

# 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@64a41c2](#) on December 2, 2021.

## Authors

---

- **Taylor Elaine Reiter**

 [0000-0002-7388-421X](#) ·  [taylorreiter](#) ·  [ReiterTaylor](#)

Department of Population Health and Reproduction, UC 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

- **Felix Reidl**

 [0000-0002-2354-3003](#) ·  [microgravitas](#) ·  [quantumgravitas](#)

Department of Computer Science and Information Systems, Birkbeck, University of London

- **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

- **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

# 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 spectrum of diseases characterized by chronic inflammation of the intestines that is likely caused by host-mediated inflammatory responses elicited in part by microorganisms [3]. IBD is cyclical with periods of active disease and remission. IBD 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 is often clinically difficult, with ramifications associated with over- or under-treatment that lead to decreased patient well-being. 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]. 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 improve insights into the dynamics of these communities. The two biggest obstacles to strain-level analysis of short read data are data getting thrown away, either because it's not in reference databases or because it doesn't assemble or bin, and resolving genomes from communities with mixed populations of closely related but distinct genomes. While long reads have made strides toward resolving the latter issue (CITE: Bickhart), in habitats like the gut where communities are dominated by single strains of microbes [14], the largest barrier to strain-level analysis is using all of the data. Here, we combine k-mer-based analysis with assembly graphs to not throw away the data.

K-mers, words of length  $k$  in nucleotide sequences, have previously been used for annotation-free characterization of sequencing data [15,16,17]. K-mers are suitable for strain-resolved metagenome analysis because they do not need to be present in reference databases to be included in analysis, they do not rely on marker genes which are largely conserved at the strain level, and they are suitable

for species- and strain-level classification [18] (CITE: gather). Investigating all k-mers in metagenomes is more computationally intensive than reference-based approaches [19], but data-reduction techniques like FracMinHash sketching make k-mer-based analysis scalable to large-scale sequence comparisons [20,21]. MinHash sketching sacrifices the fine-scaled resolution of reference-based techniques but is representative of the full sequencing sample and complete databases (CITE: gather), including strain-variable accessory elements that may be associated with diseases.

Assembly graphs complement sketch-based analysis [22,23]. 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 known functional and taxonomic annotations, recovering critical information lost through the MinHash sketching approach. While assembly graphs have been leveraged in metagenome analyses [25], their large size precluded analysis at scale. The *spacegraphcats* tool is designed to tackle this issue, encoding algorithms that scalably reduce the size of an assembly graph, enabling efficient querying and sequence retrieval [22]. These algorithms center around dominating sets, a subset of nodes that ensure that every node in the assembly graph is at most distance one from a node in the dominating set. Dominating sets partition the graph into *pieces* by assigning every node to exactly one of the closest nodes in the dominating set [22]. This simplified graph enables efficient queries: querying with a sequence that overlaps at least one k-mer in a compact de Bruijn graph (cDBG) node returns all k-mers (or all reads containing those k-mers) from the graph piece. We refer to sequences retrieved by a graph query as *neighborhoods* [22]. Genome queries often recover sequences not in reference databases or *de novo* assemblies, which disproportionately include sequences from low coverage regions or highly variable portions of the graph (e.g. sequencing reads that don't assemble or bin) [22]. When a query has a Jaccard similarity between  $10^{-2}$  and  $10^{-3}$ , 20-40% of a target genome sequence is recovered from a metagenome query [22]. This jumps to >80% when Jaccard similarity exceeds  $10^{-1}$  [22].

