Functional Activity of the Human Gut Microbiome to Classify Colorectal Cancer

Specific Aims

Changes in the taxonomic composition and metabolic activity of human microbiomes have been observed in several diseases. In the case of colorectal cancer (CRC), evidence of toxigenic activity by gut microbes implies that these changes are not only a response to disease, but may also play a role in disease etiology. Taxonomic composition is commonly defined by amplicon sequencing of the 16S rRNA gene and clustering sequences into Operational Taxonomic Units (OTUs). Previous studies have built OTU-based machine learning models to classify stool samples as normal or cancerous, to serve as a less invasive diagnostic tool for CRC than colonoscopy. Efforts to find consistent changes in taxonomic composition of microbiomes between normal and dysbiotic states have found mixed success, in part because interpersonal variability in taxonomic composition sometimes exceeds the variability between disease states. Variability of microbiome composition between individuals with the same disease status may be explained by functional redundancy, where different microbial species carry out the same functions and thus can replace each other with little effect on the overall function of the community.

Sequencing whole metagenomes to identify the genes present and annotate known gene functions is commonly used to build a profile of functional potential of the microbiome. Combining taxonomic composition from OTUs with functional potential from metagenomes allows one to characterize functional redundancy across communities, where communities with similar functional potential have different taxonomic composition. Untargeted mass spectrometry can validate the functional potential characterized from metagenomics by identifying metabolites that are active in a community, thus painting a more precise picture of active microbial functions. Here, I propose to investigate the impacts of taking functional redundancy and active metabolites into account on human stool sample classification for CRC diagnosis.

Aim 1. Assess the impact of functional redundancy of the gut microbiome on CRC classification.

Hypothesis: Using functional gene profiles instead of only taxonomic profiles improves the classification modeling of samples as CRC or non-cancerous because of functional redundancy in the gut microbiome.

A. Build taxonomic profiles with OTUs from 16S rRNA gene sequences and build profiles of functional gene potential from metagenomes.
B. Compare taxonomic composition to functional gene potential of microbiomes within and between disease states to determine presence and degree of functional redundancy.
C. Build machine learning models to classify samples as CRC or non-cancerous with taxonomic composition, functional gene potential profiles, or both as model features and compare performance.

Aim 2. Assess the impact of integrating active metabolites with functional gene potential on CRC classification.

Hypothesis: Using active metabolic pathways confirmed with mass spectrometry instead of all potential metabolic pathways from metagenomes improves the classification modeling of samples as CRC or non-cancerous.

A. Annotate compounds from untargeted mass spectrometry with the GNPS database and select those known to be products of bacterial metabolic pathways with the MetaCyc database.
B. Calculate the intersection of pathways associated with active metabolites and the pathways from functional potential profiles from metagenomes.
C. Build machine learning models to classify samples as CRC or non-cancerous with all potential metabolic pathways or only confirmed active metabolic pathways as model features and compare performance.

Dataset

Stool samples were collected from patients undergoing colonoscopy as part of the GLNE 007 study (https://clinicaltrials.gov/ct2/show/study/NCT00843375). 211 individuals were diagnosed with CRC and 223 were confirmed non-cancerous. 16S rRNA gene amplicon sequencing was performed and remaining stool was kept frozen. Part of the remaining stool will be used for whole metagenome shotgun sequencing and untargeted tandem mass spectrometry to complete these aims.

Background and Motivation

After lung cancer, colorectal cancer is the cause of the most cancer-related deaths worldwide [1]. Genetic factors explain only a small proportion of CRC cases, and lifestyle-based factors such as diet and smoking status are common to many cancer types [2]. Many studies have observed changes in the composition of the gut microbiome in CRC, implicating microbes as potential risk factors, or at least as signatures that change in response to CRC [3, 4]. Thus, there is great interest in identifying microbial biomarkers involved in CRC to improve our understanding of the disease and to develop improved diagnostic tests. Experiments in mouse models have found that treating germ-free mice with fecal matter transplants from CRC patients accelerated the progression from adenoma to carcinoma [5, 6]. Additionally, toxigenic microbial gene products such as colibactin have been associated with human CRC gut microbiomes and experimentally tested as drivers of disease progression with mouse models [7, 8, 9]. These findings imply that microbiome changes occur not only as a response to CRC, but may actually play a causative role in disease etiology.

Whether the microbiome changes in response to CRC, plays a role in disease progression, or both, microbial biomarkers hold promise as a potential diagnostic tool. Colonoscopy is currently the most effective way to detect screen-relevant neoplasias (SRNs; adenomatous and cancerous lesions), which can then be biopsied during the same colonoscopy session for diagnostic confirmation. But despite its effectiveness, the cost and highly-invasive nature of colonoscopy cause low patient compliance [10, 11]. The fecal immunochemical test (FIT), a quantitative measure of hemoglobin concentrations in stool, is currently the best non-invasive screening tool for SRNs, but has low sensitivity. OTU-based machine learning (ML) models have been developed that modestly improve sensitivity when used in complement with FIT, demonstrating the feasibility of using microbiome-based markers for diagnosis [12]. OTUs from 16S sequence data provide a lower cost way to profile taxonomic composition compared to sequencing and assembling whole metagenomes, making it a pragmatic choice for diagnostic tool devleopment.

