ON WHOLE GENOME CLASSIFIER PERFORMANCE IN RELATION TO 16S CLASSIFIERS

by

James Benjamin Johnson

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics

Charlotte

2022

Approved by:

_____
Dr. Anthony Fodor

_____
Dr. Cynthia Gibas

_____
Dr. Richard Allen White III

_____
Dr. Alex Dornburg

_____
Dr. Todd Steck

_____
Dr. Jacelyn Rice-Boayue

ABSTRACT

JAMES BENJAMIN JOHNSON. On Whole Genome Classifier Performance in Relation to 16S Classifiers. (Under the direction of DR. ANTHONY FODOR)

There is little consensus in the literature as to which approach for classification of Whole Genome Shotgun (WGS) sequences is most accurate unlike 16S classifiers that have had more time to mature. In this dissertation, two of the most popular classification algorithms, Kraken2 and Metaphlan2, were examined using four publicly available datasets. Surprisingly, Kraken2 reported not only more taxa but many more taxa that were significantly associated with metadata. By comparing the Spearman correlation coefficients of each taxa in the dataset against more abundant taxa, it was found that Kraken2, but not Metaphlan2, showed a consistent pattern of classifying low abundance taxa that were highly correlated with the more abundant taxa. Neither Metaphlan2, nor 16S sequences that were available for two of four datasets, showed this pattern. These results suggest that Kraken2 consistently misclassified high abundance taxa into the same erroneous low abundance taxa. These “phantom” taxa have a similar pattern of inference as the high abundance source. Because of the ever-increasing sequencing depths of modern WGS cohorts, these “phantom” taxa will appear statistically significant in statistical models even with a low classification error rate from Kraken2. These findings suggest a novel metric for evaluating classifier accuracy.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my advisor, Dr. Anthony Fodor. His guidance, unparalleled knowledge, and constructive criticism formed the backbone of his dissertation. Providing opportunities to participate in key projects and papers, Dr. Fodor challenged my reasoning and tested the effectiveness and efficiency of my solutions. The skills I have gained throughout my graduate studies are the direct result of his invaluable advice, wisdom, enthusiasm, and patience.

Other members of Dr. Fodor's lab that I would like to thank are Ivory Blake, Dr. Alicia Sorgen, and Aaron Yerke for providing technical support on difficult software bugs, and camaraderie throughout the pandemic and dissertation.

I also want to acknowledge Dr. Shan Sun, Dr. Matthew Tsilimigras, and Dr. Kevin Lambirth who expanded my understanding of microbiology and provided critical support on effective communication with the larger academic community. They not only helped with my academic writing but also served as great role models.

I have been most fortunate in the composition of my dissertation committee: Dr. Cynthia Gibas, Dr. Todd Steck, Dr. Alex Dornburg, Dr. Allen White III, and Dr. Jacelyn Rice-Boayue. Everyone has been exceptional in their willingness to participate and to provide critical feedback, and insightful suggestions. In spite of busy schedules, the committee made themselves available which further demonstrates their commitment to the science and the field. I sincerely thank them for their time and support.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: 1.1 Background and Introduction

Before the advent of next-generation sequencing, the mechanistic characteristics of how microbes interact with their host were difficult to study in detail. The majority of microorganisms cannot be cultured under standard laboratory conditions (Locey & Lennon, 2016),(Mohr, 2018). Improvements in sequencing technology have allowed researchers to explore the novel relationship between microbes and their environments (Payne et al., 2017), (Hultman et al., 2015), (Osman et al., 2020),(Regalado et al., 2020). As the cost of sequencing has decreased the use of non-cultured based methods in the form of 16S rRNA sequencing, and later Whole Genome Shotgun Sequencing (WGS sequencing), has become ubiquitous. With each generation of sequencing technology, the number of reads has rapidly increased requiring specialized metagenomic bioinformatics tools to process and analyze the ever increasing volumes of data. Understanding how these tools have evolved alongside improving sequencing technology can help guide our use of these tools and determine unmet computational needs for the next generation of classifiers.

Although the 16S rRNA gene was an early target of Sanger sequencing (Sanger et al., 1977), the application of the 16S rRNA gene for microbial taxonomy is credited to Pace et al in 1985 (Lane et al., 1985) who realized that the gene could be used for classification in Bacteria and Archaea while excluding Eukaryotes. The 16S rRNA gene is unique in that it has nine hypervariable regions (V1-V9) across its length of ~1500 base pairs (bp). Despite being highly conserved and facing strong selective pressure, these variable regions within the gene contain distinct nucleotide changes that can be mapped back to phylogenetic trees. The decreased cost of Sanger sequencing together with the use of cultured libraries allowed for sequence based explorations of microbial diversity beyond the 16S rRNA gene. In a landmark paper, researchers surveyed the ocean microbiome and revealed an astounding number of distinct taxa and new

gene families (Venter, 2004). The number of genes discovered in this study was 10 times what was previously known from the SwissProt database (at the time around 1.2 million) and after matching these genes to several databases the researchers argued that they had discovered 69,000 novel genes. This technique, which would be later refined as modern Metagenomic Shotgun Sequencing, would be improved by abandoning Sanger Sequencing in favor of cheaper Next Generation Sequencing (NGS).

The first NGS sequencer, the 454 Life science platform, was launched in 2005 and was rapidly adopted by the field for studying microbial genomes (Margulies et al., 2005; Rothberg & Leamon, 2008). This platform produced shorter reads with more errors than Sanger Sequencing but allowed for higher throughput of initially 25 million reads. An initial highly cited work using 454 for the human microbiome provided evidence that the gut microbiome can cause obesity or weight loss depending upon the overall ratio of *Firmicutes* and *Bacteroidetes* (Turnbaugh et al., 2006). The rapid creation of over 200 Mb of sequence data in this study enabled researchers to quickly evaluate the microbial community in mouse models through fecal transplantation from obese and lean human subjects to test the hypothesis that the microbial community controlled energy metabolism in recipient mice (Turnbaugh et al., 2006). Despite the initial human observation being largely unable to be replicated by other researchers (Sze & Schloss, 2016), these findings helped to create an increased interest in the microbiome outside of microbiology, which in turn provided the demand for applying new sequencing platforms to metagenomics.

Most modern NGS sequencing is performed on the Illumina platform, although other companies such as Ion Torrent have promoted their platforms for metagenomic use. 16S sequencing is still used extensively to this day and has become ever more affordable for individual labs to easily perform their own surveys of microbial diversity. Third generation

technologies such as Oxford Nanopore and PacBio differ from previous generations due to much longer read (currently up to a factor of 20) lengths but with fewer reads in the hundreds of thousands vs the millions of reads typically produced by Illumina (Roumpeka et al., 2017). Switching to the third generation platforms have improved the quality of both WGS and 16S sequencing. WGS sequencing has benefited disproportionately since 16S genes are only 1500 bp long while an entire bacterial genome can be up to millions of bps.

In tandem with the growth of 16S sequencing, there have been many popular algorithms for the taxonomic classification of 16S sequences. The most popular in terms of past citations, Quantitative Insights into Microbial Ecology (QIIME) (Caporaso et al., 2010), groups sequences into Operational Taxonomical Units (OTUs) that must share 97% sequence similarity in order to be clustered. Depending upon user settings, these OTUs could be found without a database (called denovo sequencing) or against reference sequences from a database. This differs from the second most popular algorithm, the Ribosomal Database Project (RDP) (Wang et al., 2007) which represents naïve Bayesian classification. Query sequences and references from the RDP database are divided into k-long segments called k-mers. These k-mers are scored through bootstrapping confidence estimates, matching nearest neighbors, and some Bayesian priors to determine the taxonomic classification for a given read. The most recent and increasingly popular 16S approach to 16S analysis called Dada2 differs from previous reference-based clustering by building Amplicon Sequence Variant (ASVs) instead of OTUs. ASVs are different from OTUs in that two sequences with a single nucleotide difference will produce two distinct ASVs while they would be clustered into OTUs with a less than 100% identity threshold. These ASVs are constructed through multiple layers of removing PCR and platform specific errors through a Bayesian modeling approach. The ASVs can then later be classified to whatever

database and taxonomic classifier users want while still preserving the original nucleotide specific data for future researchers to compare against their own data.

While 16S sequencing remains powerful, it has three significant limitations. First, it is limited to the taxonomic classification of genus and above. Second, functional information can only be inferred from taxonomy as it only targets the 16S rRNA gene. Third, the process of amplifying reads via Polymerase Chain Reaction (PCR) introduces errors. Therefore researchers have increasingly been incorporating WGS sequencing into microbiome analysis since it captures entire genomes and provides functional information about a transcript and potential proteins. At least in theory, WGS can identify taxa down to the strain levels and track horizontal gene transfer events (Parks et al., 2018). Despite its popularity, WGS sequencing remains more expensive compared with 16S studies in terms of money, time, and computational resources. Despite the reduction in the cost of the two methods, WGS samples can cost over 200 dollars per sample to run through a sequencer while 16S can cost as much as 80 dollars per sample (Zymo Research, 2019). An example of the differences between a WGS and 16S study can be shown in two of my publications on the same wastewater treatment plant samples (Clinton et al., 2020; Lambirth et al., 2018). The WGS study required terabytes of hard-disk space for storage and some of the functional tools ran on our computational cluster for months; this was in contrast with the 16S study which only require hundreds of gigabytes of storage and took less than a week to run.

Unlike 16S sequencing, WGS sequencing amplifies entire genomes and it remains an open question of how best to taxonomically classify WGS sequences. A whole host of algorithms have been proposed to address the research challenges associated with the taxonomic classification of WGS sequences. Quantifying the performance of these metagenomics tools

often relies on the authors of each new toolkit comparing their method against other published work to demonstrate superior performance. Table 1 shows how all WGS classifiers with greater than 200 citations that appear to still be in active development presented the effectiveness of their algorithms in their initial publication. By citation count, Kraken2 (Wood et al., 2019; Wood & Salzberg, 2014), Metaphlan2 (Segata et al., 2012; Truong et al., 2015), and Diamond (Buchfink et al., 2015) are three of the most popular algorithms for classification. These three algorithms, however, have very different approaches to classification. Kraken2 belongs to a subset of Sequence similarity algorithms utilizing k-mers. Query sequences and reference databases are divided into k-long segments where the aligner matches the k-long query to the k-long fragments of the database. Unique database curation, indexing, and alignment scores differentiate k-mer algorithms. Metaphlan2 belongs to an eclectic collection of marker-based classifiers that ignores most of the genome. In Metaphlan2, a curated database of informative makers is curated and reads that do not map to this small subset of genes are ignored.

Metaphlan2 relies on bowtie for read alignment before classifying clade specific marker genes that allow for it to classify somewhat novel taxa as long as it's genetically close enough to documented genus member. The marker gene does suffer from relying on a curated database preventing custom databases tailored to the biome under study. The third group of classifiers that branched out from prior marker-based classification relies on translating subsets of DNA into taxonomic specific proteins. Diamond and Kaiju (Buchfink et al., 2015; Menzel et al., 2016) are the most cited of the protein classifiers and have become quite popular despite lower scores when compared to other classifiers as shown in Table 1.

Table 1: Commonly Cited WGS Classifier Tools							
Algorithm	Method	Algorithms Benchmarks	Type of data evaluated	Type of Evaluation	Custom Databases	Number of Citations	Citation and Year
Kraken2	Sequence Similarity: k-mer	Metaphlan, Megablast, PhymmBl, NBC; Centrifuge, Clark, Kaiju KrakenUniq,	Simulated from Database & experiment;	Speed, Memory, Accuracy	Yes	2247+163 2410	Wood et al, 2014 & 2019
Clark	Sequence Similarity: k-mer	Configurations of itself	Simulated from Database & experiment	Speed, Memory, Accuracy	Yes	386	Ounit et al, 2015
Bracken	Sequence Similarity: k-mer	Kraken	Simulated from Database & experiment	Accuracy	Yes	218	Lu et al, 2017
Centrifuge	Sequence Similarity: k-mer	Kraken, Megablast	Simulated & experiment	Speed, Memory, Accuracy	Yes	471	Kim et al, 2016
MEGAN	Sequence Similarity: Local Aligner	Blast	Experiment	Accuracy, speed, memory	Yes	2604	Huson et al, 2007
mOTUs2	Marker gene	Metaphlan; Metaphlan2, Kraken	Simulated from Experiment; Simulated from Experiment	Accuracy, Statistics; Accuracy, Statistics,	No	246+37 283	Sunagawa et al 2013, Milanese 2019
Metaphlan2	Marker gene	PhymmBl, Phymm, RITA, NBC; Kraken	Simualted from experiment	Accuracy, speed	No	1206+962 2168	Segata et al, 2012, Truong etal , 2015
Diamond	Protein	RapSearch2, BlastX	Simulated from Database& Experiment	Accuracy, Speed, Memory	Yes	2184	Buchfink et al, 2015
Kaiju	Protein	Kraken, Clark	Simulated from Database & experiment, Mock Communities	Accuracy, Speed, Memory	Yes	355	Menzel 2016

In terms of how they were evaluated in the papers in which they were introduced, Kraken2 and Metaphlan2 each showed performance improvements over previous algorithms. Kraken2 itself has been a popular target for many of the newer algorithms that are seeking to improve upon it such as Kaiju, Clark (Ounit et al., 2015), Centrifuge (Kim et al., 2016), or Metaphlan2. The authors of Diamond did not attempt to compare their algorithm against other k-mer of marker gene classifiers and limited their comparison against other protein marker-based classifiers such as Kaiju. Interestingly, Clark only compared performance against different parameters of Clark but it is still highly cited and respected in the field.

An author trying to demonstrate the superiority of a new algorithm needs not only to decide which algorithms to compare against, but which datasets to use. Suitable validation datasets could be in the form of an in silico mock community, real mock community, or experimental data. **In silico mock** (simulated mock communities) comprise mock samples created from random combinations of assembled sequences from either databases or previously sequenced datasets. **In vitro mock** communities consist of known microbes with sequenced genomes grown in culture in the laboratory and mixed together in know proportions. **Naturally derived communities** are “real” datasets that consist of sequenced samples taken from the field or natural habitat of the microbiome under study with sample composition unknown prior to sequencing. The vast majority of classifier papers used a combination of in silco mock communities to refine and compare their classifier against others. The power of using in silico mock communities is that the “correct” answer is known with complete certainty. The disadvantage of this approach is that these simulated samples could have sequences that would not coexist in the wild or have an unrealistic equal distribution of sequences. Later benchmarking studies (McIntyre et al., 2017) and classifiers such as Kaiju incorporated in vitro communities

with known compositions into their benchmarks. These in vitro communities had some of the advantages of naturally derived community datasets and some of the advantages of in silico datasets by merging the randomness of sequencing from the laboratory environment with the foreknowledge of known community composition thus serving as a ground truth when comparing classifiers (Bokulich et al., 2020). As taxonomic classifier papers introduced new algorithms, authors usually asserted that their new or updated algorithm performed better than competitors and as such that researchers should use their latest new and improved tool.

A limitation of the developer of an algorithm also performing the benchmarking of that algorithm is that through conscious or unconscious bias, the developer of the algorithm will likely favor their algorithm over competing algorithms. These biases have been well documented and led to competitions such as CASP (Critical Assessment of protein Structure Prediction) and CAMI (Critical Assessment of Metagenomic Interpretation) in which the "correct" answer is not known and therefore bias towards an algorithm based on tuning to expected results is eliminated. The CASP competition has developers of protein folding algorithms submit results for unknown protein structures to determine the best algorithm and has been the focus of substantial recent algorithm development which has led some to claim that the protein folding problem is essentially solved (AlQuraishi, 2019). Because a crystal structure is a clear and "correct" answer to how a protein will fold, it is obvious how to score different predictions of that structure and this has led to the CASP competition being such a robust and successful part of the development of the theory and application of computational methods and machine learning to protein folding.

On the other hand, CAMI evaluates metagenome analysis by building broad benchmark datasets, setting standards for comparing performance metrics, and developing research questions from a wide spectrum of scientists. The metagenomics field has been slow to adopt

CAMI and its standards in comparison to CASP. Instead, many metagenomic analysts responded to the rising number of taxonomic classifier papers by independently rating the performance of selected classifiers through benchmarking papers (Table 2). Each of these papers compared commonly used tools through a large number of simulated and real mock communities from a diverse collection of environments and databases. Interestingly, the results of these benchmarking papers are not consistent and have shown conflicting results in determining the best performing taxonomic classifier. Lindgreen and Ye suggested k-mer algorithms (such as Kraken2) were better while Peabody and McIntyre suggested no method was truly superior and that such distinction should be based upon the type of dataset collected and database used in the analysis. This divide in the field is best documented in a meta-analysis of both classifiers and benchmarking papers by Gardner. This paper ignored traditional accuracy measurements and focused on the number of citations, how simulated datasets were generated, and a Pearson calculation of network meta-analysis (Gardner et al., 2019)). While the article's conclusions do not question the underlying quality of data for each of the benchmarking studies collected it does show that there is a bias towards k-mer based algorithms by microbiome experts and that for environmental datasets false negatives are more likely due to reference databases.

A recent Nature Methods paper (Z. Sun et al., 2020) suggested Metaphlan2 is the most accurate and that previous analysis mixed and matched sequenced abundance with taxonomic abundance especially when compared against Kraken2. The authors argue that sequence abundances can be undercounted or overcounted based on the size of the genome referenced for that read count. While taxonomic abundance comes from matching sequences to genomic markers and less likely to be biased. This interchanging of taxonomic abundances with sequence abundances can distort comparisons between classifiers and mislead biologists when conducting

statistical analysis. This paper argues against the previous conclusions of k-mer superiority and argues that each method has its merit in a given context.