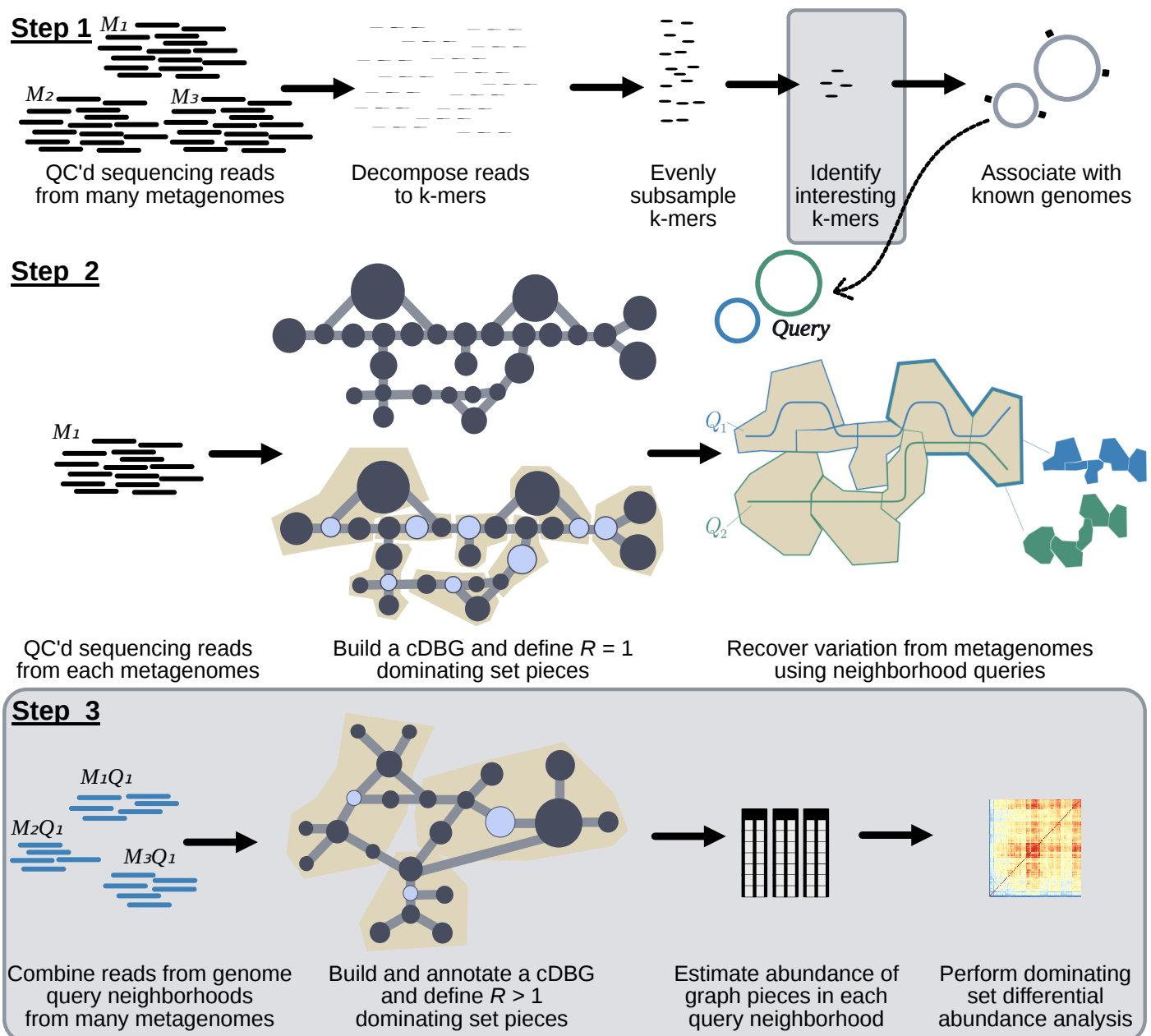
Here we develop k-mer- and assembly graph-based techniques to perform a meta-cohort analysis of six studies of IBD gut metagenome cohorts comprising 260 CD, 132 UC and 213 healthy controls (see **Table ??**) [5,8,10,12,26,27].

## SUMMARY OF TECHNIQUES WE BUILT?

1. LOOCV on FracMinHash signatures from quality controlled metagenome data
2. Identification of consistent cross-study and cross-model signatures
3. Gather to anchor to genomes
4. SGC R1 nbhd queries to recover variable regions not in reference databases, and unite all the retrieved sequences under an organizing umbrella species
5. SGC R10 metapangenome graphs
6. Dominating set differential abundance analysis

Using these approaches, we demonstrate a weak but consistent signature of IBD subtype in fecal microbiome metagenomes. Only a small subset of all k-mers are predictive of UC and CD, and these k-mers originate from a core set of microbial genomes. We find that stochastic loss of diversity in this core set of microbial genomes is a hallmark of CD, and to a lesser extent, UC. While reduced diversity is responsible for the majority of disease signatures, we find signatures of strain switches that occur in disease. Genes presumably associated with these strains occur more frequently in IBD metagenomes but are present in low abundance in nonIBD 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



**Figure 1: Overview of the metagenome analysis technique used in this paper.** 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 [22]. **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.

