Title

Mining the gut microbiome for temporal signals of inflammatory bowel disease and novel symbiont genomes

Author

Lyalina, Svetlana

Publication Date

2018

Mining the gut microbiome for temporal signals of inflammatory bowel disease and novel symbiont genomes.

by

Svetlana Lyalina

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

[redacted]

Mining the gut microbiome for temporal signals of inflammatory bowel disease and novel symbiont genomes

Svetlana Lyalina

Abstract

High-throughput sequencing has firmly established itself as the leading method for assaying the structure and functional capacity of microbial communities. With this deluge of data, care must be taken to account for technical and biological artifacts in order to produce robust candidate biomarkers. Of particular interest is the use of mixed effects models and nonlinear models to assess key differences between healthy and diseased individuals that arise over time. In my thesis work, I analyzed data from a longitudinal study of inflammatory bowel disease in mice with the aim of uncovering biological features predictive of abnormal microbiome development in the context of chronic inflammation. My analysis uncovered multiple taxa and gene families that have differential temporal trajectories, as well as a few gene families that stratify the diseased and wild type subjects early on. This investigation led to a follow-up study of the underrepresented microbial genomes present in lab mice, to expand our knowledge of the model animal's microbiome. Since the majority of microbiome studies aimed at future clinical impact are carried out in mice, it is important to know what separates human microbiomes from those of mice, in order to limit hypotheses that are not transferrable. We found that even a modest single cell sequencing effort leads to an appreciable gain in phylogenetic diversity and significantly improves the recruitment of short reads from unrelated mouse metagenomes. Overall, I have shown that robust findings are possible even with a limited set of subjects if one leverages a nuanced statistical modeling approach and undertakes targeted acquisition of new data.

Table of Contents

1 Introduction 1

1.1 The microbiome 1

1.2 Inflammatory bowel disease (IBD) 2

1.3 Model animals 3

1.4 Biological samples and data generation 4

2 Longitudinal analysis of the gut microbiome in IBD 6

2.1 Statistical modeling choices 6

2.1.1 Choice of response distribution for GLMM 8

2.1.2 Gene set testing as part of model fitting 9

2.1.3 Knowing the progression of the disease allows us to alter the statistical model to test for pre- and post-onset changes 10

2.2 An alternative to GLMMs 10

2.3 Results of GLMM tests 11

2.4 Results of FPCA-based tests 14

2.5 Results of segmented regression 16

2.6 An approach for assessing inter-species interactions 18

2.7 Interpreting the GLMM and FPCA results 21

3 Single cell sequencing: an effective approach for addressing the underrepresentation of mouse symbiotic microbes in reference databases 25

3.1 An optimization approach for prioritizing cells 25

3.2 Single cell genomes provide noticeable phylogenetic gain 28

3.3 The single cell genomes are a source of new genes and extended genomic feature assemblages 31

3.3.1 Results of annotating genome features 32

3.4 Evaluating the utility of new genomes as a reference for metagenomic classification 39

3.4.1 Results of custom database tests with MIDAS 40

3.4.2 Results of custom database tests with Sourmash-LCA 44

3.5 Conclusions from examining the genome features annotated in single cell genomes........ 47

3.6 Conclusions from evaluating the dataset as a reference 48

4 Conclusion........ 49

References 50

List of Figures

Figure 1. Smoothed abundance trajectories of species with significantly different trajectory shapes in the FPCA-based goodness-of-fit comparisons. 15

Figure 2. Permutation-based demonstration of the lack of coherent temporal signal in groups of taxonomic vectors compared to groups of functional vectors 16

Figure 3. A heatmap of the 29 segmented GLMM coefficients. 17

Figure 4. Means of interaction coefficients of the per-group Lotka-Volterra model. 20

Figure 5. Comparison of the distributions of 4 technical characteristics between the default lowest Cp proposal set, and the optimized maximal PD proposal set 27

Figure 6. Assessment of assembly quality for the single cell genomes 29

Figure 7. Improvements achieved by high coverage resequencing of select cells. 30

Figure 8. Metacoder heattree showing phylogenetic gain and concentration of single cell assembled genomes (SAGs) on select lineages in the bacterial and archaeal trees. 31

Figure 9. Venn diagrams of pairwise gene catalog intersections comparing the single cell mouse microbiome gene catalog to published catalogs (mouse, human, ocean) 35

Figure 10. Barchart of COG counts that were significantly enriched (positive counts) or depleted (negative counts) in the new gene catalog, per COG functional category 36

Figure 11. CRISPR classification and AntiSMASH predictions for single cell genomes that were successfully placed in the phylogenetic tree of GTDB genomes 37

Figure 12. Ridgeline plots of 3 MIDAS performance metrics 43

Figure 13. Ridgeline plots of 2 Sourmash-LCA performance metrics 46

List of Tables

Table 1. Significant results of testing group by time interaction coefficient in module-level GLMM fits. 13

Table 2. Significant results of testing group coefficient in module-level GLMM fits. 13

Table 3. Results of FPCA-based goodness of fit comparisons. 15

Table 4. Results of paired Mann-Whitney tests comparing 4 technical characteristics of the default and optimized sequencing proposals 28

Table 5. Summary counts of annotated features retrieved from all single cell genomes that passed technical filtering. 38

Table 6. Results of two-sided Mann-Whitney tests of MIDAS performance characteristics achieved with different reference databases. 42

Table 7. Results of two-sided Mann-Whitney tests comparing the change in performance of Sourmash-LCA when using different reference databases. 45

1 Introduction

1.1 The microbiome: an important factor in human health and disease

The complex community of bacteria, archaea, viruses, and microscopic fungi that exists in and on the human body is known as the microbiome. The presence of microbes living in close proximity to human tissues has been known since the time of Antonie van Leeuwenhoek, who was the first to study microbes in saliva and dental plaque with his newly developed microscope[1]. Throughout much of history since then, the focus has been on the disease-causing capabilities of bacteria. Generations of scientists painstakingly cultured bacterial isolates in order to study them in a controlled setting and develop disease treatments. With the advent of the polymerase chain reaction (PCR), the genetic content of bacteria became significantly easier to study, allowing microbiologists to go beyond externally observable traits. As methods for assaying the complexity of the microbiota have expanded, there has been a shift in the perception and study of human associated microbes, with more interest in obtaining a holistic picture of the ecosystem. The era of high-throughput sequencing opened the door to studying microbial diversity at a much finer scale and with lower cost. Scientists took on the challenge of characterizing not only what species live within us, but also their functional repertoires. Due to these extensive efforts that built upon prior experimental work, we now know that this complex community provides a number of beneficial services for the mammalian host, including pathogen defense[2], vitamin biosynthesis[3], production of short chain fatty acids[4], and complicated communication with the host immune system[5]. The presence of the microbiome at the interface of the immune system and environment makes it a particularly interesting subject of study in the case of autoimmune disorders.

1.2 Inflammatory bowel disease (IBD)

The focal disease for most of my graduate work has been inflammatory bowel disease, an autoimmune disorder that is thought to be at least in part caused by an exaggerated immune response to benign commensals[6]. Encompassing two major disorders, Crohn's disease and ulcerative colitis, IBD affects more than a million people in the United States alone[7]. This statistic is projected to increase[8]. Although there isn't a consensus on the explanation for this increase, one of the hypotheses offered links it with a more "Western" lifestyle – sedentary day-to-day schedules[9], a more sterile built environment[10], and more processed food choices[11][12]. These external factors are just one facet of this complicated disease. There has been a great amount of research on the genetic causes of IBD, which has uncovered more than 160 associated loci, most of them in regions related to immune function[13]. The uneven incidence of the disease between different ethnic groups also supports the presence of a hereditary component[14]. Studies of families with IBD have shown that having an affected close relative increases the likelihood of an individual being diagnosed with IBD[15]. However, twin studies and genome wide association studies (GWAS) show that genetic factors cannot fully explain IBD susceptibility: reported heritability estimates are quite high in twin studies, yet the corresponding estimates from GWAS are approximately halved[16]. While diet, exercise, smoking, and stress all have additionally been implicated in this disease in the past[17], recently the microbiome has come to the forefront as a promising new source of both predictive biomarkers and potential interventions. Since current pharmaceutical treatments for IBD do not lead to remission for all patients, and surgical interventions negatively affect quality of life, there is an unmet need for novel therapeutic approaches.

Most work that has been carried out in the IBD microbiome space has sampled human subjects undergoing some form of treatment for their disease, and the samples were taken only at one or two timepoints. Promising longitudinal IBD microbiome data have been generated by the iHMP effort[18], however even those span no more than 2 years and do not exclude samples that may be impacted by active disease management. Additionally, most prior studies have generated only 16S rRNA gene sequencing data, which provides limited taxonomic resolution. The hypervariable regions of the 16S rRNA gene are a convenient target for assessing community diversity and obtaining both reference-based and *de novo* taxonomic characterizations. However, the results of this approach can be affected by amplification bias[19], which can result in false negatives and skewed abundance estimates. Functional repertoires are also difficult to reconstruct from this kind of data, since microbial traits have variable rates of phylogenetic conservation[20]. During my PhD, I focused on moving past the common paradigm of case/control cross-sectional studies and instead examining the microbiome throughout the progression of the disease: starting at a susceptible, but uninflamed state, and tracking the evolving microbiome with higher resolution shotgun metagenomic data.

1.3 Model animals are indispensable for tackling a complex disease like IBD

It has been shown that a variety of lifestyle choices and experiences can have a compounding effect on a person's eventual diagnosis. That is why in order to truly investigate IBD with minimal confounders it is necessary to start with mice. To underscore the breadth of factors that are associated with differences in the human gut microbiome, here are some of the previously reported external influences:

1. Diet[21]
2. Medication – the most obvious being antibiotics, but other drugs also having an effect, despite not directly targeting bacteria.[22][23][24]
3. Stress (including travel)[25][26]
4. Early childhood experiences (including birth route)[27][28]
5. Smoking[29]
6. Alcohol consumption[30]

While these findings have differing levels of experimental support, they suggest that extra caution is necessary when attempting to find new microbiome-disease associations. To rule out the effects of these potential confounders, we used lab mice for our longitudinal study. Additionally, to further limit the number of uncontrolled variables, we used littermate controls. This takes genetic variability out of the equation, and addresses potential seasonality concerns that would arise with staggered cohorts.

1.4 Biological samples and data generation

IBD-susceptible mice with a **d**ominant **n**egative mutant **r**eceptor II of transforming growth factor β (referred to as DNR for the rest of the text) and their healthy control littermates (referred to as WT) were raised in collaborator Shomyseh Sanjabi's lab. The

DNR mouse model features defective TGF-β signaling in T-cells, leading to limited amounts of T regulatory cells, and eventually a pro-inflammatory phenotype[31]. While many mouse models of IBD exist, they vary in their approximation of the different aspects of the disease[32]. The DNR model represents an aspect of immune dysregulation that has been observed in human patients, namely the aberrant downstream signaling via SMAD proteins[33, 34].

Stool samples were collected from the two mouse groups (N=5 DNR, N=4 WT) at regular intervals. The collection started at weaning and ended when the health status of the DNR mice became too severe. The samples then underwent DNA extraction and library preparation as described in [35] and were sequenced at the UCSF IHG core facility. Raw sequence data were quality processed and reads originating from the host organism were removed. The clean reads were then used to generate functional and taxonomic characterizations of the mouse gut microbiome using the tools ShotMAP[36] and MIDAS[37] respectively. These two characterizations consist of abundances of KEGG[38] orthologous groups (KOs) and MIDAS genome clusters (species). Additional blood-derived data were gathered from a parallel cohort of littermates. This was used to confirm the vastly different immune profiles that develop over time in the DNR and WT mice.

2 Longitudinal analysis of the gut microbiome in IBD uncovers temporal signals in functional and taxonomic profiles

The work described in this chapter was my contribution to the published paper in *mSystems*[35]. The end result of this study was a subset of KEGG modules and MIDAS species that had significantly different abundance trajectories between the groups of interest, as well some post hoc investigations clarifying the time segment of the functional differences.

2.1 Statistical modeling choices

I took multiple novel approaches when modeling these complex high dimensional data in order to find differences over time between the DNR and WT groups of mice. The generalized linear mixed model (GLMM) approach that I started out with has been widely used in the ecology literature[39]. Using regression formula notation, the dependent and predictor variables are related as follows:

Abundance ~ group + time + time:group + kit + (1 + time|subject)

Unpacking this formula we have the following:

- *Abundance* can refer to the number of reads assigned to a functional entity (e.g. a KEGG ortholog in these data) or it can be a more complex entity, such as log counts per million (logCPM) or reads per kilobase per genome equivalent (RPKG[40]).