No single organism or gene has been indicated in all cases of CRC; rather, the activities of the microbial community as a whole seem to be implicated [14]. Taxonomic composition is often compared between ecological communities by calculating a beta diversity metric such as the Bray-Curtis dissimilarity, then plotting an ordination to visualize separation between disease states. However, patient stool samples cannot be classified as having SRNs or not using beta diversity metrics, as seen in the lack of distinct clustering of disease states shown in Fig. 1. The most likely reason is that microbiome composition is highly variable between individuals, even more so than between disease states. This finding has been observed not only across independent CRC datasets but also in other diseases including obesity and oral cancer [15, 16, 17].

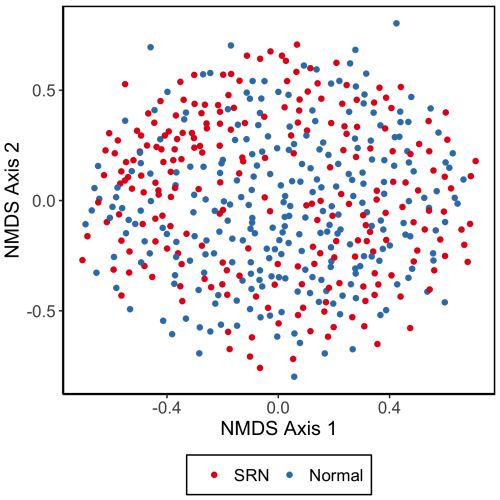

Figure 1: NMDS ordination of Bray-Curtis distances on fecal OTUs from 261 patients with normal colonoscopies and 229 patients with screen relevant neoplasias (SRN), which includes both adenoma and carcinoma. [13]

It has been proposed that this interpersonal variability can be explained by the ability of different microbial species to carry out the same functions, allowing communities with different taxonomic composition to cause the same disease [18]. Researchers are increasingly turning to whole-metagenome shotgun sequencing in order to explore the genes and gene families encoded by microbiomes and build profiles of their funtional potential. Indeed, metagenomics studies have found some associations with specific potential functional pathways that are differentially abundant in CRC [2]. Using functional potential profiles in addition to or in place of taxonomic composition may therefore improve the performance of classification models for CRC detection.

Functional redundancy is the phenomenon where different species carry out the same functions and thus can replace each other with little effect on the overall function of a community. The concept of functional redundancy has been developed and explored in the field of ecology, but few studies have directly applied it to microbial ecology specifically [19]. There does not seem to be a consensus on how to assess functional redundancy in culture-independent microbial communities, where function is often measured indirectly using meta'omic approaches [20, 21, 22, 23]. Many published studies claiming to have found functional redundancies in microbial systems lack quantitative analyses of redundancy, instead vaguely describing patterns observed by eye [24, 25, 26]. The problem is further complicated by the existence of slightly different competing definitions of functional redundancy.

Definitions from macroecology use species counts, rendering them not easily adaptable to meta'omics data which are inherently compositional. Royalty *et al.* very recently defined functional redundancy at the trait level as the evenness in relative contribution of a trait among taxa in a community [23]. Another definition which seems to be popular is that if different microbial communities have different taxonomic composition, but similar functional composition, there is functional redundancy [20]. This last heuristic has advantages: 1) it is defined by comparing different communities rather than within the same community, which easily links back to the motivating problem of interpersonal variability exceeding inter-disease state variability; 2) computational resources can be saved by comparing community compositions overall, rather than calculating a metric for every observed function in every community. Finding evidence for functional redundancy in CRC and non-cancerous gut metagenomes would support the idea that functional redundancy can explain the high degree of interpersonal variability in microbiome composition.

It is important to note that metagenomics only shows the functional *potential* of a microbiome. Metagenomics cannot discriminate between the functions that microbes can perform, and the functions they are actually performing. It would be unreasonable to expect that all potential pathways encoded by a metagenome are active at any particular point in time. Metabolomics techniques measure the metabolites, the small molecules which are inputs and outputs of chemical reactions in living systems, that are actually present in a sample. Untargeted mass spectrometry techniques such as liquid chromatography with tandem mass spectrometry (LC-MS/MS) are popular for identifying metabolites present in biological samples. Advantages of untargeted over targeted MS are that one does not need to know which metabolites to look for in advance of performing MS, and that novel metabolites can be discovered. A disadvantage is that identifying metabolites present in untargeted MS datasets is notoriously difficult; some estimate as few as 2% of mass spectra can be annotated in untargeted experiments [27]. Nevertheless, the metabolites that can be identified would provide a useful confirmation of functional potential profiles from metagenomes. Coupling metagenomics with metabolomics would also provide a way to identify which metabolites are products of bacterial metabolism and not of host metabolism or other sources such as diet [28]. Using confirmed microbial metabolites would thus paint a more precise picture of true active function for classification modeling of CRC microbiomes.