## K-mers are weakly predictive of IBD subtype

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

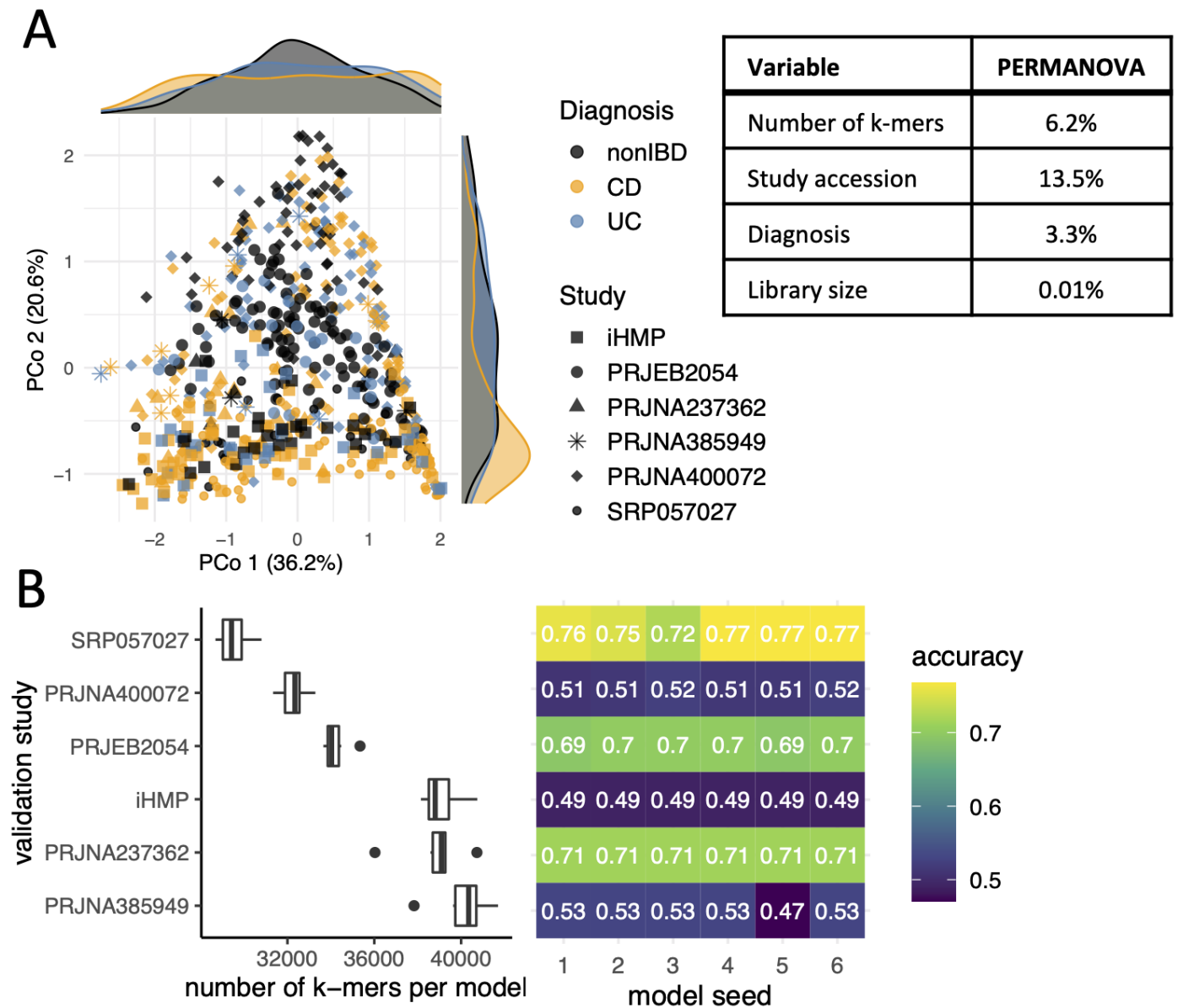
Cohort	Name	Country	Total	CD	UC	nonIBD	Reference
iHMP	IBDMDB	USA	106	50	30	26	[5]

Cohort	Name	Country	Total	CD	UC	nonIBD	Reference
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	[26]
PRJNA237362	RISK	North America	28	23	0	5	[27]
Total			605	260	132	213	

We developed a reference-free pipeline to fully characterize gut metagenomes of IBD patients (**Figure 1**). After consistent pre-processing, we used FracMinHash sketching to produce subsampled k-mer abundance profiles of metagenomes that reflect the sequence diversity in a sample [20] (CITE: Gather), 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 subsampled k-mers in a signature as *k-mers*. In total, we profiled 7,376,151 subsampled k-mers across all samples in all cohorts, representing approximately 14 billion total k-mers.