- *Group* is a binary variable, taking on the value of 1 if the observation is from the group of interest (DNR) or 0 if it is from the control group (WT). The coefficient estimated for this variable reflects the baseline differences in abundance between the groups at the start of the time series.
- *Time* is a continuous variable (unit of weeks). The coefficient estimated for this variable reflects the baseline slope of the modeled data, i.e. how much *abundance* is increasing (or decreasing if negative) per week for the WT group (and partially the DNR group, whose slope is also influenced by the *time:group* estimate).
- *Time:Group* refers to the time by group interaction, and the coefficient for this variable reflects the additional slope for the group of interest (DNR). This is the primary coefficient that we want to test. It represents the difference in temporal change of a particular biological factor. We hypothesize that this change happens alongside disease progression.
- *Kit* is a covariate that was necessary to include because our data were derived from two sequencing events, performed on biological samples that were processed with either the Qiagen or MOBIO DNA extraction kits.
- *(1 + time|subject)* reflects the random component of the GLMM, allowing for baseline and slope differences between individuals. Having a fixed and random component in a model is a useful approach for disentangling effects of interest (i.e. differences in slope between groups) from more minute inter-individual variation.

The full model is fit with one of the many GLMM packages (lme4[41], glmmTMB[42], glmmADMB[43]) available in the *R* programming language[44], and to obtain significance estimates for our coefficients of interest, we simply fit reduced models without them and perform likelihood ratio tests.

Starting with this basic GLMM specification, I additionally made a number of modeling decisions that are less common in the literature. Each customization to the general approach is outlined in a separate subsection.

2.1.1 Choice of response distribution for GLMM

Most methods that aim to fit metagenomics data generally follow the lead of the RNA-seq field and use either the negative-binomial distribution or the log-normal distribution. The negative binomial is chosen when the response variable is positive and count valued and the data show a mean-variance relationship. The log-normal distribution is chosen for positive continuously valued dependent variables, such as logCPM. Our data measured abundance as RPKG[40] – reads per kilobase of matched sequence, per number of genome equivalents, which is a continuous positive quantity. For my final analyses, I chose a promising yet less commonly used distribution known as the compound-Poisson Tweedie distribution[45]. This distribution features a power relationship between the variance and the mean, with the power coefficient adaptively determined as part of the model fitting procedure.

Actual model fitting was carried out through the cpglmm method of the cplm package[46] in R, which produces fit objects that are compatible with methods that know how to extract fields from the lmerMod class. Using the glmmTMB[42] package, I also produced comparable fits using the negative-binomial and log-normal distributions as a response, keeping the regression formula unchanged. Since the negative binomial is a discrete distribution, I used read counts as the response and library size divided by average genome size as the offset variable to ensure the same information was being provided as would be in the composite RPKG statistic. A larger number of cpglmm fits converged successfully (373 vs 86 for negative binomial and 145 for log-normal). At this point the negative binomial was no longer considered a viable option and further comparisons proceeded between the compound Poisson Tweedie and the log-normal distributions.

Since these are not nested models, the correct avenue for comparing them for difference in fit is the Vuong test[47]. Using this test, in 66 out of 145 cases the cpglmm fit was significantly better than the log-normal fit, and for the remaining non-significant cases the cpglmm fit still had a lower negative log-likelihood, although it was not statistically significant. Based on this evaluation, combined with the generally higher number of models converging successfully, I chose the cpglmm approach for the final analysis.

2.1.2 Gene set testing as part of model fitting allows for better interpretability and higher number of observation points per model

Following the interesting method proposed by Hejblum et al in their TcGSA package[48], I wanted to combine the gene trajectory grouping approach with the non-standard distribution I had chosen in subsection 2.1.1. This was relatively easy, since at its core TcGSA essentially specifies extra random effects in the formulation of the regression. For my data, this resulted in the inclusion of an extra (1+time|KO) term, and the fitting of models on whole KEGG modules instead of individual KOs. With this setup, the fixed effects estimated are for a module, and the KO level random effects allow deviations in slope and intercept for the constituent orthologs within a module. This reformulation also has the added benefit of reducing the number of tests carried out when testing the significance of the *time* and *time:group* coefficients, reducing the multiple testing burden.

2.1.3 Knowing the progression of the disease allows us to alter the statistical model to test for pre- and post-onset changes

Since immunological covariates were also assessed throughout the timecourse (albeit on a parallel cohort of mice), we knew when increased inflammation started occurring in the DNR mice. This prompted us to ask the question of whether we can not only tell what modules have different trajectories over the entire timespan, but also if those differences in slope occur before or after disease onset at week 7. I went about answering this question by inserting a “hinge” in the regression, resulting in two separate slope coefficients and two separate slope by group interaction coefficients.

2.2 An alternative to GLMMs to test differences in species trajectories

Since taxonomic abundances do not have the advantage of being grouped into coherently changing over time units, the TcGSA-style approach of effectively increasing the number of data points per model fit was not possible. Instead I chose to use a method proposed in a paper focused on a similar problem of finding overall differences in gene expression profiles[49]. I reimplemented the test using FPCA code from the refund R package[50], since the legacy code from the original publication was not maintained. This method, at its core, aims to find a set of eigenfunctions that can be used to faithfully represent all curves in a dataset and then tests whether representations learned from single group data or pooled data are better fits for individual species trajectories.

During the peer review of our manuscript, I was asked to justify why it is appropriate to group functional units into higher level blocks that can be modeled together (KOs into modules), but not to do the same with taxonomic units (species into genera). I investigated this question by constructing multiple permutations of simulated genera and simulated modules (of size appropriate for each setting) and calculating the DISCO[51] F-statistics for these constructs' longitudinal abundance trajectories. I then compared the distributions of these statistics when computed on real data versus simulated data, finding that there was a significant difference in the case of modules, but not genera. The results of Kolmogorov-Smirnov tests on these real vs. simulated comparisons are shown in Figure 2. This suggests non-random temporal coherence (as measured by distance covariance) in functional groups but not taxonomic groups.

2.3 Results of GLMM tests

Instead of blindly fitting models for all KEGG modules that had at least one constituent KO present, I chose to pre-emptively limit the candidate list by running MinPath[52]. MinPath is a simple integer linear programming approach to find the fewest gene sets (i.e. KEGG modules) that can still cover all of the individual lower-level entities (i.e. KOs) present in a dataset. This allowed me to lower the number of fits down from 394 to 373. Of those 373 modules, 29 had a significant time by group interaction (Benjamini-Hochberg[53] corrected p-value < 0.05), shown in Table 1. When testing the significance of the intercept difference, 17 modules were found to be significant, shown in Table 2.

Subsequent testing via hinge regression of the 29 modules with differences in slope between DNR and WT showed that most significant differences were in the post-onset segment, and only 2 modules had a pre-onset difference in slope.

P-value	**Interaction coefficient (time:group)**	**KEGG module**	**B-H adjusted p-value**	**Module description**
0.003526808	-0.061914095	M00031	0.045118826	Lysine biosynthesis, 2-aminoadipate => lysine
0.001289328	-0.054163057	M00252	0.023124397	Lipooligosaccharide transport system
7.85984E-14	-0.050795746	M00037	1.458E-11	Melatonin biosynthesis, tryptophan => serotonin => melatonin
0.000583901	0.001367718	M00081	0.012742783	Pectin degradation
0.001641721	0.005501227	M00096	0.02648167	C5 isoprenoid biosynthesis, non-mevalonate pathway
3.09279E-15	0.006769116	M00051	1.14743E-12	Uridine monophosphate biosynthesis, glutamine (+ PRPP) => UMP
2.47233E-08	0.008386556	M00432	2.29308E-06	Leucine biosynthesis, 2-oxoisovalerate => 2-oxoisocaproate
3.2853E-05	0.008449928	M00015	0.001354276	Proline biosynthesis, glutamate => proline
6.17762E-05	0.010710645	M00531	0.002083541	Assimilatory nitrate reduction, nitrate => ammonia
9.37223E-06	0.01250214	M00377	0.000434637	Reductive acetyl-CoA pathway (Wood-Ljungdahl pathway)
3.89476E-05	0.021300468	M00482	0.001444956	DevS-DevR (redox response) two-component regulatory system
0.001371258	0.02235754	M00532	0.023124397	Photorespiration
1.10344E-06	0.024257464	M00511	6.82292E-05	PleC-PleD (cell fate control) two-component regulatory system
0.000639329	0.029589337	M00009	0.013177285	Citrate cycle (TCA cycle, Krebs cycle)
5.65831E-08	0.033549367	M00507	4.19847E-06	ChpA-ChpB/PilGH (chemosensory) two-component regulatory system
0.001895115	0.040964063	M00515	0.028123512	FlrB-FlrC (polar flagellar synthesis) two-component regulatory system
1.7889E-09	0.053978814	M00076	2.21227E-07	Dermatan sulfate degradation
0.000488884	0.057090656	M00358	0.011335995	Coenzyme M biosynthesis
5.596E-06	0.066796574	M00538	0.000296588	Toluene degradation, toluene => benzoate
0.000238478	0.067502417	M00091	0.006319664	Phosphatidylcholine (PC) biosynthesis, PE => PC
0.003058517	0.069417949	M00079	0.040525353	Keratan sulfate degradation
0.002173789	0.070007142	M00012	0.029869476	Glyoxylate cycle
0.000134104	0.07130334	M00259	0.003827109	Heme transport system

0.001986819	0.096037367	M00334	0.028350384	Type VI secretion system
0.00028046	0.119637579	M00555	0.006936704	Betaine biosynthesis, choline => betaine
8.84876E-05	0.120030856	M00330	0.002735742	Adhesin protein transport system
0.000780192	0.2081415	M00229	0.015234268	Arginine transport system
0.001792889	0.209197878	M00332	0.02771507	Type III secretion system
0.001315562	0.209397611	M00417	0.023124397	Cytochrome o ubiquinol oxidase

Table 1. Significant results of testing group by time interaction coefficient in module-level GLMM fits. Negative coefficients reflect a reduced slope in the DNR group. P-values are obtained via likelihood ratio test against a reduced model with no interaction term

P-value	intercept coefficient (group)	KEGG module	B-H adjusted p-value	Module description
0	-0.1367	M00551	0	Benzoate degradation, benzoate => catechol / methylbenzoate => methylcatechol
0	0.0689	M00246	0	Nickel transport system
0	0.35636	M00271	0.00001	PTS system, beta-glucosides-specific II component
0	-0.2761	M00502	0.00001	GlrK-GlrR (amino sugar metabolism) two-component regulatory system
0	-0.17086	M00080	0.00005	Lipopolysaccharide biosynthesis, inner core => outer core => O-antigen
0.00003	0.00333	M00549	0.00173	Nucleotide sugar biosynthesis, glucose => UDP-glucose
0.00004	-0.0967	M00235	0.00211	Arginine/ornithine transport system
0.00006	-0.76375	M00537	0.00298	Xylene degradation, xylene => methylbenzoate
0.00026	0.30189	M00151	0.01063	Cytochrome bc1 complex respiratory unit
0.00034	-0.13058	M00211	0.01269	Putative ABC transport system
0.00075	-0.18581	M00278	0.02516	PTS system, sorbose-specific II component
0.00081	0.0352	M00535	0.02516	Isoleucine biosynthesis, pyruvate => 2-oxobutanoate
0.00098	-0.13234	M00596	0.02648	Dissimilatory sulfate reduction, sulfate => H2S
0.001	0.05319	M00159	0.02648	V-type ATPase, prokaryotes
0.00117	0.29126	M00356	0.02897	Methanogenesis, methanol => methane
0.0017	-0.01631	M00572	0.03932	Pimeloyl-ACP biosynthesis, BioC-BioH pathway, malonyl-ACP => pimeloyl-ACP
0.00182	0.57489	M00349	0.03965	Microcin C transport system

Table 2. Significant results of testing group coefficient in module-level GLMM fits. Negative coefficients reflect a reduced abundance in the DNR group at the first timepoint. P-values are obtained via likelihood ratio test against a model with the group term omitted.

