Meta-analysis of metagenomes via machine learning and assembly graphs reveals strain switches in Crohn's disease

This manuscript (permalink) was automatically generated from taylorreiter/2021-paper-ibd@c3f44d1 on May 26, 2022.

Authors

- **Taylor E. Reiter**
  0000-0002-7388-421X · taylorreiter · ReiterTaylor
  Department of Population Health and Reproduction, University of California, Davis · Funded by Grant XXXXXXXX

- **Luiz Irber**
  0000-0003-4371-9659 · luizirber · luizirber
  Graduate Group in Computer Science, UC Davis; Department of Population Health and Reproduction, University of California, Davis · Funded by Grant XXXXXXXX

- **Alicia A. Gingrich**
  0000-0002-7239-0154 · alicia-gingrich
  Department of Surgery, University of California, Davis Medical Center

- **Dylan Haynes**
  0000-0001-8986-8196 · dylan-haynes
  School of Medicine, Oregon Health & Science University

- **N. Tessa Pierce-Ward**
  0000-0002-2942-5331 · bluegenes · saltyscientist
  Department of Population Health and Reproduction, University of California, Davis · Funded by NSF 1711984

- **Phillip T. Brooks**
  0000-0003-3987-244X · brooksph · brooksph
  Department of Population Health and Reproduction, University of California, Davis

- **Yosuke Mizutani**
  · mogproject · mogproject
  School of Computing, University of Utah

- **Dominik Moritz**
  0000-0002-3110-1053 · domoritz · domoritz
  Human-Computer Interaction Institute, Carnegie Mellon University

- **Felix Reidl**
  0000-0002-2354-3003 · micrograviteas · quantumgravitas
  Department of Computer Science and Information Systems, Birkbeck, University of London

- **Amy D. Willis**
  0000-0002-2802-4317 · adw96 · AmyDWillis
  Department of Biostatistics, University of Washington

- **Blair D. Sullivan**
  0000-0001-7720-6208 · bdsullivan · BlairDSullivan
  School of Computing, University of Utah

- **C. Titus Brown**
  0000-0001-6001-2677 · ctb
  Department of Population Health and Reproduction, University of California, Davis

Abstract

Microbial strains have closely related genomes but may have different phenotypes in the same environment. Shotgun metagenomic sequencing can capture the genomes of all strains present in a community but strain-resolved analysis from shotgun sequencing alone remains difficult. We developed an approach to identify and interrogate strain-level differences in groups of metagenomes. We first developed a machine learning classifier based on compressed representations of complete metagenomes (FracMinHash sketches) and identified genomes that correlate with IBD subtype. We next we used assembly graph genome queries to recover strain variation for correlated genomes. To rescue variation that may not have been present in the sketches, we then used assembly graph genome queries to recover strain variation for correlated genomes. Lastly, we developed a novel differential abundance framework that works directly on the metagenome assembly graph to uncover all sequence variants correlated with IBD. We refer to this approach as dominating set differential abundance analysis and have implemented it in the spacegraphcats software package (github.com/spacegraphcats/spacegraphcats). We use these techniques to perform a meta-analysis of stool microbiomes from individuals with and without inflammatory bowel disease (Crohn's disease, ulcerative colitis; n = 605). We identified five bacterial strains that are associated with Crohn's disease. Our method captures variation within the entire sequencing data set, allowing for discovery of previously hidden disease associations.

Introduction

Sub-species groupings of microorganisms have functional differences that govern important genome-environment interactions across diverse ecosystems. For example, ecotypes of *Escherichia coli* have different gene complements that allow each group to thrive in diverse environments like the gut, soil, and freshwater [1]. Metagenomic sequencing data from communities of microorganisms contain information about specific strains present in a sample, but strain-resolved insights are lacking due to incomplete references or inability of current tools to retrieve such information [2]. Here we use *strain* to refer to within-species variation that generates taxonomic grouping below the species level.

Inflammatory bowel disease (IBD) is a group of disorders that are characterized by chronic inflammation of the intestines that is likely caused by host-mediated inflammatory responses elicited in part by microorganisms [3]. IBD classically manifests in three subtypes depending on clinical presentation, including Crohn's disease (CD), which presents as discontinuous patches of inflammation throughout the gastrointestinal tract, ulcerative colitis (UC), which presents as continuous inflammation isolated to the colon, and undetermined, which cannot be distinguished as CD or UC. Diagnosis can be clinically difficult, with ramifications associated with over- or under-treatment resulting in patient morbidity associated with inappropriate treatment. Detection of microbial signatures associated with IBD subtype may lead to improved diagnostic criteria and therapeutics that extend periods of remission. However, such signatures have thus far remained elusive [4].

The microbiome of CD and UC is heterogeneous, and studies that characterize the microbiome often produce conflicting results [4]. This is likely in part driven by large inter- and intra-individual variation [5], but is also attributable to non-standardized laboratory, sequencing, and analysis techniques used to profile the gut microbiome [4]. Dysbiosis is frequently observed in IBD, particularly in CD [6,7,8,9,10], however dysbiosis alone is not a signature of IBD [5]. *Dysbiosis* is defined as a decrease in gut microbial diversity that results in an imbalance between protective and harmful microorganisms, leading to intestinal inflammation [11].

Strain-level differences may account for some heterogeneity in IBD gut microbiome profiles. A recent investigation of time-series gut microbiome metagenomes found that one clade of *Ruminococcus gnavus* is enriched in CD [12]. Further, this clade produces an inflammatory polysaccharide [13,14]. While this clade is enriched in CD, its enrichment was previously masked from computational discovery by concomitant decreases in other *Ruminococcus* species in IBD [12], highlighting the need for strain-resolved analysis of metagenomic sequencing in the exploration of IBD gut microbiomes.

Given these features of the IBD gut microbiome, strain-resolved analysis may yield insights into the dynamics of these communities. The two biggest obstacles to strain-level analysis of short read data are the lack of strain representation in databases together with the challenge of haplotype-level resolution in assembly and binning. While long reads have made strides toward resolving the latter issue [15], in habitats like the gut where communities are dominated by single strains of microbes [16] the largest barrier to strain-level analysis is using data that does not match to reference databases. New data analysis techniques are needed to make full use of strain level data.

K-mers, words of length *k* in nucleotide sequences, have previously been used for annotation-free characterization of sequencing data [17,18,19]. K-mers are suitable for strain-resolved metagenome analysis because their absence in reference databases does not preclude their analysis. Moreover, k-mer analysis does not rely on marker genes which are largely conserved at the strain level, and k-mers are suitable for species- and strain-level classification [20,21]. Investigating all k-mers in metagenomes is more computationally intensive than reference-based approaches [22], but data-reduction techniques like FracMinHash sketching make k-mer-based analysis scalable to large-scale sequence comparisons [23,24]. FracMinHash sketching sacrifices the fine-scaled resolution of reference-based techniques but is representative of the full sequencing sample and can make use of all available genomes [21], thus including strain-variable accessory elements that may be associated with diseases

Assembly graphs complement sketch analysis [25,26]. (CTB: note that this is a bit prospective and might be hard to justify in the intro; I think we're the only people doing it :). While both k-mers and assembly graphs can be used to represent all sequences contained within a metagenome, assembly graphs retain important sequencing context and can aggregate known functional and taxonomic annotations, recovering critical information lost through sketching approaches. While assembly graphs have been leveraged in metagenome analyses [28], their large size precludes analysis at scale. The *spacegraphcats* tool is designed to tackle this issue, implementing algorithms that efficiently reduce the size of an assembly graph, enabling rapid querying and sequence retrieval [25]. These algorithms center around dominating sets, which partition the graph into *pieces* by assigning every node to a graph-localized neighborhood [25]. This simplified graph enables efficient queries: querying with a sequence that overlaps any k-mer in a compact de Bruijn graph (cDBG) node returns all k-mers (or all reads containing those k-mers) from the graph neighborhood. Genome queries often recover sequences not in reference databases or *de novo* assemblies, which disproportionately include sequences from both low coverage regions and highly variable portions of the graph (e.g. sequencing reads that neither assemble nor bin) [25]. When a query has a containment index between 10^{-2} and 10^{-3} with the assembly graph, 20-40% of a target genome sequence is recovered from a metagenome query, and for containment indices above 10^{-1} this increases to >80%. [25].

Here, we develop k-mer- and assembly graph-based techniques to perform a meta-analysis of stool microbiome metagenomes from individuals with (CD, UC) and without (nonIBD) IBD [5,8,10,12,29,30]. Using these approaches, we detect a consistent signature of IBD subtype in fecal microbiome metagenomes. We identify a small set of k-mers that are predictive of UC and CD, and find that these k-mers originate from a core set of microbial genomes. We find that a stochastic loss of diversity in this core set of microbial genomes is a hallmark of CD, and to a lesser extent, UC. (CTB: do we need to say "as previously shown via 16s" or something here? with citation?) While reduced diversity is responsible for the majority of disease signatures, we also find signatures of strains present in the disease state. Sequences associated with these strains occurred more frequently in IBD metagenomes

but are present in low abundance in nonIBD metagenomes as well. Our approach provides a solution for strain-level analysis of short read metagenomic data sets, and our findings provide future avenues for research into IBD therapeutics.

Results

We developed a computational approach to resolve sub-species level differences between groups of short read shotgun metagenomes (**Figure 1**). While our pipeline relies on many published algorithms, we developed two new approaches that, when combined with existing tools, generated insights into microbial sequences associated with IBD subtype. After consistent pre-processing, we used FracMinHash sketching to produce subsampled k-mer abundance profiles of metagenomes that reflected the sequence diversity in a sample [21,23], and used these profiles to perform metagenome-wide k-mer association with IBD subtype. We refer to FracMinHash sketches as *signatures*, and for simplicity, continue referring to the sub-sampled k-mers in a signature as *k-mers*. Retaining only k-mers associated with IBD, we used a minimum set cover approach to identify the genomes that best encompassed these k-mers [21].

Next, we developed an approach to perform differential abundance analysis directly on assembly graphs in order to recover all sequences that were differentially abundant in each IBD subtype when compared to nonIBD. Using the genomes identified by our k-mer association analysis, we first performed assembly graph genome queries to recover all sequences associated with a given species within a metagenome. For each genome query, we combined these sequences into a single assembly graph, which we refer to as a *metapangenome graph*. We estimated the abundance of each piece in this graph within each metagenome, and used these abundances to perform differential abundance analysis.