We detected variation due to IBD diagnosis in k-mer profiles of gut metagenomes from different cohorts. We calculated pairwise distance matrices using jaccard distance and angular distance between k-mer profiles, where jaccard distance captured sample richness and angular distance captured sample diversity. We performed principle coordinate analysis and PERMANOVA with these distance matrices (**Figure 2 A**), using the variables study accession, diagnosis, library size, and number of k-mers observed in a sample (**Figure 2 A**). Number of k-mers observed in a sample accounted for the highest variation, possibly reflecting reduced diversity in stool metagenomes of CD and UC patients (reviewed in [28]). Study accounted for the second 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-cohort analysis [29]. Diagnosis accounted for a similar amount of variation as study, 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 forests classifiers to predict CD, UC, or nonIBD subtype. Random forests is a supervised learning classification model that estimates how predictive k-mers are of IBD subtype, and weights 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 [30,31]. 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) (**fig 2 B**). Using this reduced set of k-mers, we then 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 [26].



**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 (CITE: gather). Sourmash gather determines the minimum set of genomes in database necessary to cover all of the k-mers in a query (CITE: Gather). 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. 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. We assume these genomes represent the strongest candidates for discriminating IBD subtype and focused on them for the remainder of our analyses.

## 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.

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

---

Many beneficial bacteria, in particular those that are oxygen sensitive, decreased in abundance in CD and UC. Nine of the 54 species we investigated were *Faecalibacterium prausnitzii* of some sort in GTDB (**Figure**). *F. prausnitzii* is an obligate anaerobe and a key butyrate producer in the gut, and plays a crucial role in reducing intestinal inflammation [32]. *F. prausnitzii* is extremely sensitive to oxygen, though it may be able to withstand oxygen exposure for up to 24 hours depending on the availability of metabolites for extracellular electron transfer [32]. In the original models, *Acetatifactor* sp900066565 held the largest variable importance, and dominating set pieces decreased in abundance in CD and to a lesser extent UC (**Figure**). *Acetatifactor* is a bile-acid producing bacteria associated with a healthy gut, but limited evidence has associated it with decreased abundance in IBD [33]. In UC, *Gemmiger formicilis* had the most dominating set pieces decrease in abundance compared to nonIBD (**Figure**). *G. formicilis* is a strictly anaerobic bacteria that produces both formic acid and butyric acid [34]. We also observed a decrease in other oxygen-sensitive species, including *Lachnospira eligens*. *L. eligens* is an obligate anaerobe that is unable to tolerate atmospheric oxygen for an hour [12]. Collectively, the decrease in species diversity we observed in IBD, in particular CD, is consistent with a shift toward increased oxidative stress during disease that is intolerable for many gut microbes [35].

A substantial fraction of dominating set pieces were more abundant in CD than nonIBD in the metagenome graphs of five species (**Figure**).

Only the increase in the distinct accessory genes is detectable amongst the backdrop of general loss of gene abundance due to decrease in other strains. ... Enrichment of specific metabolic pathways is consistent with functional specialization of strains in different environmental niches [36]. These four genomes annotated to *Ruminococcus gnavus* ...

*F. gnavus* is an aerotolerant anaerobe, one clade of which was recently found to be enriched in CD [12], and to produce an inflammatory polysaccharide [13]. *C. bolteae* is a member of the normal gut microbiota but is an opportunistic pathogen that exploits compromised intestinal barriers [37]. It is associated with disturbance succession and has increased gene expression during gut dysbiosis [5,38].

## Fractions of some metapangenomes are more abundant in but not exclusive to IBD

---

### ORPHEUM/PANMERS ANALYSIS FOR STRAIN SWITCHERS

Given the signatures of strain enrichment we detected in metagenomes from CD and UC, we next investigated whether there is a disease-specific metapangenome in CD or UC – i.e., whether there are genes from a species that are only observed in IBD. To perform this analysis in an assembly-free way, we used a recently developed approach that predicts open reading frames directly from short sequencing reads and then uses protein k-mers from those reads to estimate pangenomes (CITE: panmers). We generated standard pangenome metrics from reads from query genome neighborhoods.

In general, we found no evidence for disease-specific pangenomes among the XX pangenomes. Instead, in almost all pangenomes all protein k-mers are observed in at least some CD, UC, and nonIBD metagenomes. On average, XX protein k-mers were unobserved in UC, XX in CD, and XX in nonIBD per metapangenome, accounting for less than 1% of protein k-mers in the complete metapangenome (Table XX). These results in part explain heterogeneous study findings in previous IBD gut microbiome investigations.

## Methods

All code associated with our analyses is available at [www.github.com/dib-lab/2020-ibd/](https://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 [ibdbmdb.org](https://ibdbmdb.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` ) [39]. We then removed human DNA using BBDuk and a masked version of hg19 (CITE). Next, we trimmed low-abundance k-mers from sequences with high coverage using khmer's `trim-low-abund.py` [40].