2.4 Results of FPCA-based tests

The tests performed on taxonomic data from this study produced 7 MIDAS species that were found to have significantly different abundance trajectory shapes. It's important to note that significance is evaluated via a permutation-based comparison of F statistics, hence the p-values produced can be exact zeros due to limitations in the number of permutations performed. Since this test does not produce a comparison coefficient like slope in the GLMMs, I have instead included the estimated area under the smoothed abundance curve as a purely informational quantity in the results Table 3. The abundance trajectories of these significantly different species can be seen in Figure 1, where the shaded ribbons are LOESS smoothing across individuals in a group.

Justifying the necessity of this alternative approach for comparing species trajectory shapes, I show the results of comparing the DISCO F-statistic distributions (real vs permuted inputs) in modules and genera (Figure 5). The significantly different F-statistic distribution of real versus permuted longitudinal KO profiles shows that organizing low-level functional units into higher order ontology-determined groups is supported by the overall coherence in the group longitudinal development. The nonsignificant result obtained when testing taxonomic profiles in a similar manner shows that organizing species trajectories into genera does not aid in temporal coherence. An alternative explanation for the nonsignificant difference in distributions of simulated and real F-statistics is that genera in this dataset tend to have few constituent taxa present, and therefore there is a discretization effect on the F-statistic distribution (only a certain number of permutation groupings can be created).

Species ID	P-value	B-H corrected p-value	Species name	WT area under LOESS curve	DNR area under LOESS curve
54642	0	0	*Bacteroides sartorii*	0.01992	0.07699
57185	0	0	*Bacteroides xylanisolvens*	0.03051	0.02506
57318	0	0	*Bacteroides uniformis*	0.02297	0.04506
58110	0	0	*Escherichia coli* O157:H43 strain T22	5.35E−4	0.007523
59684	0.0001	0.0025	*Lachnospiraceae* bacterium COE1	0.07203	0.04213
59708	0	0	*Bacteroides rodentium*	0.0136	0.01986
61442	0.0013	0.02786	*Lachnospiraceae* bacterium A4	0.119	0.1348

Table 3. Results of FPCA-based goodness of fit comparisons. Exact zeros are generally not returned when estimating p-values with a theoretical probability distribution, but occur here due to being empirically estimated via comparison to a permutation based null distribution.

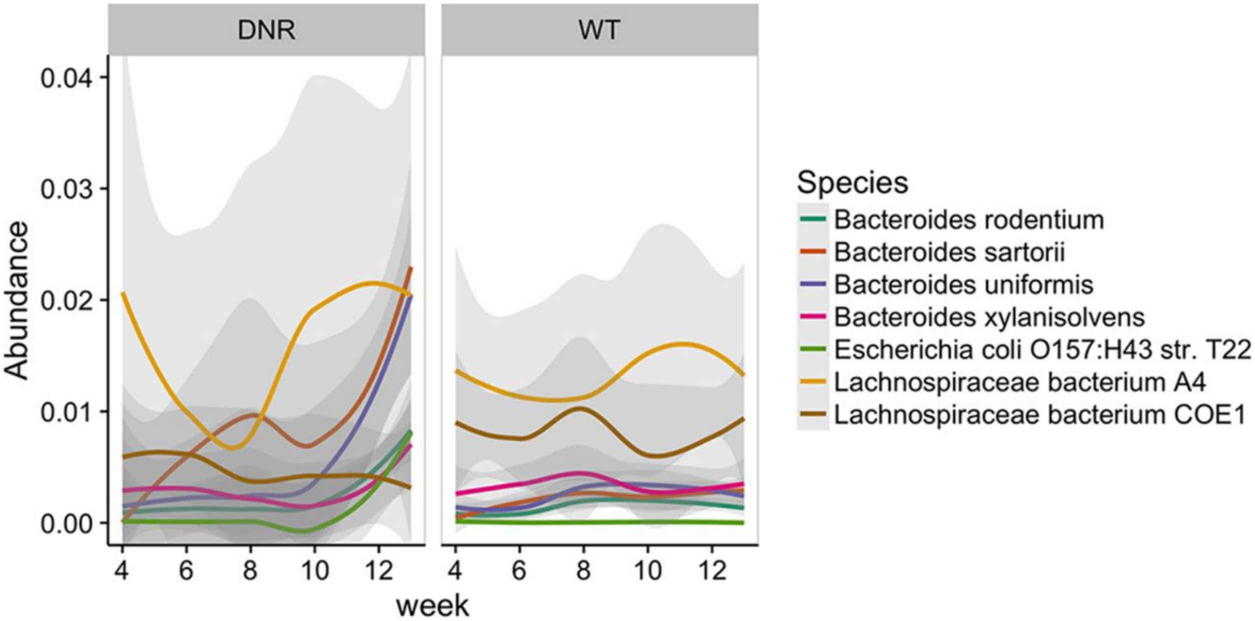

Figure 1. Smoothed abundance trajectories of species with significantly different trajectory shapes in the FPCA-based goodness-of-fit comparisons. Since the identifiable unit in MIDAS is a genome cluster, the labels shown here are those of centroid genomes.

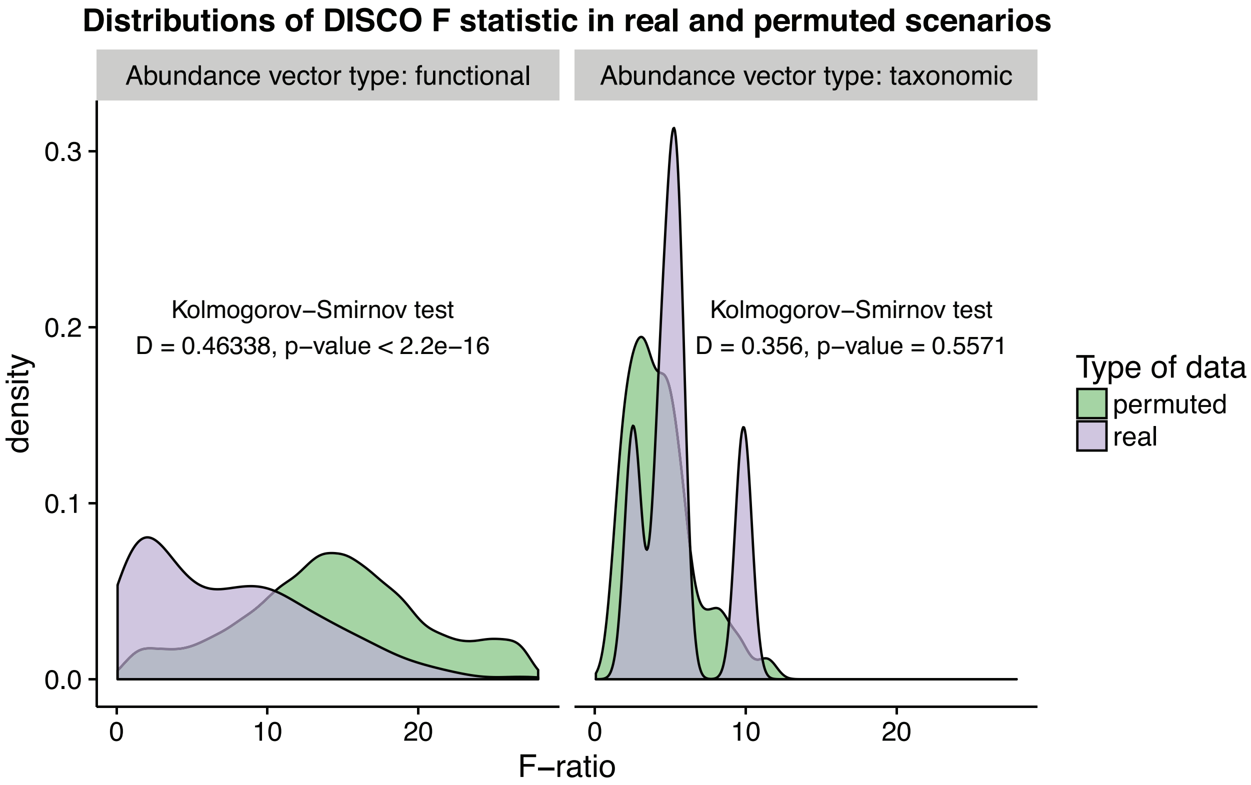

Figure 2. Permutation-based demonstration of the lack of coherent temporal signal in groups of taxonomic vectors compared to groups of functional vectors

2.5 Results of segmented regression

Since there was interest in examining the timing of the significant changes reported in Table 1, I performed a post-hoc analysis on those modules' data. This analysis was done with a modified regression formula allowing for two segments: pre-disease onset and post-disease onset. Each module's fit produced two coefficients reflecting the difference in slope between DNR and WT groups in each time segment, which I then tested for being significantly non-zero via their t-values. Only 13 of the 29 previously identified modules exhibited a significant difference in coefficients using segmented regression. A heatmap of the coefficients is presented in Figure 3 with asterisks marking coefficients that were significantly non-zero (B-H corrected p-value < 0.2). Most (11/13) of the significant segment slope differences were in the post-immune activation part of the timecourse.

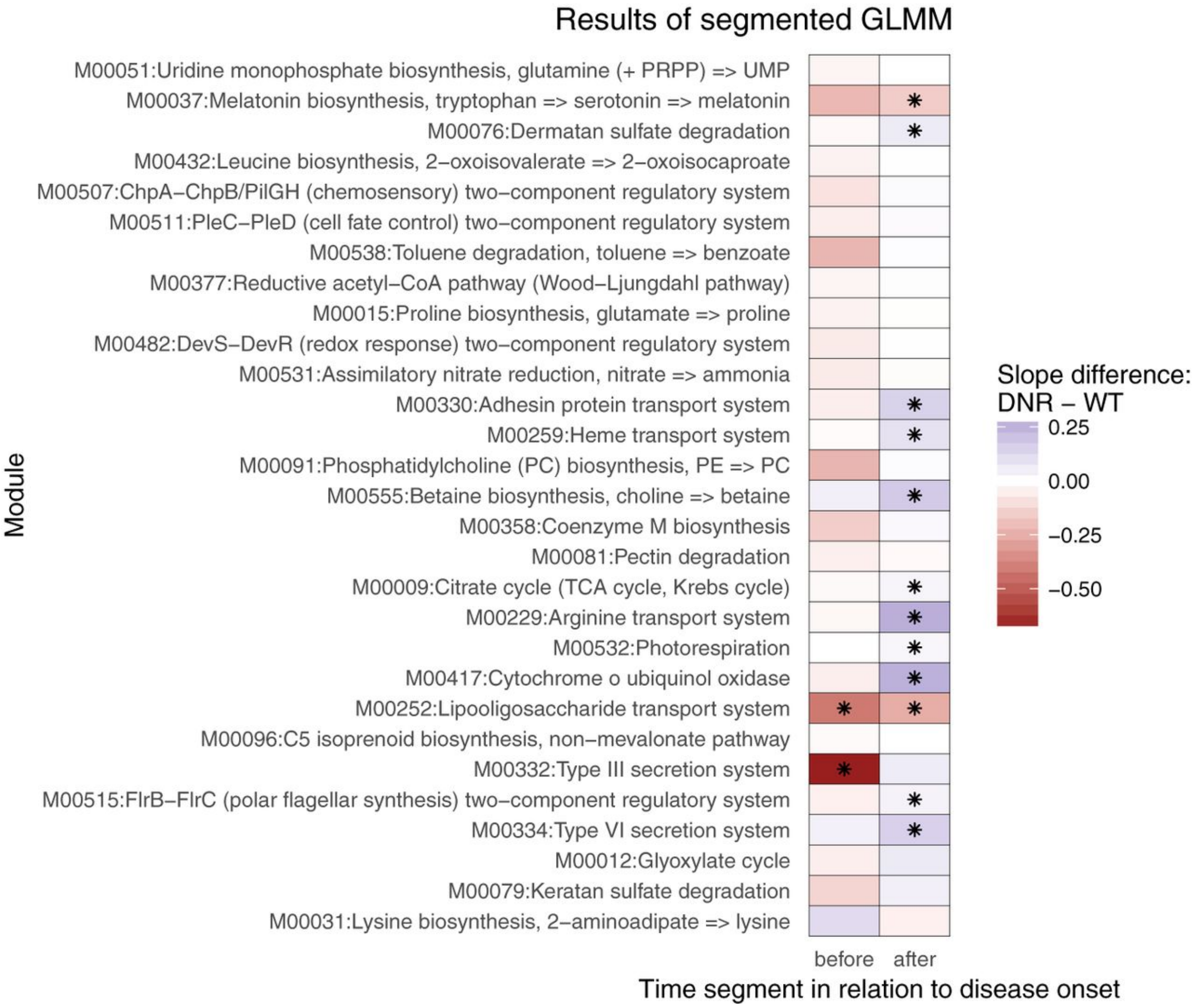

Figure 3. A heatmap of the 29 segmented GLMM coefficients. Asterisks mark significantly non-zero coefficients (B-H corrected p-value < 0.2). Color represents the estimated coefficient for group by time interaction in that segment, i.e. the extra slope of DNR.