Previous reporting on the differences between these classifiers has reached different conclusions but relied heavily on traditional testing for classifier accuracy. In this proposal, we are going to argue that differences in normalization do not explain the differences between Metaphlan2 and Kraken2 but rather that these differences are driven by differences in classification error rates. Through the mapping of Spearman correlations to taxonomic abundance and comparing the proportions of taxa abundance against each classifier, this dissertation will provide a simple but unique perspective not previously considered in the literature. By using natural derived communities, we can demonstrate issues with Kraken2 that have only been hinted at by prior studies and draw meaningful conclusions while avoiding some of the limitations of simulated communities described above.

Table 2: WGS Classifier Benchmarking Papers

Benchmarking Paper	Citations	Algorithms Benchmarked	Type of Data	Performance Evaluation	Conclusions
Peabody et al, 2015	126	CARMA3, CLARK, RAP Search2, Kraken, MEGAN, Metabin, MetaCV, Metaphyler, TACOA, PhymmBl, RITA, BLAST, GOTTCHA	In Silico Mock 2, In Vitro Mock Community 1	Accuracy, Memory	No method superior but noticed variation in false positive and true negative rates depends upon dataset selected
Lindgreen et al, 2016	281	CLARK, EBI, Genometa, GOTTCHA, Kraken, LMAT, MEGAN, Metaphlan, MetaPhyler, MG-RAST, mOTU, OneCodex, QIIME, Taxator-tk	In Silico Mock 6	Accuracy, Memory, Speed	Kraken, Clark, OneCodex were the most accurate in sensitivity, precision. OneCodex and Metaphlan were the quickest with Kraken, Clark having modest performance
McIntyre et al, 2017	193	MEGAN, GOTTCHA, Diamond, Metaphlan, Kraken, MetaFlow, LMAT, PhyloSift, Clark, NBC, BLAST	In Silico Mock 21, In Vitro Mock Community 17	Accuracy, Memory, Speed	Clark/Kraken is best for limiting False Negatives while MEGAN is Best for limiting False Positives. Metaphlan is the best for speed and memory constraints, Kraken is in the middle, and MEGAN are Clark are some of the slowest and most memory intensive. Database curation is most important for species classification than selection of classifier.
Ye 2019	151	Bracken, Centrifuge, CLARK, Kraken, Kraken2, k-SLAM, MegaBLAST, metaOthello, PathSeq, prophyle, taxMaps, DIAMOND, KAIJU, MMseqs2, Metaphlan2, mOTUs	In Silico Mock 14, In Vitro Mock Community 1	Accuracy, Memory, Speed	Mainly described the limitations of databases on how it influenced classifier performance. Metaphlan2 and mOTU were the least accurate while k-mer and protein based methods with the exception of Centrifuge were comparable in terms of sequences and beta diversity.
Sun et al, 2021	10	Bracken, Kraken2, Metaphlan2, mOTU, Kaiju, Diamond	In Silico Mock 25	Accuracy, Statistical Analysis	Protein Marker were the most inaccurate. Metaphlan2 was most accurate for taxonomic abundance and Kraken2 was most accurate for sequence abundance. Demonstrated that sequence count values were influenced by size of the referred genome.

1.2 Full Citation and Description of Published Work

I have one first-author and another co-first-author paper that has been published and that are briefly described in the next few pages. Below are the links to these papers.

1. Clinton, S., Johnson, J., Lambirth, K., Sun, S., Brouwer, C., Keen, O., Redmond, M., Fodor, A., & Gibas, C. (2020). Sediment Microbial Diversity in Urban Piedmont North Carolina Watersheds Receiving Wastewater Input. *Water*, *12*(6), 1557. https://doi.org/10.3390/w12061557
2. Johnson, J., Dellon, E., McCoy, A. N., Sun, S., Jensen, E. T., Fodor, A. A., & Keku, T. O. (2021). Lack of association of the esophageal microbiome in adults with eosinophilic esophagitis compared with non-EoE controls. *Journal of Gastrointestinal and Liver Diseases*, *30*(1), 17–24. https://doi.org/10.15403/jgld-3049

An additional paper related to my collaboration with the multidisciplinary North Carolina Urban Environmental Genomics Project where I am the second author.

3. Sorgen, A., Johnson, J., Lambirth, K., Clinton, S. M., Redmond, M., Fodor, A., & Gibas, C. (2021). Characterization of Environmental and Cultivable Antibiotic-Resistant Microbial Communities Associated with Wastewater Treatment. *Antibiotics*, *10*(4), 352. https://doi.org/10.3390/antibiotics10040352

In addition to the work above, I made minor contributions to the following two papers. These are not included in the dissertation but bring my total current publications to count during my Ph.D. to two first-authored and 3 collaborative papers.

4. Lambirth, K., Tsilimigras, M., Lulla, A., Johnson, J., Al-Shaer, A., Wynblatt, O., Sypolt, S., Brouwer, C., Clinton, S., Keen, O., Redmond, M., Fodor, A., & Gibas, C. (2018). Microbial

Community Composition and Antibiotic Resistance Genes within a North Carolina Urban Water System. *Water*, *10*(11), 1539. https://doi.org/10.3390/w10111539

5. Bajorek, S., Parker, L., Li, N., Winglee, K., Weaver, M., Johnson, J., Sioda, M., Gauthier, J., Lemas, D. J., Jobin, C., Lorca, G., Neu, J., & Fodor, A. A. (2019). Initial microbial community of the neonatal stomach immediately after birth. *Gut Microbes*, *10*(3), 289–297. https://doi.org/10.1080/19490976.2018.1520578

CHAPTER 2 (published): Lack of Association of the Esophageal Microbiome in Adults with Eosinophilic Esophagitis Compared with Non-EoE Controls

In this first author paper (Johnson et al., 2021), we sought to explore and document the esophageal microbiome for adults and to determine if any microbes were correlated to Eosinophilic Esophagitis (EoE) when compared to a relatively healthy control group. EoE is a reoccurring allergic and immune condition that is triggered by either food or some environmental allergy along the esophagus. The recent rise in the medical condition over the past few decades has led some researchers to suggest that changes to the gut microbiome and/or esophageal microbiome might be an environmental trigger. Under this hypothesis, the increased usage of various antibiotics might allow for microbes passing through the gastrolienal tract a new opportunity to colonize. Previous work had focused primarily on children suffering from this condition, but the authors of this paper wanted to document the esophageal microbiome of both healthy and EoE adult patients.

For this paper, I performed all the sequencing analyses, built most of the statistical models, and created all the figures. The study included a wide variety of covariates including a diversity of prescribed medications, dietary choices, and anatomical features in the esophagus. The results were mostly negative except significant association that was found with one of the most prescribed medications, Proton Pump Inhibitors (PPIs), which changed the overall esophageal microbiome. This strongly suggests that colonization of members of the Neisseriaceae family such as *Kingella* and *Eikenella* and other Proteobacteria along the esophagus could occur due to changes in the hydrogen ion concentration (pH). PPIs reduce the number of hydrogen ions in the stomach by inactivating hydrogen ion channels thus increasing pH levels to reduce acidity. Several papers (Freedberg et al., 2014; Hojo et al., 2018) have

documented how the gut microbiome changed as a result of this increase of pH to a less acidic state since some microbes are incapable of surviving in a harsh acidic environment. It would not be unreasonable to assume that such changes in pH would reduce the acidity of the esophagus walls as well as the number of events where stomach acid coats the esophagus. This altering of environmental conditions could allow for colonizing or predatory bacteria to remove other bacteria from their microbial niche while producing chemical irritants that symptomize in the form of allergies. Future studies will be needed to evaluate these hypotheses suggested by our study.

In this paper, we ran multiple 16S classifiers including Qiime, RDP, and DADA2 with the Silva 128 database as a reference. In addition to standard analyses such as PCoA Beta Diversity plots of the overall microbial community structure and linear models for identifying associated taxa, I conducted a power analysis to demonstrate how many individuals it would take to identify potential low abundant taxa with different effect sizes. This work could potentially be applied to future grants to conduct a larger study that could explore multiple diseases affecting the esophagus while providing healthy esophageal microbiomes as a reference for future papers.

Published results were compared with prior work to see if there were consistent differences across studies for adult and child EoE. This analysis did not find consistent results (data not shown). This was surprising as our collaborators had hypothesized that the esophageal microbiome would be altered by the presence of rings lining the esophagus or patients suffering from asthma among other documented variables. All of these studies including ours were limited by modest sample sizes. Future work and larger cohorts may have sufficient power to detect small differences in the microbiome. Taken together, my work in this field highlights issues when conducting microbial studies on the upper gastrointestinal tract as well as the dearth of

samples and studies when it comes to understanding the diverse human microbiome in medically important regions.

CHAPTER 3 (published): Sediment Microbial Diversity in Urban Piedmont North Carolina Watersheds Receiving Wastewater Input

In this paper (Clinton et al., 2020) where I am the co-first author, we performed a 16S sequence analysis exploring how stream microecology has been impacted by continuous exposure to wastewater treatment plant effluent in an urban setting. Unlike previous work, samples were collected from the water, the sediment in the stream, and the soil. This built upon prior work (Lambirth et al., 2018) looking at the microbiome of various stages of treatment in the wastewater treatment plant (WWTP) that fed into urban streams. We were able to collect samples upstream of the WWTP facility to serve as a control and document the microbiome prior to exposure to urban wastewater for water, sediment, and soil samples. This work demonstrated how the microbiomes in the water and more importantly the sediment and the soil were altered due to the chemicals released by the WTTP, the microbes that survived treatment, and the various biochemicals such as Antibiotic Resistant Genes (ARGs) vesicles.

By comparing undiscovered and poorly classified bacteria found in the WWTP facility we were able to show significant overlap with samples collected from the downstream environment. This was made possible by replacing traditional 16S classifiers that rely on Operational Taxonomic Units (OTUs) with the popular algorithm DADA2 that instead uses ASVs (Amplicon Sequence Variants). ASVs are sequences that can differ by up to a single nucleotide and allow researchers to infer the original biological sequence before the introduction of PCR amplification and other sequence errors (Callahan et al., 2016). Due to the relatively small number of samples collected, this study required a departure from standard analytical methods and instead utilized Fisher exact tests and pairwise dissimilarities distances to demonstrate statistical significance. These new methods combined with Beta Diversity plots and

DADA2 enabled the identification of several OTUs that are not classified by current taxonomic identification algorithms as ASVs that matched members of the *Neisseriaceae* and *Nocardiaceae* family.

All these taxa were found downstream of the WWTP and not upstream thus providing clear evidence that despite WWTP treating concentrated levels of antibiotics and other prescribed medications to sublethal levels, the WWTP did impact the downstream microbial community. *Neisseriaceae* includes bacteria that can be either symbiotes or parasites dependent upon the class of host animal and is responsible for several gastrointestinal diseases in humans such as meningitis and gonorrhea. *Nocardiaceae* mainly consists of native soil microbes that are typically benign but are opportunistic and infectious when ingested by humans with weak immune systems. These findings were of relevance to our main sponsor, Charlotte Water, who is in charge of water quality for the city of Charlotte. For this paper, I performed all the sequencing analysis, built the majority of the statistical models, and created the majority of the figures for both the main paper and the appendix.

CHAPTER 4 (published): Characterization of Environmental and Cultivable Antibiotic-Resistant Microbial Communities Associated with Wastewater Treatment

This paper (Sorgen et al., 2021), which I am a co-author, was an experimental continuation of studying the bacterial communities in the Charlotte Urban streams of Mallard creek and Sugar creek along with the Wastewater Treatment Plants (WWTPs) that deposit their treated water into each stream system. The primary author, Alicia Sorgen, who at the time was a Ph.D. student in the biology department, grew heterotrophic plates using water samples collected from not only multiple stages of WTTPs and upstream and downstream from those facilities but also from Hospitals and Residential sewage that fed into the sewer lines that are routed to the WWTPs. In order to best simulate antibiotic resistance in a laboratory setting, several plates were grown with varying combinations of two different growth media, room, and human body temperatures, and 4 different antibiotics with an additional control group having no antibiotics added to the plates. One antibiotic, Doxycycline, was effective at hampering bacterial growth on the plates with only a few cultures surviving on average. Every culturable bacteria taxon discovered in this urban stream system was resistant to at least one of the four antibiotics.

Additional findings included higher temperatures leading to a higher volume of bacteria present overall with little change in microbial diversity. Nutrient poor growth media such as Reasoner's 2A agar (R2A) in relation to the other experimented growth media Lysogeny broth (LB), led to significantly higher levels of microbial diversity indicating that such harsh conditions enable a more diverse community to arise. Of interest, the WWTP significantly reduced the number of culturable antibiotic-resistant bacteria to levels observed upstream of the WWTPs after treatment in the Aeration Tanks.

These findings were possible thanks to the primary author who grew the plates and tediously measured the number of cultures. My role in this paper was more of an active advisory position where I collaborated by writing custom software and selecting the initial statistical models to evaluate the dataset. Later work involved taking a less active role and providing feedback on how to interpret data. My final contribution was integrating data analysis from previous UEGP papers to compare with 16S sequencing from the plates for the primary author while critiquing the models and figure design. This paper provided valuable experience in collaborating with scientists and understanding the requirements of preparing microbial sequence data for publication.

CHAPTER 5 (Preprint): Systematic Classification Error Profoundly Impacts Inference in High-depth Whole Genome Shotgun Sequencing Datasets (Johnson et al., 2022)

(Parts of this chapter have been deposited in the bioarchvie at https://www.biorxiv.org/content/10.1101/2022.04.04.487034v2.abstract)

5.1 Abstract

There is little consensus in the literature as to which approach for the classification of Whole Genome Shotgun (WGS) sequences is best. In this paper, we examine two of the most popular algorithms, Kraken2 and Metaphlan2 utilizing four publicly available datasets. As expected from previous literature, we found that Kraken2 reports more overall taxa while Metaphlan2 reports fewer taxa while classifying fewer overall reads. To our surprise, however, Kraken2 reported not only more taxa but many more taxa that were significantly associated with metadata. This implies that either Kraken2 is more sensitive to taxa that are biologically relevant and are simply missed by Metaphlan2, or that Kraken2's classification errors are generated in such a way as to impact inference. To discriminate between these two possibilities, we compared the Spearman correlation coefficients of each taxa against each taxa with higher abundance from the same dataset. We found that Kraken2, but not Metaphlan2, showed a consistent pattern of classifying low abundance taxa that generated high correlation coefficients with higher abundance taxa. Neither Metaphlan2 nor 16S sequences that were available for two of our four datasets showed this pattern. Simple simulations based on a variable Poisson error rate sampled from the uniform distribution with an average error rate of 0.0005 showed strikingly strong concordance with the observed correlation patterns from Kraken2. Our results suggest that Kraken2 consistently misclassifies high abundance taxa into the same erroneous low abundance taxa creating "phantom" taxa that have a similar pattern of inference as the high abundance

source. Because of the large sequencing depths of modern WGS cohorts, these “phantom” taxa will appear statistically significant in statistical models even with a low overall rate of classification error from Kraken2. Our simulations suggest that this can occur with average error rates as low as 1 in 2,000 reads. These data suggest a novel metric for evaluating classifier accuracy and suggest that the pattern of classification errors should be considered in addition to the overall classification error rate since consistent classification errors have a more profound impact on inference compared to classification errors that do not always result in assignment to the same erroneous taxa. This work highlights fundamental questions on how classifiers function and interact with large sequencing depths and statistical models that still need to be resolved for WGS, especially if correlation coefficients between taxa are to be used to build covariance networks. Our work also suggests that despite its limitations, 16S rRNA sequencing may still be useful as neither of the two most popular 16S classifiers showed these patterns of inflated correlation coefficients between taxa.

5.2 Introduction

Improvements in sequencing technology have allowed researchers to explore the novel relationship between microbes and their environments (Hultman et al., 2015; Osman et al., 2020; Payne et al., 2017; Regalado et al., 2020). In tandem with the growth of 16S sequencing, two of the most popular classification algorithms in terms of past citations include Quantitative Insights into Microbial Ecology (QIIME) (Caporaso et al., 2010), which is often used for the creation of closed-reference or open-referenced OTUs, and the Ribosomal Database Project (RDP) (Wang et al., 2007) which represents naïve Bayesian classification. While 16S sequencing remains powerful, it has several limitations: it can only reliably generate taxonomy at genus and above, functional information can only be inferred from taxonomy, it targets only a single gene, and the process of amplifying reads via Polymerase Chain Reaction (PCR) can introduce chimeras and other errors. Therefore, researchers have increasingly been incorporating WGS sequencing into microbiome analysis since it provides functional information about potential proteins, and can at least in theory, identify taxa down to the strain levels while tracking horizontal gene transfer events (Parks et al., 2018). Unlike 16S sequencing, WGS sequencing amplifies entire genomes and it remains an open question of how best to taxonomically classify WGS sequences.