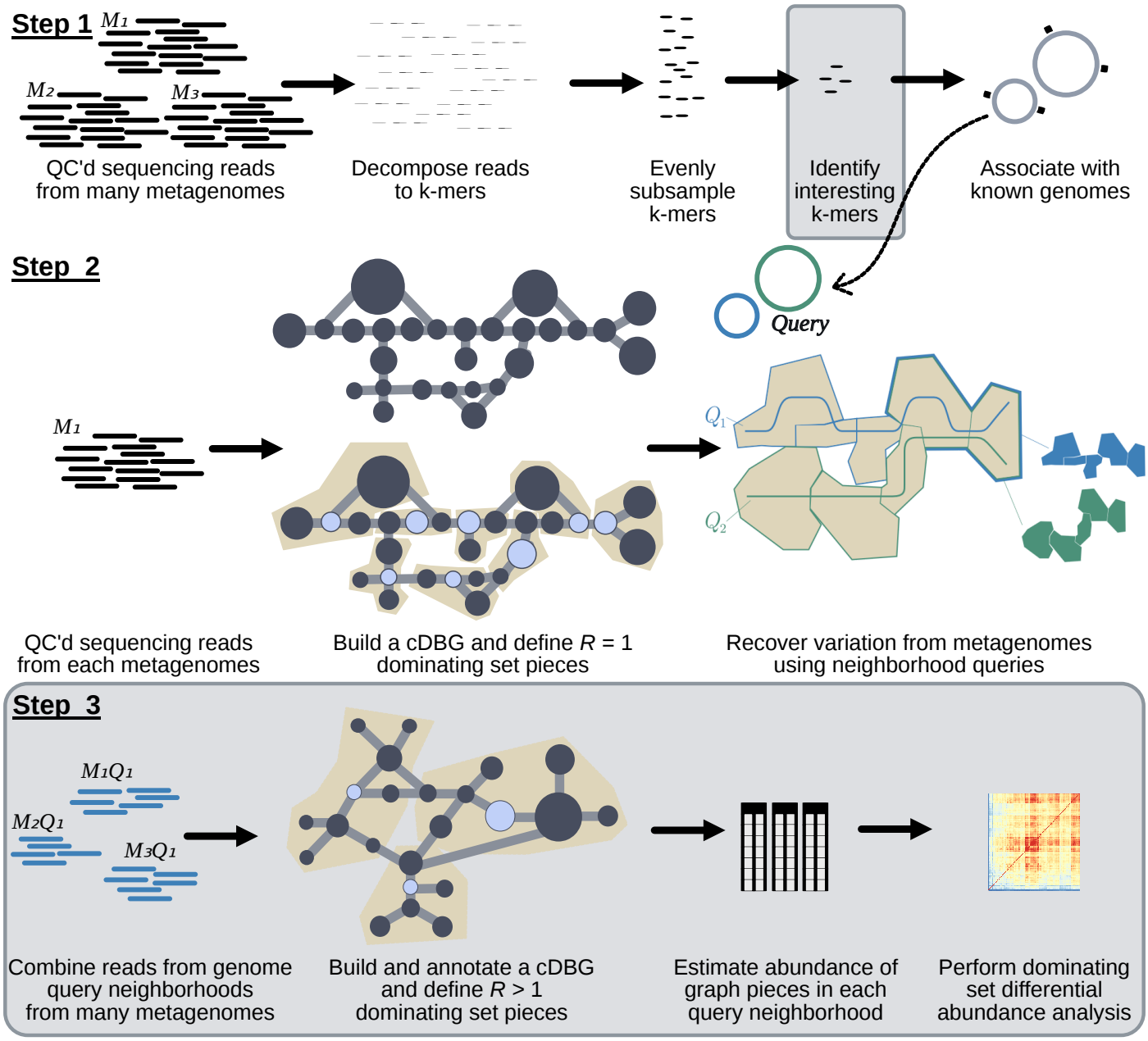


Figure 1: Overview of the metagenome analysis technique. Steps that are outlined in grey were developed in this paper. **Step 1:** Using quality controlled sequencing reads from many metagenomes, we decomposed reads into k-mers and subsample these k-mers using FracMinHash, thereby selecting k-mers that evenly represent the sequence diversity within a sample. We then identified interesting k-mers using random forests, and associate these k-mers with genomes in reference databases. **Step 2:** For each metagenome, we constructed a compact de Bruijn assembly graph that contains all k-mers from a metagenome. We used dominating sets to carve the graph into pieces. We queried this graph with genomes associated with interesting k-mers identified in Step 1, recovering sequence diversity nearby in the assembly graph. We refer to these sequences as genome query neighborhoods. Step 2 is the workflow published in [25]. **Step 3:** We combined genome query neighborhoods for a single genome from all metagenomes. We constructed a compact de Bruijn assembly graph from these sequences, and used a dominating set with a large radius to carve the graph into large pieces. Here, we diagram construction of R=2 dominating set pieces, but in practice we used R=10. We estimated the abundance of k-mers in each metagenome for each dominating set piece, and used these abundances to perform differential abundance analysis.

We applied this approach to the analysis of IBD gut microbiomes via meta-analysis. Meta-analyses have recently shown success in improving detection of microbial signatures of colorectal cancer [31,32,33]. To this end, we identified studies that performed metagenomic sequencing of individuals with CD, UC, or nonIBD and combined these to perform a meta-analysis (**Table 1**). All studies profiled fecal gut microbiomes via Illumina shotgun metagenome sequencing. Individuals were from five distinct countries and seven cohorts (**Table 1**). In many studies, samples were taken in time series to profile disease progression or individual response to treatment. In these cases we included only the

first sample in the time series so organized interventions would not skew our results. In addition, many of the nonIBD samples, particularly those from the iHMP, profiled sick individuals that were not diagnosed with IBD, meaning some of these samples are not healthy controls.

Table 1: Six IBD shotgun metagenome sequencing cohorts used in this meta-cohort analysis.

Cohort	Name	Country	Total	CD	UC	nonIBD	Reference
iHMP	IBDMDB	USA	106	50	30	26	[5]
PRJEB2054	MetaHIT	Denmark, Spain	124	4	21	99	[10]
SRP057027	NA	Canada, USA	112	87	0	25	[8]
PRJNA385949	PRISM, STiNKi	USA	17	9	5	3	[12]
PRJNA400072	PRISM, LLDeep, and NLIBD	USA, Netherlands	218	87	76	55	[29]
PRJNA237362	RISK	North America	28	23	0	5	[30]
Total			605	260	132	213	

K-mers are weakly predictive of IBD subtype

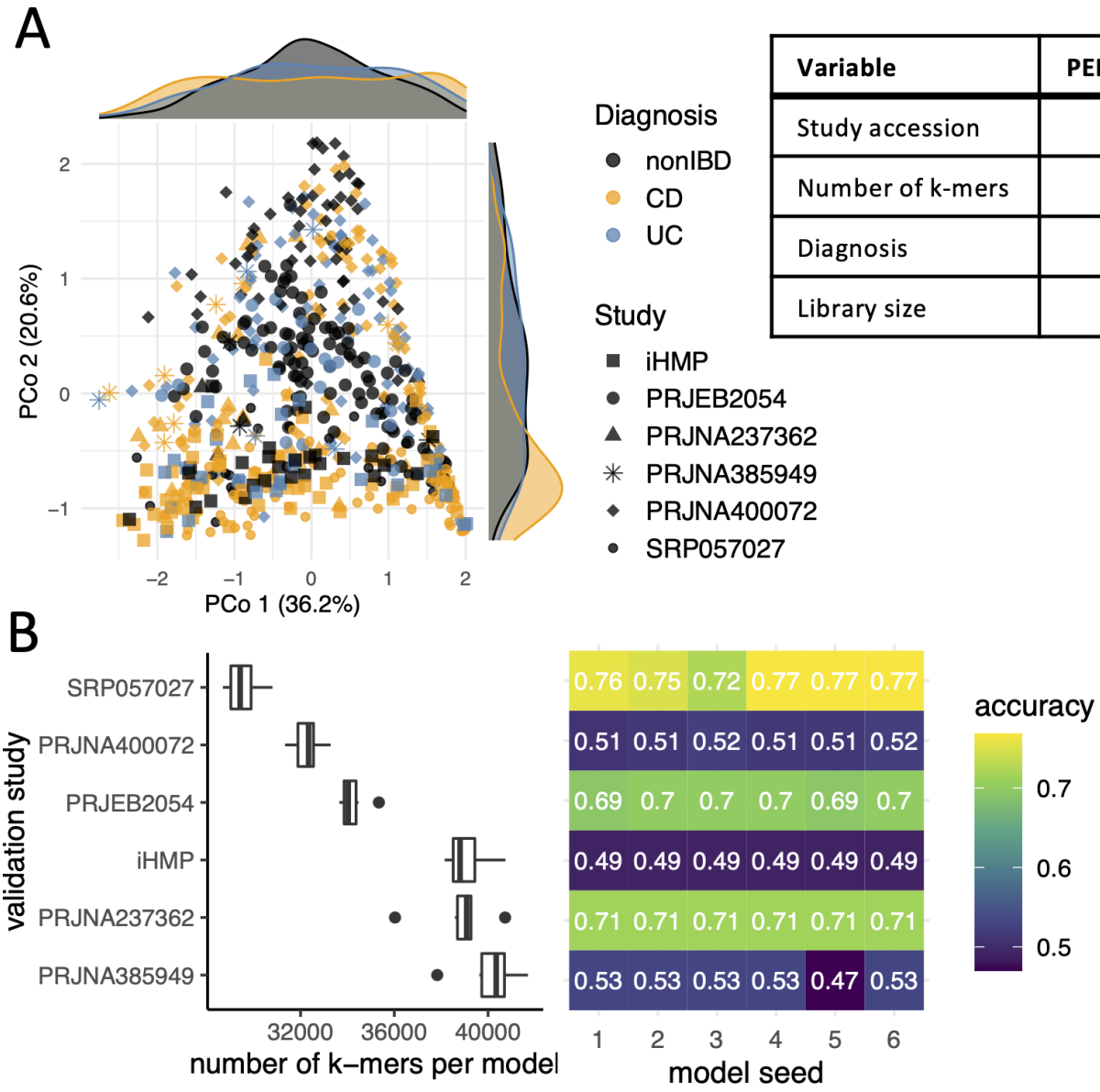
We first sought an approach to compare many metagenomes without relying on reference databases, *de novo* assembly, or annotations. We reasoned that FracMinHash sketches randomly subsample k-mers to allow comparisons, which may provide an unbiased approach to quickly compare across many metagenomes. In total, we profiled 7,376,151 subsampled k-mers across all samples in all cohorts, representing approximately 14 billion distinct k-mers in the original samples.

We detected variation correlated with IBD diagnosis in k-mer profiles of gut metagenomes from different cohorts. We calculated a pairwise distance matrix using angular distance between k-mer abundance profiles to assess sample diversity. We performed principle coordinate analysis and PERMANOVA with this distance matrix (**Figure 2 A**), using the variables study accession, diagnosis, library size, and number of k-mers observed in a sample (**Figure 2 A**). Study accounted for highest variation, emphasizing that technical artifacts can introduce strong signals that may influence heterogeneity in results across IBD microbiome studies but that can be mitigated through meta-analysis [31]. The number of k-mers observed in a sample accounted for the second highest variation, possibly reflecting reduced diversity in stool metagenomes of CD and UC patients (reviewed in [34]). Diagnosis accounted for a substantial amount of variation as well, indicating that there is a small but detectable signal of IBD subtype in stool metagenomes.

To evaluate whether the variation captured by diagnosis is predictive of IBD subtype, we built random forest classifiers to predict CD, UC, or nonIBD subtype. Random forest techniques are a supervised learning classification model that estimates how predictive k-mers are of IBD subtype, and weight individual k-mers as more or less predictive using a metric called variable importance. To assess whether disease signatures generalize 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. We built each model six times, using a different random seed each time, to hone in on cross-study and cross-model signal. Given the high-dimensional structure of this data set (e.g. many more k-mers than metagenomes), we first used variable selection to narrow the set of predictive k-mers in the training set [35,36]. Variable selection reduced the number of k-mers used in each model by two orders of magnitude, from 7,376,151 to 28,684-41,701 (mean = 35,673.1, sd = 4090.3) (**Figure 2 B**).

Using this reduced set of k-mers, we optimized each random forests classifier on the training set, producing 36 optimized models. We validated each model on the left-out study. The accuracy on the

validation studies ranged from 49%-77% (**Figure 2 B**), outperforming a previously published model built on metagenomic data alone [29].