2.6 An approach for assessing inter-species interactions

One of the analyses that was not included in the published *mSystems* manuscript was focused on gauging whether species interactions in the DNR and WT mice were also different, and if they were affected by the changing abundances of active immune cells. The generalized Lotka-Volterra model has been a common approach in the ecological literature used to characterize interspecies interactions within a community[54][55]. When applied to microbiome data, this model has not produced many successful inferences so far. For the cases where coefficients were obtained with high confidence, the researchers were able to produce interesting mechanistic hypotheses for community alteration[56][57].

In the context of our small mouse IBD study, fitting the necessary system of ordinary differential equations (ODEs) was a daunting task even when armed with the information sharing approach proposed in the work of Chung et al[58]. Briefly, the approach taken in a typical generalized Lotka-Volterra model involves finding optimal coefficients for the system ODEs specified by the following formula (with coefficients as they appear in reference [56]):

$$\frac{d}{dt}x_i(t) = \mu_i x_i(t) + x_i(t)\sum_{j=1}^{L} M_{ij}x_j(t) + x_i(t)\sum_{l=1}^{P} \varepsilon_{il}u_l(t)$$

To unpack this formula, each species' rate of abundance change is a function of:

- an inherent growth parameter $\boldsymbol{\mu}$ and current abundance $\boldsymbol{x_i}$

- the sum of interspecies interaction effects, calculated by multiplying the other species' abundance ($\boldsymbol{x_j}$) by the focal species' abundance ($\boldsymbol{x_i}$) and their interaction coefficient ($\boldsymbol{M_{ij}}$)
- the sum of species-host interaction effects, calculated by multiplying the host quantity's abundance ($\boldsymbol{u_l}$) by the focal species' abundance ($\boldsymbol{x_i}$) and their interaction coefficient ($\boldsymbol{\varepsilon_{il}}$)

The $\mathbf{M}$ and $\boldsymbol{\varepsilon}$ matrices are conceptually similar, with the exception that the host effects are unidirectional (microbes are not permitted to affect host-related abundances in this model). The host-originating actors in this case were immune cell subpopulations.

Since the data we have are quite sparse, I selected only the top 10 prevalent species and interpolated the time series data using splines within the Amelia II multiple imputation package[59]. To obtain cohort-level confidence intervals, all samples from a cohort were used in the two step inference of ODE coefficients via modFit and modMCMC functions in the FME package[60]. This approach produces an initial set of coefficients via conventional gradient-based methods, and then uses the Hessian from the first step for the proposal distribution in the MCMC part.

Initially I saw promising results in the interaction networks that emerged for the two groups because they appeared quite different and could have potentially explained some of the complex interplay happening in the background of IBD development (summary of coefficients presented in Figure 4). Upon closer inspection, I was left skeptical of the results when I examined the MCMC trace plots and calculated convergence diagnostics using the R package coda[61]. Ultimately, we decided that the uncertainty was too large to make effective comparisons. Such an over-parameterized approach needs more data than our small study or simpler communities with fewer actors. Recent research has suggested that Lotka-Volterra approaches may be fundamentally incapable of describing the complexity of ecological interactions between microbes [62].

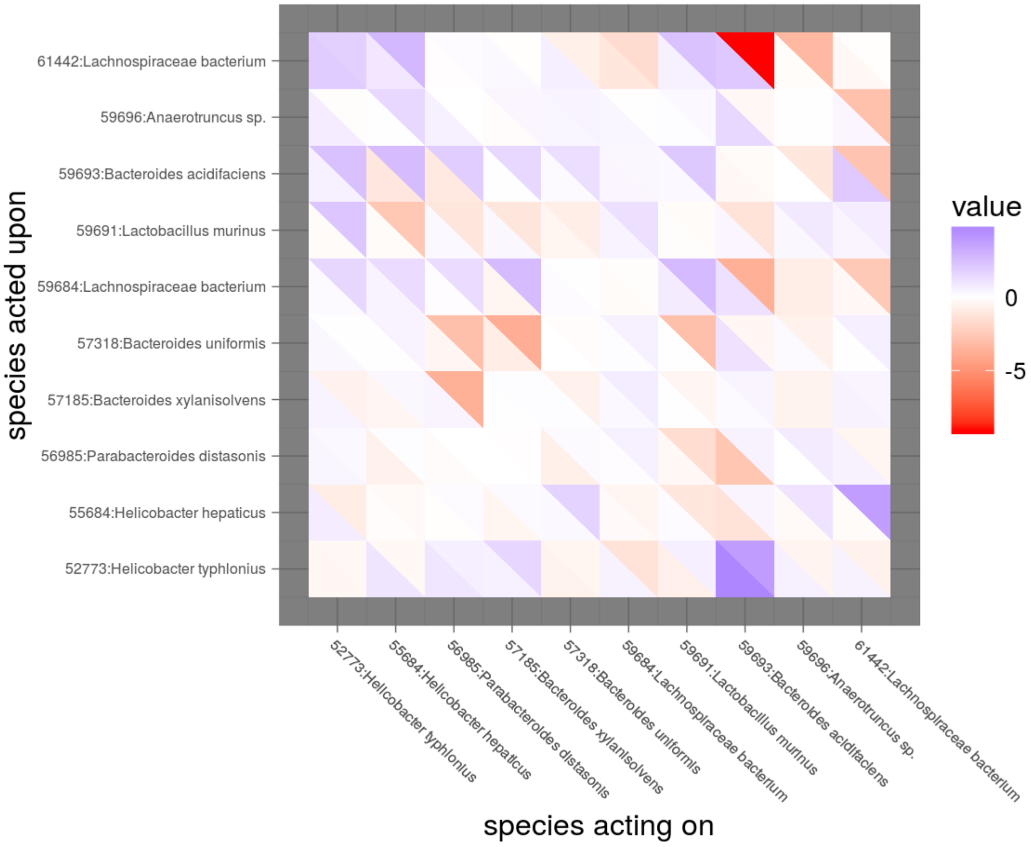

Figure 4. Means of interaction coefficients of the per-group Lotka-Volterra model. Upper triangles are the DNR estimates, lower triangles are WT estimates.

2.7 Interpreting the GLMM and FPCA results in the context of IBD and the changing intestinal environment

The 29 modules with different time trends between DNR and WT mice suggest a shifting ecological landscape with perturbed physicochemical properties and the accompanying response of the microbiota to these altered conditions. This shift is most evident from the presence of multiple two-component signaling, chemotactic, and redox homeostasis related modules. Bacteria capable of moving to a more favorable microenvironment within the gut or neutralizing reactive oxygen species are more likely to survive the localized effects of the inflammatory response[63]. Species capable of taking advantage of inflammation-related host metabolites also have a survival advantage[64].

Another overarching theme is of increased pathogenic potential, seen in the increase of adhesion related modules, Type III and VI secretion systems, keratan and dermatan degradation, and heme transport. Bacterial secretion systems generally carry out the function of injecting proteins into a target host or competitor cell[65][66]. While the increase of both systems is indicative of increased capacity to inject toxins into host cells, Type VI secretion additionally suggests that other microbiota members may also be targeted[67]. Keratan and dermatan sulfate are components of mucin, and while many non-harmful bacteria use them as an energy source[68], the increase in bacteria focused on extracting energy from this defensive barrier is suggestive of increased access to the intestinal epithelium[69]. Heme transport is another activity that regularly occurs within almost all bacteria[70], but the increase in host-heme “theft” is primarily observed in pathogenic bacteria[71][72]. Since bacteria seldom have access to free iron, they have acquired many adaptations to effectively trap sequestered iron[73].

Of extra interest are the 3 negative coefficient modules: melatonin biosynthesis, lysine biosynthesis, and lipooligosaccharide transport. Melatonin has a dual effect on the immune system[74], potentially acting in a stimulatory manner in the context of infection, and in an immunomodulatory manner in some cases of chronic inflammation[75]. The sharp decrease in melatonin synthesis as the inflammation progresses suggests that it is interpreted as the mark of a pathogen in this scenario, leading to the eradication of genomes that contain this function. The decrease in lysine biosynthesis indirectly leads to a presumed decrease in the synthesis of short chain fatty acids like butyrate[76], which normally serves an anti-inflammatory and colonocyte nurturing function[77]. The decrease in lipooligosaccharide transport seems puzzling at first glance, since lipooligosaccharides from Gram-negative bacteria have been shown to have a pro-inflammatory effects rather conclusively[78]. However, upon closer inspection of the KEGG references for this specific module, we find that it is sparsely characterized, and most references are to export activities in rhizobial bacteria[79]. Therefore this particular candidate biomarker still needs to be examined, as it has promising signal in my results (the only module to have significantly different slope in both pre- and post-onset segments), but lacks a clear mechanistic explanation.

From testing the intercept coefficient of GLMMs, we see that the microbiomes of DNR and WT mice already have significant differences early in development. The 17 modules that were significantly different at the earliest time point (week 4) represent a broad diversity of biosynthetic, signaling, and methanogenesis modules. This underscores the importance of adjusting for these pre-weaning differences when focusing on temporal changes. Additionally these intercept effects could represent the influences of cage effects, since the two groups of animals were housed separately, but were not single-caged.

It is notable that in the more detailed post-hoc segmented regression analysis of candidate modules we see almost universally that the significant intergroup slope differences occur after disease onset, suggesting that at least in this model of IBD, the microbiome primarily changes in response to the inflammation. There are however two modules which have a significant slope-group interaction in the pre-onset segment (type III secretion and lipooligosaccharide), making them potential candidate biomarkers for future more in-depth studies to test.

While the results of this work have primarily garnered leads from the functional characterization of the microbiome, a few species signals also emerge. Out of the 7 species that had significantly different abundance trajectories in the DNR group, 4 are members of the *Bacteroides* genus. Commensal *Bacteroides* species have been shown in the past to induce colitis in mouse models of the disease[80]. However, *Bacteroides* bacteria are also common members of most human and mouse microbiomes, therefore we need further experimental data before suggesting these bacterial species are singularly responsible for disease. Particularly because most of the significant trajectories share a sharp late upswing pattern, this suggests more of a response role and not a causative role.

It is important to note that all these findings were obtained from metagenomic data, hence they reflect the functional potential of the community, but not necessarily the transcriptional activity or protein and metabolite abundances. While DNA sequencing is currently much more widespread in the microbiome field, there are increasing numbers of integrative approaches that tackle metatranscriptomic and metaproteomic data as well. The work I have presented in this chapter would serve as a jumping-off point for in-depth hypothesis testing with these more complex approaches. Final confirmation of the effects of certain pathways or species would still need to come from carefully designed experimental perturbations of the microbial community.

3 Single cell sequencing: an effective approach for addressing the underrepresentation of mouse symbiotic microbes in reference databases

While many efforts are currently underway to characterize more members of the human microbiome, little data exists that could serve as a high-quality genome reference for mouse microbiome studies. To try and address this database bias, we used two biological samples from WT and DNR mice to generate preliminary low coverage and more thorough high coverage sequencing data for more than 700 individual cells. I evaluated the novelty of the newly sequenced genomes, both in terms of phylogenetic placement as well as the genomic features that could be retrieved. I also assessed the utility of these genomes in serving as a custom reference for two taxonomic classifier methods.

3.1 An optimization approach for prioritizing cells for high coverage sequencing

The single cell sequencing service provided by the Bigelow SCGC follows a two-stage pattern. First, low coverage genomes are generated and assessed for technical quality as well as relevance to the researcher's biological question. Second, the researcher chooses what cells from the plate they would like to get sequenced with higher coverage, and a second sequencing run is performed at greater depth. An agnostic approach for making this second choice relies on picking the first 150 cells per plate that had the lowest values of a technical parameter referred to as the critical point (Cp). Cp is the time needed by a particular well's reaction to reach the inflection point in its amplification curve. The sequencing center has determined empirically that this gives the highest chance of getting good quality data from the resequencing.