Many algorithms have been proposed to address the research challenges associated with the taxonomic classification of WGS sequences (Buchfink et al., 2015; Huson et al., 2007; Kim et al., 2016; Lu et al., 2017; Menzel et al., 2016; Milanese et al., 2019; Ounit et al., 2015; Segata et al., 2012; Sunagawa et al., 2013; Truong et al., 2015; Wood et al., 2019; Wood & Salzberg, 2014). Attempts to benchmark these algorithms have led to conflicting results (Lindgreen et al., 2016; McIntyre et al., 2017; Peabody et al., 2015; Z. Sun et al., 2020; Ye et al., 2019). One paper by Peabody et al explored the precision and accuracy of WGS classifiers but argued that

performance varied widely between different classifiers and advocated for the use of standardized datasets. In 2016, Lindgreen et al examined datasets they argued were more complex and more realistic than previous comparisons and found a high degree of variability between different tools. In 2017, McIntyre et al found that Kraken2 was among the tools with the highest read-level precision. In 2019, Ye et al argued that explicit consideration of reference databases should be an important consideration in evaluating and improving classifier performance. More recently in 2021, Sun et al argued that explicit consideration of sequence abundance vs. taxonomic abundance and how to reference genome size was incorporated into normalization schemes is a crucial and underappreciated aspect of classifier performance.

All of the benchmarking papers referenced make extensive utilization of **in silico mock** (simulated mock) communities which comprise of samples created from random combinations of assembled sequences from either databases or previously sequenced datasets. These datasets have the obvious advantage that the "correct" answer is known but the obvious disadvantage is that the mock communities might not resemble "real", or naturally derived communities, in community structure and distribution. Moreover, choices made in assembling the mock communities can bias benchmarking results in favor of one or another class of algorithms. Similar concerns can be raised about **in vitro mock** communities which consist of known microbes with sequenced genomes grown in culture in the laboratory and mixed together in know proportions. In vitro mock communities are in some sense more "real" than in silico mock communities as they involve sequencing of actual microbes, but it is still unclear how choices made in assembling the mock community could influence benchmark results or how in vitro mock communities might miss important features of complex "real" communities. Ideally, benchmarking papers could utilize **naturally derived communities**, or "real" datasets, that

consists of sequenced samples taken from the field or natural habitat of the microbiome under study with sample composition unknown prior to sequencing. While “real” samples are appropriately complex and do not require the choices involved in constructing mock communities, the obvious disadvantage of using these samples is that the correct answer is not known making it difficult to benchmark different algorithms.

In this manuscript, we consider Kraken2 and Metaphaln2, the two most widely used classification algorithms, two traits of classification performance on real datasets beyond “correct classification” that have not previously been extensively examined in benchmarking papers. The first of these is inference. For four “real” (naturally derived communities) datasets we examine the pattern of inference and ask how many significant taxa each classifier finds. In addition, we consider the structure of Spearman correlation patterns between the reported taxa for each classifier. We argue that even with an overall very low error rate and standard sequencing depth, Kraken2, but not Metaphlan2 or any 16S classifiers, reliably produces a high percentage of “phantom” taxa. This reflects systematic classification error, and can be significantly associated with metadata under study.

5.3 Methods

The four datasets that have been chosen for this comparison (Table 1) include three human gut whole genome studies (Jones et al., 2018; Schirmer et al., 2018; S. Sun et al., 2021; Winglee et al., 2017) and one pig gut whole genome classifier study (Xiao et al., 2016). All four of the datasets are available on NCBI: Vanderbilt project is PRJNA693850 (WGS) and PRJNA397450 (16S), Pig Gut study is PRJEB11755 (WGS), IBD study is PRJNA389280 (WGS), and the China study is PRJNA349463 (WGS and 16S). Samples that did not have corresponding metadata were not included in the analysis.

Reverse reads were ignored and only forward reads were included. Contaminant reads were removed from the forward fastq files using kneaddata, package in bioBakery (McIver et al., 2018), which automatically calls on trimmomatic (Bolger et al., 2014) for trimming and bowtie (Langmead & Salzberg, 2012) to align reads to contaminant genomes. Two classifiers were tested in this study due to popularity and ease of use which were Metaphlan2 and Kraken2. These classifiers constitute the two most highly cited and used forms of classification of whole genome shotgun data with Metaphlan2 representing marker-gene based classifiers and Kraken2 representing k-mer based classifiers. This study used the default database for Metaphlan2. For Kraken2, we exclusively used the prokaryotic combination of refseq databases downloaded from https://benlangmead.github.io/aws-indexes/k2 that only included archaea and bacteria as downloaded on April 2020.

Table 3: Metadata and Sequencing Depth at Genus Level							
Dataset	**Metadata**	**Sample Size WGS**	**Average Sequence Depth Kraken2**	**Average Sequence Depth Metaphlan2**	**Sample Size 16S**	**Average Sequence Depth RDP**	**Average Sequence Depth Qiime**
China	Rural vs Urban	40	3,668,438	3,463,600	40	69,021.02	40,830.70
Vanderbilt	Sample Type (stool vs. swab vs. tissue)	1,203	2,746,762	2,043,168	240	134,289.20	107,769.70
IBD	Disease vs Control	341	5,637,535	3,049,583	NA	NA	NA
Pig Gut	3 countries (China, France, Denmark)	281	1,476,263	1,017,447	NA	NA	NA

Statistical Analysis started with normalizing taxa tables from all classifiers through the Normalization Formula, $log_{10}\left(\frac{RC}{n} * 10^6 + 1\right)$, and subdividing by phylogenetic level. RC represents the raw counts' value of a given taxon for a given sample, n is the total number of sequence counts for that sample, multiplying by a million normalized the proportions to the same sequence depth to allow for comparisons across different datasets, and a pseudo count of 1 was chosen to make zeros appear as zeros after log transformation. This formula is designed to minimize differences in the impact of the pseudo count and differences caused by different datasets with different sampling depths. Nonparametric Kruskal Wallis tests were employed to highlight the differences in the microbiome for the variable with the strongest effect size in each dataset. In the case of the Vanderbilt dataset, this was the type of method for collecting samples either swabbing, tissue, or using a patient's stool. The Pig Gut dataset used the geographical location of the pig which consisted of three locations China, Denmark, and France. The IBD dataset was simply whether the patient suffered from Intestinal Bowel Disease or not. Lastly, the China dataset looked at whether the patient lived in a rural community or an urban community. All coefficients of determination and p-values reported came from these variables modeled to their datasets. FDR correction was via the Benjamini-Hochberg (Benjamini & Hochberg, 1995) method at a 5% FDR threshold.

Additional statistical tests consisted of calculating the highest Spearman correlation Rho values for each taxon when compared to every other taxa with higher abundance in the dataset. In this calculation, we sort each taxon based on its mean relative abundance and, starting with the 2nd most abundant taxon, report the highest Spearman correlation for each taxa with higher abundance. So, for the 2nd most abundant taxon, we report the Spearman correlation with the 1st most abundant taxon. For the 3rd most abundant taxon, we report the higher of the Spearman

correlation coefficients between the 3rd and the 1st taxa and the 3rd and the 2nd. And so forth until each taxa (except the most abundant) is assigned such a value. Spearman correlation coefficients were calculated on un-normalized data to avoid artifacts associated with normalization. Performing the results on log-normalized data, however, yielded nearly identical results (data not shown). Prevalence calculations counted the number of samples with non-zero values for a given taxon and reported them as a percentage of the total number of samples that contained that taxon in a given dataset. All Poisson models used the top ten most abundant taxa for a given dataset and converted them to proportions for the model. In the Poisson model, the lambda is selected by sampling from the uniform distribution in a range from zero to a maximum error range which is indicated in the results and figure legends for the different models.

5.4 Results

Kraken2 finds both more taxa and more taxa significantly associated with metadata than Metaphlan2

In this analysis, we compared average relative abundance across all samples for four previously published publicly available datasets using Metaphlan2 and Kraken2 classifiers. In addition, for each dataset, we performed inference with relevant metadata using non-parametric tests (see methods). When comparing Metaphlan2 and Kraken2 (Figures 1-2), we noticed a strikingly similar pattern across all 4 datasets with the following features: (i) Both classifiers found taxa that were not found by the other classifier, but Kraken2 reported many more of these than Metaphlan2 at both the phyla (Fig.1) and genus (Fig. 2) level; (ii) for high abundance taxa, the two classifiers agreed very strongly on the average relative log-normalized abundances for all datasets; (iii) for all datasets there were a number of low abundance taxa which Kraken2 consistently reported a lower log-normalized relative abundance than Metaphlan2; and (iv) Kraken2 not only reports more taxa but more taxa that are significantly associated with metadata at both the phyla and genus level.

Patterns of Spearman correlation between taxa suggests that many classifications from Kraken2 are "phantom" taxa that represent systematic misclassification of high-abundance taxa

The fact that Kraken2 reports not only more taxa but also more taxa significantly associated with metadata suggests two alternative hypotheses. The first is that Kraken2 is more sensitive and is reporting biologically important taxa that Metaphlan2 simply misses. An alternative, not necessarily mutually exclusive hypothesis, is that Kraken2 is misclassifying some reads in a non-random way that impacts inference. In order to discriminate between these two hypotheses, we asked for each taxon what was the highest Spearman correlation coefficient with any more abundant taxa (see methods). Surprisingly we discovered that the vast majority of the

taxa that Kraken2 identified appeared to be highly correlated with at least one more abundant taxa and therefore shared that taxa's distribution across samples (Fig. 3-6). Metaphlan2 did not show a similar correlation structure (Figs. 3-6).

Figure 1

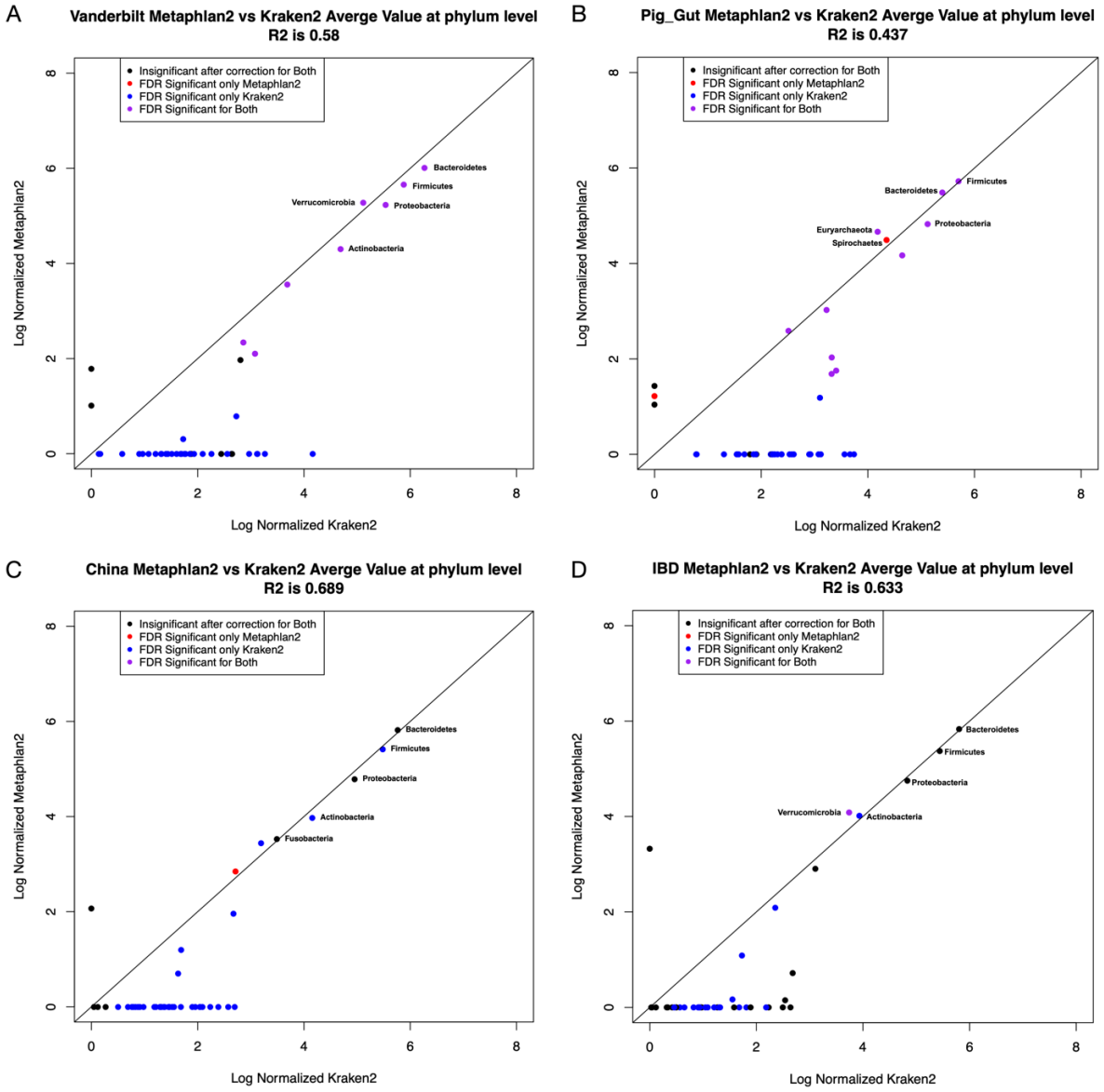


Figure 1: Kraken2 and Metaphlan2 normalized sample counts at the phylum level for our four publicly available datasets. Colors indicate the results of statistical tests measuring the association of each taxa with metadata for each dataset at a 5% FDR (see methods).

Figure 2

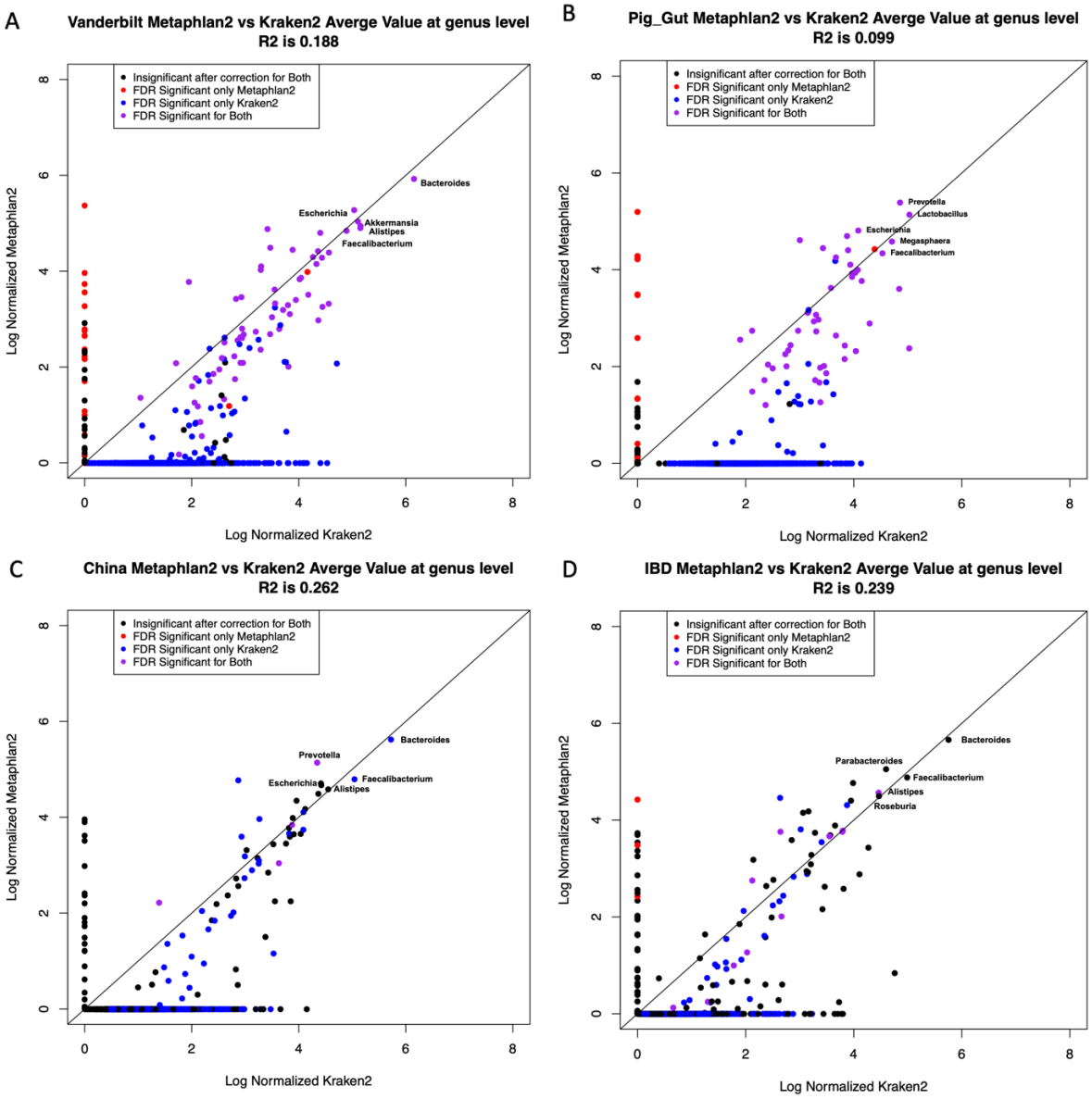


Figure 2: Kraken2 and Metaphlan2 normalized sample counts at the genus level for our four publicly available datasets. Colors indicate the results of statistical tests measuring the association of each taxa with metadata for each dataset at a 5% FDR (see methods).