Variable	PERMANOVA
Study accession	13.5%
Number of k-mers	6.2%
Diagnosis	3.3%
Library size	0.01%

Figure 2: Long nucleotide k-mers retain information about IBD subtype classification. A. Principal coordinate analysis of distance matrices obtained from comparing FracMinHash signatures with abundances and PERMANOVA results that explain the variance. Number of k-mers refers to the number of k-mers in a signature, while library size refers to the number of raw reads per sample. All test were significant at $p < .001$. **B.** Random forests models built on FracMinHash signatures predicted IBD subtype better than chance. Variable selection reduced the number of k-mers used to build each model, and model performance varied by validation study.

To understand which genomes were responsible for disease signatures detected by our models, we anchored k-mers in the models against genomes in reference databases using sourmash gather [21]. Sourmash gather determines the minimum set of genomes in a database necessary to cover all of the k-mers in a query [21]. We used the GTDB rs202 representatives database, which contains bacterial and archaeal genomes, and the GenBank viral, fungal, and protozoa databases. We found that a substantial fraction of genomes were shared between models, indicating there is a consistent biological signal captured among classifiers: of 3,889 total genomes detected across all classifiers, 360 genomes were shared between all classifiers (**Figure 3**, **Figure 7**). The presence of shared k-mers between classifiers indicates that there is a weak but consistent biological signal in metagenomes for IBD subtype between cohorts.

K-mers that anchored to these shared genomes represented 65% of all k-mers used to build the optimized classifiers, but accounted for an outsize proportion of variable importance in the optimized classifiers. After normalizing variable importance across classifiers, 76% of the total variable importance was held by these k-mers. These k-mers contribute a large fraction of predictive power for classification of IBD subtype, and the genomes in which they are found represent a microbial core that contains predictive power in IBD subtype classification.

Given that 360 genomes anchored the majority of k-mers and variable importance across all models, we were curious whether a smaller number of genomes could still retain the majority of variable importance. Limiting genomes to those that could hold at least 1% of the normalized variable importance, we found that 54 genomes accounted for 50% of the variable importance (**Figure 3**, **Figure S 7**, **Figure S 8**). We assumed these genomes represent the strongest candidates for discriminating IBD subtype and focused on them for the remainder of our analyses.

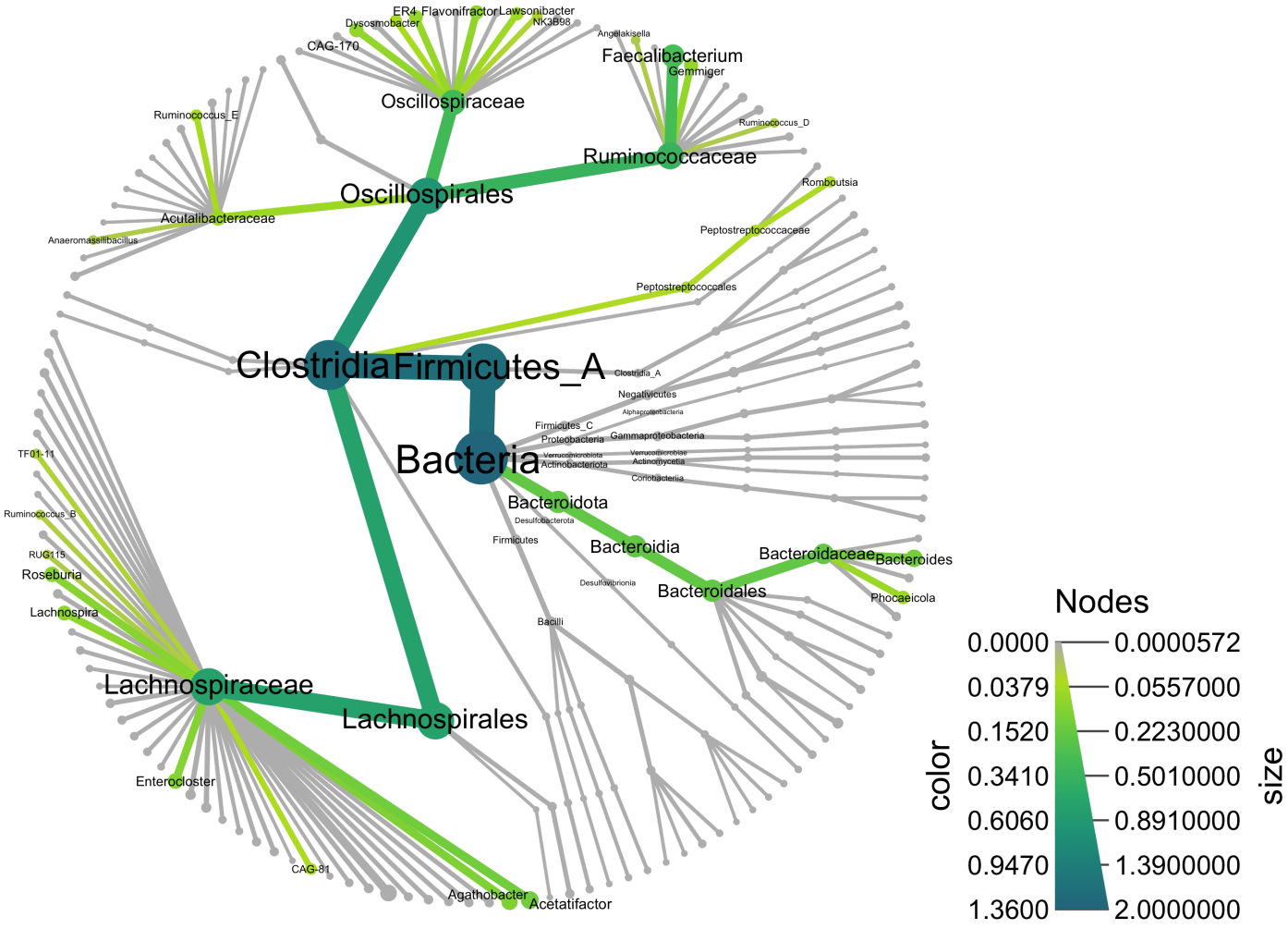


Figure 3: Tree of bacterial species that were predictive of IBD subtype in all models. Nodes are summarized to the genus level. All taxa up to the class level are labelled. Taxa that could account for at least 1% of the normalized variable importance across Random forests models are colored and labelled. Node size and node color reflects potential variable importance.

Genome queries into metagenome assembly graphs recover neighborhoods of sequence variation and establish species umbrellas

While we were able to identify the majority of k-mers that were important for predicting IBD subtype, 26% of k-mers remained unannotated. We hypothesized that these k-mers represented strain variable

sequences not in reference databases but belonging to species represented by annotated k-mers. To test this hypothesis, we performed genome queries on assembly graphs of each metagenome using the 54 candidate genomes that discriminated IBD subtype (**Figure 1**). Assembly graph genome queries recover sequences in a metagenome that match the query, as well as those that are nearby in the assembly graph, making queries akin to but more sensitive than read mapping against reference genomes (**Figure 1**) [25]. The resulting genome query neighborhood represents a species-level umbrella that contains sequence variation from the metagenome associated with a query.

After performing genome queries, we re-anchored k-mers against the resulting query neighborhoods as well as the databases used previously. We observed that the percent of unassigned k-mers decreased from 26% to 8% (**Figure 4**), supporting our hypothesis that many of these k-mers are sequence variants belonging to species identified in k-mers important for predicting IBD subtype. We further observed that many other k-mers previously anchored by other genomes were reassigned to the genome query neighborhoods (**Figure 4**). This suggests that the genome queries create species umbrellas that represent sequence variation for the query genome itself as well as other closely related genomes that occur within a metagenome.

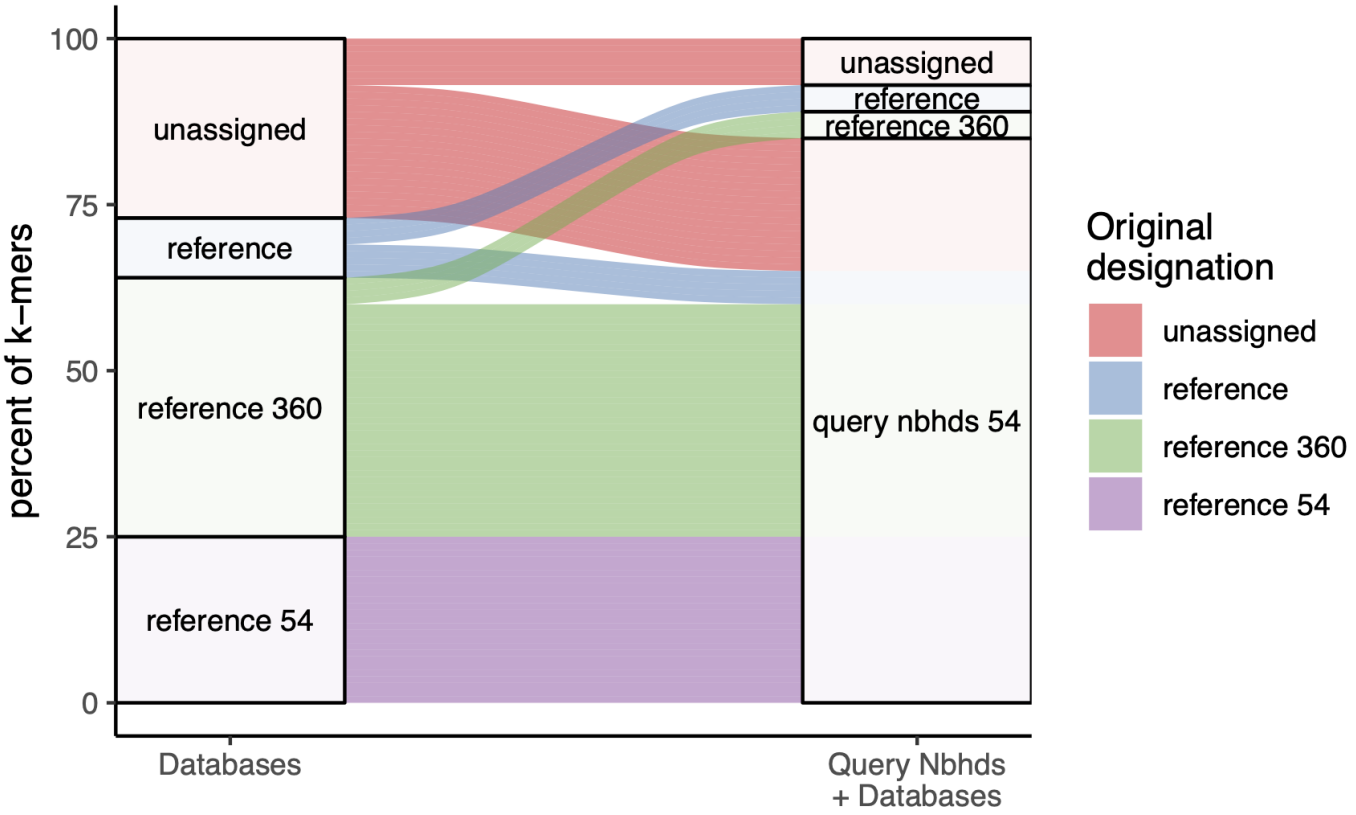


Figure 4: Alluvial plot depicting the set of genomes that anchored k-mers that were important for predicting IBD subtype. The blocks on the left represent the breakdown of k-mer assignations from greedy exact matching against databases alone, while the blocks on the right represent k-mer assignations after metagenome assembly graph queries.