This approach of taking the most easily sequenceable samples is a good idea for diverse understudied environments like the marine microbiome, where virtually all new genomes are likely to be from organisms with limited representation. Since our motivating reason was the underrepresentation of mouse-specific microbes, I chose to alter the prioritization scheme to disfavor genomes that already have well characterized close relatives. If we had proceeded with the default proposal set, 44 of the 300 samples would have been from genomes with average nucleotide identities of more than 95% with a RefSeq genome, as determined by FastANI[81]. To generate a new set of proposed cells, I set the optimization objective to maximize the total branch length of selected tree tips, with the constraint that the sum of Cp values of the chosen tips must still stay under 1.2 times the minimal possible sum of Cp values. I additionally adjusted the costs of known undesirable samples (positive and negative sequencing controls, samples with known close representatives in RefSeq, samples that had poor assembly at the low coverage stage) by artificially inflating their Cp value.

The problem of maximizing phylogenetic diversity (PD) while limiting cost has been discussed in the ecology literature, where it is relevant for species conservation efforts and is generally solved using a greedy algorithm[82]. In theoretical discussions of the runtime of any feasible solution to this problem it has been linked to the broader computational task of maximizing set coverage, which is an NP-hard problem[83]. Despite this theoretical limitation on efficient computation, it’s still possible to get an approximate solution by using a mixed integer programming approach. I implemented this optimization problem using the PuLP[84] package in python and generated two sets of 150 samples that had increased phylogenetic diversity when compared to the default set achieved by just minimizing total Cp (total branch length of 60.24 vs 52.27, respectively). The distributions of a number of key technical characteristics that we already had from the low coverage data (Figure 5) were not significantly different when compared by the Mann-Whitney test (Table 4) suggesting that aside from the unavoidable shift in amplification efficiency, no other serious obstacles to high coverage sequencing should occur.

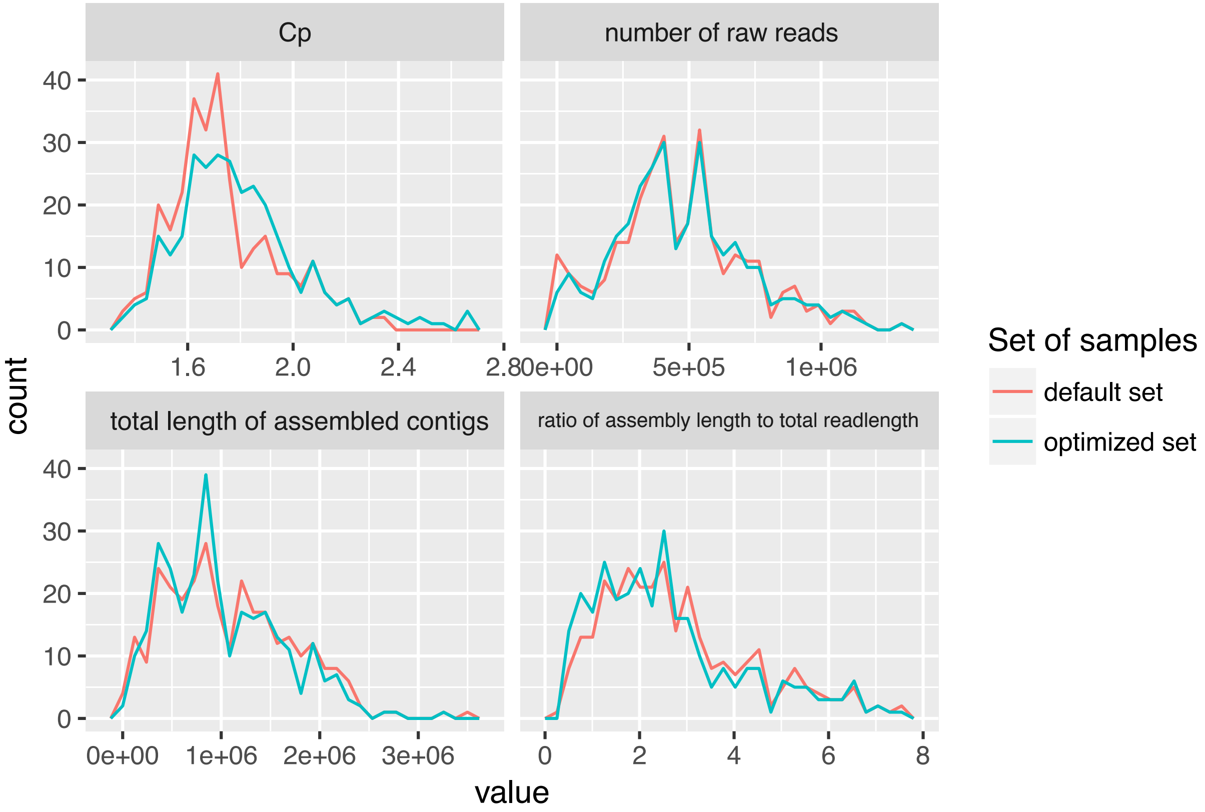

Figure 5. Comparison of the distributions of 4 technical characteristics between the default lowest Cp proposal set, and the optimized maximal PD proposal set

variable	**U statistic**	**P-value**	**B-H adjusted p-value**
Cp	52011	0.0009598	0.003839345
number of raw reads	45441	0.8356	0.835633779
total length of assembled contigs	42271	0.1987	0.264981305
ratio of assembly length to total read length	41073	0.06439	0.128789857

Table 4. Results of paired Mann-Whitney tests comparing 4 technical characteristics of the default and optimized sequencing proposals

3.2 Single cell genomes provide noticeable phylogenetic gain

The data processing pipeline followed by the sequencing center involves general QC of the short reads for contaminants and technical artifacts, generation of genome assemblies by SPAdes[85], and profiling of these assemblies by CheckM[86]. Compiling the CheckM results of all the cells profiled, we see the distributions shown in Figure 6. With follow-up high coverage sequencing of a subset of 300 cells, a significant improvement is observed in multiple assembly criteria. This includes completeness, maximum contig length, number of contigs, and total assembled sequence length (Figure 7). Since one of the primary goals of this sequencing endeavor was the to expand the diversity present in the tree of life, I used the GTDB-Tk python package[87] to place the new assemblies in the GTDB (release 80) genome tree[88]. I then calculated the phylogenetic gain for all named clades with GenomeTreeTk[89] and visualized it as a color gradient on the taxonomy tree with metacoder[90] (Figure 8).