Significance: No studies to date have directly assessed functional redundancy in the human gut microbiome in CRC, nor coupled the functional potential of metagenomes with metabolomes in CRC gut microbiomes. There also does not exist a sufficiently large dataset including both metagenomic and untargeted metabolomics data from CRC and non-cancerous gut microbiomes for machine learning, although there are a handful datasets which include 16S sequence and metabolomics data. If ML models using functional potential outperform those using only OTUs as model features, that would implicate the importance of functional redundancy in the biological processes underpinning CRC. If ML models using confirmed active metabolic pathways outperform those using all potential pathways, that would underscore the need to consider true function and not just functional potential when investigating microbiomes. Overall, improving the performance of CRC classification models would represent a step in the right direction for developing non-invasive methods for CRC diagnosis.

Research Design and Methods

Aim 1. Functional redundancy of the gut microbiome

1A) Build profiles of taxonomic composition and functional potential. 16S rRNA gene sequencing was previously performed on stool samples from patients in the GLNE 007 cohort for classification modeling to detect CRC and adenomas [12]. Since then, additional samples have been collected and sequenced, bringing the total dataset to 211 CRC and 223 non-cancerous samples. Sequences will be processed with mothur according to the MiSeq SOP [29, 30]. Briefly, processing steps include filtering for quality, removing chimeric sequences, clustering sequences into OTUs using the *de novo* OptiClust method with a similarity threshold of 97%, and generating a table of OTU abundances by samples [31]. Abundances will be rarefied and converted to relative abundances to circumvent biases in sampling depth across samples and to represent the inherently compositional nature of next-generation sequencing data. This final OTU abundance table will serve as the taxonomic composition profiles of the gut microbial communities.

Whole metagenome shotgun sequencing will be performed and metagenomes will be processed with HUMAnN2 [18] to characterize functional potential of the CRC and non-cancerous microbial communities. Sequences will be trimmed for quality and reads aligning to the human reference genome will be filtered out prior to processing

with HUMAnN2. HUMAnN2 uses MetaPhlan2 to screen sequences against a curated reference of 400,000 clade-specific marker genes to detect the microbial species present in each sample [32]. This strategy is assembly-free and saves considerable computational resources over assembly-based methods. Next, sequences are mapped to annotated reference genomes to identify the gene families defined by Uniref90 and the metabolic pathways defined by MetaCyc [33] that are encoded by each community. The MetaCyc database contains pathways involving both primary and secondary metabolism and can be filtered by the domain of life. MinPath pares down the list of metabolic pathways to the minimum set that can be explained by the genes encoded in each metagenome in order to avoid overestimating the pathways present [34]. The end result is a conservative table of metabolic pathways encoded by each microbial community and their abundances. As with OTU abundances, pathway abundances will be converted to relative abundances. This table of pathway abundances will serve as the functional potential profiles of the gut microbial communities.

1B) Functional redundancy in CRC and non-cancerous gut microbiomes. Beta diversity is the difference in taxonomic composition between communities. The Bray-Curtis dissimilarity index will be calculated on OTU relative abundances for all pairwise comparisons of samples as a measure of beta diversity [35]. For relative abundances, where the OTU abundances sum to 1 in each sample, the Bray-Curtis dissimilarity between a pair of samples is as follows:

$$b_{ii'} = \frac{1}{2}\sum_{j}^{J} |r_{ij} - r_{i'j}|$$

where the relative abundance of OTU j in sample i is r_{ij} and the Bray-Curtis index between samples i and i' is $b_{ii'}$. This can also be expressed as 1 minus the sum of the lesser abundances of OTUs that the samples have in common, which is equivalent to the above formula when applied to relative abundances [36]. The range is 0 to 1, with 0 meaning the samples share all the same OTUs at equal abundance and 1 meaning the samples share no OTUs. The Bray-Curtis index is preferred over the Jaccard index because the Jaccard index only takes presence and absence of OTUs into account, while Bray-Curtis also uses abundance data such that OTUs of higher abundance have greater influence over the diversity index than rarer OTUs. A Non-metric Multidimensional Scaling (NMDS) ordination plot will be created to visualize the Bray-Curtis dissimilarities.

