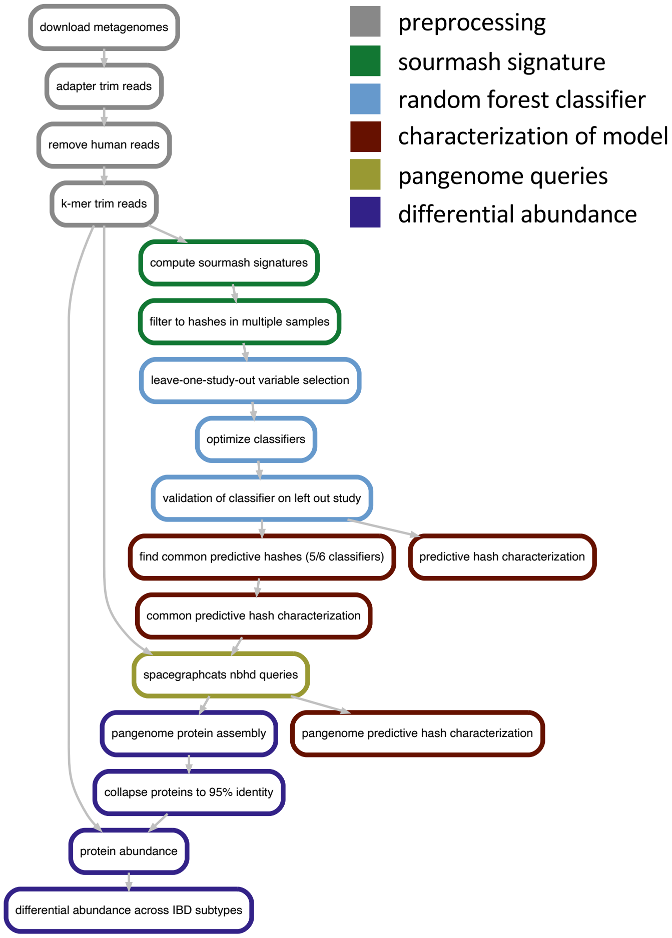
download metagenomes
adapter trim reads
remove human reads
k-mer trim reads
compute sourmash signatures
filter to hashes in multiple samples
leave-one-study-out variable selection
optimize classifiers
validation of classifier on left out study
find common predictive hashes (5/6 classifiers)
predictive hash characterization
common predictive hash characterization
spacegraphcats nbhd queries
pangenome protein assembly
pangenome predictive hash characterization
collapse proteins to 95% identity
protein abundance
differential abundance across IBD subtypes
preprocessing
sourmash signature
random forest classifier
characterization of model
pangenome queries
differential abundance

Figure 1: 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.