Figure 3

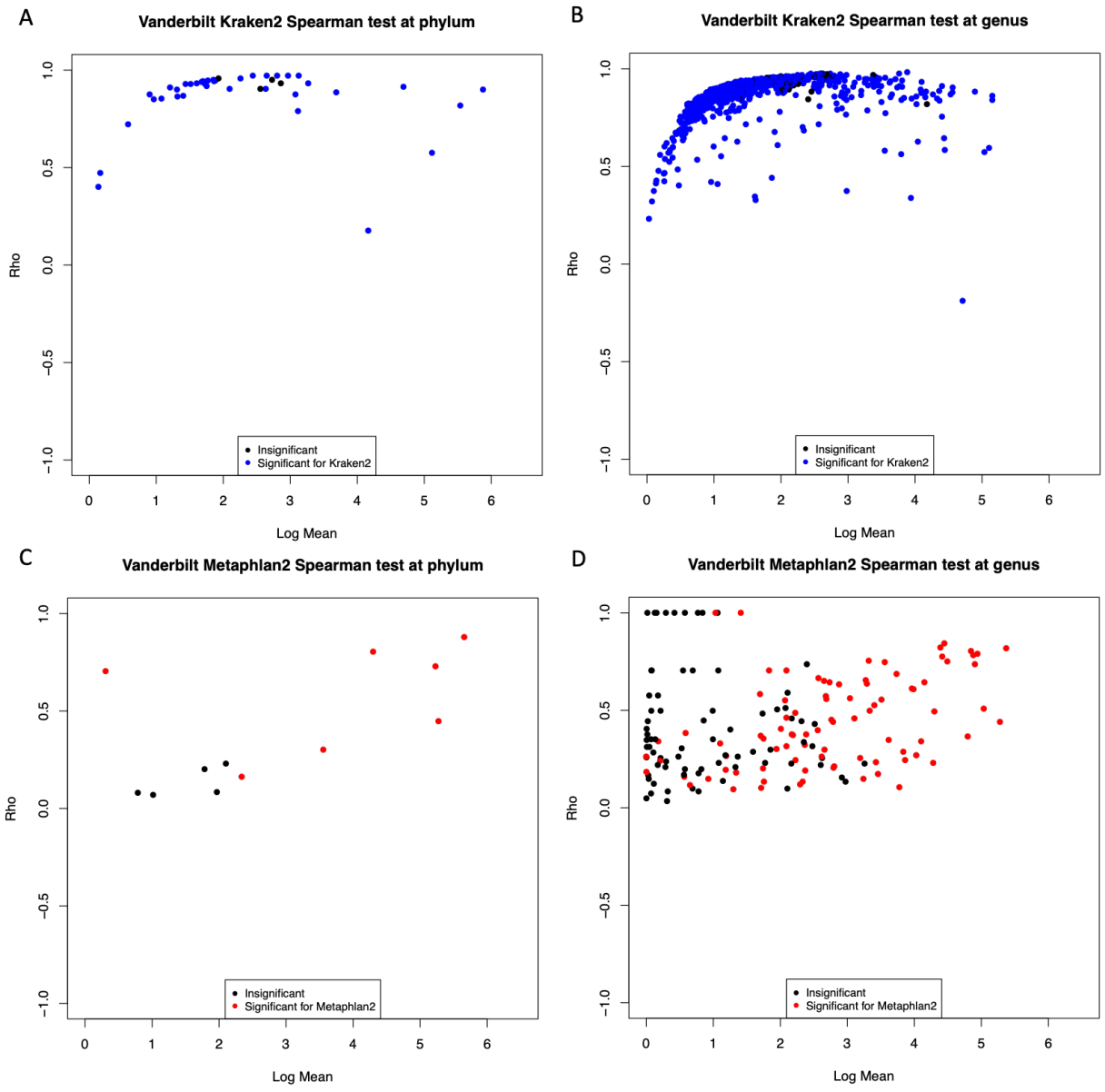


Figure 3: For each taxa, the highest Spearman correlation coefficient (y-axis) with any other higher abundance taxa in the Vanderbilt dataset. The x-axis shows the normalized log-abundance for each taxa. Colors indicate the results of statistical tests measuring the association of each taxa with differences in testing sites for the Vanderbilt dataset at a 5% FDR (see methods).

Figure 4

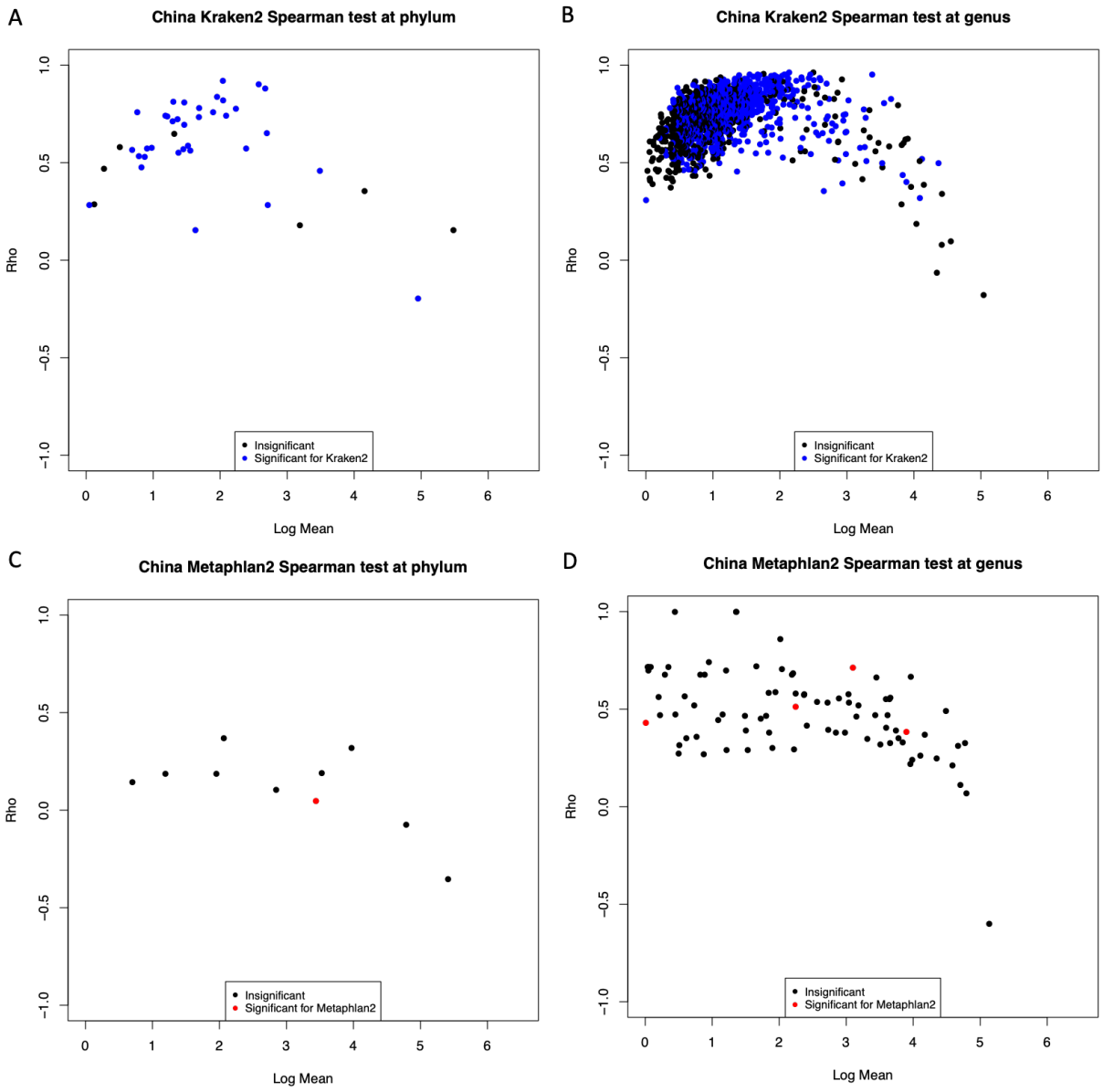


Figure 4: For each taxa, the highest Spearman correlation coefficient (y-axis) with any other higher abundance taxa in the China dataset. The x-axis shows the normalized log-abundance for each taxa. Colors indicate the results of statistical tests measuring the association of each taxa with differences in geographical urban density reduced to a binary choice of urban or rural for the China dataset (see methods).

Figure 5

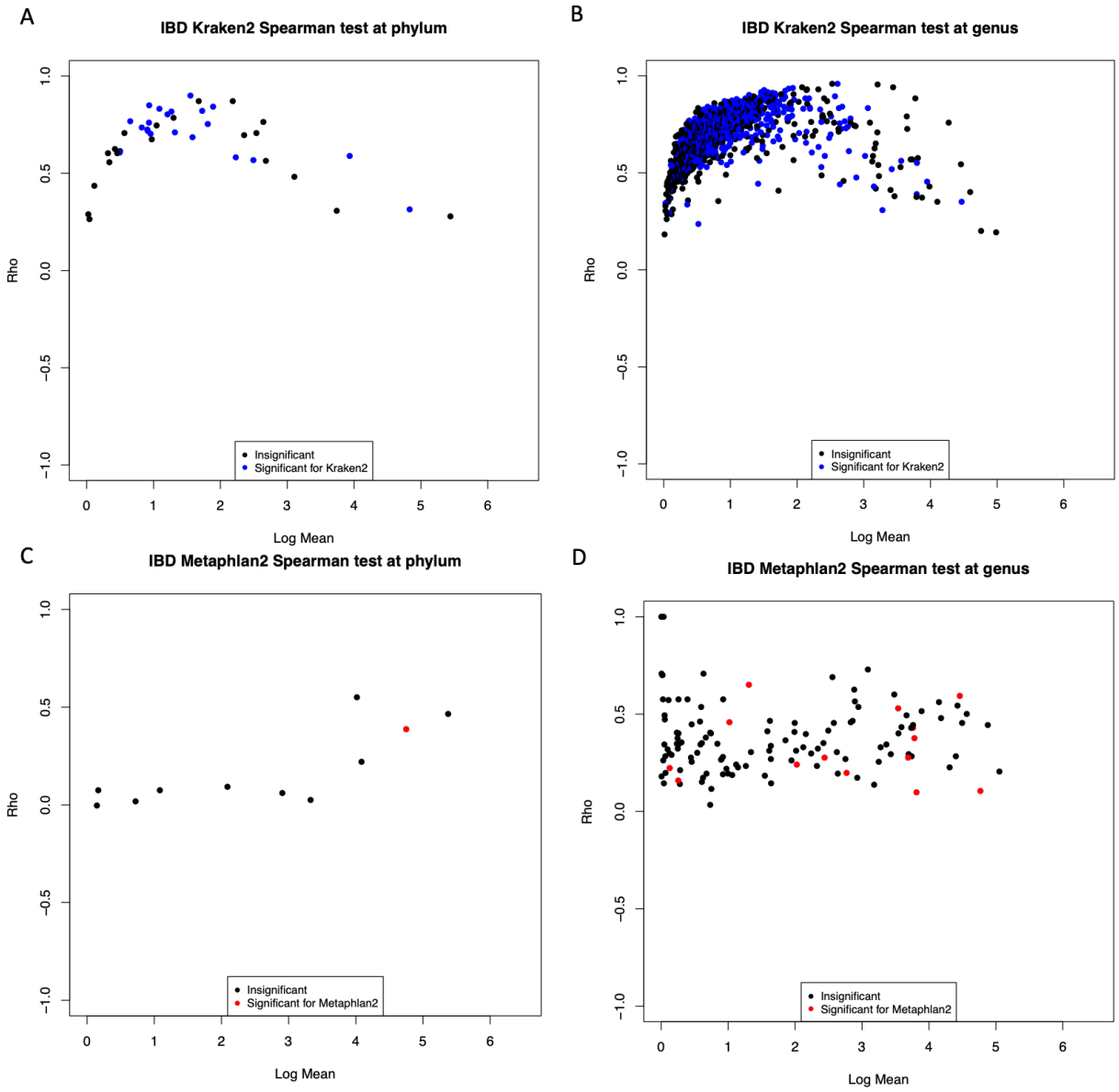


Figure 5: For each taxa, the highest Spearman correlation coefficient (y-axis) with any other higher abundance taxa in the IBD dataset. The x-axis shows the normalized log-abundance for each taxa. Colors indicate the results of statistical tests measuring the association of each taxa with differences in case/control for IBD (see methods).

Figure 6

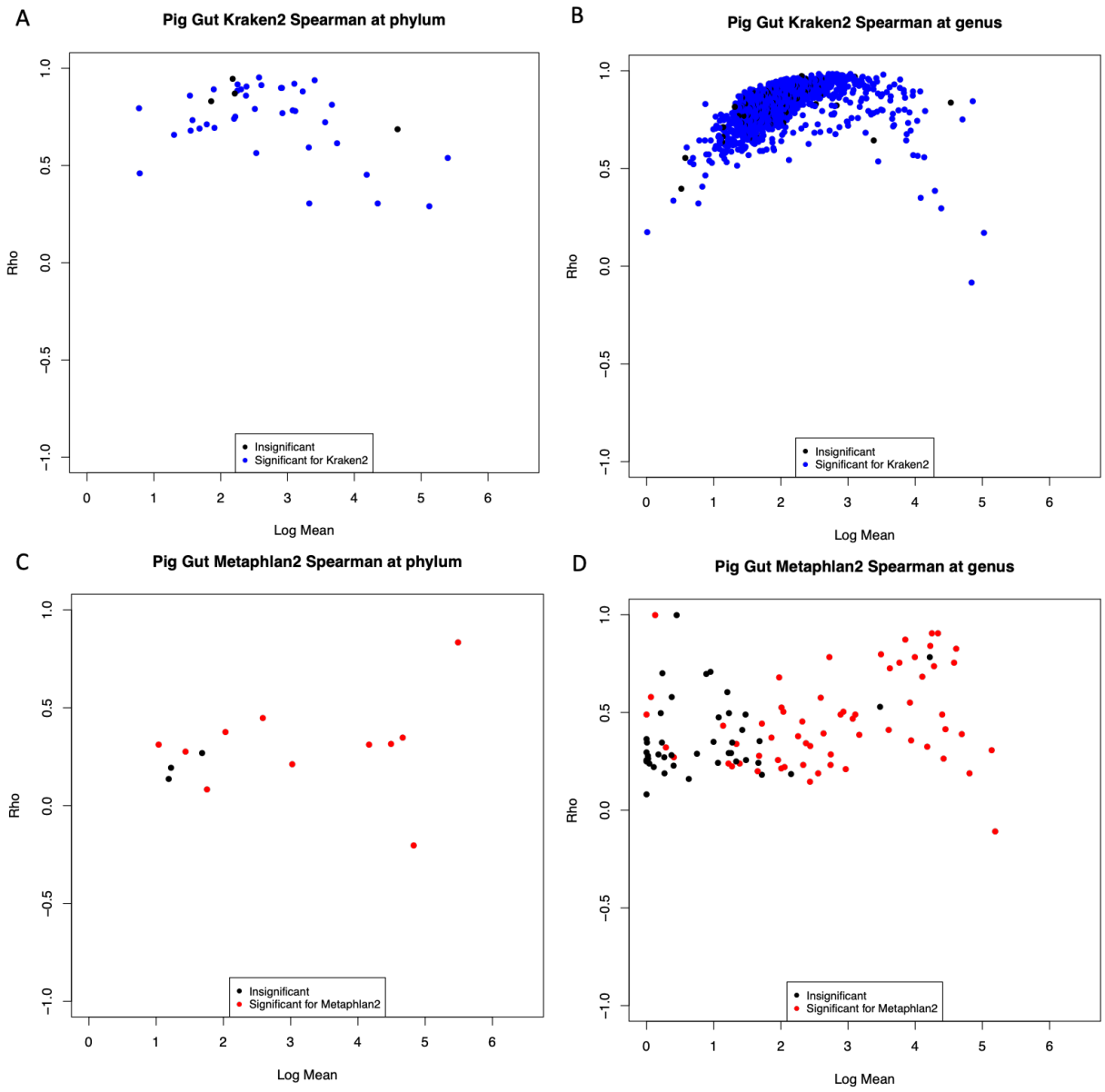


Figure 6: For each taxa, the highest Spearman correlation coefficient (y-axis) with any other higher abundance taxa in the Pig Gut dataset. The x-axis shows the normalized log-abundance for each taxa. Colors indicate the results of statistical tests measuring the association of each taxa with differences in the geographical location of the pig gut microbiomes observed for the Pig Gut dataset (see methods).

One possible explanation for this observation is that there exists in these samples many low abundance taxa that are truly correlated with the high abundance taxa that Kraken2 can classify but that are ignored by the more conservative Metaphlan2. To explore this possibility, we examined matched 16S sequences which were available for two of our four datasets. We classified the 16S sequences with RDP and QIIME and found that like WGS classifiers, the two 16S classifiers showed reasonable concordance for high abundance taxa but distinct differences for many lower abundance taxa (Fig. 7). However, when we calculated for each taxa the maximum Spearman correlation coefficient with another higher abundance taxa, we did not observe the high degree of correlations that we found for Kraken2 (Fig. 8-9). Despite Kraken2 and Metaphlan2 tending to agree with both 16S classifiers for the high abundance taxa (Fig. 10-11), Kraken2 again detects more lower-abundance taxa than Metaphlan2 or either of the 16S classifiers. If the highly correlated low-abundance taxa detected by Kraken2 were truly present, it is reasonable to assert that the correlation structure would be detected by both WGS and 16S. The fact that we did not observe this leads us to argue that the low abundance taxa detected by Kraken2 are better explained as a classification error.