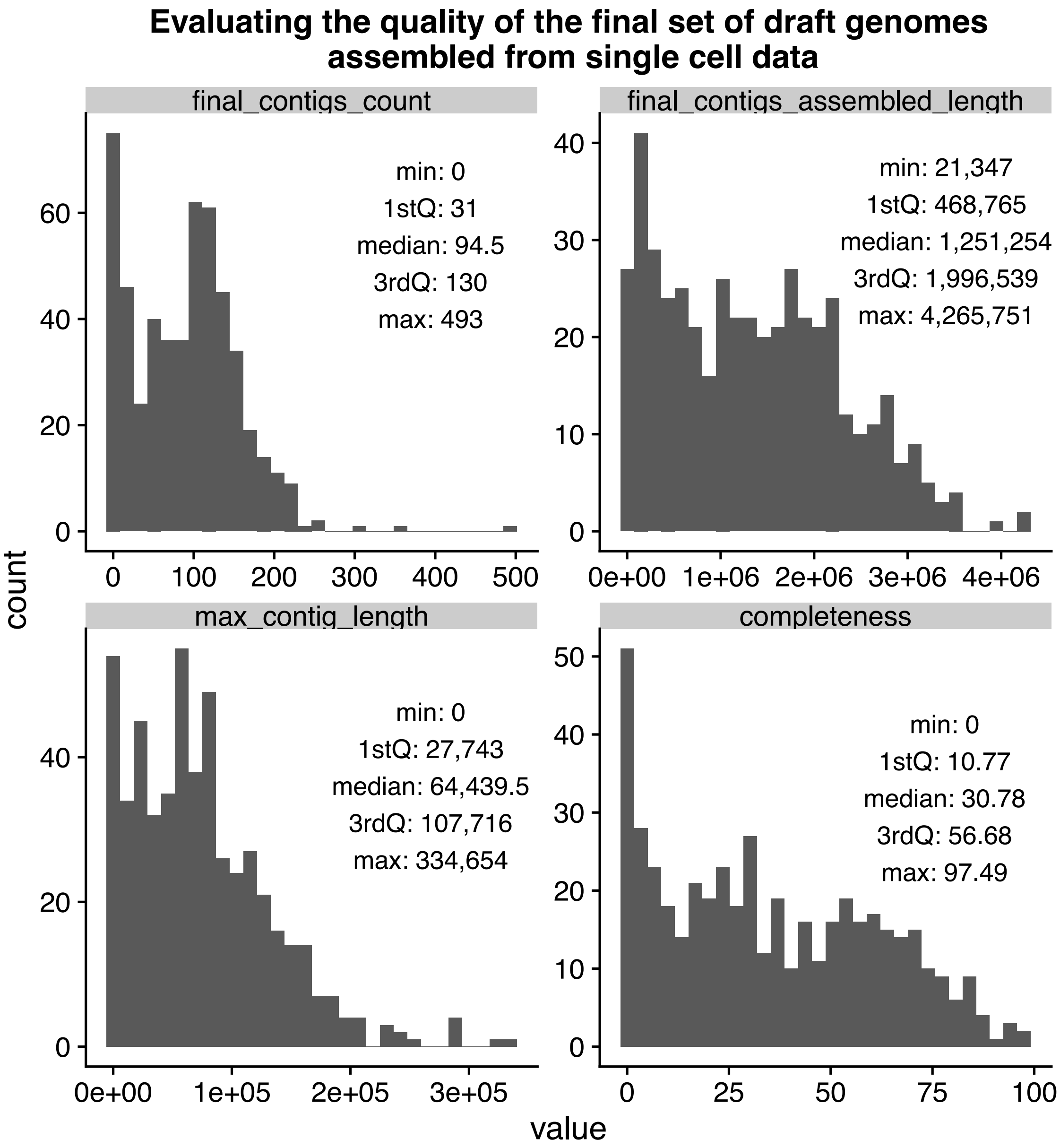

Figure 6. Assessment of assembly quality for the single cell genomes

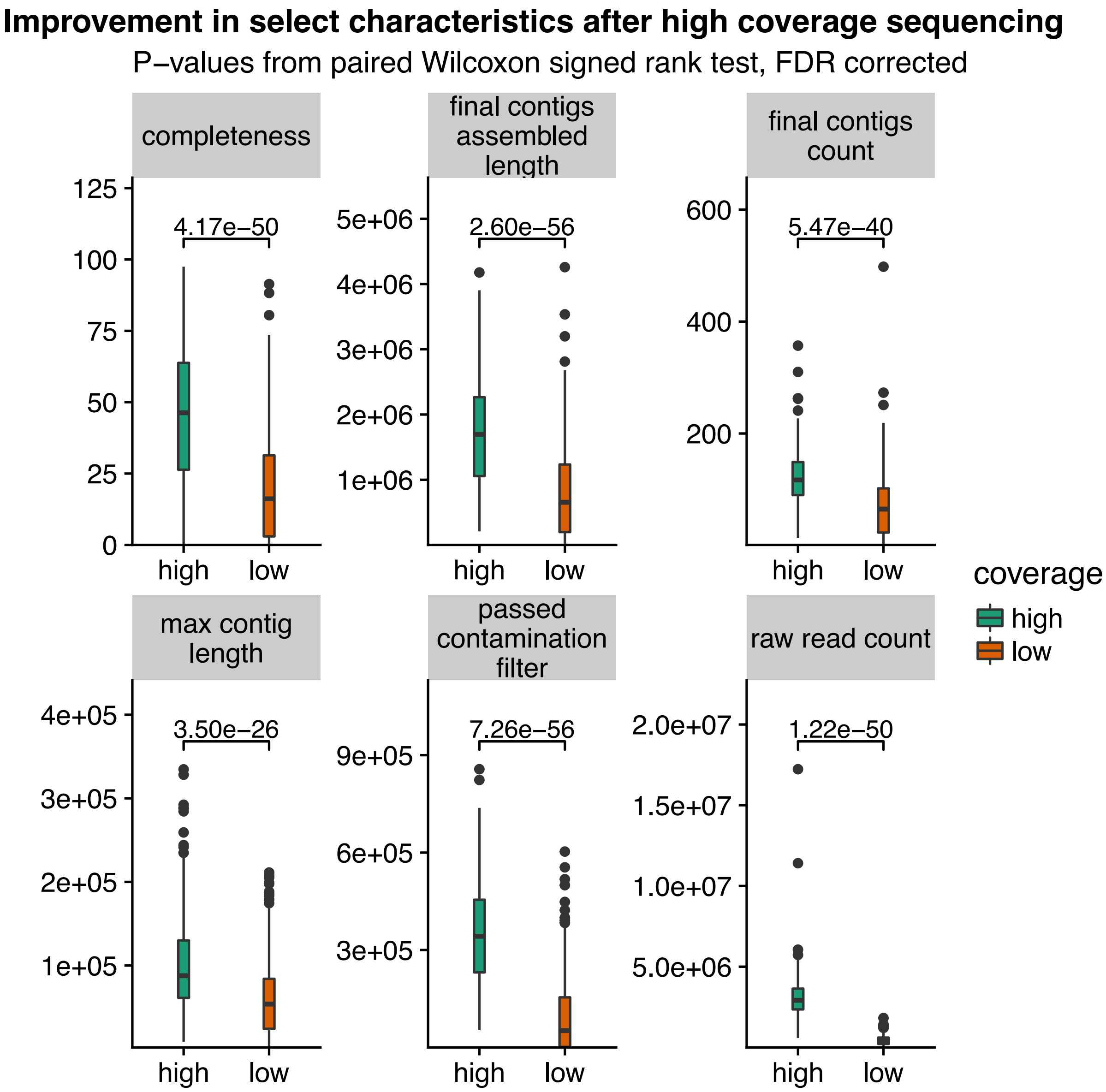

Figure 7. Improvements achieved by high coverage resequencing of select cells.

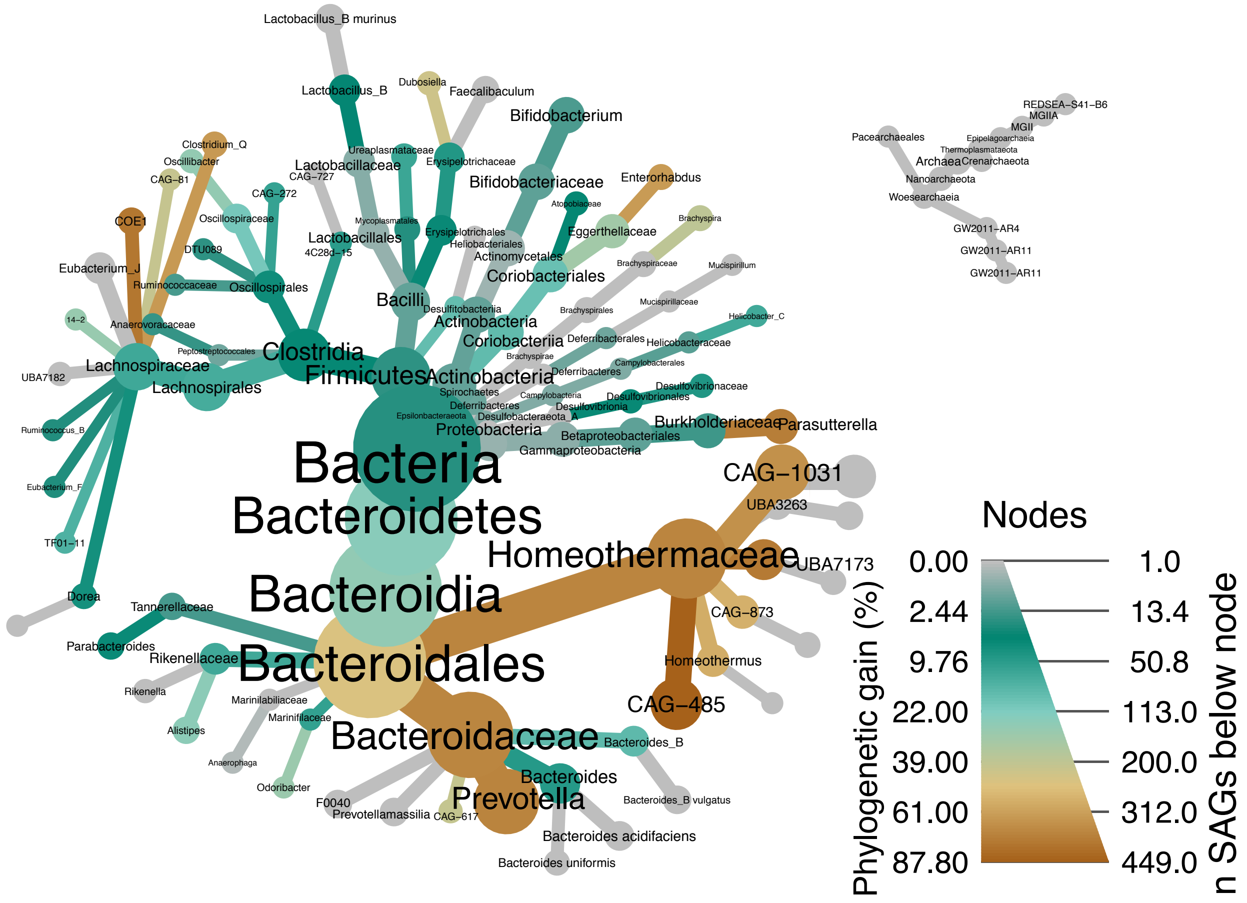

Figure 8. Metacoder heattree showing phylogenetic gain and concentration of single cell assembled genomes (SAGs) on select lineages in the bacterial and archaeal trees. This tree is generated from lineage strings assigned by GTDB-Tk's *classify* workflow, and hence this is not a true phylogenetic tree, as the branch lengths are not meaningful

3.3 The single cell genomes are a source of new genes and extended genomic feature assemblages

To show that we not only increased phylogenetic diversity but had also increased the collection of putative genes, I created a gene catalog from all predicted genes generated by checkM (which in turn uses Prodigal[91] for gene calling). To make this sequence collection comparable with other published gene catalogs, I reduced redundancy by running CD-HIT-EST[92] greedy clustering. I then annotated the remaining non-redundant sequences with

EggNOG's emapper.py utility[93], which outputs predicted membership of the query sequences in databases such as COG[94]. To compare the gene catalog I had generated with those that had been previously published, I used CD-HIT-EST-2D with settings "–r 1 –c 0.95 –n 8" to cluster three pairs of databases: the new SCG gene catalog against (1) the human metagenome gene catalog[95], (2) the mouse metagenome gene catalog[96], and (3) the Tara Oceans gene catalog[97]. To further probe the enrichment/depletion of various COG annotations present only in the single cell genomes gene catalog when compared against the mouse metagenome gene catalog, I performed a series of Fisher's exact tests assessing the relationship between a gene being labeled a certain COG category versus it being considered novel.

One of the advantages of uncontaminated single source genomic data is that we no longer have the question of whether a collection of functions truly coexists in a single closely spaced environment. This assurance of reasonable physical proximity allows us to investigate two kinds of interesting genomic features – biosynthetic clusters and CRISPR-Cas systems. I annotated biosynthetic gene clusters (BGCs) within the single cell draft genomes with AntiSMASH[98]. I found predicted CRISPR arrays with metaCRT[99] and classified the CRISPR-Cas types and subtypes with CRISPRdisco[100].

3.3.1 Results of annotating genome features

When analyzing the data from the gene catalog perspective, I found that despite the mouse metagenome catalog being a much larger set of sequences, over half of our predicted genes were not represented in it. The intersections with the human and ocean microbiome catalogs were even smaller, with the ocean dataset serving as the expected most dissimilar comparison. The counts of the set intersections can be seen in Figure 9.

Taking the catalog pairing with the largest intersection (mouse metagenome catalog), I wanted to investigate whether the genes that had not been cataloged before were enriched or depleted for a certain function. The results in Figure 10 are a barchart summarizing the collection of Fisher's exact tests performed on contingency tables relating the novelty of a gene versus it's annotation as a particular COG. Overall 144 COGs were significant in this analysis. Nearly every functional category had COGs enriched or depleted, with the exception of A(RNA processing and modification), B(Chromatin structure and dynamics), W(Extracellular structures), and Y(Nuclear structure). I further filtered the significant results by the value of inter-catalog ratios of COG proportions, and subsequently tallied the number of COGs per category with absolute value of the ratio greater than 4. The more salient functional groups after this filtering step are G(Carbohydrate transport and metabolism), which has more hits for the published catalog, and C(Energy production and conversion), which has more COGs enriched for the new catalog.

To represent the classified CRISPR-Cas types and the predicted biosynthetic gene clusters I plotted this information with ggtree[101], restricting the displayed genomes to the 449 that were placed successfully onto the phylogenetic tree by GTDB-Tk (Figure 11). Not all assembled genomes had a sufficiently complete set of single copy marker genes, which caused them to be dropped from the multiple sequence alignment performed by GTDB-Tk. The total counts of CRISPR-Cas types and subtypes, as well as secondary metabolite gene clusters by category can be seen in Table 5. The findings from this analysis show a phylogenetic separation in CRISPR-Cas types, a broad presence of two biosynthetic gene cluster types (saccharide and fatty acid), and a more narrow phylogenetic distribution of rarer types of BGCs (bacteriocin and resorcinol). These results are further discussed in section 1.1.

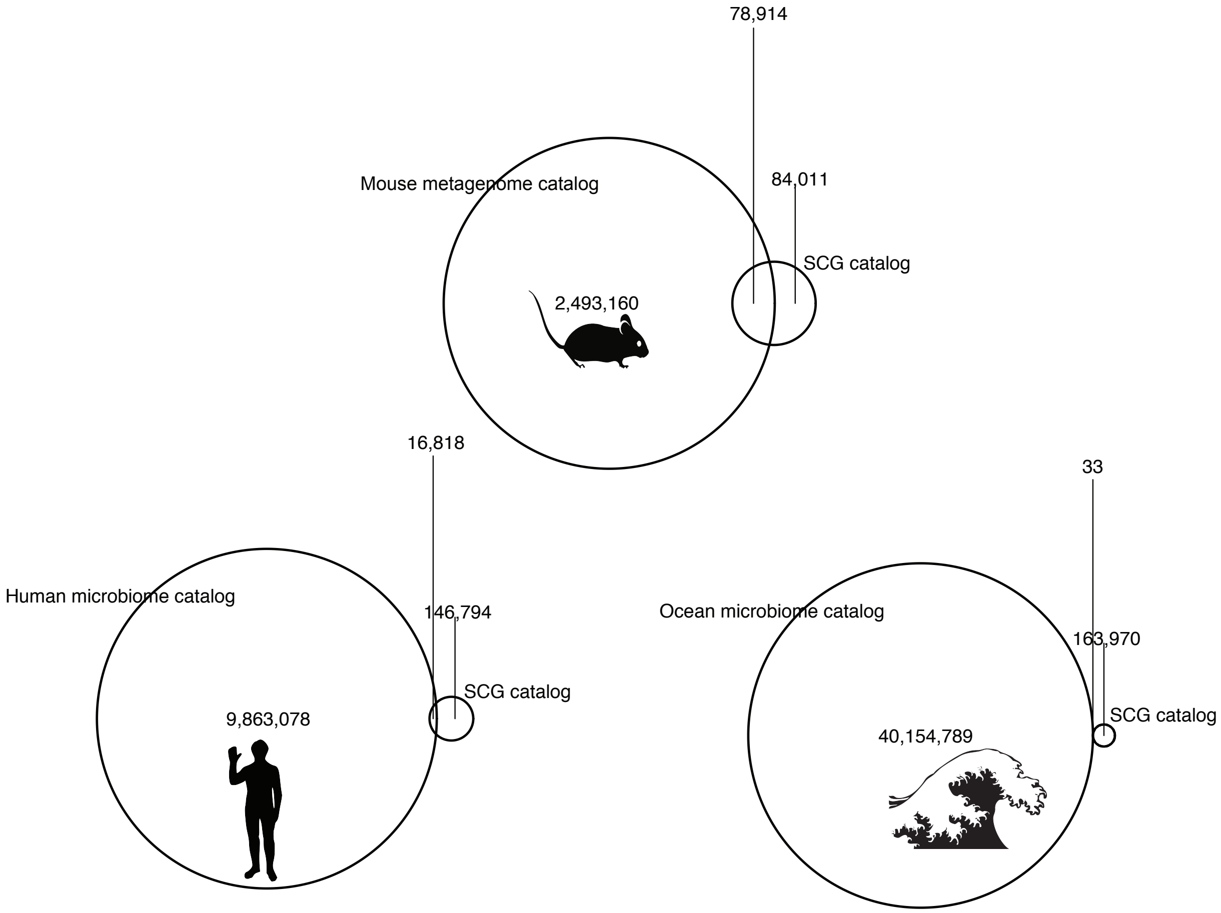

Figure 9. Venn diagrams of pairwise gene catalog intersections comparing the single cell mouse microbiome gene catalog to published catalogs (mouse, human, ocean). The 3 diagrams are not comparable between each other area-wise since they are scaled to be equal despite having an order of magnitude difference in total genes.