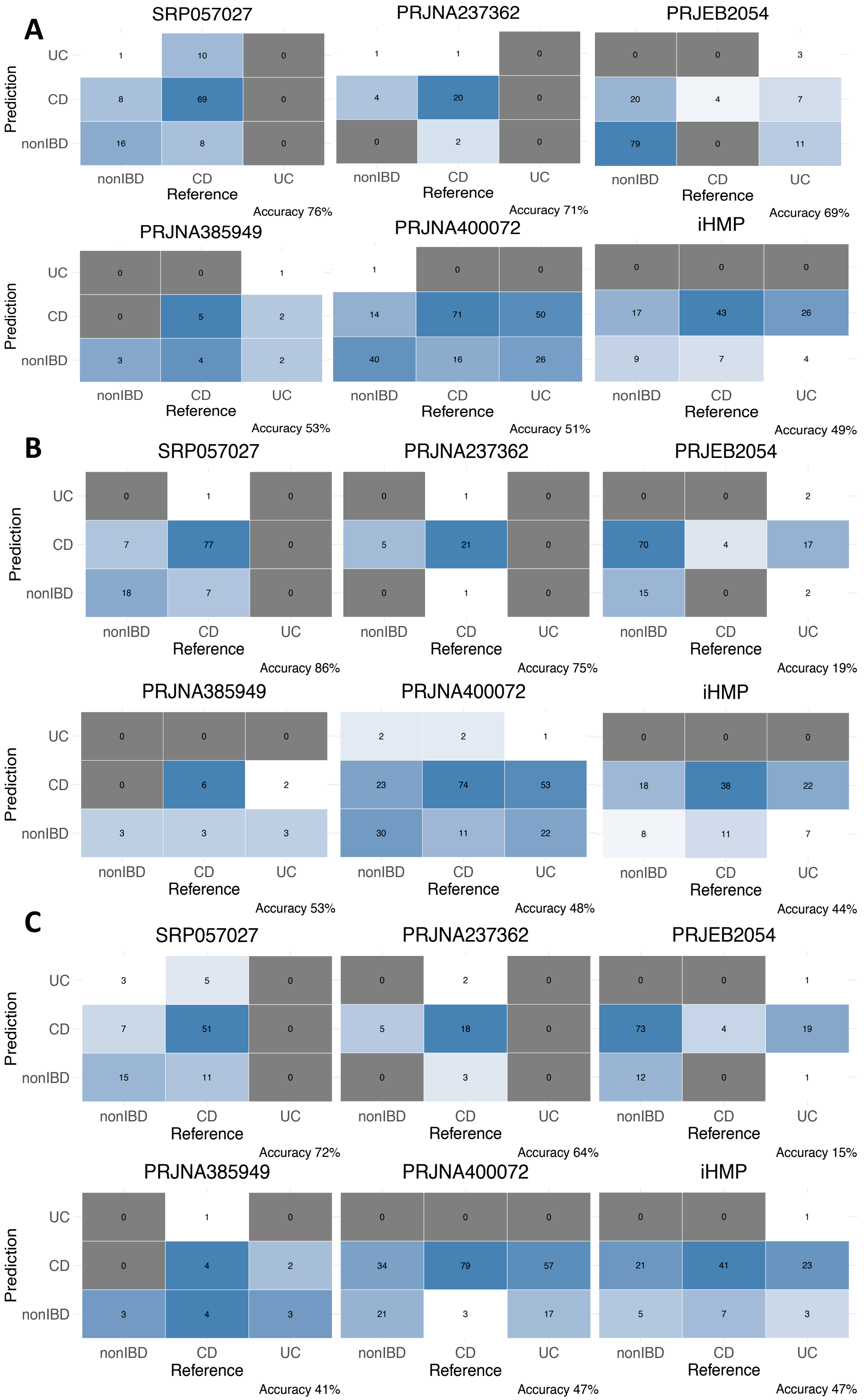

Figure 2: Confusion matrices for random forest model evaluated on the validation set. **A** hash model. **B** marker gene model. **C** hash model of marker genes.

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

41 genome accessions and taxonomy

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. UBA2087
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 VE202-03
GCF_001405615.1_13414_6_47_genomic.fna	s__Agathobacter faecis	Roseburia faecis strain 2789STDY5608863
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

genome	GTDB	NCBI
LiSS_2016__FAT_DON_8-22-0-0__bin.28.fa	s__CAG-170 sp000432135	Firmicutes bacterium CAG:124
LoombaR_2017__SID1050_bax__bin.11.fa	s__Anaeromassilibacillus sp002159845	Anaeromassilibacillus sp. An250
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; Clostridium sp. 42_12
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:65; clostridium
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:94
ZeeviD_2015__PNP_Main_232__bin.27.fa	s__Blautia_A sp900066165	uncultured Blautia sp.; Ruminococcus sp. Marseille-P328

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

Contamination in 41 shared genomes

We identified 41 genomes that were important for IBD subtype classification across six models. We used sourmash lca classify to assign 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

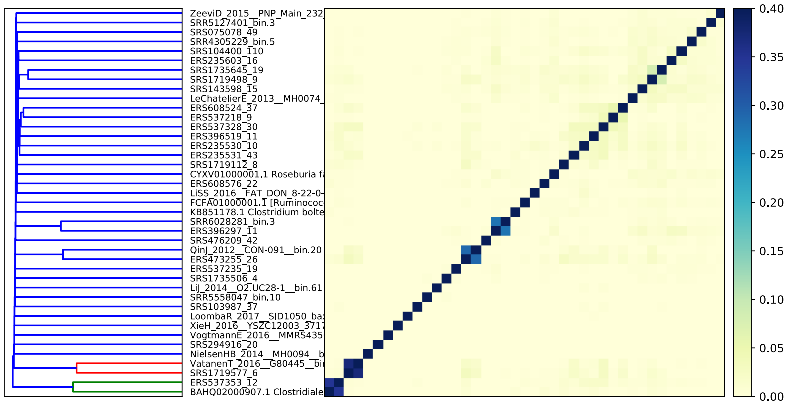

Figure 3: Jaccard similarity between 41 genomes. The highest similarity between genomes is 0.37 and is shared by genomes of the same species, while most genomes have no similarity. This indicates that each genome represents distinct nucleotide sequence.

we take the Jaccard index of these 41 genomes, we observe little similarity despite contamination (**Figure 3**). Therefore, we proceeded with analysis with the idea that each of the 41 genomes is a self-contained entity that captures distinct biology.