IBD gut microbiomes have decreased diversity in strict anaerobes that is punctuated by strain switches for some facultative anaerobes

After recovering all sequences in metagenomes in the neighborhoods of the species that discriminate IBD subtypes, we next sought to determine the specific genome segments that were differentially abundant in IBD. Differential abundance analysis is a common step in metagenome analysis, however it is typically applied to gene counts [37,38], which requires assembly or mapping prior to abundance

estimation. To avoid assembly or mapping and the accompanying loss of reads [39], we developed an abundance estimation approach that works directly on the assembly graphs, enabling differential abundance analysis from the assembly graph. Our abundance estimation approach was based on *R*-dominating sets, an algorithm introduced in [25] that efficiently computes the dominating nodes in a cDBG so that every node is no more that distance *R* from a dominator. The dominating set is used to carve the graph into pieces, each of which contains one dominating node. Here, we first build a species-level assembly graph that contains neighborhood sequences for a given genome across all metagenomes, which we call a *metapangenome graph*. We then partition the graph into pieces using a large radius ($R = 10$). The large radius carves the graph into pieces that average 103 k-mers in size. We next estimated the abundance of each piece within each metagenome using average k-mer abundance. We also annotated the graph pieces using XXX. Using this information, we performed dominating set differential abundance analysis using corncob [40], a statistical package that tests for differential relative abundance in the presence of variable sequencing depth and excessive zeroes for unobserved observations, conditions which occur in abundances from dominating sets.

We applied this method for each of our genome queries, building 54 metapangenome graphs and performing dominating set differential abundance analysis on each. We tested for differential abundance in pieces that occurred in at least 100 metagenomes, since we sought differences that characterized the majority of our samples within a group. Note that corncob fits a model for each dominating set piece and therefore does not require abundance information for all pieces [40]. On average, this condition was met in 6.4% of dominating set pieces. Focusing on pieces that occurred in many metagenomes increased the average piece size to 1088 k-mers, which is similar to the average bacterial gene length of approximately 1000 base pairs [41] and should enable biologically meaningful comparisons across groups.

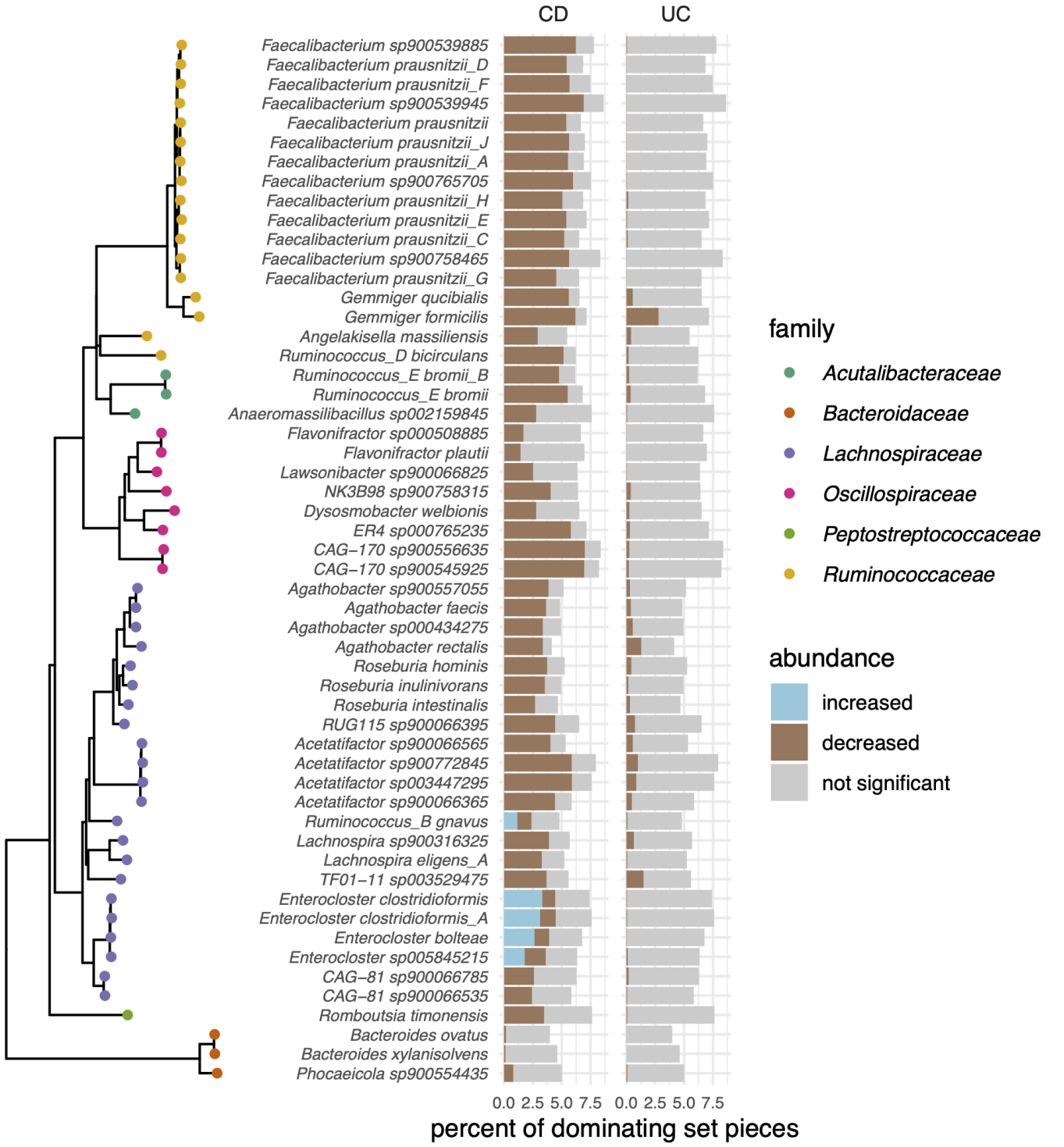


Figure 5: Dominating set differential abundance analysis revealed genome segments that were significantly different in CD and UC compared to nonIBD. Results are organized by GTDB taxonomy, with a tree representing the 54 species and colored by family on the far left. The percent of dominating set pieces tested is labelled in grey, and the percent of significantly differentially abundant pieces are colored by increased (blue) or decreased (brown) abundance.

We found that the majority of species decreased in abundance in CD, and to a lesser extent, UC (**Figure 5**). Many of these species are generally regarded as beneficial bacteria. For example, nine of the 54 genomes we investigated were *Faecalibacterium prausnitzii*, the phylogroups of which are separated in the GTDB taxonomy but combined into a single species in the NCBI taxonomy. *F. prausnitzii* is a key butyrate producer in the gut and plays a crucial role in reducing intestinal inflammation [42]. Similarly, *Acetatifactor* is a bile-acid producing bacteria associated with a healthy gut, but limited evidence has associated it with decreased abundance in IBD [43]. These species, as well as others that decreased in abundance in IBD, are strictly anaerobic (**Figure 5**), so these observed

trends are consistent with a shift toward oxidative stress during disease that is intolerable for many gut microbes [44].

A substantial fraction of dominating set pieces were more abundant in CD than nonIBD in five metapangenome graphs (**Figure 5**). These graphs represented sequences from species *R. gnavus*, *Enterocloster bolteae*, *Enterocloster sp005845215*, *Enterocloster clostridioformis*, and *Enterocloster clostridioformis_A*. We posit that the increased abundance in some genomic segments amid the decrease in abundance of others represents strain switching that occurs in CD.

In support of this, when we annotated the differentially abundant pieces using PFAM orthologs, we found that in some cases pieces that were increased in abundance and pieces that were decreased in abundance were annotated with the same ortholog (average XX per graph, SD). These genes likely represent the portion of the core genome shared by the strain(s) that is more abundant in CD and the strain(s) that is more abundant in nonIBD, but that is encoded by distinct sequences (PULL OUT MARKER GENES, MAKE FIGURE).

Many orthologs were only annotated among the pieces that were increased in abundance in CD. Among all five metapangenome graphs, XX orthologs were annotated... XX pathways were enriched (NOS/ROS/abx res). Enrichment of specific metabolic pathways is consistent with functional specialization of strains in different environmental niches [45].

Table 2: Maximum containment between sequences that were increased in abundance in CD and isolate genomes.

Metapangenome graph species	Closest strain match	Maximum containment
Enterocloster clostridioformis	*Enterocloster clostridioformis* MSK.2.78	0.71
Enterocloster bolteae	*[Clostridium] bolteae* 90A5	0.68
Ruminococcus_B gnavus	*[Ruminococcus] gnavus* RJX1122	0.66
Enterocloster clostridioformis_A	*[Clostridium] clostridioforme* AGR2157	0.61
Enterocloster sp005845215	*Enterocloster clostridioformis* MSK.2.78	0.50

While dominating set differential abundance analysis identified genomic sequences that were more abundant in CD, the nature of short shotgun metagenomic sequencing reads precludes haplotype phasing or lineage resolution [15], meaning our results likely represent genomic variants from many distinct genomes that would not all naturally occur together in a single isolate genome. Therefore, to identify isolate genomes that contain the genomic sequences that were more abundant in CD, we searched the GTDB rs202 database with the significantly differentially abundant sequences. On average, the top matching isolate genome contained 63% of the sequences that were more abundant in CD (Table 2).

One aerotolerant clade of *R. gnavus* was previously identified as being enriched in CD [12], and has been shown to produce a polysaccharide that induces the inflammatory cytokine TNF-alpha [13]. The three isolate genomes we identified as containing the highest amount of sequences that were increased in abundance in CD were among those that have been shown to induce TNF-alpha secretion (RJX1122, RJX1127, RJX1128) [14]. This suggests our method identified the same strain switch previously discovered to occur in IBD [12,13,14]. In further support of this, we found that 17 of the 23 genes in the operon that encodes the proteins responsible for producing the inflammatory polysaccharide were annotated in the dominating set pieces that were more abundant in CD. These genes were encoded across 66 dominating set pieces, with multiple neighboring genes in the operon annotated in 6 of these dominating set pieces.

For two of the four *Enterocloster* species, the top matching isolate sequence was the same (*Enterocloster clostridioformis* MSK.2.78). This points to overlap in the genomic sequences we identified as differentially abundant across these metapangenome graphs. Indeed, the average Jaccard similarity between the sequences that were increased in CD in the *Enterocloster* graphs was 0.53, while the average max containment was 0.74. Given that a Jaccard similarity of 0.1 is required to recover at least 80% of a genome via assembly graph query, which approximately corresponds to an average nucleotide identity of 93% (CITE: TESSA), and that species boundaries in GTDB are drawn at 95% average nucleotide identity [46], the metapangenome graphs likely store genomic sequences associated with both the query genome species and closely related species. However, the metapangenome graphs presented here, as well as the differentially abundant sequences, contain both common and distinct nucleotide sequences, suggesting that multiple closely related *Enterocloster* species/genomes are associated with CD. Taken together, our ability to recover a previously validated sub-species association with IBD (*R. gnavus*) suggests that the three new *Enterocloster* isolates we identified should be further investigated for their potential role in eliciting CD-like symptoms in the gut.

Genomic sequences that are differentially abundant in IBD are not exclusive to IBD

Since genome sequences belonging to many species were differentially abundant from nonIBD in CD and UC, we next investigated whether there was a disease-specific microbiome in CD or UC – i.e., whether there are sequences from a species that were only observed in IBD. Using FracMinHash sketches from the differentially abundant sequences, we identified the differentially abundant sequences in each metagenome and compared their occurrence and distribution across diagnoses.

In general, we found no evidence for disease-specific sequences among the 54 species we investigated. Instead, we observed almost all sequences in at least some CD, UC, and nonIBD metagenomes (**Figure 6**, **Figure S 6**). Across all species, an average of 14.9% differentially abundant k-mers were unobserved in either CD, UC, or nonIBD. These results in part explain heterogeneous study findings in previous IBD gut microbiome investigations.