Analysis of Similarity (ANOSIM) will be performed on the pairwise dissimilarity matrix of samples to test whether the similarity in taxonomic composition within disease states is greater than the similarity within disease states [37]. To perform ANOSIM, all pairwise dissimilarities are ranked in order from most to least similar. The average rank similarities between samples ($\bar{r}_B$) and within samples ($\bar{r}_W$) are computed, and the test statistic R is then calculated as:

$$R = \frac{\bar{r}_B - \bar{r}_W}{\frac{1}{4}n(n-1)}$$

with n as the total number of samples. The R statistic ranges from -1 to 1, with 1 representing greatest similarity within samples, -1 representing greatest similarity between samples, and 0 representing no difference. The null hypothesis is that the disease states are interchangeable, i.e. there is no significant difference in taxonomic composition between disease states. A permutation test will then be performed to determine the significance. The sample disease states will be permuted and R will be calculated again for a random sample of 1000 possible permutations. Sampling permutations is necessary because there are $(kn)!/[(n!)^k k!]$ possible permutations of k disease states for n samples each; for 2 disease states with 200 samples each that would be 5×10^{118} permutations. The permutated samples gives an estimate of the R distribution under the null hypothesis, thus the fraction of permutation Rs that are greater than or equal to the observed R is the P value. Previous studies have not found significant differences in beta diversity between CRC and non-cancerous communities, and that result is expected here as well (see fig. 1) [15, 38, 13].

While diversity metrics are traditionally applied to taxa abundance data, they can also be applied to other types of abundance data such as functional potential profiles [18]. Bray-Curtis dissimilarity will be calculated on metabolic pathway relative abundances for all pairwise comparisons of samples to measure the beta diversity of functional potential. As with taxonomic beta diversity, significance of functional diversity will be assessed with ANOSIM and an NMDS plot will be created to visualize the dissimilarities. A lack of significant difference in taxonomic beta diversity between disease states while there is a significant difference in functional beta diversity between disease states would imply the presence of functional redundancy *within* disease states. This concords

with the heuristic definition of functional redundancy as different microbial communities having different taxonomic composition but similar functional composition.

1C) CRC classification models with taxonomic composition or functional potential. Binary random forest models will be built to classify samples as CRC or non-cancerous using OTU abundances, metabolic pathway abundances, or both as model features. The random forest method has been found to perform well for microbiome-based classification problems because it can be used for non-linear data and accounts for interactions between features [12]. A more recent study comparing modeling methods for OTU-based classification of SRNs found that random forest performed better than other methods including logistic regression, but not significantly so (see Fig. 2) [13]. Prior to training, features will be filtered to remove any OTUs and pathways with near-zero variance, as these are not likely to be informative and would only increase training time. The dataset will be randomly split into 80% training and 20% testing sets, stratified to maintain the proportion of CRC to non-cancerous samples. The mtry hyperparameter, which is the number of features used in each tree split, will be tuned to maximize the mean area under the receiver operating characteristic curve (AUROC) with 5-fold cross-validation. ROC plots the true positive rate over the false positive rate, while AUROC is interpreted as the chance of classification accuracy. Each model will then be trained with the best mtry value and the test AUROC will be calculated with the held-out test data. These steps will be repeated for 100 iterations with a different training/testing data split each time, and the test AUROCs will be recorded. The statistical significance of differences in mean AUROCs between the three types of models will be evaluated with a pairwise Wilcoxon test with Bonferroni-corrected P values for comparisons among the three models [38]. Permutation importance will be performed to determine which features (OTUs and pathways) have the greatest influence over model performance, which implies their importance as CRC biomarkers. These methods have been lauded as best practices for building OTU-based machine learning models [13] and are currently being implemented in an R package (https://github.com/SchlossLab/mikRopML). To reduce the runtime, these tasks will be run in parallel where possible on the Great Lakes HPC cluster.

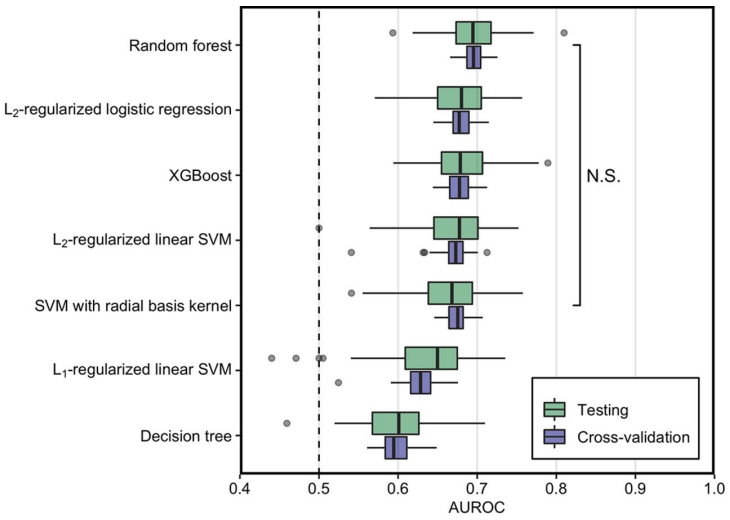

Figure 2: Performance of ML models using AUROC values of cross-validation and testing performances for classifying individuals with SRN (adenoma or carcinoma) using OTU abundances. The performance of random forest was higher than other models, but not significantly (P > 0.05). The vertical line at 0.5 depicts an AUROC which performs no better than random. [13]