Characterization of unknown but predictive hashes through cDBG queries

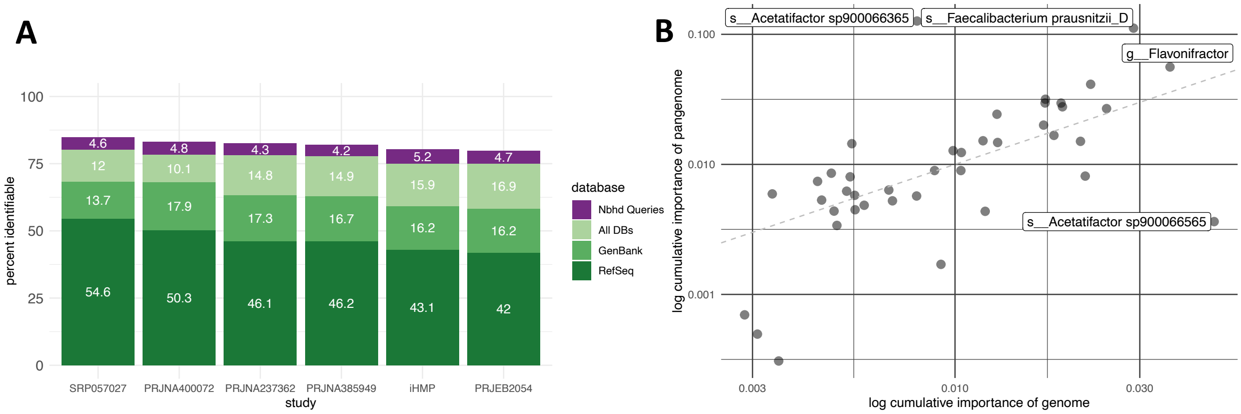

Figure 4: **A** Some hashes anchor to known genomes in RefSeq, GenBank, or human microbiome metagenome-assembled genome databases. An additional approximately 5% of hashes anchor to metapangenome of the 41 shared genomes. **B** Metapangenome neighborhoods generated with compact de Bruijn graph queries recover strain variation that is important for predicting IBD subtype. While the variable importance attributable to some genomes does not change with cDBG queries, other genomes increase by more than 7%.

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 compact de Bruijn graph queries into each metagenome sample with the 41 genomes that contained shared hashes (Brown et al. 2020), producing a pangenome for each query genome within each metagenome sample. Combining pangenomes from all metagenome, 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 novel 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 4**). The metapangenomes also captured 74.5% of all variable importance, a 24% increase over the 41 genomes alone. This indicates that strain variation contributes substantial predictive power toward IBD subtype classification.

Recovery of metapangenomic variation disproportionately impacts the variable importance attributable to specific genomes (**Figure 4**). 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).

Comparing IBD metagenome analysis by assembly

Table 1: Accuracy of model on each validation set.

Validation Study	Hash model	Ribosomal model	Gene model
SRP057027	75.9	86.4	44
PRJNA237362	71.4	75	NA
PRJEB2054	69.4	19.1	NA
PRJNA385949	52.9	52.9	35.3
PRJNA400072	50.9	48.1	50
iHMP	49.1	44.2	44.3

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 in each metapangenome through assembly and annotation. To build a gene catalogue for each metapangenome, we asesmbled each pangenome individually and extracted open reading frames (ORFs). We then clustered ORFs and ORF fragments from pangeomes in the metapangenome at 90% identity.

While the reads from all metapangenomes contain 90.9% of shared hashes, the metapangenome gene catalogues 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 1**). 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).

Given that many important hashes do not assemble and are therefore difficult to annotate, compared our shared hashes against the assemblies to determine which hashes assembled and which did not. Using gene neighborhood queries from the 41 shared genomes as described in the main text, we looked at the identity of unassembled hashes, many 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).