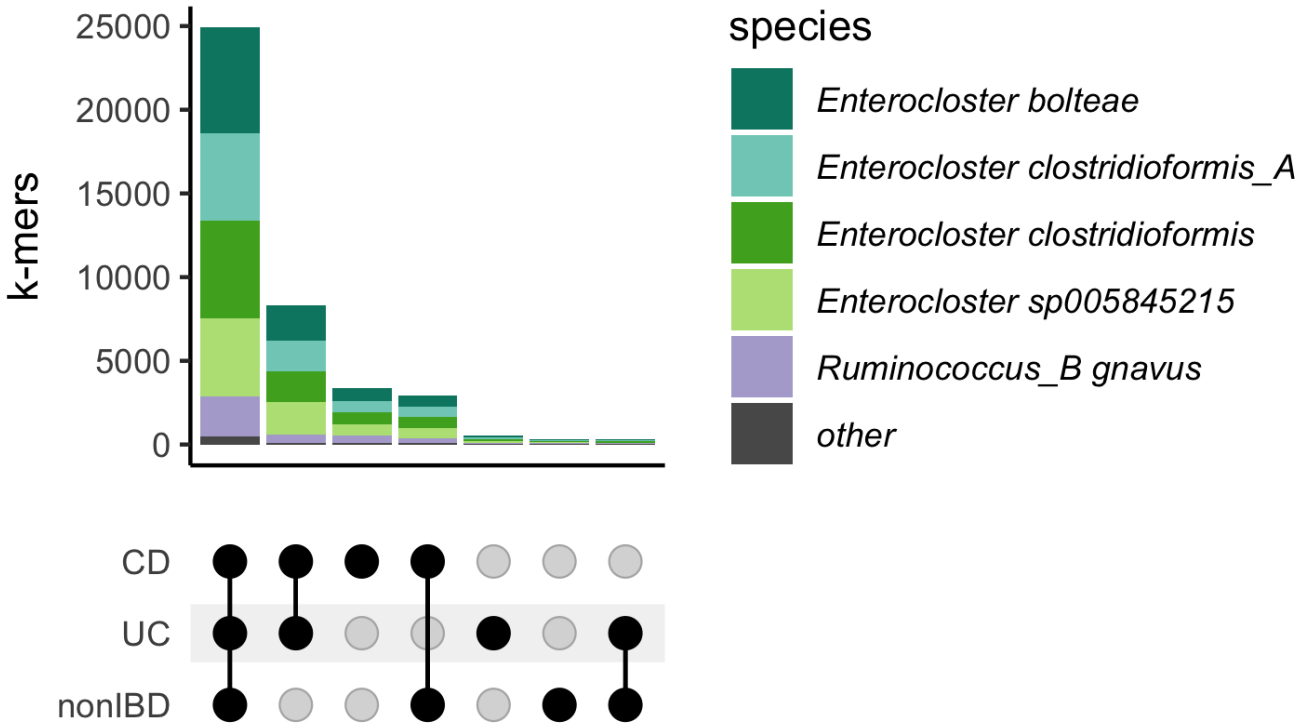


Figure 6: Most differentially abundant sequences occur in metagenomes of individuals diagnosed with CD, UC and non-IBD. Upset plot of k-mers that were increased in abundance in CD and their occurrence in CD, UC, and nonIBD metagenomes.

Discussion

In this paper, we present an assembly-free metagenome analysis framework for group association discovery that is minimally reliant on reference databases. Our approach uses k-mers to discover genomes associated with groups of metagenomes, and then recovers sequence variation from those genomes and closely related genomes in the metagenomes. These sequences are organized in a "metapangenome graph" which is then used to perform differential abundance analysis to discover specific genomic sequences that differ between groups.

We applied this method to perform a meta-analysis of fecal gut microbiome metagenomes from individuals with CD, UC, and nonIBD and uncovered cross-study microbial signatures of IBD subtype. The underlying etiology of IBD remains poorly understood with inconsistent microbiological findings produced from different studies [4]. The signatures we identified demonstrate consistent loss of diversity of specific microorganisms, particularly in CD. Among the background of generalized loss of diversity, we observed that some genomic sequences increased in abundance while others decreased in abundance for five species in CD. This pattern is consistent with strain switching, where one strain is more abundant in CD and another is more abundant in nonIBD. For one species we identified, *R. gnavus*, this strain switching behavior was previously discovered via isolate sequencing and metagenome mapping [12]. Our recovery of this pattern demonstrates the utility of our approach for discovering sub-species level associations from metagenomic sequences alone, and opens the door for additional discovery. Indeed, we identified four additional species where strain switching occurred in CD.

While our approach identified genomic sequences that were more abundant in CD than nonIBD, the nature of short read sequences precludes haplotype or lineage resolution directly from the metagenomic data analyzed here. To circumvent this issue, we identified isolate genomes that

encoded all of the genomic sequences that were more abundant in CD. These isolate genomes represent candidate organisms for further research into the microbial component of CD pathophysiology. As high fidelity long read sequencing of microbiomes becomes increasingly common [15], long reads can be integrated into the approaches introduced here, enabling lineage-resolved association detection directly from sequencing data.

While we found conserved signatures in IBD subtype, we found no evidence for disease-specific sequences within the gut microbiome. The observation that almost all differentially abundant sequences for a given species occur in CD, UC, and nonIBD suggests the presence of ecotypes – subspecies that are adapted to different environments – rather than pathotypes – subspecies associated with a specific disease. 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 loss of diversity [4].

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 [5]. We selected the first sample in each time series for this analysis, and given the relatively poor performance of these models, this may suggest that disease onset is a distinct biological process. However, the inclusion of the iHMP cohort in this analysis insured that not all nonIBD samples were healthy controls and some fraction were symptomatic cases that did not result in an IBD diagnosis [5].

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 diseases such as colorectal cancer, where a recent meta-analysis using a marker gene approach was successful in classifying colorectal samples from healthy controls [31]. Our method may bring strain-level resolution and generate hypotheses for further research.

The methods we used to perform the k-mer association analysis are modular and may be improved by substituting parts of the pipeline with different approaches. For example, we used abundances from long nucleotide k-mers ($k = 31$) – which capture species-level sequence similarity [20] – as our features and achieved model accuracies that were too low to be clinically relevant. K-mers constructed from protein or other reduced alphabets may improve accuracy, as we would expect more shared sequence content between metagenomes as well as a better representation of functional content (CITE: metapangenomes). While this may improve classification accuracy, switching to reduced alphabet k-mers may not be desirable in the context of strain-specific differences which may be obscured by these degenerate representations. Similarly, while we used random forests to to perform k-mer association analysis, other machine learning or statistical techniques may improve classification accuracy. These approaches remain areas of future research.

The first part of the pipeline is disconnectable from the second part of the pipeline – that is, the discovery of discriminatory genomes between groups is not a prerequisite for dominating set differential abundance analysis as query genomes could be selected arbitrarily. Therefore, the assembly graph differential abundance approach presented here could be applied to metagenomes for samples originating from diverse environments. The requirements for the application of dominating set differential abundance are threefold. First, there must be sufficient samples for statistical testing (e.g., a minimum of three cases and three controls, with the typical caveats for small sample sizes [47]). Second, we must have a genome with which to query the graph. And third, we must have sufficient compute resources to run spacegraphcats [25]. These requirements make the application of dominating set differential abundance analysis available to metagenomes from diverse environments, not just the well-studied human gut microbiome.

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 fast 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`) [48]. We then removed human DNA using BBMap and a masked version of hg19 [49]. Next, we trimmed low-abundance k-mers from sequences with high coverage using khmer's `trim-low-abund.py` [50].

Using these trimmed reads, we generated FracMinHash signatures for each library using sourmash (k-size 31, scaled 2000, abundance tracking on) [51]. FracMinHash sketching produces compressed representations of k-mers in a metagenome while retaining the sequence diversity in a sample [21,23]. This approach creates a consistent set of k-mers across samples by retaining the same k-mers when the same k-mers were observed. This enables comparisons between metagenomes. We refer to FracMinHash sketches as *signatures*, and to each sub-sampled k-mer in a signature as a *k-mer*. 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 [20]. We retained all k-mers that were present in multiple samples.

Principle Coordinates Analysis

We used Jaccard distance and angular similarity as implemented in `sourmash compare` to pairwise compare FracMinHash signatures. We then used the `dist()` function in base R to compute distance matrices. We used the `cmdscale()` function to perform principle coordinate analysis [52]. We used ggplot2 and ggMarginal to visualize the principle coordinate analysis [53]. To test for sources of variation in these distance matrices, we performed PERMANOVA using the `adonis` function in the R vegan package [54]. The PERMANOVA was modeled as `~ diagnosis + study accession + library size + number of k-mers`.

Random forests classifiers

We built random forests classifiers to predict CD, UC, and non-IBD status using FracMinHash signatures. We transformed sourmash signatures into a k-mer (hash) abundance table where each metagenome was a sample, each k-mer was a feature, and abundances were recorded for each k-mer for each sample. We normalized abundances by dividing by the total number of k-mers in each FracMinHash 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. We built each model six times, each time using a different random seed. To build each model, we first performed vita variable

selection on the training set as implemented in the Pomona and ranger packages [36,55]. Vita variable selection reduces the number of variables (e.g. k-mers) to a smaller set of predictive variables through selection of variables with high cross-validated permutation variable importance [35]. 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 [35]. This approach retains important variables that are correlated [35,56], 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 [57,58]. Using this smaller set of k-mers, we then built an optimized random forests model using tuneRanger [59]. We evaluated each validation set using the optimal model, and extracted variable importance measures for each k-mer 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 k-mers in a model and the total number of models.

Anchoring predictive k-mers to genomes

We used sourmash `gather` with parameters `k 31` and `--scaled 2000` to anchor predictive k-mers to known genomes [51]. Sourmash `gather` searches a database of known k-mers for matches with a query [21]. We used the sourmash GTDB rs202 representatives data base (https://osf.io/w4bcm/download). 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 k-mers contained within its genome. These k-mers were then removed, and we repeated the process for the genome with the next largest fraction of predictive k-mers. To genomes that were predictive in all models, we took the union of predictive genomes from the 36 models. We filtered this set of genomes to contain only those genomes with a cumulative normalized variable importance greater than 1%.

R dominating sets

The original spacegraphcats publication defined the dominating set as a set of nodes in the cDBG such that every node is a distance-1 neighbor of a node in the dominating set [25]. However, the algorithms as implemented allow this distance to be flexible and tunable [25]. We refer to the largest distance that any node may be from a member of the dominating set as the *radius*, *R*. Increasing the radius increases the average piece size while reducing the total number of pieces in the graph.

Genome neighborhood queries with spacegraphcats

To recover sequence variation associated with genomes that were correlated with IBD subtype, we used spacegraphcats `search` to retrieve k-mers in the compact de Bruijn graph neighborhood of each genomes (k = 31, R = 1) [25]. 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.

Construction of the metapangenome graph

After retrieving genome neighborhood sequences from each metagenome, we combined these sequences to build a single metapangenome graph (R = 10, k = 31). We increased the radius size of the metapangenome graph to produce larger level 1 dominating set pieces and to overcome highly articulated cDBGs resulting from an abundance of sequencing data. While working with single-species metapangenome graphs from many metagenomes reduced the graph size compared working with

complete metapangenome graphs, we performed two preprocessing steps prior to the metapangenome graph generation. We combined all genome query neighborhood reads and performed digital normalization and then truncated reads at k-mer that was not present in the data set at least 4 times. These are heuristic steps that we believe are unlikely to remove biologically important sequences.