Using these trimmed reads, we generated scaled MinHash signatures for each library using sourmash (k-size 31, scaled 2000, abundance tracking on) [41]. Scaled MinHash sketching produces compressed



representations of k-mers in a metagenome while retaining the sequence diversity in a sample [20]. 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 scaled MinHash 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 [18]. We retained all k-mers that were present in multiple samples.

## Principle Coordinates Analysis

---

We used jaccard distance and cosine distance implemented in `sourmash compare` to pairwise compare scaled MinHash signatures. We then used the `dist()` function in base R to compute distance matrices. We used the `cmdscale()` function to perform principle coordinate analysis [42]. We used `ggplot2` and `ggMarginal` to visualize the principle coordinate analysis [43]. To test for sources of variation in these distance matrices, we performed PERMANOVA using the `adonis` function in the R `vegan` package [44]. 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 scaled MinHash 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 scaled MinHash 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 variable selection on the training set as implemented in the `Pomona` and `ranger` packages [31,45]. 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 [30]. 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 [30]. This approach retains important variables that are correlated [30,46], 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 [47,48]. Using this smaller set of k-mers, we then built an optimized random forests model using `tuneRanger` [49]. 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 [41]. `Sourmash gather` searches a database of known k-mers for matches with a query (CITE: gather paper). 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 [22]. However, the algorithms as implemented allow this distance to be flexible and tunable [22]. 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 ( $k = 31$ ,  $R = 1$ ) [22]. 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 dataset 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 metapangenom 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 [50]. 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 [50]. 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 `bbdm1()` 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.

## Metapangenome analysis

---

TBD

## References

---

1. **Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species**  
C Luo, ST Walk, DM Gordon, M Feldgarden, JM Tiedje, KT Konstantinidis  
*Proceedings of the National Academy of Sciences* (2011-04-11) <https://doi.org/cvhvmz>  
DOI: [10.1073/pnas.1015622108](https://doi.org/10.1073/pnas.1015622108) · PMID: [21482770](https://pubmed.ncbi.nlm.nih.gov/21482770/) · PMCID: [PMC3084108](https://pubmed.ncbi.nlm.nih.gov/PMC3084108/)
2. **Multiple levels of the unknown in microbiome research**  
Andrew Maltez Thomas, Nicola Segata  
*BMC Biology* (2019-06-12) <https://doi.org/gnm4t7>  
DOI: [10.1186/s12915-019-0667-z](https://doi.org/10.1186/s12915-019-0667-z) · PMID: [31189463](https://pubmed.ncbi.nlm.nih.gov/31189463/) · PMCID: [PMC6560723](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1053/j.gastro.2014.02.009) · PMID: [24560869](https://pubmed.ncbi.nlm.nih.gov/24560869/) · PMCID: [PMC4034132](https://pubmed.ncbi.nlm.nih.gov/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* (2019-12-13) <https://doi.org/gnm4t8>  
DOI: [10.1186/s12967-019-02174-1](https://doi.org/10.1186/s12967-019-02174-1) · PMID: [31836022](https://pubmed.ncbi.nlm.nih.gov/31836022/) · PMCID: [PMC6909475](https://pubmed.ncbi.nlm.nih.gov/PMC6909475/)
5. **Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases**  
Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, Colin J Brislawn, ... IBDMDB Investigators  
*Nature* (2019-05-29) <https://doi.org/ggd6wc>  
DOI: [10.1038/s41586-019-1237-9](https://doi.org/10.1038/s41586-019-1237-9) · PMID: [31142855](https://pubmed.ncbi.nlm.nih.gov/31142855/) · PMCID: [PMC6650278](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1002/ibd.21319) · PMID: [20848492](https://pubmed.ncbi.nlm.nih.gov/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 Arijis, 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](https://doi.org/10.1136/gutjnl-2013-304833) · PMID: [24021287](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1016/j.chom.2015.09.008) · PMID: [26468751](https://pubmed.ncbi.nlm.nih.gov/26468751/) · PMCID: [PMC4633303](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1038/ctg.2017.58) · PMID: [29345635](https://pubmed.ncbi.nlm.nih.gov/29345635/) · PMCID: [PMC5795019](https://pubmed.ncbi.nlm.nih.gov/PMC5795019/)