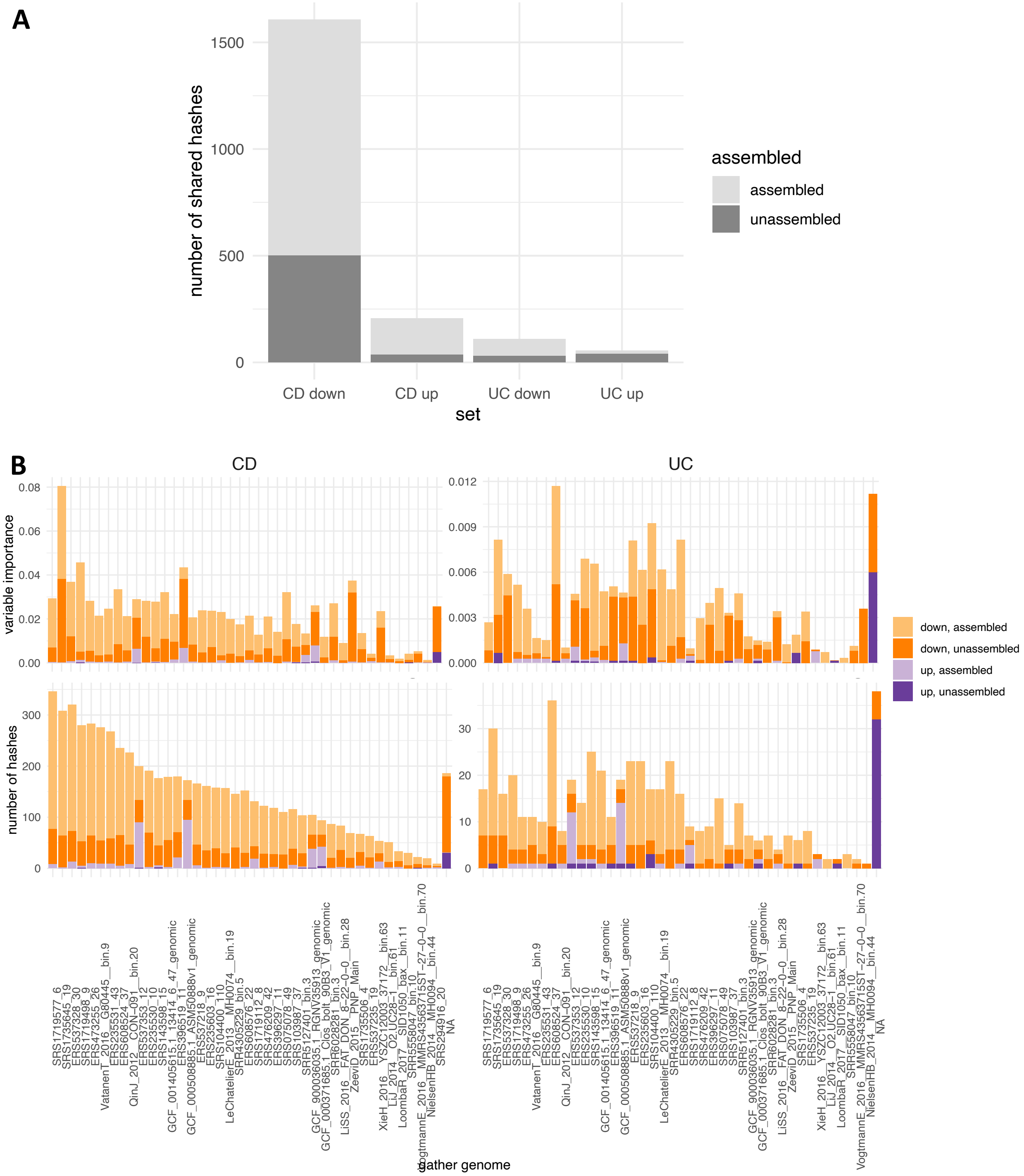

Figure 5: **A** A large fraction of shared hashes do not assemble. The largest fraction segregates to those that are less abundant in CD than nonIBD. **B** Unassembled shared hashes are distributed across the 41 shared genomes.

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 sepcies in the gut microbiome is a hallmark of IBD (CITATIONS), we first investigated whether the diversity of metapangeome 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 non-IBD (ANOVA $p < .05$, Tukey's HSD $p < .05$). 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* also 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 using read mapping information for the metapangenome gene catalogues. While our assemblies were incomplete, we reasonsed that by investigating the same set of genes for all samples. For most metapangenomes, the majority of genes are observed in CD, UC, and nonIBD. This in part explains heterogenous study findings in IBD gut microbiome investigations (CITATIONS) and underscores that IBD is a spectrum of diseases characterized by intermittent health and dysbiosis.

ADD A GENE ACCUMULATION CURVE PANEL

One of 41 metapangenome accumulation curves did not saturate for UC, while 10 did not saturate for CD. *C. bolteae* does not saturate in UC. One hundred seventy-one of 16,822 genes were not observed in UC, many of which had no annotated function.

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

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).

Supplementary 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.

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.

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.

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, 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.

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.

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.

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.

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.

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.

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.

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.

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

Woodcroft, B. 2018. "Singlem."

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