Figure 7

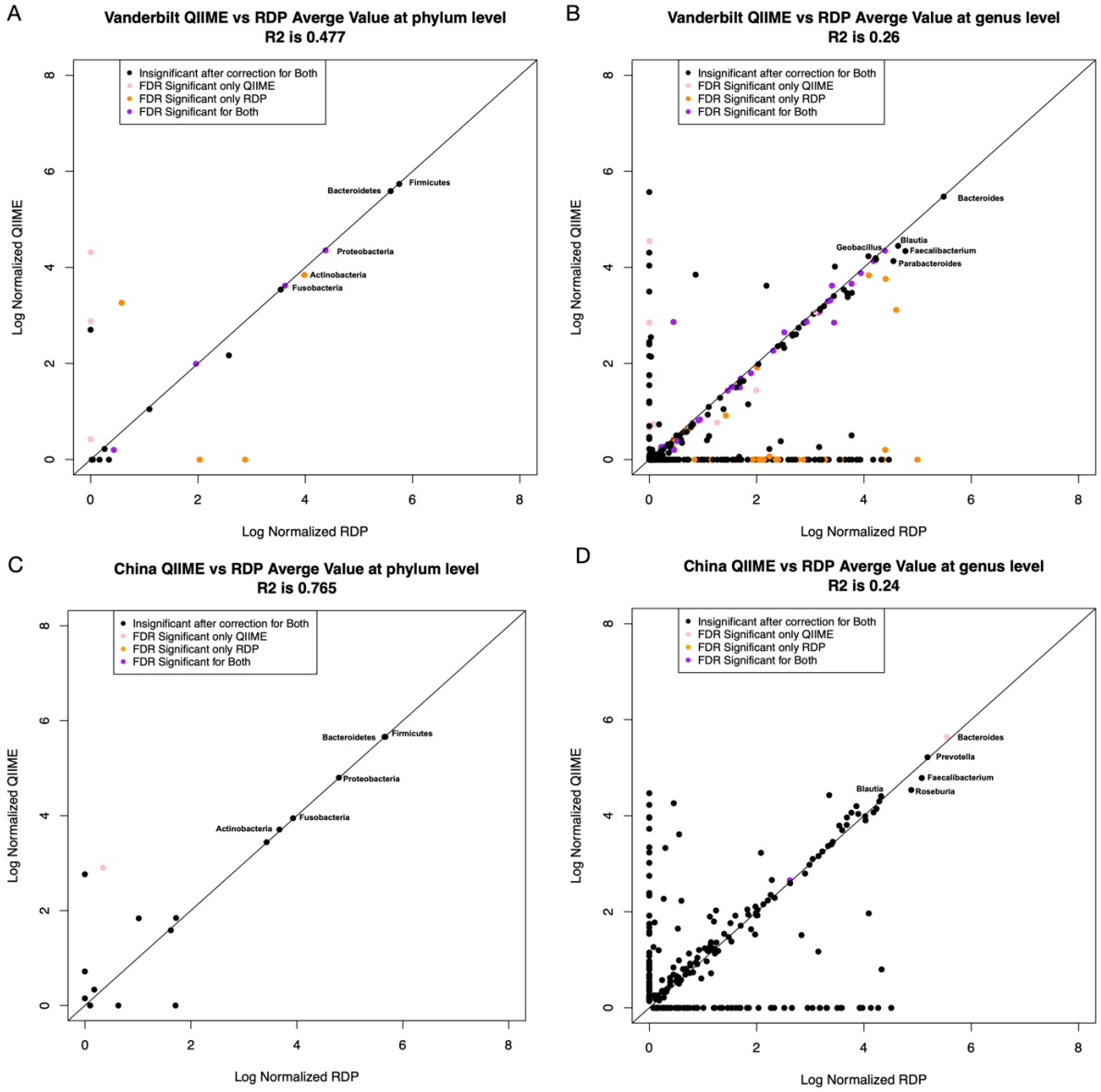


Figure 7: Normalized sample counts at the phylum and genus level were averaged by taxa for both 16S classifiers RDP on the x-axis and QIIME on the y-axis. Phylum level is on the left column and the Genus level is on the right column. A) Vanderbilt Phylum B) Vanderbilt Genus C) China Phylum D) China genus.

Figure 8

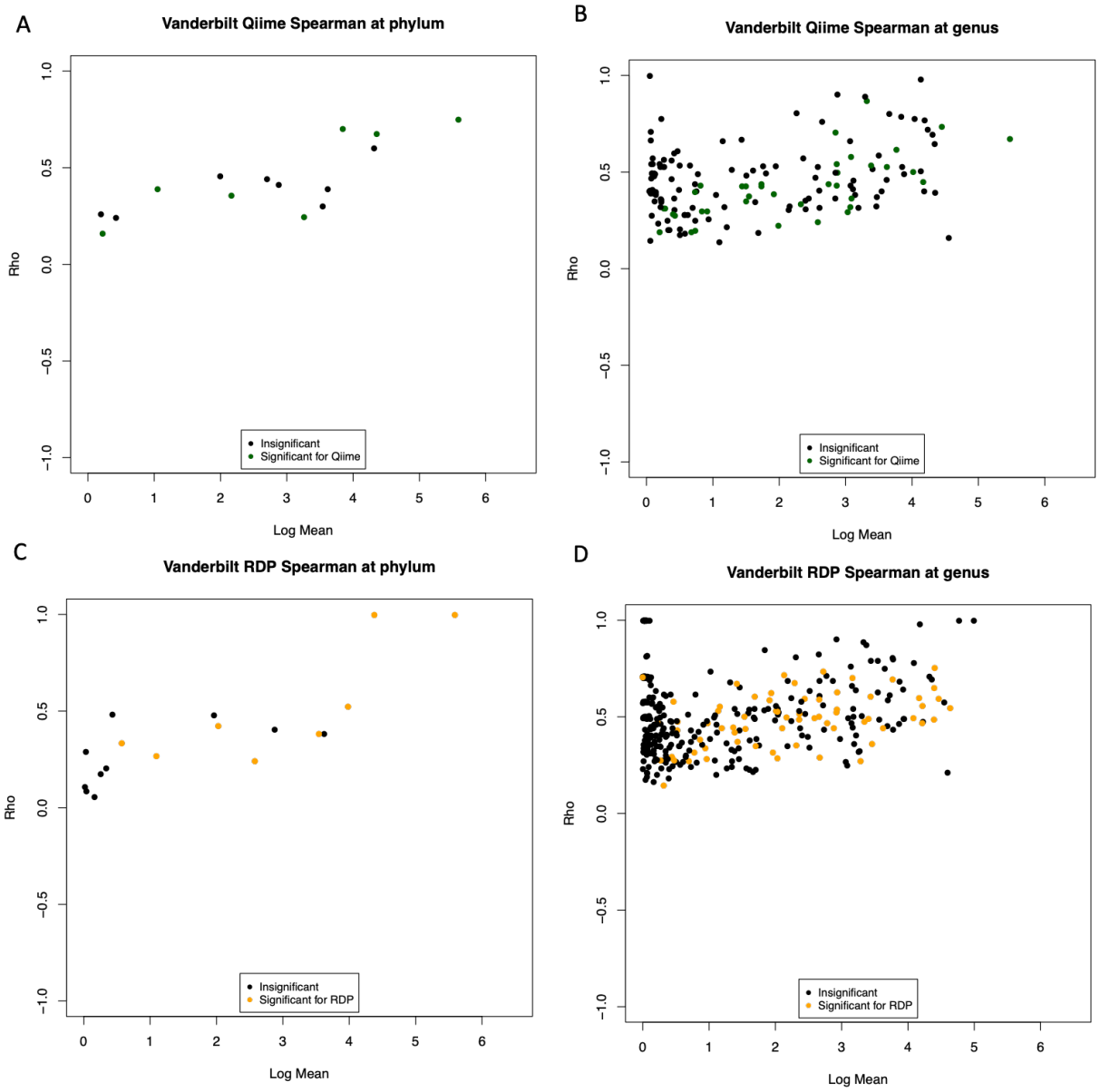


Figure 8: Spearman Correlations (the highest correlation coefficient (y-axis) with any other taxa in each dataset) for QIIME and RDP for the Vanderbilt dataset plotted against the log average mean for each taxa. Figures (A-B) are for QIIME and Figures (C-D) are for RDP. Phylum level is on the left column (A and C) and the Genus level is on the right column (B and D).

Figure 9

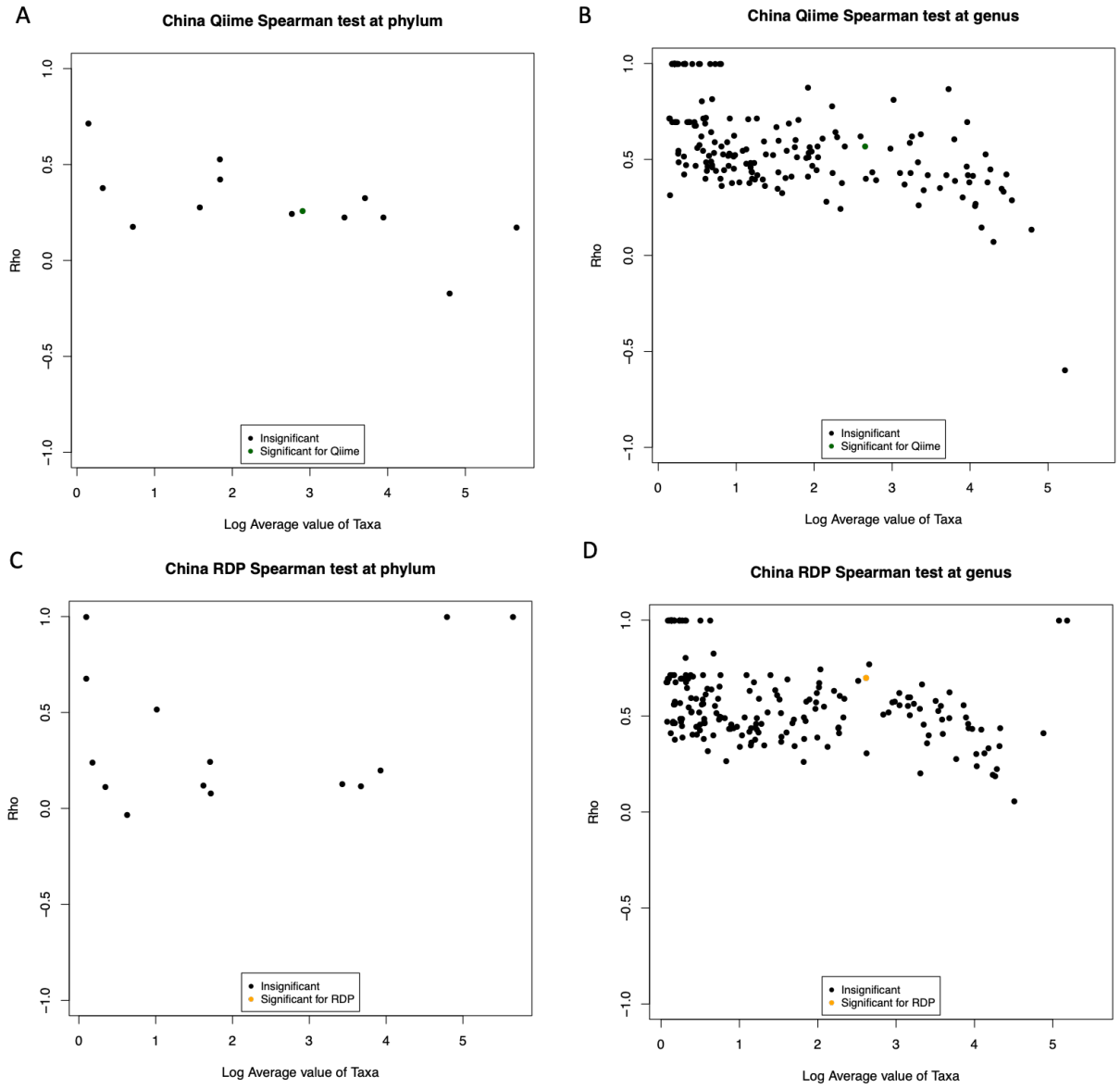
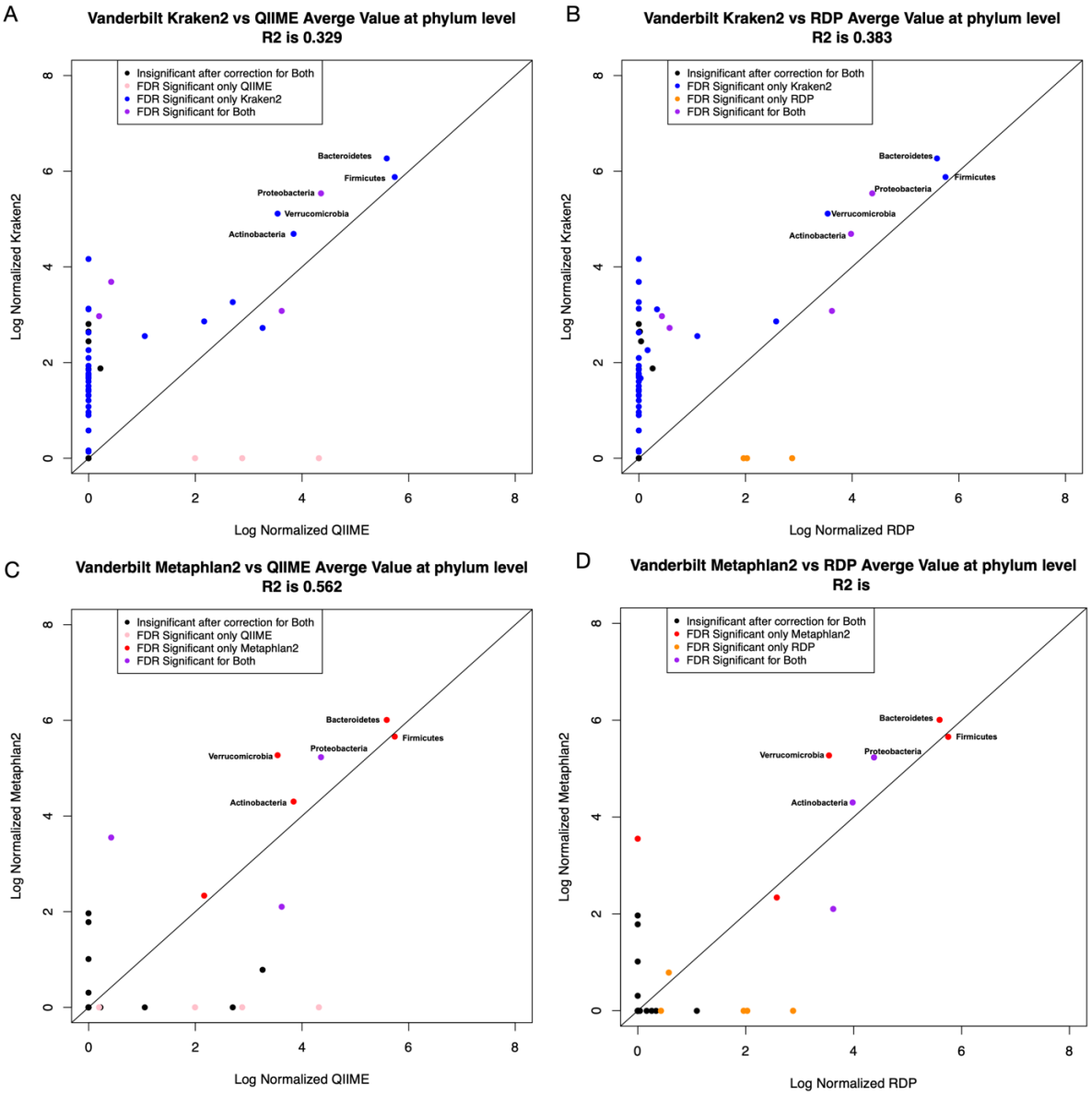


Figure 9: Spearman Correlations (the highest correlation coefficient (y-axis) with any other taxa in each dataset) for QIIME and RDP for the China dataset plotted against the log average mean for each taxa. Figures (A-B) are for QIIME and Figures (C-D) are for RDP. Phylum level is on the left column (A and C) and the Genus level is on the right column (B and D).