Aim 2. Integrating active metabolites with functional gene potential

2A) Annotate known products of bacterial metabolism from untargeted mass spectrometry. Untargeted liquid chromatography tandem mass spectrometry (LC-MS/MS) will be performed on stool samples to determine the functions actively performed by the bacterial community. LC-MS/MS spectra will be processed with Global Natural Products Social Molecular Networking (GNPS), a popular web-based tool for processing, annotating, and sharing tandem mass spectrometry data [39]. GNPS queries spectra against all reference spectra accumulated in GNPS libraries to find near-exact matches and annotate known compounds at Level 2 or 3 (Level 1 is only possible by confirming with commerical standards) [40]. As of 2016, GNPS had 18,163 compounds in its database, and trained users can contribute new spectra created from high quality standards [41]. The spectral search outputs structures of known annotated metabolites represented by spectra, which will be converted to International Chemical Identifiers (InChi) through the GNPS API for compatibility with the MetaCyc database. It is important to note that stool samples contain metabolites which can be derived from host metabolism, microbial metabolism, both, or neither. The functional potential from metagenomes provides an avenue to identify which metabolites are likely products of microbial metabolism.

2B) Find overlapping pathways from active metabolites and functional potential profiles. The IDs of metabolic products of all pathways encoded in the metagenomes of each microbial community (functional potential profiles) will be queried from the MetaCyc database to create a set of potential metabolites. The set of

potential metabolites will be intersected with the set of metabolites annotated in LC-MS/MS. The intersection of these sets represents known metabolites which are 1) known to be products of bacterial metabolism in general and 2) capable of being produced by members of these specific microbial communities. This set intersection would exclude metabolites that are not known to be capable of being produced by microbes, i.e. any metabolites that are only produced by human metabolism or from outside sources such as the host diet (Fig. 3). This method is inspired by AMON, which uses KEGG KOs rather than the MetaCyc database to putatively annotate the origins of metabolites in integrated metagenomic and metabolomic experiments [28]. MetaCyc will be used here because it is already integrated with the HUMAnN2 tool for profiling functional potential, it contains more metabolic pathways than the KEGG database, and the KEGG database can no longer be downloaded in its entirety for free [42].

2C) CRC classification models with potential or confirmed active pathways. Binary random forest models to classify samples as CRC or non-cancerous will be built in a similar manner as described in Aim 1C, but with model features as potential metabolic pathways identified by HUMAnN2 or using only confirmed active metabolic pathways confirmed with LC-MS/MS. Features will be coded as binary variables with 1 for pathway presence and 0 for pathway absence. Presence/absence will be used rather than relative abundance because abundaces of potential pathways as determined by HUMAnN2 are based on gene sequence abundances, which is fundamentally different from metabolite quantitation performed in mass spectrometry. Best practices for model training and evaluation will be performed as described above including splitting training and testing data, tuning the mtry hyperparameter with 5-fold cross validation, calculating AUROCs of each model on the held-out test data, and repeating these steps for 100 iterations. The statistical significance of differences in mean AUROC between the two types of models will be evaluated with a Wilcoxon test. Finally, permutation importance will be performed to determine which metabolic pathways were most important for classification model performance. If the mean AUROC is significantly higher for models using only confirmed active pathways than those using functional pathways, functional potential on its own is not sufficient to characterize true community function in CRC, pointing to the importance of validating functions with metabolomics.

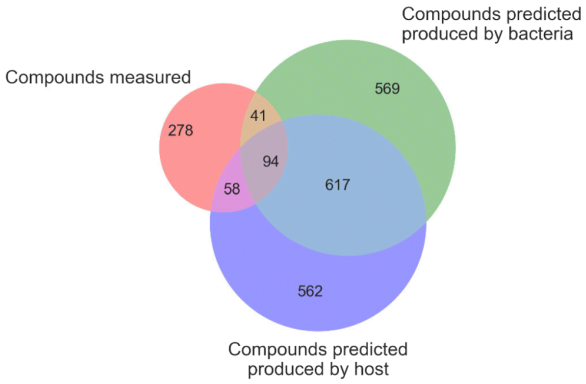

Figure 3: A case study of applying AMON to identify sources of metabolites in a microbiome sample. The method proposed here will use pathways from MetaCyc identified with HUMAnN2 rather than KEGG, and would result in confirming $41+94=135$ active microbial metabolites if this example dataset were used. [28]

Potential Outcomes and Conclusions

If the performance of the models using functional potential profiles is significantly better than the model using only taxonomic composition profiles, that would support the idea that functional redundancy explains the phenomenon of high interpersonal variability in microbiome composition obscuring differences between disease states. However, if no evidence for functional redundancy is found in Aim 1B, alternative explanations will be required. Performing feature permutation importance will then be especially necessary in order to identify which functions and/or OTUs are most important in accurate classification and to explain model performance. If models from 1C with functional potential perform no better or worse than taxonomic models, a possible explanation is that functional potential is not a close enough approximation to true function to discriminate disease states. In that scenario, validating potential functions with active metabolites will be especially important. On the other hand, if models from Aim 2C using potential pathways perform at least as well as models using only confirmed active pathways, functional potential may compensate for metabolites missed by mass spectrometry.

