Manuscript Title

This manuscript (permalink) was automatically generated from taylorreiter/2021-paper-ibd@284109d on November 24, 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, UC Davis · Funded by Grant XXXXXXXX

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

- **Amy D. Willis**
  0000-0002-2802-4317 · ctb · AmyDWillis
  Department of Biostatistics, UW

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

Abstract

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 [PNAS. 2011;108: 7200–7205]. 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 (CITATIONS). 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 [**kostic2014?**]. 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 a signature has thus far remained elusive [**kumar2019integrating?**].

The microbiome of CD and UC is heterogeneous, and studies that characterize the microbiome often produce conflicting results. This is likely in part driven by large inter- and intra-individual variation [**lloyd2019?**], but is also attributable to non-standardized laboratory, sequencing, and analysis techniques used to profile the gut microbiome [**kumar2019integrating?**]. Dysbiosis is frequently observed in IBD, particularly in CD [**kang2010dysbiosis?**,**machiels2014decrease?**,**lewis2015?**,**moustafa2018?**,**qin2010?**], however dysbiosis alone is not a signature of IBD [**lloyd2019?**]. Dysbiosis is defined as a decrease in gut microbial diversity that results in an imbalance between protective and harmful microorganisms, leading to intestinal inflammation [**weiss2017mechanisms?**].

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 [**hall2017?**]. Further, this clade produces an inflammatory polysaccharide [**henke2019ruminococcus?**]. While this clade is enriched in CD, its enrichment was previously masked from computational discovery by concomitant decreases in other *Ruminococcus* species in IBD [**hall2017?**], 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 XXX 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 (CITE: bork lab paper), 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 [sheppard2013genome?,dubinkina2016assessment?,standage2019kevlar?]. 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 (CITE: metapallete, gather). Investigating all k-mers in metagenomes is more computationally intensive than reference-based approaches [benoit2016multiple?]., by data-reduction techniques like scaled MinHash sketching make k-mer-based analysis scalable to large-scale sequence comparisons [pierce2019?,rowe2019levee?]. MinHash sketching sacrifices the fine-scaled resolution of reference-based techniques but is representative of the full sequencing sample and complete databases, including strain-variable accessory elements that may be associated with diseases.

Assembly graphs complement sketch-based analysis [brown2020exploring?,jaillard2018fast?]. 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 [12, 13], 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 (CITE SGC). 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 (CITE SGC). 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* (CITE SGC). 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 (CITE SGC). 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 (CITE SGC). This jumps to >80% when Jaccard similarity exceeds 10-1 (CITE SGC).

Here we use 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 @ref(tab:cohorts)**) [lloyd2019?,lewis2015?,hall2017?,franzosa2019?,gevers2014?,qin2010?]. SUMMARY OF TECHNIQUES WE BUILT? 1. LOOCV on scaled MinHash 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 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 enrichment 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 findings highlight the need for strain-level analysis of metagenomic data sets, and provide future avenues for research into IBD therapeutics.

Results

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	[lloyd2019?]
PRJEB2054	MetaHIT	Denmark, Spain	124	4	21	99	[qin2010?]
SRP057027	NA	Canada, USA	112	87	0	25	[lewis2015?]
PRJNA385949	PRISM, STiNKi	USA	17	9	5	3	[hall2017?]
PRJNA400072	PRISM, LLDeep, and NLIBD	USA, Netherlands	218	87	76	55	[franzosa2019?]
PRJNA237362	RISK	North America	28	23	0	5	[gevers2014?]
Total			605	260	132	213	

We developed a reference-free pipeline to fully characterize gut metagenomes of IBD patients (**Figure ??**). After consistent pre-processing, we use scaled MinHash sketching to produce subsampled k-mer abundance profiles of metagenomes that reflect the sequence diversity in a sample [pierce2019?], and use these profiles to perform metagenome-wide k-mer association with IBD subtype. We refer to scaled MinHash sketches as *signatures*, and for simplicity, continue referring to the sub-sampled 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 ??**), using the variables study accession, diagnosis, library size, and number of k-mers observed in a sample (**Table 2**). 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 [schirmer2019microbial?]). Study accounted for the second highest variation, emphasizing that technical artifacts can introduce strong signals that may influence heterogeneity in IBD microbiome studies but that can be mitigated through meta-cohort analysis [wirbel2019?]. 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.

Table 2: Results from PERMANOVA performed on Jaccard and Angular distance matrices. 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.

Variable	Jaccard distance	Angular distance
Number of k-mers	9.9%	6.2%
Study accession	6.6%	13.5
Diagnosis	6.2%	3.3%
Library size	0.009%	0.01%

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 separate 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 [janitza2018?,degenhardt2017?]. Variable selection reduced the number of k-mers used in each model by two orders of magnitude, from 7,376,151 to 29,264-41,701 (**Table @ref(tab:varselhashes)**). 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 ??**), outperforming a previously published model built on metagenomic data alone [franzosa2019?].

To understand which species were responsible for disease signatures detected by our models, we anchored k-mers in the models against genomes in the GTDB rs202 representatives database using sourmash 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 found that a substantial fraction of genomes were shared between models, indicating there is a consistent biological signal captured among classifiers. Of XX total genomes detected across all classifiers, 360 genomes were shared between all classifiers (supplemental upset figure? will be atrocious bc 36 is a lot of sets). 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 XX% 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, XX% 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.

INCLUDE DETAILS OF HOW WE GOT FROM 360 GENOMES TO 54.

nbhds

To recover strain variation in the metagenomes but not in the reference databases for these 54 genomes…

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`) [**bolger2014?**]. We then removed human DNA using BBMap and a masked version of hg19 [**bushnell2014?**]. Next, we trimmed low-abundance k-mers from sequences with high coverage using khmer's `trim-low-abund.py` [**crusoe2015?**].

Using these trimmed reads, we generated scaled MinHash signatures for each library using sourmash (k-size 31, scaled 2000, abundance tracking on) [**brown2016?**]. Scaled MinHash sketching produces compressed representations of k-mers in a metagenome while retaining the sequence diversity in a sample [**pierce2019?**]. 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 [**koslicki2016?**]. 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 [**gower1966?**]. We used ggplot2 and ggMarginal to visualize the principle coordinate analysis [**wickham2019?**]. To test for sources of variation in these distance matrices, we performed PERMANOVA using the `adonis` function in the R vegan package [**oksanen2010?**]. 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 vita variable selection on the training set as implemented in the Pomona and ranger packages [**degenhardt2017?**,**wright2015?**]. 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 [**janitza2018?**]. 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 [**janitza2018?**]. This approach retains important variables that are correlated [**janitza2018?**,**seifert2019?**], 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 [**stuart2003gene?**,**sabatti2002co?**]. Using this smaller set of k-mers, we then built an optimized random forests model using tuneRanger [**probst2019?**]. 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 [brown2016?]. 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 (CITE: SGC). However, the algorithms as implemented allow this distance to be flexible and tunable (CITE: SGC). 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) [brown2020exploring?]. 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 time 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 [18]. 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 [18]. 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.

Metapangenome analysis

TBD

References