Figure 10


A
Vanderbilt Kraken2 vs QIIME Averge Value at phylum level
R2 is 0.329
Insignificant after correction for Both
FDR Significant only QIIME
FDR Significant only Kraken2
FDR Significant for Both
Bacteroidetes
Firmicutes
Proteobacteria
Verrucomicrobia
Actinobacteria
Log Normalized Kraken2
Log Normalized QIIME
B
Vanderbilt Kraken2 vs RDP Averge Value at phylum level
R2 is 0.383
Insignificant after correction for Both
FDR Significant only Kraken2
FDR Significant only RDP
FDR Significant for Both
Bacteroidetes
Firmicutes
Proteobacteria
Verrucomicrobia
Actinobacteria
Log Normalized Kraken2
Log Normalized RDP
C
Vanderbilt Metaphlan2 vs QIIME Averge Value at phylum level
R2 is 0.562
Insignificant after correction for Both
FDR Significant only QIIME
FDR Significant only Metaphlan2
FDR Significant for Both
Bacteroidetes
Firmicutes
Proteobacteria
Verrucomicrobia
Actinobacteria
Log Normalized Metaphlan2
Log Normalized QIIME
D
Vanderbilt Metaphlan2 vs RDP Averge Value at phylum level
R2 is
Insignificant after correction for Both
FDR Significant only Metaphlan2
FDR Significant only RDP
FDR Significant for Both
Bacteroidetes
Firmicutes
Proteobacteria
Verrucomicrobia
Actinobacteria
Log Normalized Metaphlan2
Log Normalized RDP

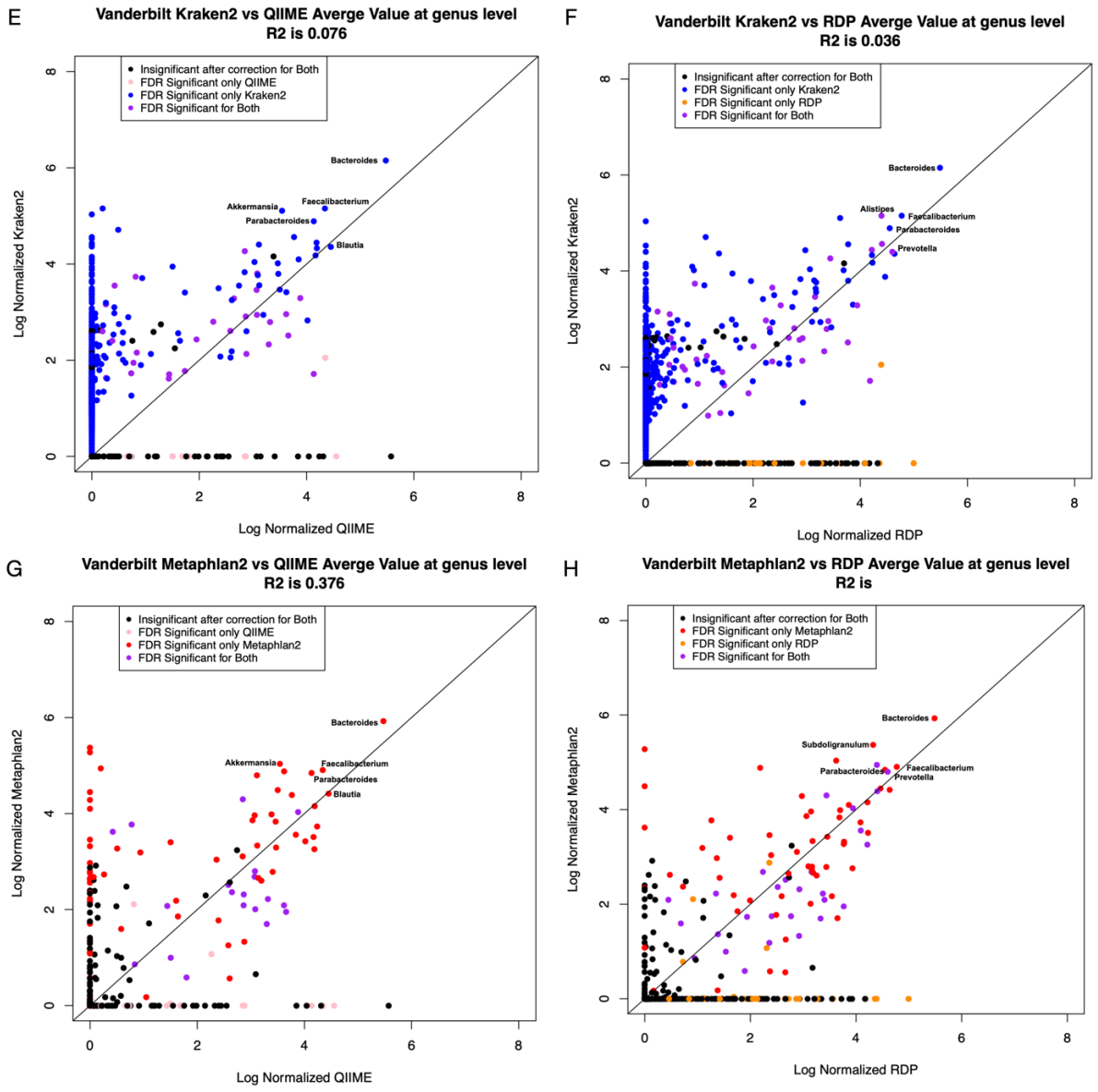
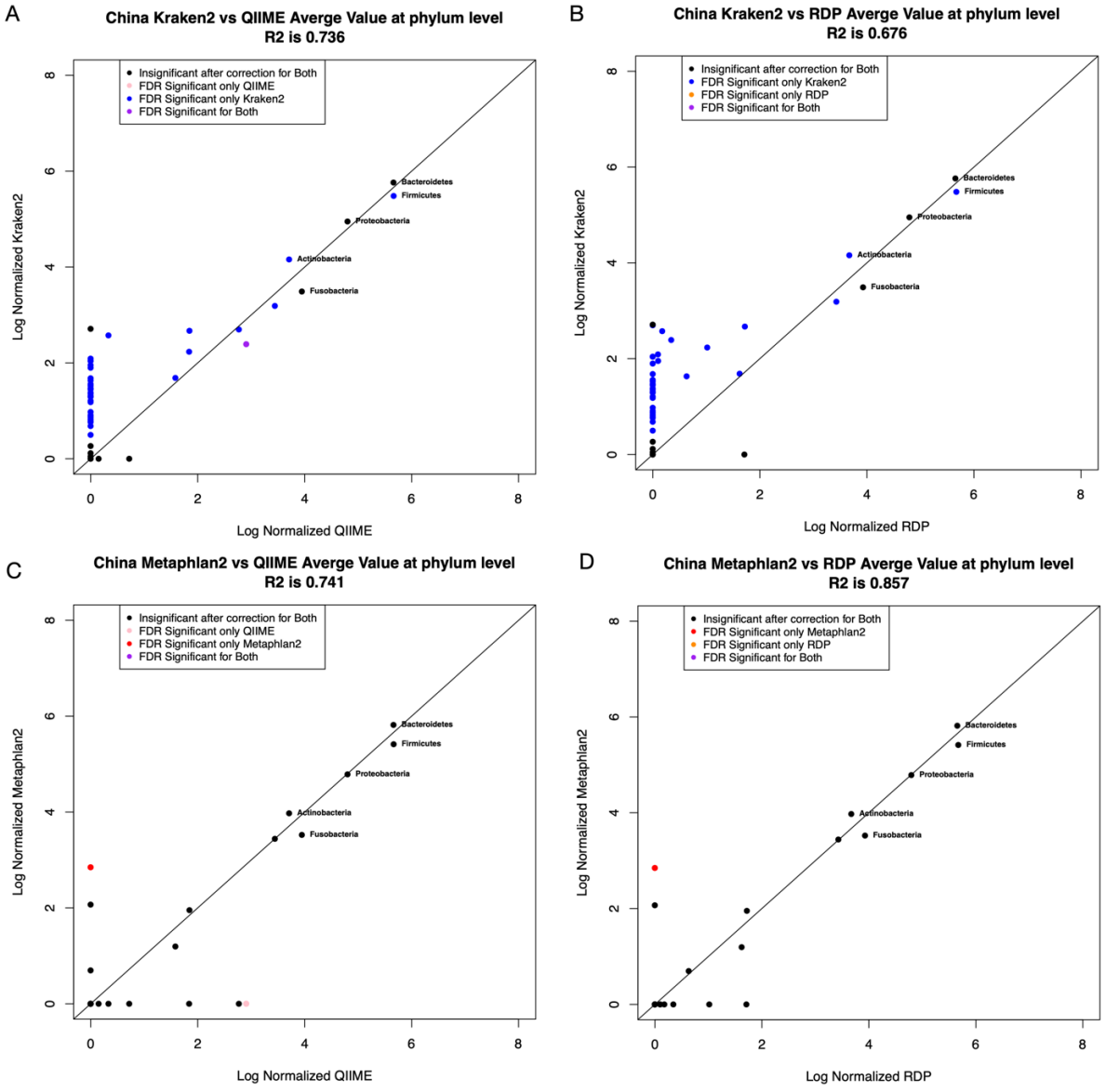


Figure 10: Normalized sample counts at the phylum and genus level were averaged by taxa for both 16S classifiers and WGS Classifiers in the Vanderbilt dataset. Figures A-D are at the phylum level while Figures E-H are at the genus level. Figures A-B and Figures E-F compare Kraken2 against QIIME (A and E) and RDP (B and F) while Figures C-D and Figures G-H compare Metaphlan2 to QIIME (C and G) and RDP (D and H).

Figure 11


A
China Kraken2 vs QIIME Averge Value at phylum level
R2 is 0.736
Insignificant after correction for Both
FDR Significant only QIIME
FDR Significant only Kraken2
FDR Significant for Both
Bacteroidetes
Firmicutes
Proteobacteria
Actinobacteria
Fusobacteria
Log Normalized Kraken2
Log Normalized QIIME
B
China Kraken2 vs RDP Averge Value at phylum level
R2 is 0.676
Insignificant after correction for Both
FDR Significant only Kraken2
FDR Significant only RDP
FDR Significant for Both
Bacteroidetes
Firmicutes
Proteobacteria
Actinobacteria
Fusobacteria
Log Normalized Kraken2
Log Normalized RDP
C
China Metaphlan2 vs QIIME Averge Value at phylum level
R2 is 0.741
Insignificant after correction for Both
FDR Significant only QIIME
FDR Significant only Metaphlan2
FDR Significant for Both
Bacteroidetes
Firmicutes
Proteobacteria
Actinobacteria
Fusobacteria
Log Normalized Metaphlan2
Log Normalized QIIME
D
China Metaphlan2 vs RDP Averge Value at phylum level
R2 is 0.857
Insignificant after correction for Both
FDR Significant only Metaphlan2
FDR Significant only RDP
FDR Significant for Both
Bacteroidetes
Firmicutes
Proteobacteria
Actinobacteria
Fusobacteria
Log Normalized Metaphlan2
Log Normalized RDP

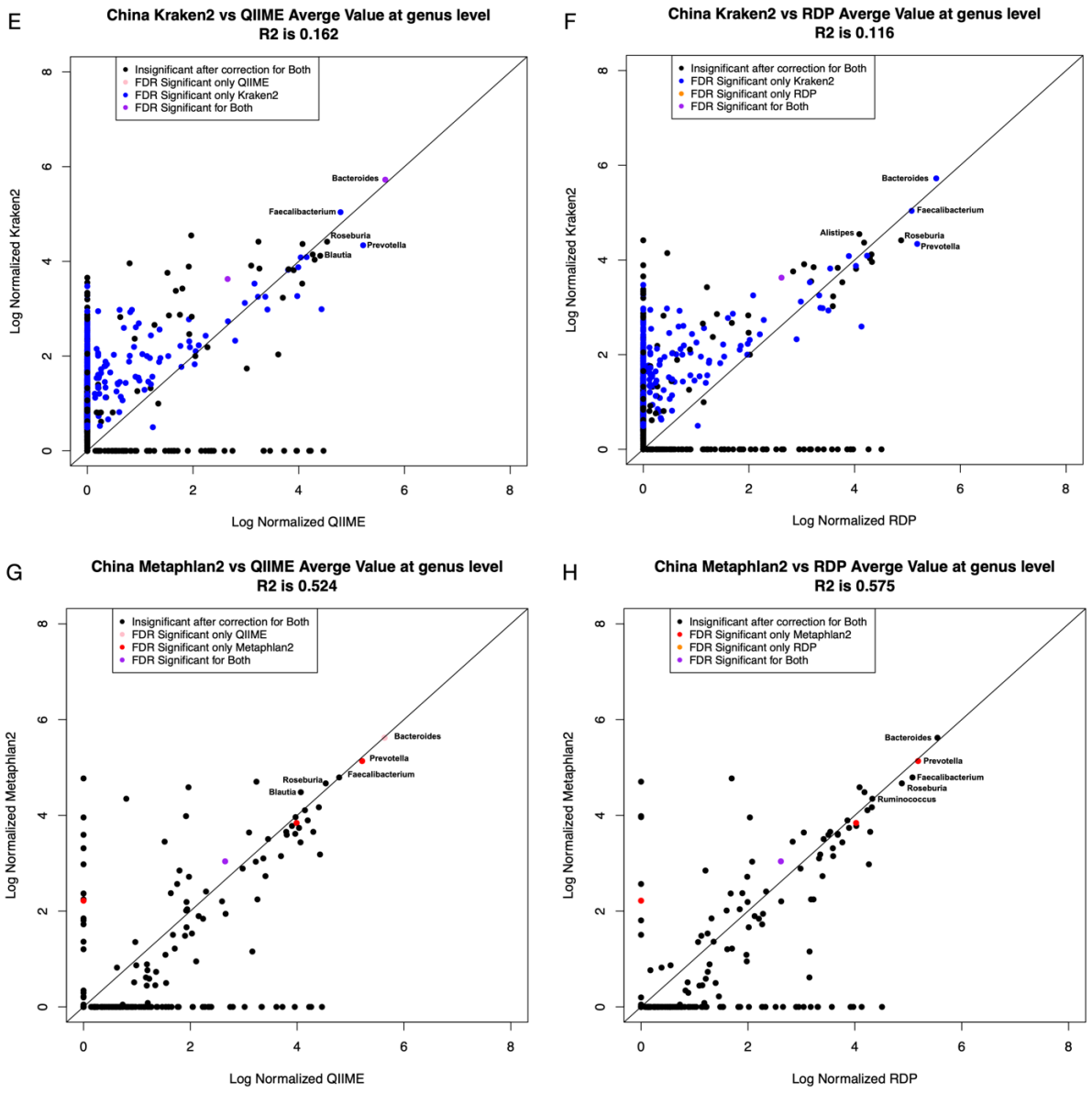
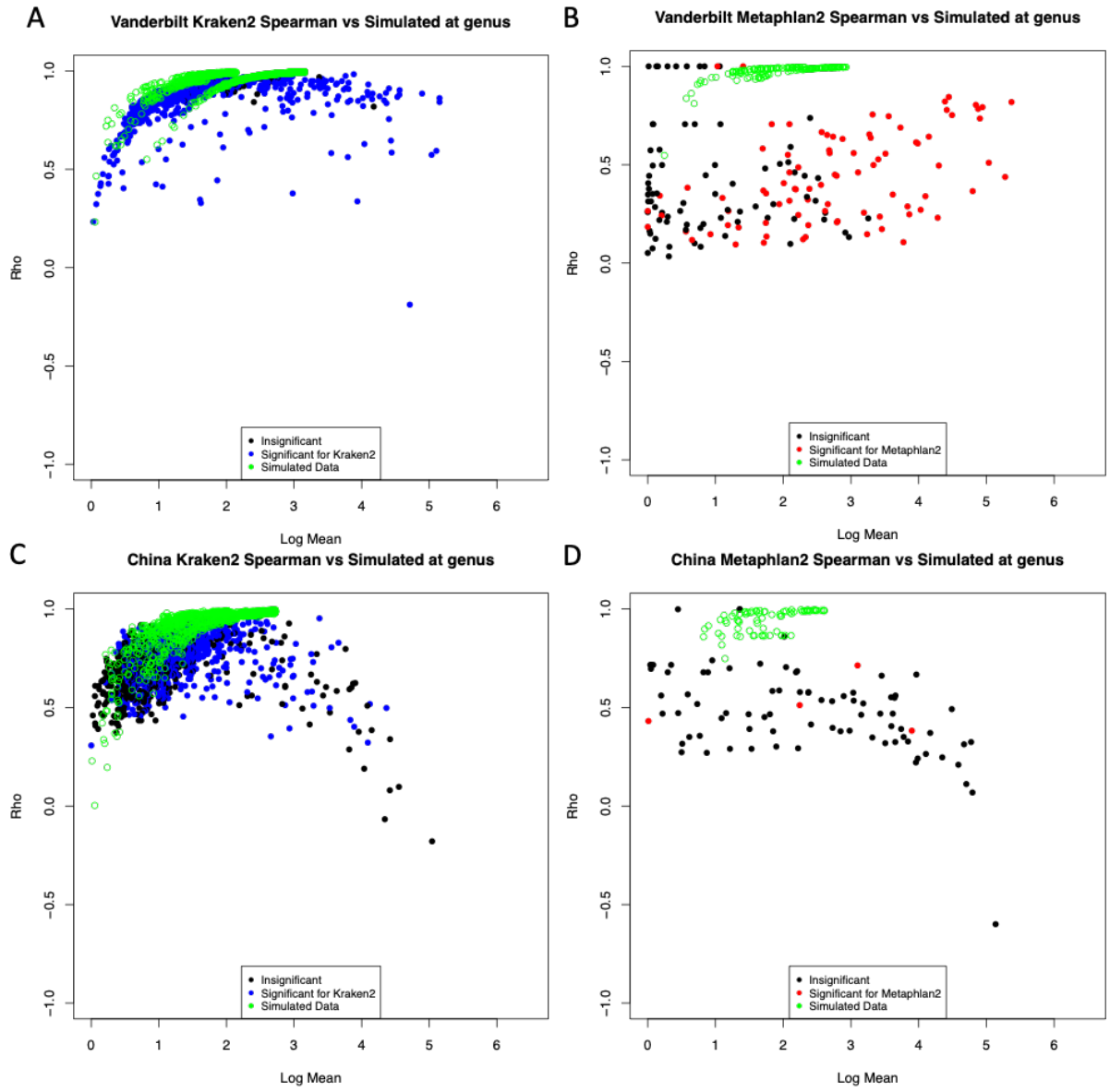


Figure 11: Normalized sample counts at the phylum and genus level were averaged by taxa for both 16S classifiers and WGS Classifiers in the China dataset. Figures A-D are at the phylum level while Figures E-H are at the genus level. Figures A-B and Figures E-F compare Kraken2 against QIIME (A and E) and RDP (B and F) while Figures C-D and Figures G-H compare Metaphlan2 to QIIME (C and G) and RDP (D and H).