10. **A human gut microbial gene catalogue established by metagenomic sequencing**  
Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, ... MetaHIT Consortium  
*Nature* (2010-03) <https://doi.org/dpw2s3>  
DOI: [10.1038/nature08821](https://doi.org/10.1038/nature08821) · PMID: [20203603](https://pubmed.ncbi.nlm.nih.gov/20203603/) · PMCID: [PMC3779803](https://pubmed.ncbi.nlm.nih.gov/PMC3779803/)
11. **Mechanisms and consequences of intestinal dysbiosis**  
GAdrienne Weiss, Thierry Hennet  
*Cellular and Molecular Life Sciences* (2017-03-28) <https://doi.org/gj9fx>  
DOI: [10.1007/s00018-017-2509-x](https://doi.org/10.1007/s00018-017-2509-x) · PMID: [28352996](https://pubmed.ncbi.nlm.nih.gov/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* (2017-11-28) <https://doi.org/gnm4t9>  
DOI: [10.1186/s13073-017-0490-5](https://doi.org/10.1186/s13073-017-0490-5) · PMID: [29183332](https://pubmed.ncbi.nlm.nih.gov/29183332/) · PMCID: [PMC5704459](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1073/pnas.1904099116) · PMID: [31182571](https://pubmed.ncbi.nlm.nih.gov/31182571/) · PMCID: [PMC6601261](https://pubmed.ncbi.nlm.nih.gov/PMC6601261/)
14. **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* (2012-12-05) <https://doi.org/j5d>  
DOI: [10.1038/nature11711](https://doi.org/10.1038/nature11711) · PMID: [23222524](https://pubmed.ncbi.nlm.nih.gov/23222524/) · PMCID: [PMC3536929](https://pubmed.ncbi.nlm.nih.gov/PMC3536929/)
15. **Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter***  
SK Sheppard, X Didelot, G Meric, A Torralbo, KA Jolley, DJ Kelly, SD Bentley, MCJ Maiden, J Parkhill, D Falush  
*Proceedings of the National Academy of Sciences* (2013-07-01) <https://doi.org/f4562b>  
DOI: [10.1073/pnas.1305559110](https://doi.org/10.1073/pnas.1305559110) · PMID: [23818615](https://pubmed.ncbi.nlm.nih.gov/23818615/) · PMCID: [PMC3718156](https://pubmed.ncbi.nlm.nih.gov/PMC3718156/)
16. **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* (2016-01-16) <https://doi.org/gk7ks3>  
DOI: [10.1186/s12859-015-0875-7](https://doi.org/10.1186/s12859-015-0875-7) · PMID: [26774270](https://pubmed.ncbi.nlm.nih.gov/26774270/) · PMCID: [PMC4715287](https://pubmed.ncbi.nlm.nih.gov/PMC4715287/)
17. **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](https://doi.org/10.1016/j.isci.2019.07.032) · PMID: [31377530](https://pubmed.ncbi.nlm.nih.gov/31377530/) · PMCID: [PMC6682328](https://pubmed.ncbi.nlm.nih.gov/PMC6682328/)

18. **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](https://doi.org/10.1128/msystems.00020-16) · PMID: [27822531](https://pubmed.ncbi.nlm.nih.gov/27822531/) · PMCID: [PMC5069763](https://pubmed.ncbi.nlm.nih.gov/PMC5069763/)
19. **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](https://doi.org/10.7717/peerj-cs.94)
20. **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](https://doi.org/10.12688/f1000research.19675.1) · PMID: [31508216](https://pubmed.ncbi.nlm.nih.gov/31508216/) · PMCID: [PMC6720031](https://pubmed.ncbi.nlm.nih.gov/PMC6720031/)
21. **When the levee breaks: a practical guide to sketching algorithms for processing the flood of genomic data**  
Will PM Rowe  
*Genome Biology* (2019-09-13) <https://doi.org/gf8bfj>  
DOI: [10.1186/s13059-019-1809-x](https://doi.org/10.1186/s13059-019-1809-x) · PMID: [31519212](https://pubmed.ncbi.nlm.nih.gov/31519212/) · PMCID: [PMC6744645](https://pubmed.ncbi.nlm.nih.gov/PMC6744645/)
22. **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* (2020-07-06) <https://doi.org/d4bb>  
DOI: [10.1186/s13059-020-02066-4](https://doi.org/10.1186/s13059-020-02066-4) · PMID: [32631445](https://pubmed.ncbi.nlm.nih.gov/32631445/) · PMCID: [PMC7336657](https://pubmed.ncbi.nlm.nih.gov/PMC7336657/)
23. **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](https://doi.org/10.1371/journal.pgen.1007758) · PMID: [30419019](https://pubmed.ncbi.nlm.nih.gov/30419019/) · PMCID: [PMC6258240](https://pubmed.ncbi.nlm.nih.gov/PMC6258240/)
24. **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* (2018-02-23) <https://doi.org/gdms93>  
DOI: [10.1038/s41396-018-0081-5](https://doi.org/10.1038/s41396-018-0081-5) · PMID: [29476141](https://pubmed.ncbi.nlm.nih.gov/29476141/) · PMCID: [PMC5955982](https://pubmed.ncbi.nlm.nih.gov/PMC5955982/)
25. **MetaCherchant: analyzing genomic context of antibiotic resistance genes in gut microbiota**  
Evgenii I Olekhovich, 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](https://doi.org/10.1093/bioinformatics/btx681) · PMID: [29092015](https://pubmed.ncbi.nlm.nih.gov/29092015/)
26. **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* (2018-12-10) <https://doi.org/gf9727>