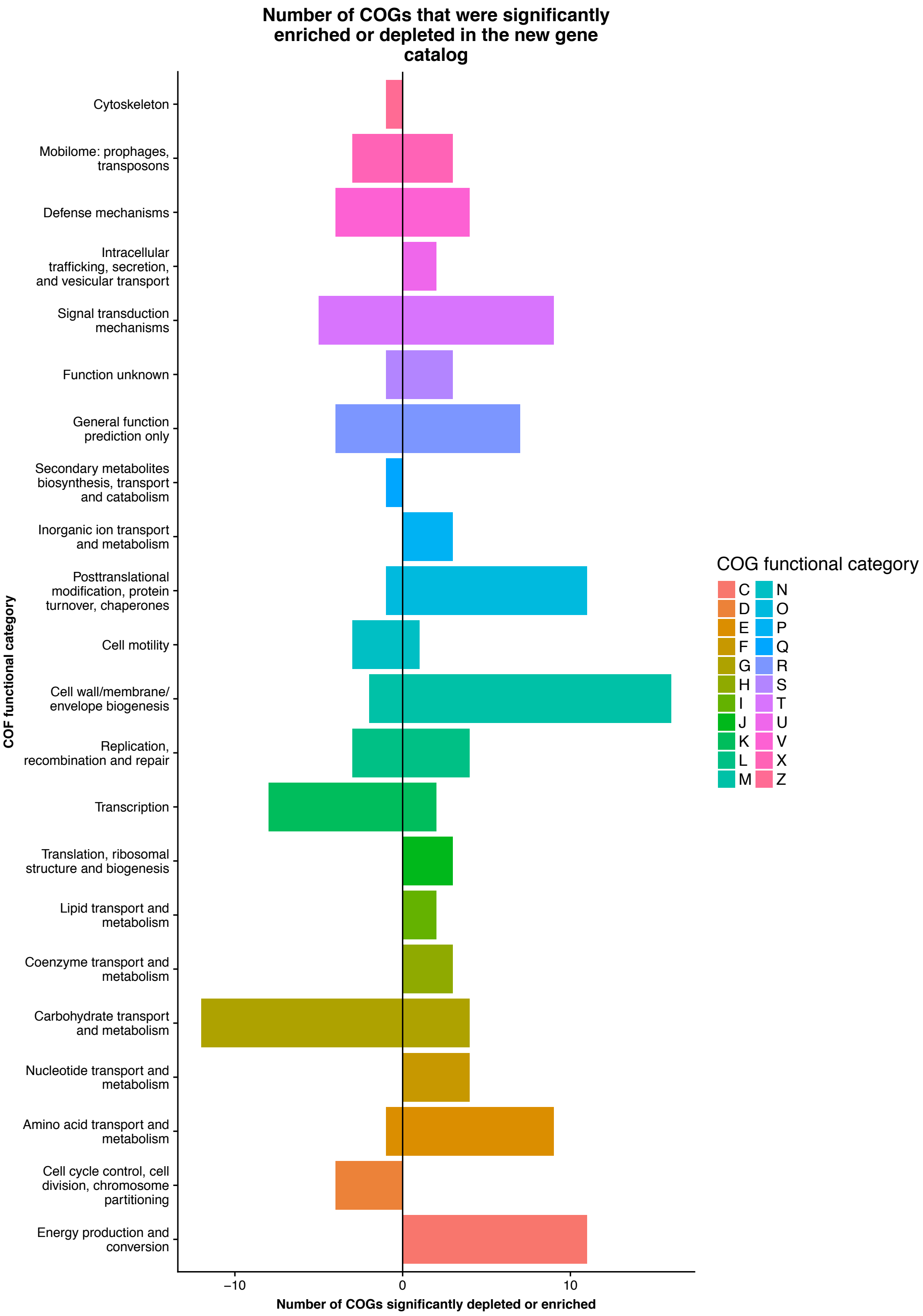

Figure 10. Barchart of COG counts that were significantly enriched (positive counts) or depleted (negative counts) in the new gene catalog, per COG functional category

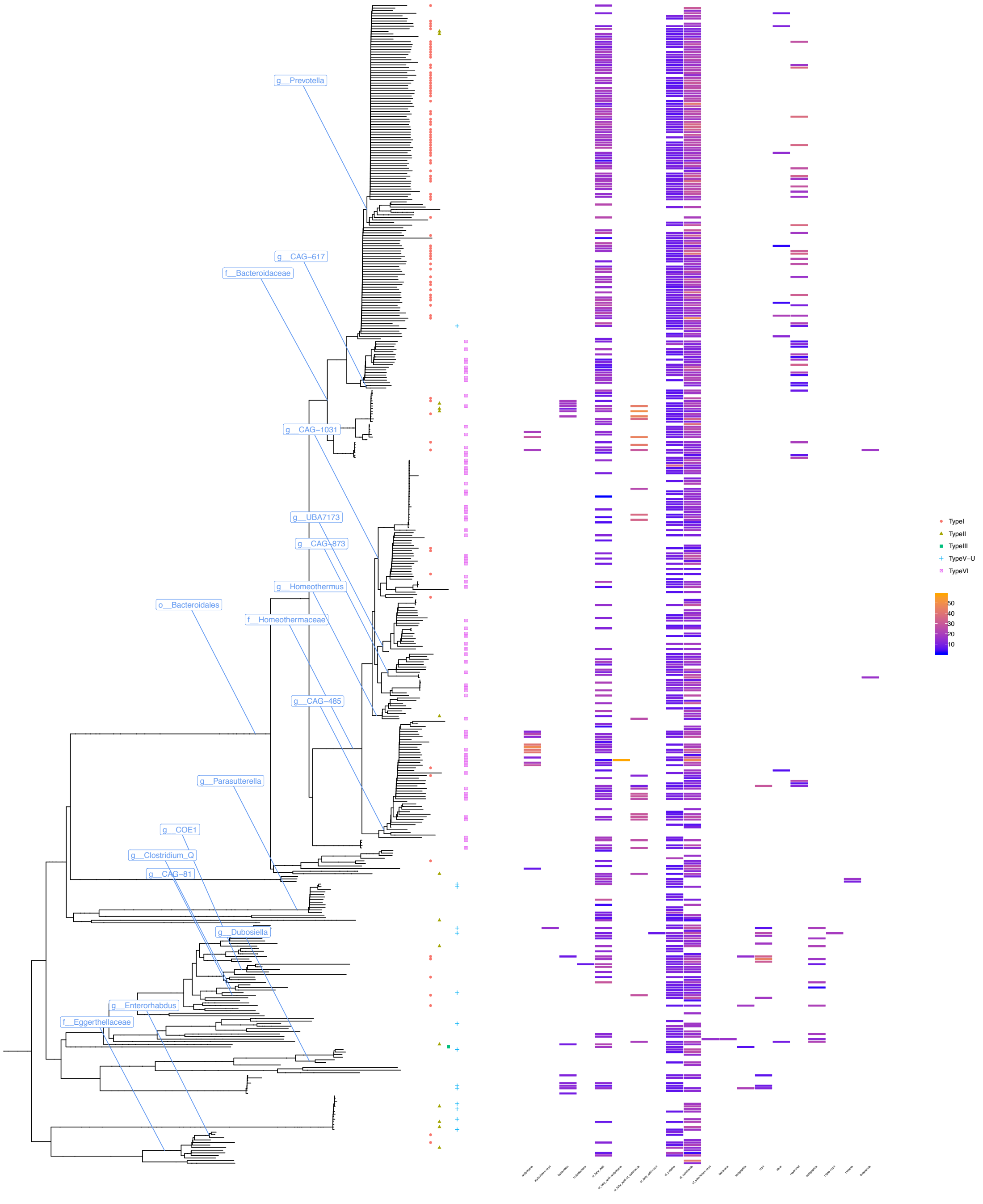

Figure 11. CRISPR classification and AntiSMASH predictions for single cell genomes that were successfully placed in the phylogenetic tree of GTDB genomes. The points near the tree tips represent CRISPR-Cas type classifications, while the heatmap on the right shows median number of genes per biosynthetic cluster.

Gene clusters:	
1416	cf_saccharide
653	cf_putative
334	cf_fatty_acid
42	resorcinol
26	cf_fatty_acid-cf_saccharide
17	arylpolyene
14	nrps
13	bacteriocin
12	sactipeptide
9	other
5	lantipeptide
2	terpene
2	thiopeptide
1	arylpolyene-nrps
1	butyrolactone
1	cf_fatty_acid-arylpolyene
1	cf_fatty_acid-nrps
1	cf_saccharide-nrps
1	ladderane
1	t1pks-nrps

CRISPR-Cas Types:	
96	TypeI
78	TypeVI
18	TypeII
16	TypeV-U
2	TypeIII

CRISPR-Cas Subtypes:	
78	TypeVI-B
44	TypeI-B
19	TypeI-C
16	TypeVU-4
13	TypeII-C
4	TypeII-B
2	TypeVU-2
1	TypeII-A

Table 5. Summary counts of annotated features retrieved from all single cell genomes that passed technical filtering.

3.4 Evaluating the utility of new genomes as a reference for metagenomic classification

To show that the draft genomes are a useful resource for our research and for others aiming to classify metagenomic data, I created custom reference databases for two taxonomic classification tools (MIDAS[37] and Sourmash[102]). I evaluated their performance on mouse metagenomes (derived from mice from the same lines housed in our facility, mice of various lines in other labs[96], a wild mouse population[103]), human metagenomes[104][105], and ocean metagenomes[97]. The expected results from this experiment were that the custom genomes would improve the classification of metagenomes from related mice at our facility, and hopefully other lab mice as well. The performance on the ocean dataset was expected to be poor, as there is very little similarity between free-living ocean microbes and host-associated mouse symbionts. Performance on human samples was expected to be intermediate, as there is some degree of similarity in conditions and taxonomic makeup within the guts of warm-blooded mammals.

For both taxonomic classification tools, three reference databases were used:

1. Default reference – created from RefSeq published bacterial genomes (MIDAS db v1.2; Sourmash LCA db created from genomes in GTDB release 80)
2. SCG-only reference – a reference created from the newly sequenced draft mouse microbiome genomes
3. Combined reference - a combined database incorporating both our new single cell genomes and the existing genomes that make up the default reference

The two tools used were chosen as representatives of two major paradigms in the taxonomic classification of metagenomes: the marker gene-based approach (MIDAS) and the least common ancestor-aware kmer approach (Sourmash). It should be noted that Sourmash technically does not operate on the full set of all computed kmers from a sequence, but instead uses a locality sensitive hashing algorithm called MinHash to generate much more compact signatures that still retain approximately the same nucleotide comparison properties as the full kmer feature vector.