A very simple Poisson model of classification error with a low error rate sampled from the uniform distribution shows good concordance with Kraken's correlation structure

In order to test the idea that "phantom" taxa could be generated from very simple models of misclassification, we created a simple Poisson-based model. In this model, we assumed that the 10 most abundant taxa in a sample are "real". Then for each taxa not in the top ten, we estimate an abundance in each sample with an assumption that the abundance was only caused by classification error. For this process, we choose one of the 10 most abundant taxa in a weighted way. For example, if the most abundant taxa represented 50% of all the reads of the 10 most abundant taxa, we would choose those taxa 50% of the time. If the next most abundant taxa were present 30% of the time, we would choose those taxa 30% of the time and so forth. We then assume an error rate that is different for each "phantom" taxa sampled from the uniform distribution from 0 to 0.001 (for an average error rate of 1 in 2000 reads). For each taxa in each sample, we use the Poisson distribution with lambda set to the number of reads of the chosen "real" taxa in that sample times the error rate sampled from the uniform distribution. In this way, we generate a counts table in which the top 10 taxa are what was observed from the sequencer and all the other taxa represent classification error. As can be seen in Figures 12-13, for Kraken2 the correlation structure of these tables in which we graph the correlation coefficient of each "phantom" taxa to its "real" parent (green dots in Figs. 12-13) is a very reasonable match to the actual correlation structure where we graph the maximum correlation coefficient of each taxa to all other taxa in the sample (non-green dots in Figs. 12-13). This model, however, fails to match the correlation patterns of either Metaphlan2 or either 16S classifiers.

Figure 12


A
Vanderbilt Kraken2 Spearman vs Simulated at genus
Rho
Log Mean
Insignificant
Significant for Kraken2
Simulated Data
B
Vanderbilt Metaphlan2 Spearman vs Simulated at genus
Rho
Log Mean
Insignificant
Significant for Metaphlan2
Simulated Data
C
China Kraken2 Spearman vs Simulated at genus
Rho
Log Mean
Insignificant
Significant for Kraken2
Simulated Data
D
China Metaphlan2 Spearman vs Simulated at genus
Rho
Log Mean
Insignificant
Significant for Metaphlan2
Simulated Data

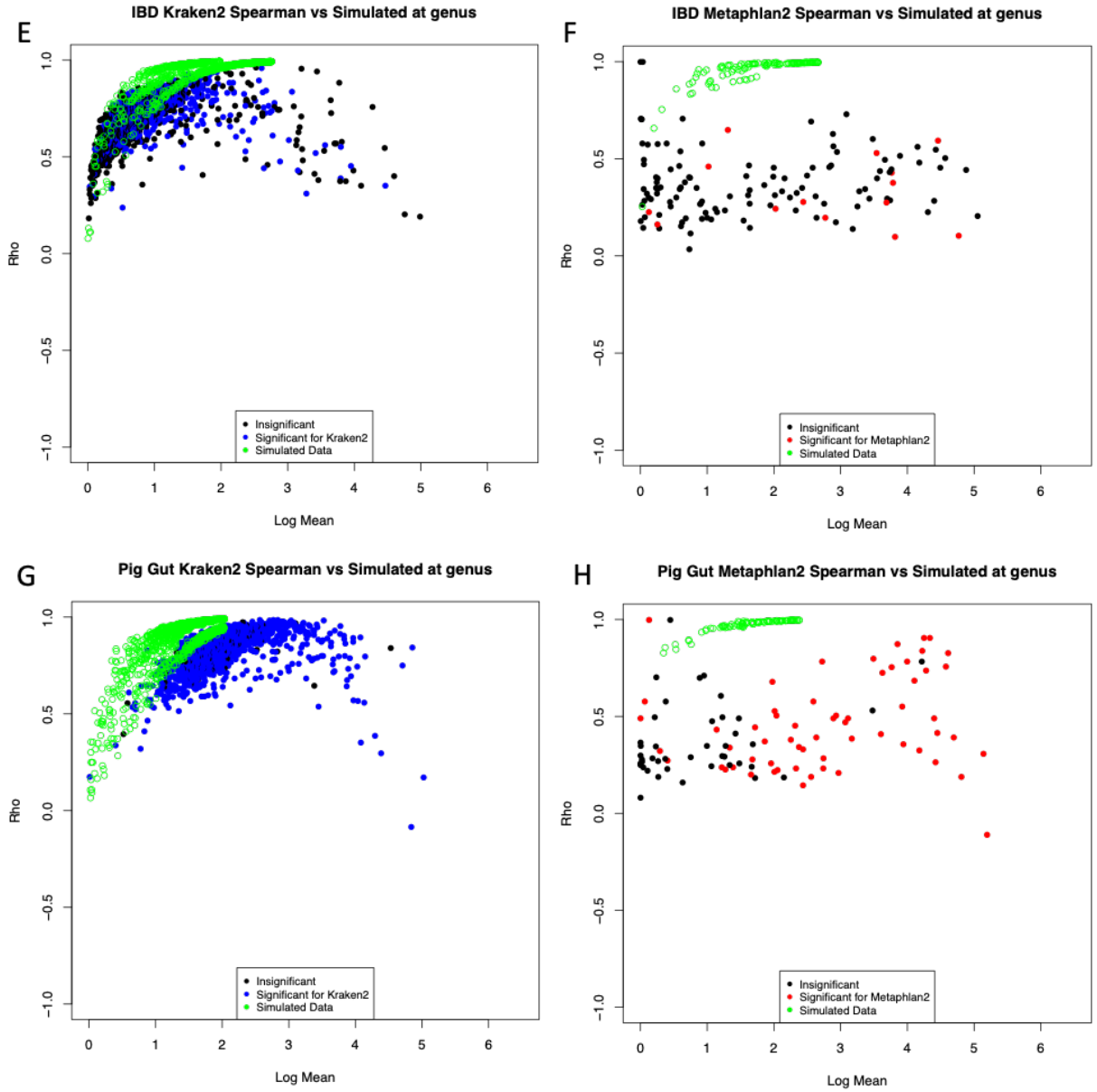


Figure 12: Normalized sample counts averaged by taxa at the genus level were compared against a simulated binomial distribution of phantom taxa labeled in green. Kraken2 results are shown in the left column (A,C,E,G) while Metaphlan2 results are shown in the right column (B,D,F,H). The y-axis is the highest Spearman correlation against each other high abundance taxa.

Figure 13

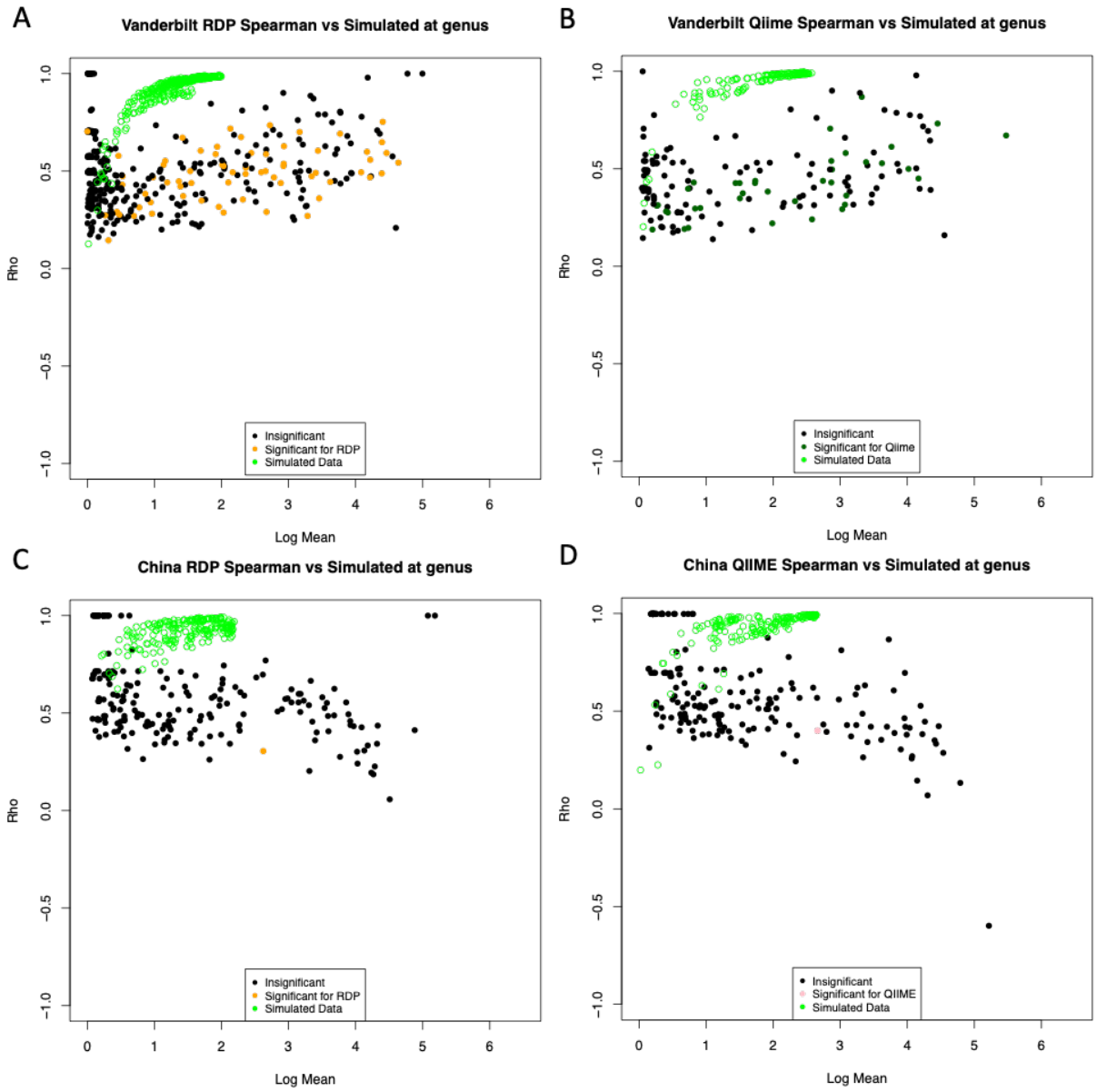


Figure 13: 16S normalized sample counts averaged by taxa at the genus level were compared against a simulated binomial distribution of phantom taxa labeled in green. RDP results are shown in the left column (A,C) and QIIME results are shown in the right column (B,D). The y-axis is the highest Spearman correlation against each other high abundance taxa.

Poisson Error model fits Kraken2 reporting of Non-zero count values by sample for every taxonomy in the dataset

As another test of our simple Poisson model, we examined the prevalence (or % of samples that are not zero) for each taxa as a function of the abundance of each taxa (Figure 14). Metaphlan2 results are strongly skewed towards having many zeros in the spreadsheet, as can be seen from the fact that Metaphlan2 reports very low prevalence when compared to 16S sequences for the two datasets for which we have 16s data. By contrast, Kraken2 reports a very high prevalence as if each taxa were biologically present in each sample. The relationship between the Poisson Error Model results and Kraken2 is strong but requires some changes in the error rate. The error rate was set to 1/200 for both the Pig Gut and Vanderbilt while the error rate was set to 1/50,000 for both the China and IBD dataset.

Figure 14

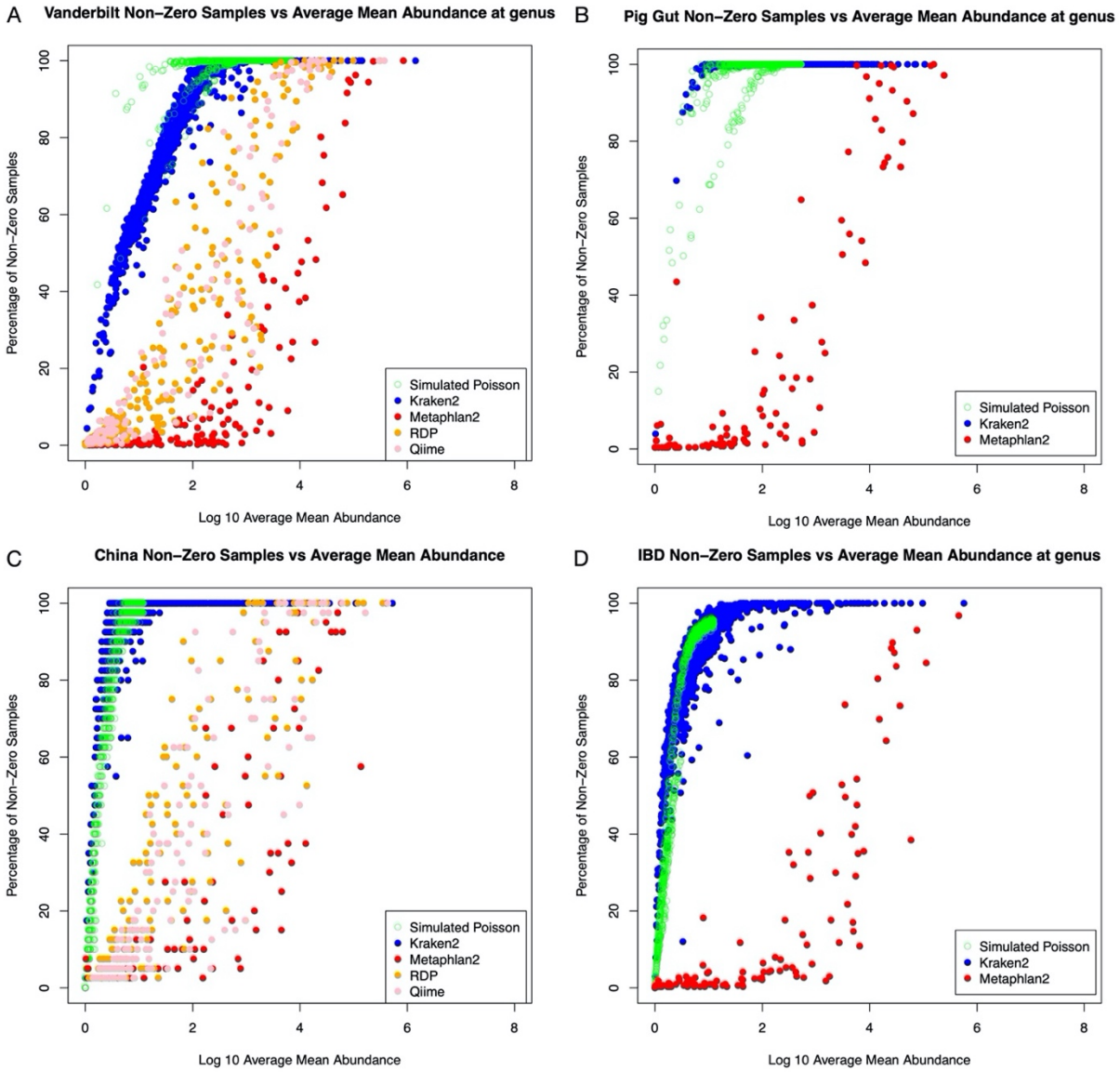


Figure 14: WGS and 16S normalized sample counts averaged by taxa at the genus level were compared against the prevalence of non-zero counts for that taxa at the genus level on the y-axis (A,C). While only both WGS classifiers were compared in the right column since those datasets lacked companion 16S data (B,D). The Simulated Poisson error model of phantom taxa was plotted in green after calculating their respective prevalence of non-zero counts as a percentage of all samples for given phantom taxa.

5.5 Discussion

Our analysis has immediate implications for biologists who wish to use taxonomic patterns for inference based on WGS data. For the most highly abundant taxa, the use of Kraken2 or Metaphlan2, or for that matter analysis of 16S sequences with QIIME or RDP, will yield highly similar answers. If there are no high abundance taxa that are significantly associated with metadata variables of interest, then likewise the choice of sequencing technology or algorithm will likely have a limited impact on inference. However, as is often the case in biological datasets and appears to be the case for the four datasets that we have chosen, if there is a high abundance taxa that is associated with metadata, and a sufficient sequencing depth, there is a possibility that Kraken2, but not Metaphlan2 or 16S classifiers, will have a small but systematic classification error in which members of the high abundance taxa are mischaracterized as a different taxon. Evidence for this assertion can be seen in our observation that Kraken2, but not Metaphlan2 or 16S classifiers, routinely produces many taxa that have very high correlation coefficients with more abundant taxa. This leads to the danger that these "phantom" taxa that share a distribution with a "real" more abundant taxon that is correlated with metadata, will produce a spuriously significant result in inference. Such a result lead to the conclusion of a significant association with a taxon that is either not present in the samples or present but with a distribution that has been distorted by a mix of true reads and misclassifications from a higher abundance taxa.