DOI: [10.1038/s41564-018-0306-4](https://doi.org/10.1038/s41564-018-0306-4) · PMID: [30531976](https://pubmed.ncbi.nlm.nih.gov/30531976/) · PMCID: [PMC6342642](https://pubmed.ncbi.nlm.nih.gov/PMC6342642/)

27. **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](https://doi.org/10.1016/j.chom.2014.02.005) · PMID: [24629344](https://pubmed.ncbi.nlm.nih.gov/24629344/) · PMCID: [PMC4059512](https://pubmed.ncbi.nlm.nih.gov/PMC4059512/)
28. **Microbial genes and pathways in inflammatory bowel disease**  
Melanie Schirmer, Ashley Garner, Hera Vlamakis, Ramnik J Xavier  
*Nature Reviews Microbiology* (2019-06-27) <https://doi.org/gf8tk6>  
DOI: [10.1038/s41579-019-0213-6](https://doi.org/10.1038/s41579-019-0213-6) · PMID: [31249397](https://pubmed.ncbi.nlm.nih.gov/31249397/) · PMCID: [PMC6759048](https://pubmed.ncbi.nlm.nih.gov/PMC6759048/)
29. **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 Pallega, Ruby Ponnudurai, ... Georg Zeller  
*Nature Medicine* (2019-04-01) <https://doi.org/gfxrv9>  
DOI: [10.1038/s41591-019-0406-6](https://doi.org/10.1038/s41591-019-0406-6) · PMID: [30936547](https://pubmed.ncbi.nlm.nih.gov/30936547/) · PMCID: [PMC7984229](https://pubmed.ncbi.nlm.nih.gov/PMC7984229/)
30. **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* (2016-11-29) <https://doi.org/gdj8zn>  
DOI: [10.1007/s11634-016-0276-4](https://doi.org/10.1007/s11634-016-0276-4)
31. **Evaluation of variable selection methods for random forests and omics data sets**  
Frauke Degenhardt, Stephan Seifert, Silke Szymczak  
*Briefings in Bioinformatics* (2019-03) <https://doi.org/gdz6nz>  
DOI: [10.1093/bib/bbx124](https://doi.org/10.1093/bib/bbx124) · PMID: [29045534](https://pubmed.ncbi.nlm.nih.gov/29045534/) · PMCID: [PMC6433899](https://pubmed.ncbi.nlm.nih.gov/PMC6433899/)
32. **Faecalibacterium prausnitzii: from microbiology to diagnostics and prognostics**  
Mireia Lopez-Siles, Sylvia H Duncan, Ljesús Garcia-Gil, Margarita Martinez-Medina  
*The ISME Journal* (2017-01-03) <https://doi.org/f9kfz3>  
DOI: [10.1038/ismej.2016.176](https://doi.org/10.1038/ismej.2016.176) · PMID: [28045459](https://pubmed.ncbi.nlm.nih.gov/28045459/) · PMCID: [PMC5364359](https://pubmed.ncbi.nlm.nih.gov/PMC5364359/)
33. **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* (2018-05-21) <https://doi.org/gkx66p>  
DOI: [10.1002/hep.29857](https://doi.org/10.1002/hep.29857) · PMID: [29486523](https://pubmed.ncbi.nlm.nih.gov/29486523/) · PMCID: [PMC6111007](https://pubmed.ncbi.nlm.nih.gov/PMC6111007/)
34. **Gemmiger formicilis, n.gen., n.sp., an Anaerobic Budding Bacterium from Intestines**  
J GOSSLING, WEC MOORE  
*International Journal of Systematic Bacteriology* (1975-04-01) <https://doi.org/d2kmpp>  
DOI: [10.1099/00207713-25-2-202](https://doi.org/10.1099/00207713-25-2-202)
35. **Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis**  
Lionel Rigottier-Gois  
*The ISME Journal* (2013-05-16) <https://doi.org/f43frf>  
DOI: [10.1038/ismej.2013.80](https://doi.org/10.1038/ismej.2013.80) · PMID: [23677008](https://pubmed.ncbi.nlm.nih.gov/23677008/) · PMCID: [PMC3695303](https://pubmed.ncbi.nlm.nih.gov/PMC3695303/)
36. **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* (2017-12-14) <https://doi.org/gcpgk4k>  
DOI: [10.15252/msb.20177589](https://doi.org/10.15252/msb.20177589) · PMID: [29242367](https://pubmed.ncbi.nlm.nih.gov/29242367/) · PMCID: [PMC5740502](https://pubmed.ncbi.nlm.nih.gov/PMC5740502/)