Annotating the metapangenome graph

TBD on if the PFAM stuff works.

Calculating abundances metagenome abundances of dominating set nodes in the metapangenome graph

We calculated k-mer abundances for each graph piece in the level 1 dominating set.

Performing dominating set differential abundance analysis

We used Corncob to perform dominating set differential abundance analysis [40]. Corncob tests for differential relative abundance in the presence of variable sequencing depth and excessive zeroes for unobserved observations, conditions which occur in abundances from dominating sets [40]. To focus on the most common sequencing variants and to reduce runtimes, we first filtered to dominating set pieces that were present in at least 100 (16.5%) metagenomes; corncob fits a model to each dominating set piece, so it does not require abundance information for all pieces. We performed differential abundance testing using the `bbdml()` function using a likelihood ratio test with `formula = ~ study_accession + diagnosis` and `formula_null = ~study_accession`. We estimated the number of k-mers in the quality controlled metagenome reads using ntcard and used this as the denominator. We performed Bonferroni p value correction and used a significance cut off of 0.05.

Searching for isolates that contained differentially abundant genomic sequences

To identify isolate genomes that contained sequences that were in CD, we searched the GTDB rs202 database. We generated FracMinHash signatures (k = 31, scaled = 2000) of differentially abundant sequences using sourmash `sketch`. We searched GTDB rs202 using sourmash `search`, using parameter `--max-containment`. We filtered results to only include isolate genome sequences (e.g., removed metagenome-assembled genomes) and selected the top match as the best match.

Searching for metagenomes that contained differentially abundant genome sequences

We intersected FracMinHash signatures (k = 31, scaled = 2000) of differentially abundant sequences and query neighborhoods for each genome query, producing hashes that were differentially abundant and observed within each metagenome. We combined these hashes across diagnosis conditions (CD, UC, and nonIBD) and used the complexUpset R package to visualize the intersection size across conditions.

References

1. **Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species**
Chengwei Luo, Seth T Walk, David M Gordon, Michael Feldgarden, James M Tiedje, Konstantinos T Konstantinidis
Proceedings of the National Academy of Sciences (2011-04-26) https://doi.org/cvhvmz
DOI: 10.1073/pnas.1015622108 · PMID: 21482770 · PMCID: PMC3084108

2. **Multiple levels of the unknown in microbiome research**
Andrew Maltez Thomas, Nicola Segata
BMC Biology (2019-12) https://doi.org/gnm4t7
DOI: 10.1186/s12915-019-0667-z · PMID: 31189463 · PMCID: PMC6560723

3. **The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead**
Aleksandar D Kostic, Ramnik J Xavier, Dirk Gevers
Gastroenterology (2014-05) https://doi.org/f2rggj
DOI: 10.1053/j.gastro.2014.02.009 · PMID: 24560869 · PMCID: PMC4034132

4. **Integrating omics for a better understanding of Inflammatory Bowel Disease: a step towards personalized medicine**
Manoj Kumar, Mathieu Garand, Souhaila Al Khodor
Journal of Translational Medicine (2019-12) https://doi.org/gnm4t8
DOI: 10.1186/s12967-019-02174-1 · PMID: 31836022 · PMCID: PMC6909475

5. **Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases**
IBDMDB Investigators, Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, … Curtis Huttenhower
Nature (2019-05) https://doi.org/ggd6wc
DOI: 10.1038/s41586-019-1237-9 · PMID: 31142855 · PMCID: PMC6650278

6. **Dysbiosis of fecal microbiota in Crohn's disease patients as revealed by a custom phylogenetic microarray:**
Seungha Kang, Stuart E Denman, Mark Morrison, Zhongtang Yu, Joel Dore, Marion Leclerc, Chris S McSweeney
Inflammatory Bowel Diseases (2010-12) https://doi.org/ckh8bd
DOI: 10.1002/ibd.21319 · PMID: 20848492

7. **A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis**
Kathleen Machiels, Marie Joossens, João Sabino, Vicky De Preter, Ingrid Arijs, Venessa Eeckhaut, Vera Ballet, Karolien Claes, Filip Van Immerseel, Kristin Verbeke, … Séverine Vermeire
Gut (2014-08) https://doi.org/f59nf3
DOI: 10.1136/gutjnl-2013-304833 · PMID: 24021287

8. **Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease**
James D Lewis, Eric Z Chen, Robert N Baldassano, Anthony R Otley, Anne M Griffiths, Dale Lee, Kyle Bittinger, Aubrey Bailey, Elliot S Friedman, Christian Hoffmann, … Frederic D Bushman
Cell Host & Microbe (2015-10) https://doi.org/f7zp6n
DOI: 10.1016/j.chom.2015.09.008 · PMID: 26468751 · PMCID: PMC4633303

9. **Genetic risk, dysbiosis, and treatment stratification using host genome and gut microbiome in inflammatory bowel disease**
Ahmed Moustafa, Weizhong Li, Ericka L Anderson, Emily HM Wong, Parambir S Dulai, William J Sandborn, William Biggs, Shibu Yooseph, Marcus B Jones, Craig J Venter, … Brigid S Boland
Clinical and Translational Gastroenterology (2018-01) https://doi.org/gctt4v
DOI: 10.1038/ctg.2017.58 · PMID: 29345635 · PMCID: PMC5795019

10. **A human gut microbial gene catalogue established by metagenomic sequencing**
MetaHIT Consortium, Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, … Jun Wang
Nature (2010-03) https://doi.org/dpw2s3
DOI: 10.1038/nature08821 · PMID: 20203603 · PMCID: PMC3779803

11. **Mechanisms and consequences of intestinal dysbiosis**
GAdrienne Weiss, Thierry Hennet
Cellular and Molecular Life Sciences (2017-08) https://doi.org/gj9fxf
DOI: 10.1007/s00018-017-2509-x · PMID: 28352996

12. **A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients**
Andrew Brantley Hall, Moran Yassour, Jenny Sauk, Ashley Garner, Xiaofang Jiang, Timothy Arthur, Georgia K Lagoudas, Tommi Vatanen, Nadine Fornelos, Robin Wilson, … Curtis Huttenhower
Genome Medicine (2017-12) https://doi.org/gnm4t9
DOI: 10.1186/s13073-017-0490-5 · PMID: 29183332 · PMCID: PMC5704459

13. **Ruminococcus gnavus , a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide**
Matthew T Henke, Douglas J Kenny, Chelsi D Cassilly, Hera Vlamakis, Ramnik J Xavier, Jon Clardy
Proceedings of the National Academy of Sciences (2019-06-25) https://doi.org/ghzzmg
DOI: 10.1073/pnas.1904099116 · PMID: 31182571 · PMCID: PMC6601261

14. **Capsular polysaccharide correlates with immune response to the human gut microbe Ruminococcus gnavus**
Matthew T Henke, Eric M Brown, Chelsi D Cassilly, Hera Vlamakis, Ramnik J Xavier, Jon Clardy
Proceedings of the National Academy of Sciences (2021-05-18) https://doi.org/gnq78g
DOI: 10.1073/pnas.2007595118 · PMID: 33972416 · PMCID: PMC8157926

15. **Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities**
Derek M Bickhart, Mikhail Kolmogorov, Elizabeth Tseng, Daniel M Portik, Anton Korobeynikov, Ivan Tolstoganov, Gherman Uritskiy, Ivan Liachko, Shawn T Sullivan, Sung Bong Shin, … Timothy PL Smith
Nature Biotechnology (2022-05) https://doi.org/gn3fkx
DOI: 10.1038/s41587-021-01130-z · PMID: 34980911

16. **Genomic variation landscape of the human gut microbiome**
Siegfried Schloissnig, Manimozhiyan Arumugam, Shinichi Sunagawa, Makedonka Mitreva, Julien Tap, Ana Zhu, Alison Waller, Daniel R Mende, Jens Roat Kultima, John Martin, … Peer Bork
Nature (2013-01) https://doi.org/j5d
DOI: 10.1038/nature11711 · PMID: 23222524 · PMCID: PMC3536929

17. **Genome-wide association study identifies vitamin B ₅ biosynthesis as a host specificity factor in Campylobacter**

Samuel K Sheppard, Xavier Didelot, Guillaume Meric, Alicia Torralbo, Keith A Jolley, David J Kelly, Stephen D Bentley, Martin CJ Maiden, Julian Parkhill, Daniel Falush
Proceedings of the National Academy of Sciences (2013-07-16) https://doi.org/f4562b
DOI: 10.1073/pnas.1305559110 · PMID: 23818615 · PMCID: PMC3718156

18. **Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis**
Veronika B Dubinkina, Dmitry S Ischenko, Vladimir I Ulyantsev, Alexander V Tyakht, Dmitry G Alexeev
BMC Bioinformatics (2016-12) https://doi.org/gk7ks3
DOI: 10.1186/s12859-015-0875-7 · PMID: 26774270 · PMCID: PMC4715287

19. **Kevlar: A Mapping-Free Framework for Accurate Discovery of De Novo Variants**
Daniel S Standage, CTitus Brown, Fereydoun Hormozdiari
iScience (2019-08) https://doi.org/ghfc63
DOI: 10.1016/j.isci.2019.07.032 · PMID: 31377530 · PMCID: PMC6682328

20. **MetaPalette: a k -mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation**
David Koslicki, Daniel Falush
mSystems (2016-06-28) https://doi.org/gg3gbd
DOI: 10.1128/msystems.00020-16 · PMID: 27822531 · PMCID: PMC5069763

21. **Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers**
Luiz Irber, Phillip T Brooks, Taylor Reiter, NTessa Pierce-Ward, Mahmudur Rahman Hera, David Koslicki, CTitus Brown
Bioinformatics (2022-01-12) https://doi.org/gn34zt
DOI: 10.1101/2022.01.11.475838

22. **Multiple comparative metagenomics using multiset k -mer counting**
Gaëtan Benoit, Pierre Peterlongo, Mahendra Mariadassou, Erwan Drezen, Sophie Schbath, Dominique Lavenier, Claire Lemaitre
PeerJ Computer Science (2016-11-14) https://doi.org/gnm4vb
DOI: 10.7717/peerj-cs.94

23. **Large-scale sequence comparisons with sourmash**
NTessa Pierce, Luiz Irber, Taylor Reiter, Phillip Brooks, CTitus Brown
F1000Research (2019-07-04) https://doi.org/gf9v84
DOI: 10.12688/f1000research.19675.1 · PMID: 31508216 · PMCID: PMC6720031

24. **When the levee breaks: a practical guide to sketching algorithms for processing the flood of genomic data**
Will PM Rowe
Genome Biology (2019-12) https://doi.org/gf8bfj
DOI: 10.1186/s13059-019-1809-x · PMID: 31519212 · PMCID: PMC6744645

25. **Exploring neighborhoods in large metagenome assembly graphs using spacegraphcats reveals hidden sequence diversity**
CTitus Brown, Dominik Moritz, Michael P O'Brien, Felix Reidl, Taylor Reiter, Blair D Sullivan
Genome Biology (2020-12) https://doi.org/d4bb
DOI: 10.1186/s13059-020-02066-4 · PMID: 32631445 · PMCID: PMC7336657

26. **A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events**

Magali Jaillard, Leandro Lima, Maud Tournoud, Pierre Mahé, Alex van Belkum, Vincent Lacroix, Laurent Jacob
PLOS Genetics (2018-11-12) https://doi.org/gjjs4c
DOI: 10.1371/journal.pgen.1007758 · PMID: 30419019 · PMCID: PMC6258240