One limitation of this study is that stool samples are only proxies for the actual gut environment. Not all microbes or metabolites in the gut make it to the stool. However, stool is preferable because it is far less invasive to provide a stool sample than to undergo colonoscopy. Also, the data are not longitudinal. Each patient only provides one stool sample, and we do not have information on the timing of the bowel movement such as following fasting, eating, sleep, or time of day. Microbial metabolism is likely to fluctuate depending on some or all of these factors. Metabolites could be capable of being produced by microbes, but actually weren't being produced in the community at the time of sampling. Fluctuations in metabolite production would be entirely unknown, and different stool samples could be provided under very different conditions. This could negatively impact the ability to identify metabolic markers of CRC and reduce model performance.

Another limitation is that there are numerous microbial genes with unknown functions which would be missed by this study. Similarly to unknown genes, the vast majority of spectra from untargeted mass spectrometry have unknown identity. As few as 2% of spectra can be annotated in untargeted mass spectrometry experiments [27]. This is a major unsolved problem in untargeted metabolomics because confirming spectra with known commercial standards is time-consuming. Additionally, LC-MS/MS with data-dependent acquisition (the only type supported by GNPS) captures only the most abundant metabolites that cross a specified threshold [43]. This study would miss all metabolites that cannot be annotated via spectral search with the GNPS database, that are not abundant enough to be captured by LC-MS/MS, or that participate in pathways that are encoded by unknown genes. The pathways that these unknown genes and metabolites participate in could be important in CRC etiology and/or classification, and excluding them could negatively impact model performance as well as obscure our understanding of the underlying biology. If models using only active pathways do not perform better than those using all potential pathways in Aim 2C, the importance of unknown metabolites in CRC is a possible explanation. Directly using all spectra instead of collapsing into known pathways and metabolites could circumvent these problems, but that would greatly increase the number of ML features and would also necessarily include metabolites of non-microbial origin.

Finally, cost is a plausible explanation for the lack of large datasets containing both metagenomics and metabolomics data. Depending on the rates charged by metagenomics sequencing and metabolomics core services, this study could cost anywhere from 50 thousand to 200 thousand dollars just to generate the omics data for all 434 stool samples. That does not even take into account the cost of analyzing the data, training complex machine learning models on a high performance computing cluster, and the salaries of the researchers who will perform the study. Despite the cost, a large, high-quality dataset of CRC and non-cancerous stool samples with 16S sequence, metagenomics, and untargeted metabolomics would be incredibly valuable for the field.

References

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer*, 136(5):E359–386, March 2015. PMID: 25220842.

[2] A. M. Thomas, P. Manghi, F. Asnicar, E. Pasolli, F. Armanini, M. Zolfo, F. Beghini, S. Manara, N. Karcher, C. Pozzi, S. Gandini, D. Serrano, S. Tarallo, A. Francavilla, G. Gallo, M. Trompetto, G. Ferrero, S. Mizutani, H. Shiroma, S. Shiba, T. Shibata, S. Yachida, T. Yamada, J. Wirbel, P. Schrotz-King, C. M. Ulrich, H. Brenner, M. Arumugam, P. Bork, G. Zeller, F. Cordero, E. Dias-Neto, J. C. Setubal, A. Tett, B. Pardini, M. Rescigno, L. Waldron, A. Naccarati, and N. Segata. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine*, 25(4):667–678, April 2019.

[3] A. D. Kostic, D. Gevers, C. S. Pedamallu, M. Michaud, F. Duke, A. M. Earl, A. I. Ojesina, J. Jung, A. J. Bass, J. Tabernero, J. Baselga, C. Liu, R. A. Shivdasani, S. Ogino, B. W. Birren, C. Huttenhower, W. S. Garrett, and M. Meyerson. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.*, 22(2):292–298, February 2012. PMID: 22009990.

[4] H.-M. Chen, Y.-N. Yu, J.-L. Wang, Y.-W. Lin, X. Kong, C.-Q. Yang, L. Yang, Z.-J. Liu, Y.-Z. Yuan, F. Liu, J.-X. Wu, L. Zhong, D.-C. Fang, W. Zou, and J.-Y. Fang. Decreased dietary fiber intake and structural alteration of gut microbiota in patients with advanced colorectal adenoma. *Am J Clin Nutr*, 97(5):1044–1052, May 2013.