Our very simple Poisson-based classification error model was able to capture much of the pattern of the correlation structure of Kraken2, but not Metaphlan2 or the 16S classifiers. The model shows that even with a very low average classification error (an average error rate of 1 read in 2,000), this type of behavior can produce many spurious correlations given adequate

sequencing depth. It is somewhat remarkable that our model with sampling from the uniform distribution with a single parameter (the average error rate of 0.001) was able to describe the pattern of correlation for all the datasets so well with only the Pig Gut dataset at the genus level suggesting our model had underestimated the error rate. These results suggest an overall fairly consistent pattern of error across datasets collected from host-associated microbiomes under very different biological circumstances. It would be of interest in future work to see if environmental datasets such as soil or ocean, which we did not consider in the current work, had an overall higher error rate when viewed through this prism of the correlation structure. Our results suggest that benchmarking studies should consider not only absolute error rate but patterns of how misclassifications occur. A classifier that randomly misclassifies a read to one of a large number of taxa with a higher overall error rate will not generate this pattern as much as a classifier that systematically misclassifies reads even with a much lower error rate.

Ironically, this problem is made more serious by the very high sequencing depths made possible by technological advancements in the last few years. As sequencing depth has increased to the range of billions of reads per sample, even a relatively low error rate can produce quantifiable problems in terms of classification, as our Poisson modeling demonstrates. Consider a hypothetical example. Given an error rate of 1% on a taxon with a relative abundance of 10% in a dataset with 100,000 reads per sample, the taxon would have on average 10,000 reads per sample and misclassification would only produce 100 reads. This might not be enough to survive filters for low abundance taxa or successfully perform inference across samples. However, with modern technology, it is easy to achieve 100 million reads per sample. At this sequencing depth, the taxon would be present at 10 million reads and a 1% error rate would produce 100,000 reads devoted to a “phantom” taxon. This of course will be highly correlated across samples with the

"real" highly abundant taxa so that we would expect a common pattern of inference between the phantom and the real taxa. The dependence of these problems on sequencing depths means that the patterns that we report here may not have been observable in previous evaluation papers that used older sequence datasets with less sequencing depth. We note as a limitation to our analysis that the 16S sequencing depths in our datasets are much lower than the WGS sequencing depths (Table 1). This may explain why we fail to see the same pattern of correlation coefficients in the 16S datasets that we saw for Kraken2 and would require datasets with more varied sequencing depth to properly test. We also note that these patterns of correlation structure would not be detectable in typical MDS or PCoA analysis as all of the highly correlated taxa would be well represented by individual principle components or coordinates. However, alpha diversity estimates, especially richness, have the potential to be profoundly impacted by this sort of systematic classification error.

A recent benchmarking paper (Z. Sun et al., 2020), argues that due to the difference in taxonomic abundance vs sequence abundance researchers should proceed with extreme caution in comparing classifiers such as Kraken2 with algorithms such as Metaphlan2. We find that for high abundance taxa there is very strong agreement between Kraken2 and Metaphlan2 despite the differences between the normalization schemes. We argue based on the differences in correlation structure that differences between Metaphlan2 and Kraken2 are in part explained by differences in patterns of classification error in addition to the differences in normalization. Our results suggest that caution must be used in the construction of network diagrams that explicitly depend on the correlation coefficient. The fragility of these correlation coefficients in relative abundance data has been well studied and many algorithms attempt to correct for compositional artifacts (Gloor et al., 2017; Tsilimigras & Fodor, 2016). We note that correlations produced by

systematic classification errors of high abundance taxa are another source of potential error. We anticipate the development of algorithms that can potentially correct for these sources of error as a priority for bioinformatics especially as sequencing depths continue to improve with the development of new technology.

5.6 Conclusions

While 16S classifiers do not report the exact same taxa especially at lower taxonomic levels such as genus, the results of 16S classifiers are generally similar enough to make classifier selection more a choice about performance or personal preference. Since the results reported by 16S classifiers and Metaphlan2 are so similar it provides an interesting contrast to the unique identification of taxonomies reported by Kraken2 especially at lower taxonomic levels. We argue in this dissertation that there is strong evidence that Kraken2 generates "phantom" taxa from the misclassification of high abundance taxa. These taxa share approximately the same distribution of reads across samples leading to sharing correlations with the same metadata. If the relationship between the dominant taxa and the metadata is insignificant it will be safely ignored and not reported in the literature, but many studies have a least one dominant taxon correlated to the biological variable under study. As sequencing depth increases, the number of phantom taxa should increase as well leading to more falsely identified correlations between low abundance taxa and reported variables of the study. This will likely lead to spurious or irreproducible results. The fact that this error can be reproduced when constructing a simple Poisson error model supports a simple pattern of classification error as the cause of the phantom taxa. When combined with the fact that another WGS classifier completely avoids this problem points to the error arising from the classifier and not the data collection or sequencing processing steps.

This work has focused on Kraken2 systematic classification error when comparing WGS classifiers, but this does not preclude some downsides on the Metaphlan2 curated marker database. The conservative nature of Metaphlan2's classification could be ignoring some true low abundance taxa that share a strong relationship with metadata variables. It was surprising to see how few additional taxa that Metaphlan2 identified over 16S classifiers despite the deeper

sequencing depth. Our work suggests that further development in the area of WGS classifiers is warranted and that there are many unsolved problems in this area. Given this, we argue that 16S sequencing still has some utility, especially when paired with Whole Genome Shotgun Sequencing as it allows for informative comparisons with WGS classifiers. As sequencing costs continue to decline, future studies may be able to perform high-depth 16S sequencing as a control paired with high-depth WGS. Such paired studies could provide confidence that the observed correlation between taxa was not the result of systematic classification error.

REFERENCES

AlQuraishi, M. (2019). AlphaFold at CASP13. *Bioinformatics*, *35*(22), 4862–4865. https://doi.org/10.1093/bioinformatics/btz422

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bokulich, N. A., Ziemski, M., Robeson, M. S., & Kaehler, B. D. (2020). Measuring the microbiome: Best practices for developing and benchmarking microbiomics methods. *Computational and Structural Biotechnology Journal*, *18*, 4048–4062. https://doi.org/10.1016/j.csbj.2020.11.049

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Clinton, S., Johnson, J., Lambirth, K., Sun, S., Brouwer, C., Keen, O., Redmond, M., Fodor, A., & Gibas, C. (2020). Sediment Microbial Diversity in Urban Piedmont North Carolina Watersheds Receiving Wastewater Input. *Water*, *12*(6), 1557. https://doi.org/10.3390/w12061557

Freedberg, D. E., Lebwohl, B., & Abrams, J. A. (2014). The Impact of Proton Pump Inhibitors on the Human Gastrointestinal Microbiome. *Clinics in Laboratory Medicine*, *34*(4), 771–785. https://doi.org/10.1016/j.cll.2014.08.008

Gardner, P. P., Watson, R. J., Morgan, X. C., Draper, J. L., Finn, R. D., Morales, S. E., & Stott, M. B. (2019). Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. *PeerJ*, *7*, e6160. https://doi.org/10.7717/peerj.6160

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, *8*, 2224. https://doi.org/10.3389/fmicb.2017.02224

Hojo, M., Asahara, T., Nagahara, A., Takeda, T., Matsumoto, K., Ueyama, H., Matsumoto, K., Asaoka, D., Takahashi, T., Nomoto, K., Yamashiro, Y., & Watanabe, S. (2018). Gut Microbiota Composition Before and After Use of Proton Pump Inhibitors. *Digestive Diseases and Sciences*, *63*(11), 2940–2949. https://doi.org/10.1007/s10620-018-5122-4

Hultman, J., Waldrop, M. P., Mackelprang, R., David, M. M., McFarland, J., Blazewicz, S. J., Harden, J., Turetsky, M. R., McGuire, A. D., Shah, M. B., VerBerkmoes, N. C., Lee, L. H., Mavrommatis, K., & Jansson, J. K. (2015). Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature*, *521*(7551), 208–212. https://doi.org/10.1038/nature14238

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386. https://doi.org/10.1101/gr.5969107

Johnson, J., Dellon, E., McCoy, A. N., Sun, S., Jensen, E. T., Fodor, A. A., & Keku, T. O. (2021). Lack of association of the esophageal microbiome in adults with eosinophilic esophagitis compared with non-EoE controls. *Journal of Gastrointestinal and Liver Diseases*, *30*(1), 17–24. https://doi.org/10.15403/jgld-3049

Johnson, J., Sun, S., & Fodor, A. A. (2022). *Systematic classification error profoundly impacts inference in high-depth Whole Genome Shotgun Sequencing datasets* [Preprint]. Bioinformatics. https://doi.org/10.1101/2022.04.04.487034

Jones, R. B., Zhu, X., Moan, E., Murff, H. J., Ness, R. M., Seidner, D. L., Sun, S., Yu, C., Dai, Q., Fodor, A. A., Azcarate-Peril, M. A., & Shrubsole, M. J. (2018). Inter-niche and inter-individual variation in gut microbial community assessment using stool, rectal swab, and mucosal samples. *Scientific Reports*, *8*(1), 4139. https://doi.org/10.1038/s41598-018-22408-4

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729. https://doi.org/10.1101/gr.210641.116

Lambirth, K., Tsilimigras, M., Lulla, A., Johnson, J., Al-Shaer, A., Wynblatt, O., Sypolt, S., Brouwer, C., Clinton, S., Keen, O., Redmond, M., Fodor, A., & Gibas, C. (2018). Microbial Community Composition and Antibiotic Resistance Genes within a North Carolina Urban Water System. *Water*, *10*(11), 1539. https://doi.org/10.3390/w10111539

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., & Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings*

of the National Academy of Sciences, *82*(20), 6955–6959. https://doi.org/10.1073/pnas.82.20.6955

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lindgreen, S., Adair, K. L., & Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, *6*(1), 19233. https://doi.org/10.1038/srep19233

Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, *113*(21), 5970–5975. https://doi.org/10.1073/pnas.1521291113

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science*, *3*, e104. https://doi.org/10.7717/peerj-cs.104

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

McIntyre, A. B. R., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., Minot, S. S., Danko, D., Foox, J., Ahsanuddin, S., Tighe, S., Hasan, N. A., Subramanian, P., Moffat, K., Levy, S., Lonardi, S., Greenfield, N., Colwell, R. R., Rosen, G. L., & Mason, C. E. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology*, *18*(1), 182. https://doi.org/10.1186/s13059-017-1299-7

McIver, L. J., Abu-Ali, G., Franzosa, E. A., Schwager, R., Morgan, X. C., Waldron, L., Segata, N., & Huttenhower, C. (2018). bioBakery: A meta'omic analysis environment. *Bioinformatics*, *34*(7), 1235–1237. https://doi.org/10.1093/bioinformatics/btx754

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*(1), 11257. https://doi.org/10.1038/ncomms11257

Milanese, A., Mende, D. R., Paoli, L., Salazar, G., Ruscheweyh, H.-J., Cuenca, M., Hingamp, P., Alves, R., Costea, P. I., Coelho, L. P., Schmidt, T. S. B., Almeida, A., Mitchell, A. L., Finn, R. D., Huerta-Cepas, J., Bork, P., Zeller, G., & Sunagawa, S. (2019). Microbial abundance, activity and population genomic profiling with mOTUs2. *Nature Communications*, *10*(1), 1014. https://doi.org/10.1038/s41467-019-08844-4

Mohr, K. (2018). Diversity of Myxobacteria—We Only See the Tip of the Iceberg. *Microorganisms*, *6*(3), 84. https://doi.org/10.3390/microorganisms6030084

Osman, E. O., Suggett, D. J., Voolstra, C. R., Pettay, D. T., Clark, D. R., Pogoreutz, C., Sampayo, E. M., Warner, M. E., & Smith, D. J. (2020). Coral microbiome composition along the northern Red Sea suggests high plasticity of bacterial and specificity of endosymbiotic dinoflagellate communities. *Microbiome*, *8*(1), 8. https://doi.org/10.1186/s40168-019-0776-5

Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, *16*(1), 236. https://doi.org/10.1186/s12864-015-1419-2

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny

substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996–1004. https://doi.org/10.1038/nbt.4229

Payne, J. T., Millar, J. J., Jackson, C. R., & Ochs, C. A. (2017). Patterns of variation in diversity of the Mississippi river microbiome over 1,300 kilometers. *PLOS ONE*, *12*(3), e0174890. https://doi.org/10.1371/journal.pone.0174890

Peabody, M. A., Van Rossum, T., Lo, R., & Brinkman, F. S. L. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics*, *16*(1), 362. https://doi.org/10.1186/s12859-015-0788-5

Regalado, J., Lundberg, D. S., Deusch, O., Kersten, S., Karasov, T., Poersch, K., Shirsekar, G., & Weigel, D. (2020). Combining whole-genome shotgun sequencing and rRNA gene amplicon analyses to improve detection of microbe–microbe interaction networks in plant leaves. *The ISME Journal*, *14*(8), 2116–2130. https://doi.org/10.1038/s41396-020-0665-8

Rothberg, J. M., & Leamon, J. H. (2008). The development and impact of 454 sequencing. *Nature Biotechnology*, *26*(10), 1117–1124. https://doi.org/10.1038/nbt1485

Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I., & Watson, M. (2017). A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Frontiers in Genetics*, *8*. https://doi.org/10.3389/fgene.2017.00023

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Schirmer, M., Franzosa, E. A., Lloyd-Price, J., McIver, L. J., Schwager, R., Poon, T. W., Ananthakrishnan, A. N., Andrews, E., Barron, G., Lake, K., Prasad, M., Sauk, J.,

Stevens, B., Wilson, R. G., Braun, J., Denson, L. A., Kugathasan, S., McGovern, D. P. B., Vlamakis, H., … Huttenhower, C. (2018). Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nature Microbiology*, *3*(3), 337–346. https://doi.org/10.1038/s41564-017-0089-z

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, *9*(8), 811–814. https://doi.org/10.1038/nmeth.2066

Sorgen, A., Johnson, J., Lambirth, K., Clinton, S. M., Redmond, M., Fodor, A., & Gibas, C. (2021). Characterization of Environmental and Cultivable Antibiotic-Resistant Microbial Communities Associated with Wastewater Treatment. *Antibiotics*, *10*(4), 352. https://doi.org/10.3390/antibiotics10040352

Sun, S., Zhu, X., Huang, X., Murff, H. J., Ness, R. M., Seidner, D. L., Sorgen, A. A., Blakley, I. C., Yu, C., Dai, Q., Azcarate-Peril, M. A., Shrubsole, M. J., & Fodor, A. A. (2021). On the robustness of inference of association with the gut microbiota in stool, rectal swab and mucosal tissue samples. *Scientific Reports*, *11*(1), 14828. https://doi.org/10.1038/s41598-021-94205-5

Sun, Z., Huang, S., Zhang, M., Zhu, Q.-Y., Haiminen, N., Carrieri, A.-P., Vázquez-Baeza, Y., Parida, L., Kim, H.-C., Knight, R., & Liu, Y.-Y. (2020). *Challenges in Benchmarking Metagenomic Profilers* [Preprint]. Bioinformatics. https://doi.org/10.1101/2020.11.14.382994

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., Coelho, L. P., Arumugam, M., Tap, J., Nielsen, H. B., Rasmussen, S., Brunak, S., Pedersen, O., Guarner, F., de Vos, W. M., Wang, J., Li, J., Doré, J., Ehrlich, S. D., …

Bork, P. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*, *10*(12), 1196–1199. https://doi.org/10.1038/nmeth.2693

Sze, M. A., & Schloss, P. D. (2016). Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *MBio*, *7*(4). https://doi.org/10.1128/mBio.01018-16

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., & Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, *12*(10), 902–903. https://doi.org/10.1038/nmeth.3589

Tsilimigras, M. C. B., & Fodor, A. A. (2016). Compositional data analysis of the microbiome: Fundamentals, tools, and challenges. *Annals of Epidemiology*, *26*(5), 330–335. https://doi.org/10.1016/j.annepidem.2016.03.002

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, *444*(7122), 1027–1031. https://doi.org/10.1038/nature05414

Venter, J. C. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, *304*(5667), 66–74. https://doi.org/10.1126/science.1093857

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261. https://doi.org/10.1128/AEM.00062-07

Winglee, K., Howard, A. G., Sha, W., Gharaibeh, R. Z., Liu, J., Jin, D., Fodor, A. A., & Gordon-Larsen, P. (2017). Recent urbanization in China is correlated with a Westernized microbiome encoding increased virulence and antibiotic resistance genes. *Microbiome*, *5*(1), 121. https://doi.org/10.1186/s40168-017-0338-7

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

Xiao, L., Estellé, J., Kiilerich, P., Ramayo-Caldas, Y., Xia, Z., Feng, Q., Liang, S., Pedersen, A. Ø., Kjeldsen, N. J., Liu, C., Maguin, E., Doré, J., Pons, N., Le Chatelier, E., Prifti, E., Li, J., Jia, H., Liu, X., Xu, X., … Wang, J. (2016). A reference gene catalogue of the pig gut microbiome. *Nature Microbiology*, *1*(12), 16161. https://doi.org/10.1038/nmicrobiol.2016.161

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, *178*(4), 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Zymo Research. (2019, November 13). 16S Sequencing vs Shotgun Metagenomic Sequencing [Corporate]. *Zymo Research*. https://www.zymoresearch.com/blogs/blog/16s-sequencing-vs-shotgun-metagenomic-sequencing