37. **Comparative genomics of *Clostridium bolteae* and *Clostridium clostridioforme* reveals species-specific genomic properties and numerous putative antibiotic resistance determinants**  
Pierre Dehoux, Jean Christophe Marvaud, Amr Abouelleil, Ashlee M Earl, Thierry Lambert, Catherine Dauga  
*BMC Genomics* (2016-10-21) <https://doi.org/gkdb3s>  
DOI: [10.1186/s12864-016-3152-x](https://doi.org/10.1186/s12864-016-3152-x) · PMID: [27769168](https://pubmed.ncbi.nlm.nih.gov/27769168/) · PMCID: [PMC5073890](https://pubmed.ncbi.nlm.nih.gov/PMC5073890/)
38. **Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts**  
Catherine Lozupone, Karoline Faust, Jeroen Raes, Jeremiah J Faith, Daniel N Frank, Jesse Zaneveld, Jeffrey I Gordon, Rob Knight  
*Genome Research* (2012-10) <https://doi.org/f4bhq4>  
DOI: [10.1101/gr.138198.112](https://doi.org/10.1101/gr.138198.112) · PMID: [22665442](https://pubmed.ncbi.nlm.nih.gov/22665442/) · PMCID: [PMC3460192](https://pubmed.ncbi.nlm.nih.gov/PMC3460192/)
39. **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](https://doi.org/10.1093/bioinformatics/btu170) · PMID: [24695404](https://pubmed.ncbi.nlm.nih.gov/24695404/) · PMCID: [PMC4103590](https://pubmed.ncbi.nlm.nih.gov/PMC4103590/)
40. **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 Edverson, Scott Fay, ... C Titus Brown  
*F1000Research* (2015-09-25) <https://doi.org/9qp>  
DOI: [10.12688/f1000research.6924.1](https://doi.org/10.12688/f1000research.6924.1) · PMID: [26535114](https://pubmed.ncbi.nlm.nih.gov/26535114/) · PMCID: [PMC4608353](https://pubmed.ncbi.nlm.nih.gov/PMC4608353/)
41. **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](https://doi.org/10.21105/joss.00027)
42. **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](https://doi.org/10.2307/2333639)
43. **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](https://doi.org/10.21105/joss.01686)
44. **vegan: Community Ecology Package**  
Jari Oksanen, FG Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R Minchin, RB O'Hara, Gavin L Simpson, Peter Solymos, ... Helene Wagner  
(2020-11-28) <https://CRAN.R-project.org/package=vegan>
45. **<b>ranger</b> : A Fast Implementation of Random Forests for High Dimensional Data in <i>C++</i> and <i>R</i>**  
Marvin N Wright, Andreas Ziegler

*Journal of Statistical Software* (2017) <https://doi.org/b8q3>  
DOI: [10.18637/jss.v077.i01](https://doi.org/10.18637/jss.v077.i01)

46. **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](https://doi.org/10.1093/bioinformatics/btz149) · PMID: [30824905](https://pubmed.ncbi.nlm.nih.gov/30824905/) · PMCID: [PMC6761946](https://pubmed.ncbi.nlm.nih.gov/PMC6761946/)
47. **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](https://doi.org/10.1126/science.1087447) · PMID: [12934013](https://pubmed.ncbi.nlm.nih.gov/12934013/)
48. **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](https://doi.org/10.1093/nar/gkf388) · PMID: [12087173](https://pubmed.ncbi.nlm.nih.gov/12087173/) · PMCID: [PMC117043](https://pubmed.ncbi.nlm.nih.gov/PMC117043/)
49. **Hyperparameters and tuning strategies for random forest**  
Philipp Probst, Marvin N Wright, Anne-Laure Boulesteix  
*WIREs Data Mining and Knowledge Discovery* (2019-01-28) <https://doi.org/gf3sz2>  
DOI: [10.1002/widm.1301](https://doi.org/10.1002/widm.1301)
50. **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-aos1283](https://doi.org/10.1214/19-aos1283) · PMID: [32983313](https://pubmed.ncbi.nlm.nih.gov/32983313/) · PMCID: [PMC7514055](https://pubmed.ncbi.nlm.nih.gov/PMC7514055/)