27. **Genome-resolved metagenomics identifies genetic mobility, metabolic interactions, and unexpected diversity in perchlorate-reducing communities**
Tyler P Barnum, Israel A Figueroa, Charlotte I Carlström, Lauren N Lucas, Anna L Engelbrektson, John D Coates
The ISME Journal (2018-06) https://doi.org/gdms93
DOI: 10.1038/s41396-018-0081-5 · PMID: 29476141 · PMCID: PMC5955982

28. **MetaCherchant: analyzing genomic context of antibiotic resistance genes in gut microbiota**
Evgenii I Olekhnovich, Artem T Vasilyev, Vladimir I Ulyantsev, Elena S Kostryukova, Alexander V Tyakht
Bioinformatics (2018-02-01) https://doi.org/gcg2gv
DOI: 10.1093/bioinformatics/btx681 · PMID: 29092015

29. **Gut microbiome structure and metabolic activity in inflammatory bowel disease**
Eric A Franzosa, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J Haiser, Stefan Reinker, Tommi Vatanen, ABrantley Hall, Himel Mallick, Lauren J McIver, … Ramnik J Xavier
Nature Microbiology (2019-02) https://doi.org/gf9727
DOI: 10.1038/s41564-018-0306-4 · PMID: 30531976 · PMCID: PMC6342642

30. **The Treatment-Naive Microbiome in New-Onset Crohn's Disease**
Dirk Gevers, Subra Kugathasan, Lee A Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, … Ramnik J Xavier
Cell Host & Microbe (2014-03) https://doi.org/f5vq7x
DOI: 10.1016/j.chom.2014.02.005 · PMID: 24629344 · PMCID: PMC4059512

31. **Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer**
Jakob Wirbel, Paul Theodor Pyl, Ece Kartal, Konrad Zych, Alireza Kashani, Alessio Milanese, Jonas S Fleck, Anita Y Voigt, Albert Palleja, Ruby Ponnudurai, … Georg Zeller
Nature Medicine (2019-04) https://doi.org/gfxrv9
DOI: 10.1038/s41591-019-0406-6 · PMID: 30936547 · PMCID: PMC7984229

32. **Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation**
Andrew Maltez Thomas, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, Serena Manara, Nicolai Karcher, Chiara Pozzi, … Nicola Segata
Nature Medicine (2019-04) https://doi.org/gfxrv6
DOI: 10.1038/s41591-019-0405-7 · PMID: 30936548

33. **A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome**
Courtney R Armour, Stephen Nayfach, Katherine S Pollard, Thomas J Sharpton
mSystems (2019-08-27) https://doi.org/ggrn32
DOI: 10.1128/msystems.00332-18 · PMID: 31098399 · PMCID: PMC6517693

34. **Microbial genes and pathways in inflammatory bowel disease**
Melanie Schirmer, Ashley Garner, Hera Vlamakis, Ramnik J Xavier
Nature Reviews Microbiology (2019-08) https://doi.org/gf8tk6

DOI: 10.1038/s41579-019-0213-6 · PMID: 31249397 · PMCID: PMC6759048

35. **A computationally fast variable importance test for random forests for high-dimensional data**
Silke Janitza, Ender Celik, Anne-Laure Boulesteix
Advances in Data Analysis and Classification (2018-12) https://doi.org/gdj8zn
DOI: 10.1007/s11634-016-0276-4

36. **Evaluation of variable selection methods for random forests and omics data sets**
Frauke Degenhardt, Stephan Seifert, Silke Szymczak
Briefings in Bioinformatics (2019-03-25) https://doi.org/gdz6nz
DOI: 10.1093/bib/bbx124 · PMID: 29045534 · PMCID: PMC6433899

37. **Comparison of normalization methods for the analysis of metagenomic gene abundance data**
Mariana Buongermino Pereira, Mikael Wallroth, Viktor Jonsson, Erik Kristiansson
BMC Genomics (2018-12) https://doi.org/gdmzhp
DOI: 10.1186/s12864-018-4637-6 · PMID: 29678163 · PMCID: PMC5910605

38. **Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics**
Viktor Jonsson, Tobias Österlund, Olle Nerman, Erik Kristiansson
BMC Genomics (2016-12) https://doi.org/f3p6xv
DOI: 10.1186/s12864-016-2386-y · PMID: 26810311 · PMCID: PMC4727335

39. **Exploring neighborhoods in large metagenome assembly graphs using spacegraphcats reveals hidden sequence diversity**
CTitus Brown, Dominik Moritz, Michael P O'Brien, Felix Reidl, Taylor Reiter, Blair D Sullivan
Genome Biology (2020-12) https://doi.org/d4bb
DOI: doi.org/10.1186/s13059-020-02066-4

40. **Modeling microbial abundances and dysbiosis with beta-binomial regression**
Bryan D Martin, Daniela Witten, Amy D Willis
The Annals of Applied Statistics (2020-03-01) https://doi.org/gg6825
DOI: 10.1214/19-aoas1283 · PMID: 32983313 · PMCID: PMC7514055

41. **MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes**
H Noguchi, T Taniguchi, T Itoh
DNA Research (2008-10-17) https://doi.org/frz7bm
DOI: 10.1093/dnares/dsn027 · PMID: 18940874 · PMCID: PMC2608843

42. **Faecalibacterium prausnitzii: from microbiology to diagnostics and prognostics**
Mireia Lopez-Siles, Sylvia H Duncan, LJesús Garcia-Gil, Margarita Martinez-Medina
The ISME Journal (2017-04) https://doi.org/f9kfz3
DOI: 10.1038/ismej.2016.176 · PMID: 28045459 · PMCID: PMC5364359

43. **Intestine farnesoid X receptor agonist and the gut microbiota activate G-protein bile acid receptor-1 signaling to improve metabolism**
Preeti Pathak, Cen Xie, Robert G Nichols, Jessica M Ferrell, Shannon Boehme, Kristopher W Krausz, Andrew D Patterson, Frank J Gonzalez, John YL Chiang
Hepatology (2018-10) https://doi.org/gkx66p
DOI: 10.1002/hep.29857 · PMID: 29486523 · PMCID: PMC6111007

44. **Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis**
Lionel Rigottier-Gois

The ISME Journal (2013-07) https://doi.org/f43frf
DOI: 10.1038/ismej.2013.80 · PMID: 23677008 · PMCID: PMC3695303

45. **Subspecies in the global human gut microbiome**
Paul I Costea, Luis Pedro Coelho, Shinichi Sunagawa, Robin Munch, Jaime Huerta-Cepas, Kristoffer Forslund, Falk Hildebrand, Almagul Kushugulova, Georg Zeller, Peer Bork
Molecular Systems Biology (2017-12) https://doi.org/gcpk4k
DOI: 10.15252/msb.20177589 · PMID: 29242367 · PMCID: PMC5740502

46. **GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy**
Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, Philip Hugenholtz
Nucleic Acids Research (2022-01-07) https://doi.org/gm97d8
DOI: 10.1093/nar/gkab776 · PMID: 34520557 · PMCID: PMC8728215

47. **How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?**
Nicholas J Schurch, Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G Simpson, Tom Owen-Hughes, … Geoffrey J Barton
RNA (2016-06) https://doi.org/f8mrmk
DOI: 10.1261/rna.053959.115 · PMID: 27022035 · PMCID: PMC4878611

48. **Trimmomatic: a flexible trimmer for Illumina sequence data**
Anthony M Bolger, Marc Lohse, Bjoern Usadel
Bioinformatics (2014-08-01) https://doi.org/f6cj5w
DOI: 10.1093/bioinformatics/btu170 · PMID: 24695404 · PMCID: PMC4103590

49. **Introducing RemoveHuman: Human Contaminant Removal [Archive] - SEQanswers**
http://seqanswers.com/forums/archive/index.php/t-42552.html

50. **The khmer software package: enabling efficient nucleotide sequence analysis**
Michael R Crusoe, Hussien F Alameldin, Sherine Awad, Elmar Boucher, Adam Caldwell, Reed Cartwright, Amanda Charbonneau, Bede Constantinides, Greg Edvenson, Scott Fay, … CTitus Brown
F1000Research (2015-09-25) https://doi.org/9qp
DOI: 10.12688/f1000research.6924.1 · PMID: 26535114 · PMCID: PMC4608353

51. **sourmash: a library for MinHash sketching of DNA**
C Titus Brown, Luiz Irber
The Journal of Open Source Software (2016-09-14) https://doi.org/ghdrk5
DOI: 10.21105/joss.00027

52. **Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis**
JC Gower
Biometrika (1966-12) https://doi.org/ch3msp
DOI: 10.2307/2333639

53. **Welcome to the Tidyverse**
Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, … Hiroaki Yutani
Journal of Open Source Software (2019-11-21) https://doi.org/ggddkj
DOI: 10.21105/joss.01686

54. **vegan: Community Ecology Package**
Jari Oksanen, Gavin L Simpson, FGuillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R Minchin, RB O'Hara, Peter Solymos, MHenry H Stevens, Eduard Szoecs, ... James Weedon (2022-04-17) https://CRAN.R-project.org/package=vegan

55. **ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R**
Marvin N Wright, Andreas Ziegler
Journal of Statistical Software (2017) https://doi.org/b8q3
DOI: 10.18637/jss.v077.i01

56. **Surrogate minimal depth as an importance measure for variables in random forests**
Stephan Seifert, Sven Gundlach, Silke Szymczak
Bioinformatics (2019-10-01) https://doi.org/gmmrnk
DOI: 10.1093/bioinformatics/btz149 · PMID: 30824905 · PMCID: PMC6761946

57. **A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules**
Joshua M Stuart, Eran Segal, Daphne Koller, Stuart K Kim
Science (2003-10-10) https://doi.org/bkfpd8
DOI: 10.1126/science.1087447 · PMID: 12934013

58. **Co-expression pattern from DNA microarray experiments as a tool for operon prediction**
C Sabatti
Nucleic Acids Research (2002-07-01) https://doi.org/d8zkdv
DOI: 10.1093/nar/gkf388 · PMID: 12087173 · PMCID: PMC117043

59. **Hyperparameters and tuning strategies for random forest**
Philipp Probst, Marvin N Wright, Anne-Laure Boulesteix
WIREs Data Mining and Knowledge Discovery (2019-05) https://doi.org/gf3sz2
DOI: 10.1002/widm.1301