[5] L. Li, X. Li, W. Zhong, M. Yang, M. Xu, Y. Sun, J. Ma, T. Liu, X. Song, W. Dong, X. Liu, Y. Chen, Y. Liu, Z. Abla, W. Liu, B. Wang, K. Jiang, and H. Cao. Gut microbiota from colorectal cancer patients enhances the progression of intestinal adenoma in Apcmin/+ mice. *EBioMedicine*, 48:301–315, October 2019.

[6] I. Sobhani, E. Bergsten, S. Couffin, A. Amiot, B. Nebbad, C. Barau, N. de'Angelis, S. Rabot, F. Canoui-Poitrine, D. Mestivier, T. Pédron, K. Khazaie, and P. J. Sansonetti. Colorectal cancer-associated microbiota contributes to oncogenic epigenetic signatures. *PNAS*, 116(48):24285–24295, November 2019. PMID: 31712445.

[7] A. Cougnoux, G. Dalmasso, R. Martinez, E. Buc, J. Delmas, L. Gibold, P. Sauvanet, C. Darcha, P. Déchelotte, M. Bonnet, D. Pezet, H. Wodrich, A. Darfeuille-Michaud, and R. Bonnet. Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut*, 63(12):1932–1942, December 2014. PMID: 24658599.

[8] A. Gagnaire, B. Nadel, D. Raoult, J. Neefjes, and J.-P. Gorvel. Collateral damage: insights into bacterial mechanisms that predispose host cells to cancer. *Nature Reviews Microbiology*, 15(2):109–128, February 2017.

[9] B. K. Thakur, Y. Malaisé, and A. Martin. Unveiling the Mutational Mechanism of the Bacterial Genotoxin Colibactin in Colorectal Cancer. *Molecular Cell*, 74(2):227–229, April 2019.

[10] B. S. Ling, M. A. Moskowitz, D. Wachs, B. Pearson, and P. C. Schroy. Attitudes toward colorectal cancer screening tests. *J Gen Intern Med*, 16(12):822–830, December 2001. PMID: 11903761.

[11] R. M. Jones, K. J. Devers, A. J. Kuzel, and S. H. Woolf. Patient-reported barriers to colorectal cancer screening: a mixed-methods analysis. *Am J Prev Med*, 38(5):508–516, May 2010. PMID: 20409499.

[12] N. T. Baxter, M. T. Ruffin, M. A. M. Rogers, and P. D. Schloss. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med*, 8(1):37, December 2016.

[13] B. D. Topçuoğlu, N. A. Lesniak, M. T. Ruffin, J. Wiens, and P. D. Schloss. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. *mBio*, 11(3), June 2020. PMID: 32518182.

[14] P. Louis, G. L. Hold, and H. J. Flint. The gut microbiota, bacterial metabolites and colorectal cancer. *Nature Reviews Microbiology*, 12(10):661–672, October 2014.

[15] T. L. Weir, D. K. Manter, A. M. Sheflin, B. A. Barnett, A. L. Heuberger, and E. P. Ryan. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS ONE*, 8(8):e70803, 2013. PMID: 23940645.

[16] M. M. Finucane, T. J. Sharpton, T. J. Laurent, and K. S. Pollard. A Taxonomic Signature of Obesity in the Microbiome? Getting to the Guts of the Matter. *PLOS ONE*, 9(1):e84689, January 2014.

[17] M. Perera, N. Al-hebshi, I. Perera, D. Ipe, G. Ulett, D. Speicher, T. Chen, and N. Johnson. Inflammatory Bacteriome and Oral Squamous Cell Carcinoma. *J Dent Res*, 97(6):725–732, June 2018.

[18] E. A. Franzosa, L. J. McIver, G. Rahnavard, L. R. Thompson, M. Schirmer, G. Weingart, K. S. Lipson, R. Knight, J. G. Caporaso, N. Segata, and C. Huttenhower. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods*, 15(11):962–968, November 2018.

[19] C. Ricotta, F. d. Bello, M. Moretti, M. Caccianiga, B. E. L. Cerabolini, and S. Pavoine. Measuring the functional redundancy of biological communities: a quantitative guide. *Methods in Ecology and Evolution*, 7(11):1386–1395, 2016.

[20] S. Louca, M. F. Polz, F. Mazel, M. B. N. Albright, J. A. Huber, M. I. O'Connor, M. Ackermann, A. S. Hahn, D. S. Srivastava, S. A. Crowe, M. Doebeli, and L. W. Parfrey. Function and functional redundancy in microbial systems. *Nature Ecology & Evolution*, 2(6):936–943, June 2018.

[21] A. Heintz-Buschart and P. Wilmes. Human Gut Microbiome: Function Matters. *Trends in Microbiology*, 26(7):563–574, July 2018. PMID: 29173869.