Supplementary information

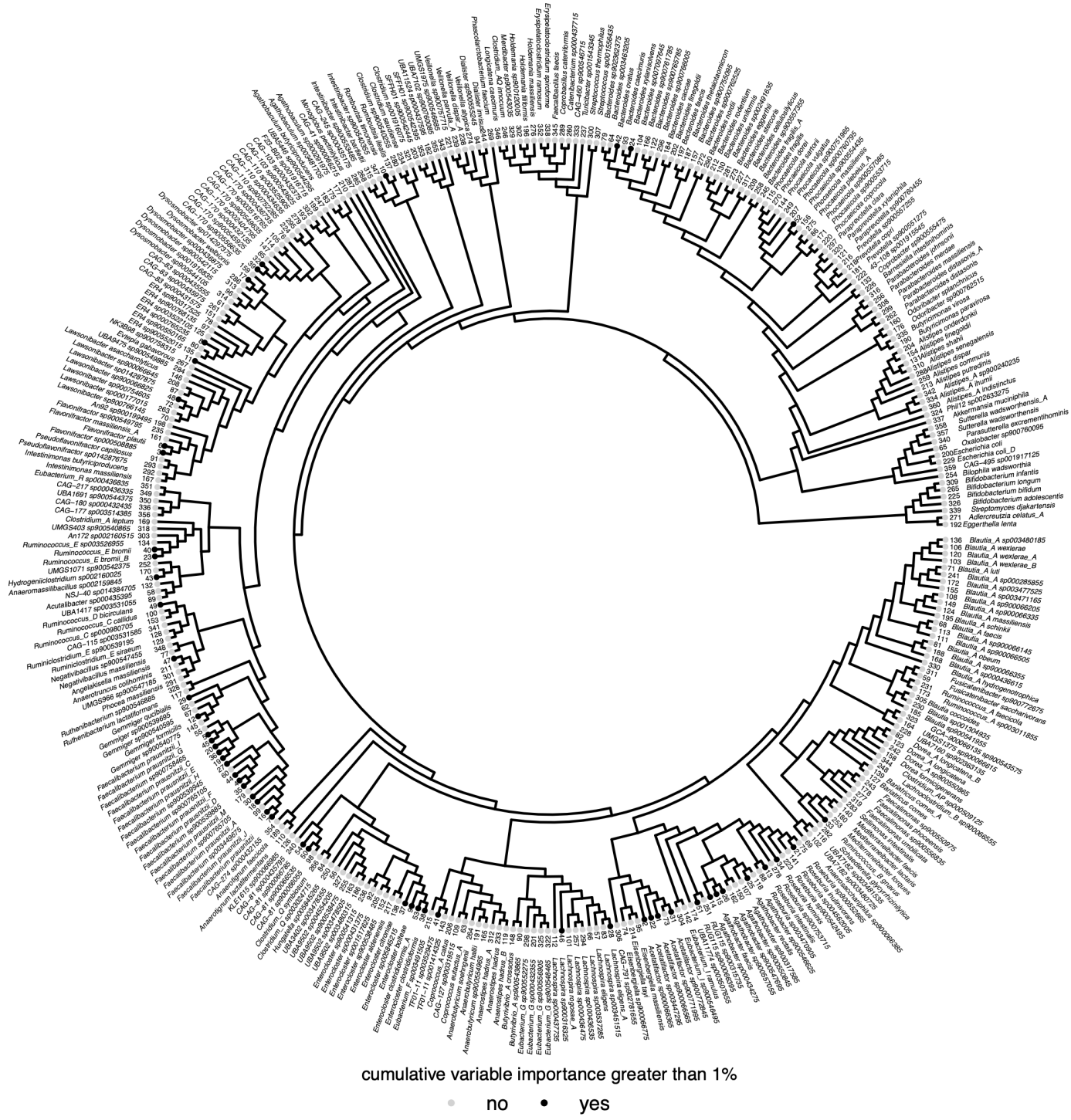


Figure 7: Phylogenetic tree of 360 bacterial species that were predictive of IBD subtype in all models. Tree was built from the GTDB rs202 tree with all tips except those represented by the 360 genomes removed. Tree tips are labelled by genomes that anchored at least 1% of the normalized variable importance. The inner ring annotates the rank of the genomes, with the genome holding the most normalized variable importance across models ranked as 1. The outer ring is the species name within the GTDB database.

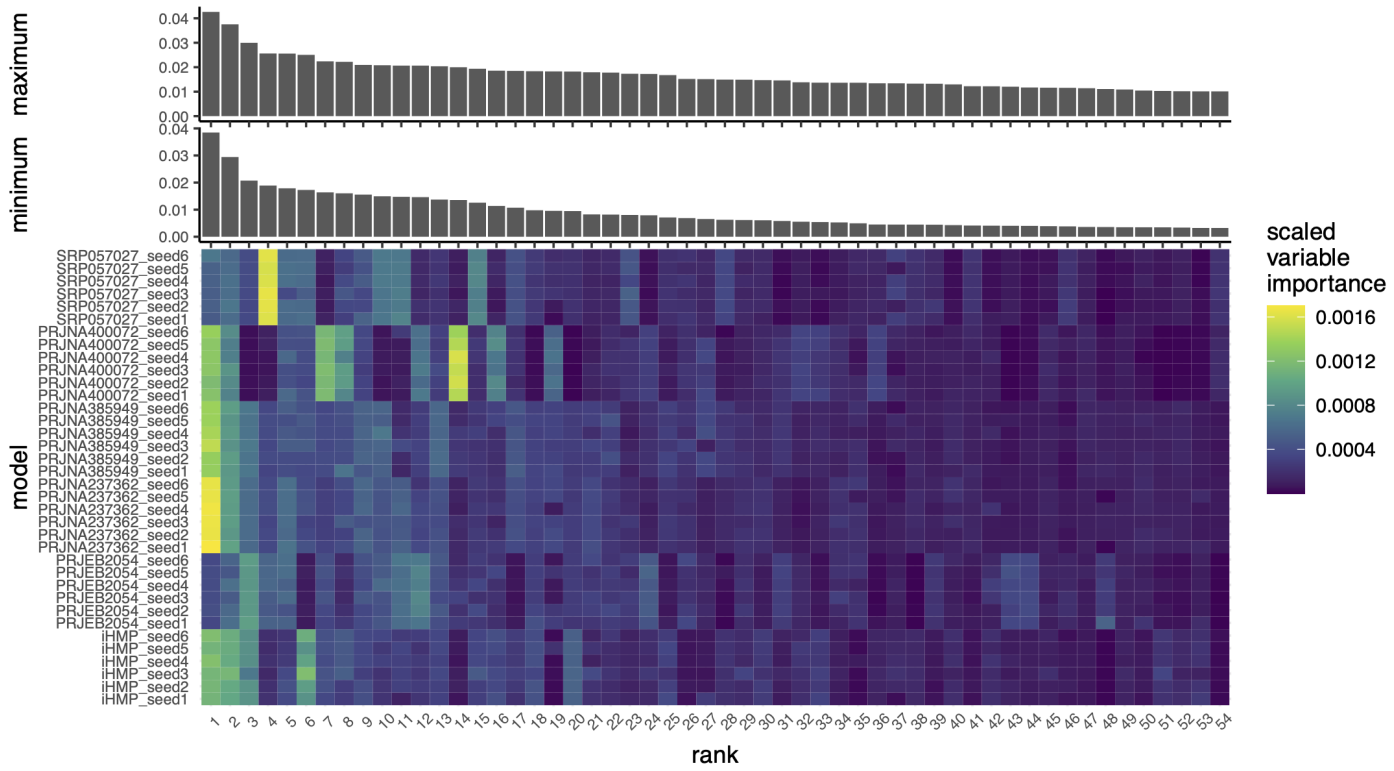


Figure 8: Fifty-four genomes are important across models and anchor the majority of variable importance.. The bottom panel depicts a heat map of the scale variable importance contributed by k-mers that anchored to each of the top 54 genomes that were important for predicting IBD subtype. Models are labelled by the validation study and by the random seed used to build the model.Rank corresponds to the genome that anchored the most variable importance. Rank:species can be decoded using the tree in **Figure S** 7. The top panels depict bar charts that correspond to the minimum (lower) or maximum (upper) variable importance a genome could anchor. The minimum variable importance was estimated following the sourmash gather algorithm, where each important k-mer was assigned to only one genome, and the genome it was assigned to was determined by a greedy winner-takes-all approach. Therefore, in the minimum bar chart, variable importance attributable to a k-mer was only summed once per k-mer, even if that k-mer occurred in multiple genomes. The maximum variable importance was estimated by allowing k-mers to be anchored to multiple genomes, so all k-mers were assigned to all possible genomes even if that meant a k-mer was assigned multiple times.

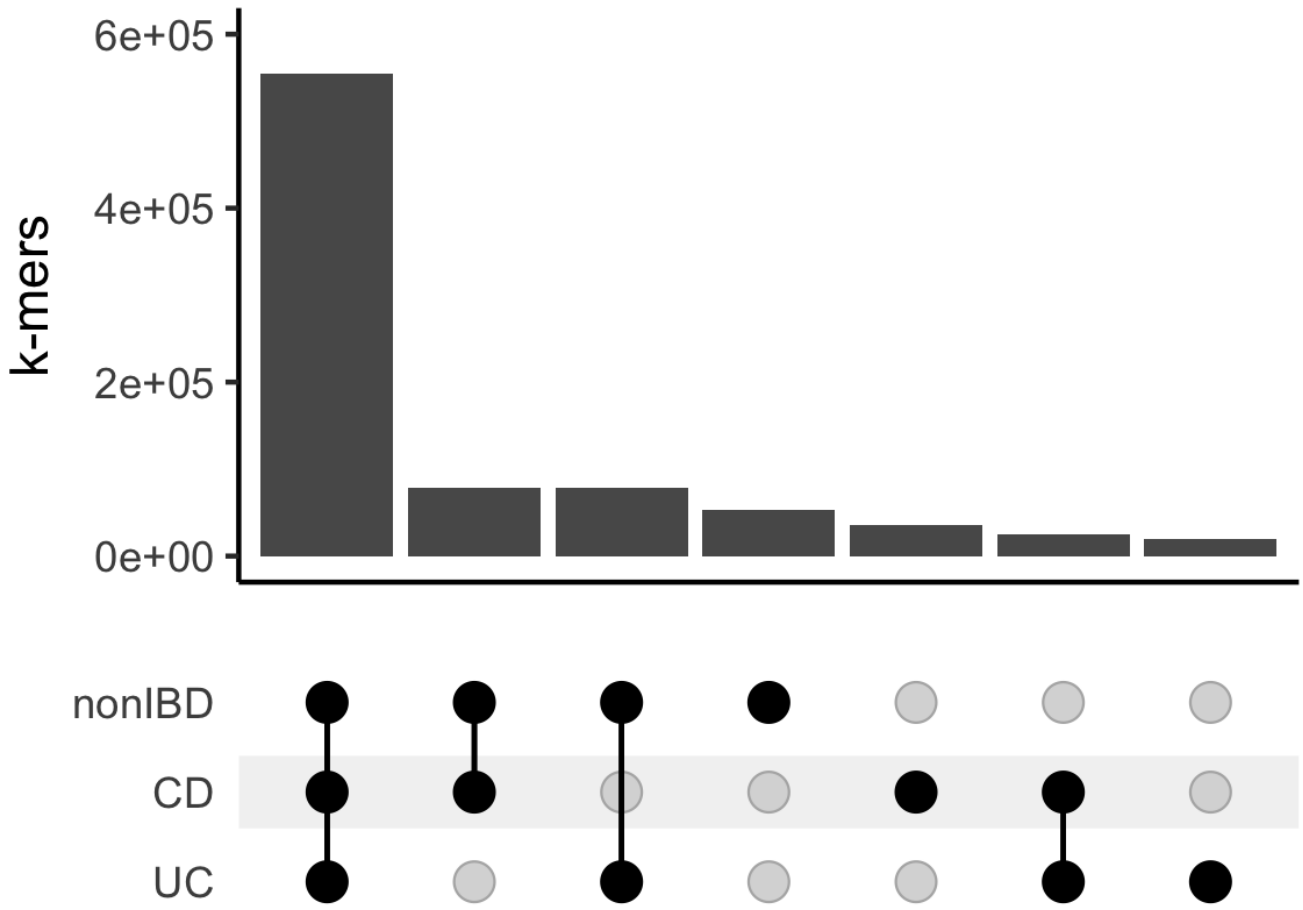


Figure 9: Most differentially abundant sequences occur in metagenomes of individuals diagnosed with CD, UC and non-IBD. Upset plot of k-mers that were decreased in abundance in CD and their occurrence in CD, UC, and nonIBD metagenomes.