[22] B. J. Tully, C. G. Wheat, B. T. Glazer, and J. A. Huber. A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer. *The ISME Journal*, 12(1):1–16, January 2018.

[23] T. M. Royalty and A. D. Steen. A quantitative measure of functional redundancy in microbial ecosystems. *bioRxiv*, page 2020.04.22.054593, April 2020.

[24] R. C. Souza, M. Hungria, M. E. Cantão, A. T. R. Vasconcelos, M. A. Nogueira, and V. A. Vicente. Metagenomic analysis reveals microbial functional redundancies and specificities in a soil under different tillage and crop-management regimes. *Applied Soil Ecology*, 86:106–112, February 2015.

[25] M. Ferrer, A. Ruiz, F. Lanza, S.-B. Haange, A. Oberbach, H. Till, R. Bargiela, C. Campoy, M. T. Segura, M. Richter, M. v. Bergen, J. Seifert, and A. Suarez. Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. *Environmental Microbiology*, 15(1):211–226, 2013.

[26] D. Galambos, R. E. Anderson, J. Reveillaud, and J. A. Huber. Genome-resolved metagenomics and metatranscriptomics reveal niche differentiation in functionally redundant microbial communities at deep-sea hydrothermal vents. *Environmental Microbiology*, 21(11):4395–4410, November 2019.

[27] R. R. da Silva, P. C. Dorrestein, and R. A. Quinn. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci USA*, 112(41):12549–12550, October 2015.

[28] M. Shaffer, K. Thurimella, K. Quinn, K. Doenges, X. Zhang, S. Bokatzian, N. Reisdorph, and C. A. Lozupone. AMON: annotation of metabolite origins via networks to integrate microbiome and metabolome data. *BMC Bioinformatics*, 20(1):614, November 2019.

[29] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, 2009.

[30] J. J. Kozich, S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Appl. Environ. Microbiol.*, 79(17):5112–5120, September 2013.

[31] S. L. Westcott and P. D. Schloss. OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. *mSphere*, 2(2):e00073–17, March 2017.

[32] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8):811–814, August 2012.

[33] R. Caspi, R. Billington, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, P. E. Midford, Q. Ong, W. K. Ong, S. Paley, P. Subhraveti, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res*, 46(D1):D633–D639, January 2018.

[34] Y. Ye and T. G. Doak. A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes. *PLOS Computational Biology*, 5(8):e1000465, August 2009.

[35] J. R. Bray and J. T. Curtis. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, 27(4):325–349, February 1957.

[36] M. Greenacre and R. Primicerio. *Multivariate analysis of ecological data*. Fundación BBVA, Bilbao, 2014. OCLC: ocn870408454.

[37] K. R. Clarke. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18(1):117–143, 1993.

[38] G. D. Hannigan, M. B. Duhaime, M. T. Ruffin, C. C. Koumpouras, and P. D. Schloss. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. *mBio*, 9(6), December 2018. PMID: 30459201.

[39] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. Boya P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. □. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein, and N. Bandeira. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol*, 34(8):828–837, August 2016.

[40] A. T. Aron, E. C. Gentry, K. L. McPhail, L.-F. Nothias, M. Nothias-Esposito, A. Bouslimani, D. Petras, J. M. Gauglitz, N. Sikora, F. Vargas, J. J. J. van der Hooft, M. Ernst, K. B. Kang, C. M. Aceves, A. M. Caraballo-Rodríguez, I. Koester, K. C. Weldon, S. Bertrand, C. Roullier, K. Sun, R. M. Tehan, C. A. B. P, M. H. Christian, M. Gutiérrez, A. M. Ulloa, J. A. Tejeda Mora, R. Mojica-Flores, J. Lakey-Beitia, V. Vásquez-Chaves, Y. Zhang, A. I. Calderón, N. Tayler, R. A. Keyzers, F. Tugizimana, N. Ndlovu, A. A. Aksenov, A. K. Jarmusch, R. Schmid, A. W. Truman, N. Bandeira, M. Wang, and P. C. Dorrestein. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nature Protocols*, 15(6):1954–1991, June 2020.

[41] F. Vargas, K. C. Weldon, N. Sikora, M. Wang, Z. Zhang, E. C. Gentry, M. W. Panitchpakdi, M. Caraballo, P. C. Dorrestein, and A. K. Jarmusch. Protocol for Community-created Public MS/MS Reference Library Within the GNPS Infrastructure. preprint, Bioinformatics, October 2019.

[42] R. Caspi, R. Billington, I. M. Keseler, A. Kothari, M. Krummenacker, P. E. Midford, W. K. Ong, S. Paley, P. Subhraveti, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res*, 48(D1):D445–D453, January 2020.

[43] J. F. Xiao, B. Zhou, and H. W. Ressom. Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trends Analyt Chem*, 32:1–14, February 2012. PMID: 22345829.