3.4.1 Results of custom database tests with MIDAS

To assess the improvement in performance when using the SCG-only or combined reference with MIDAS's *run_midas.py species* command, I looked at three metrics of classification success: marker gene mean coverage, marker gene median coverage, and species prevalence. These three metrics can be qualitatively compared per dataset in the ridgeline plots in Figure 12, and are quantitatively compared with two-sided Mann-Whitney tests, the results of which are in Table 6.

dataset	**variable**	**refA**	**refB**	**medianA**	**medianB**	**p.value**	**p.adjusted**
dnr	mean_coverage	midas_db_combined	midas_db_scg_only	15.74	14.57	0.44812999	0.733303619
dnr	mean_coverage	midas_db_combined	midas_db_v1.2	15.74	7.015	5.70E-05	6.15E-04
dnr	mean_coverage	midas_db_scg_only	midas_db_v1.2	14.57	7.015	2.54E-04	0.002281843
dnr	median_coverage	midas_db_combined	midas_db_scg_only	5.22	2.08	0.53599105	0.785401455
dnr	median_coverage	midas_db_combined	midas_db_v1.2	5.22	3.495	0.597459928	0.786898441
dnr	median_coverage	midas_db_scg_only	midas_db_v1.2	2.08	3.495	0.261915016	0.471447028
dnr	prevalence	midas_db_combined	midas_db_scg_only	42	34	0.198740044	0.397480087
dnr	prevalence	midas_db_combined	midas_db_v1.2	42	44.5	0.429642592	0.725021874
dnr	prevalence	midas_db_scg_only	midas_db_v1.2	34	44.5	0.07355344	0.208075713
hmp	mean_coverage	midas_db_combined	midas_db_scg_only	20.875	9.18	0.245789638	0.457677258
hmp	mean_coverage	midas_db_	midas_db_	20.875	18.08	0.650985378	0.829919145

		combined	v1.2				
hmp	mean_coverage	midas_db_scg_only	midas_db_v1.2	9.18	18.08	0.34633574	0.603294514
hmp	median_coverage	midas_db_combined	midas_db_scg_only	2.615	4.46	0.218098389	0.420618322
hmp	median_coverage	midas_db_combined	midas_db_v1.2	2.615	2.28	0.736107221	0.883328665
hmp	median_coverage	midas_db_scg_only	midas_db_v1.2	4.46	2.28	0.139878478	0.314726575
hmp	prevalence	midas_db_combined	midas_db_scg_only	142	173	0.077065079	0.208075713
hmp	prevalence	midas_db_combined	midas_db_v1.2	142	131	0.507906583	0.783627299
hmp	prevalence	midas_db_scg_only	midas_db_v1.2	173	131	0.026956845	0.090979351
humanT1D	mean_coverage	midas_db_combined	midas_db_scg_only	4.89	2.13	0.183102723	0.39120868
humanT1D	mean_coverage	midas_db_combined	midas_db_v1.2	4.89	4.89	0.85559304	0.943737821
humanT1D	mean_coverage	midas_db_scg_only	midas_db_v1.2	2.13	4.89	0.188359735	0.39120868
humanT1D	median_coverage	midas_db_combined	midas_db_scg_only	2.035	1.07	0.590910697	0.786898441
humanT1D	median_coverage	midas_db_combined	midas_db_v1.2	2.035	2.02	0.913555212	0.970190443
humanT1D	median_coverage	midas_db_scg_only	midas_db_v1.2	1.07	2.02	0.557333175	0.786898441
humanT1D	prevalence	midas_db_combined	midas_db_scg_only	22	18	0.538145441	0.785401455
humanT1D	prevalence	midas_db_combined	midas_db_v1.2	22	23	0.916290974	0.970190443
humanT1D	prevalence	midas_db_scg_only	midas_db_v1.2	18	23	0.490195555	0.778545882
lab_mouse	mean_coverage	midas_db_combined	midas_db_scg_only	6.19	6.155	0.568679866	0.786898441
lab_mouse	mean_coverage	midas_db_combined	midas_db_v1.2	6.19	1.205	1.62E-07	2.91E-06
lab_mouse	mean_coverage	midas_db_scg_only	midas_db_v1.2	6.155	1.205	1.79E-06	2.41E-05
lab_mouse	median_coverage	midas_db_combined	midas_db_scg_only	0	0.015	0.660861541	0.829919145
lab_mouse	median_coverage	midas_db_combined	midas_db_v1.2	0	0.17	0.014817621	0.057153682
lab_mouse	median_coverage	midas_db_scg_only	midas_db_v1.2	0.015	0.17	0.044084809	0.132254426
lab_mouse	prevalence	midas_db_combined	midas_db_scg_only	11	7	0.759251269	0.891294968
lab_mouse	prevalence	midas_db_combined	midas_db_v1.2	11	2	7.31E-08	1.97E-06
lab_mouse	prevalence	midas_db_scg_only	midas_db_v1.2	7	2	3.05E-08	1.65E-06
tara	mean_coverage	midas_db_combined	midas_db_scg_only	1.86	0	0.001258474	0.006946059
tara	mean_coverage	midas_db_combined	midas_db_v1.2	1.86	1.81	0.963298771	1
tara	mean_coverage	midas_db_scg_only	midas_db_v1.2	0	1.81	0.001286307	0.006946059
tara	median_coverage	midas_db_combined	midas_db_scg_only	0.03	0	0.002161396	0.010610489
tara	median_coverage	midas_db_combined	midas_db_v1.2	0.03	0.03	0.990764315	1
tara	median_coverage	midas_db_scg_only	midas_db_v1.2	0	0.03	0.004128568	0.018578554
tara	prevalence	midas_db_combined	midas_db_scg_only	4	0	0.001114472	0.006946059

tara	prevalence	midas_db_combined	midas_db_v1.2	4	4	0.856354689	0.943737821
tara	prevalence	midas_db_scg_only	midas_db_v1.2	0	4	0.001135264	0.006946059
wild_mouse	mean_coverage	midas_db_combined	midas_db_scg_only	1.63	0.885	0.004647395	0.019304565
wild_mouse	mean_coverage	midas_db_combined	midas_db_v1.2	1.63	1.11	0.126207095	0.296312309
wild_mouse	mean_coverage	midas_db_scg_only	midas_db_v1.2	0.885	1.11	0.701766439	0.861258812
wild_mouse	median_coverage	midas_db_combined	midas_db_scg_only	0.64	0.39	0.042821646	0.132254426
wild_mouse	median_coverage	midas_db_combined	midas_db_v1.2	0.64	0.555	0.779618551	0.895731953
wild_mouse	median_coverage	midas_db_scg_only	midas_db_v1.2	0.39	0.555	0.121881795	0.296312309
wild_mouse	prevalence	midas_db_combined	midas_db_scg_only	3	2	0.01667653	0.060035508
wild_mouse	prevalence	midas_db_combined	midas_db_v1.2	3	2	0.104136244	0.267778914
wild_mouse	prevalence	midas_db_scg_only	midas_db_v1.2	2	2	1	1

Table 6. Results of two-sided Mann-Whitney tests of MIDAS performance characteristics achieved with different reference databases. Yellow shaded rows mark tests where the B-H adjusted p-value was less than 0.1. Green shaded rows also fulfill that condition, but additionally signify that the comparison is not trivial, i.e. it involves the SCG-only or combined database outperforming the default GTDB database.

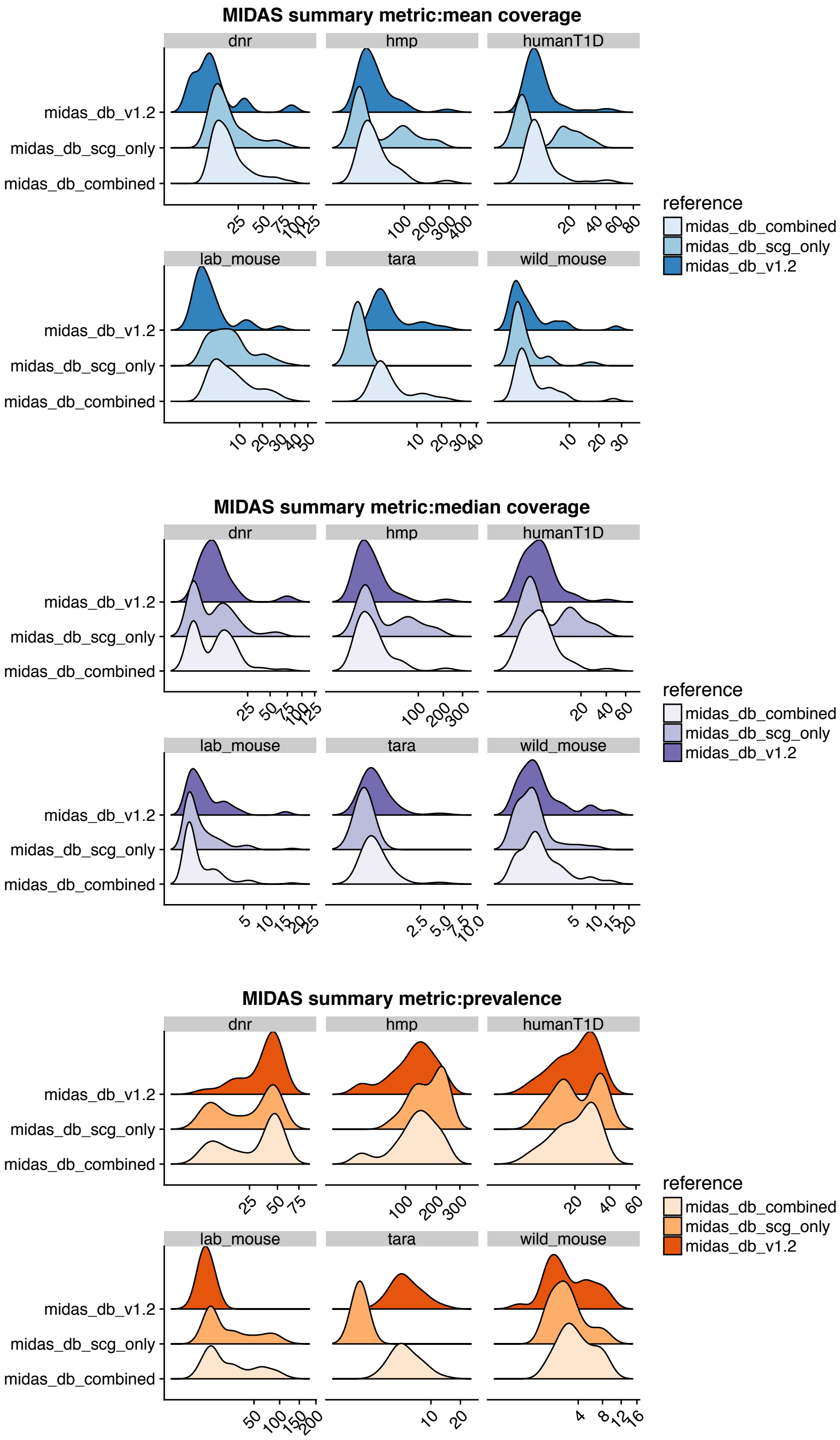

Figure 12. Ridgeline plots of 3 MIDAS performance metrics (mean coverage, median coverage, prevalence), plotted with each reference type per line, facetted by test dataset.

3.4.2 Results of custom database tests with Sourmash-LCA

To assess the improvement in performance when using the SCG-only or combined reference with Sourmash's *lca summarize* command, I looked at two metrics of metagenomic sequence recruitment: total number of hashes assigned to a sample and number of species-level nodes in the taxonomy tree that had more than 5 hashes assigned to them (a proxy for species prevalence). These two metrics can be qualitatively compared per dataset in the ridgeline plots in Figure 13, and are quantitatively compared with two-sided Mann-Whitney tests, the results of which are in Table 7. While many of the comparisons are significant, it should be noted that comparisons involving "scg_ref" are of secondary importance, since they seldom detect unusual cases where the much smaller SCG-only reference outperforms the other two contenders. Most of the focus of this series of tests is in showing the increased performance of the combined reference.

dataset	variable	refA	refB	medianA	medianB	p.value	p.adjusted
dnr	total hashes assigned	combined_ref	gtdb_ref	86334	36425	1.04E-20	3.33E-20
dnr	total hashes assigned	combined_ref	scg_ref	86334	52214	1.02E-16	2.51E-16
dnr	total hashes assigned	gtdb_ref	scg_ref	36425	52214	5.67E-13	1.01E-12
dnr	species with count > 5	combined_ref	gtdb_ref	26	29	0.014081245	0.018774993
dnr	species with count > 5	combined_ref	scg_ref	26	5	1.67E-16	3.60E-16
dnr	species with count > 5	gtdb_ref	scg_ref	29	5	1.69E-16	3.60E-16
hmp	total hashes assigned	combined_ref	gtdb_ref	88886.5	88263	0.791983639	0.873990843
hmp	total hashes assigned	combined_ref	scg_ref	88886.5	4929	1.54E-75	1.65E-74
hmp	total hashes assigned	gtdb_ref	scg_ref	88263	4929	2.25E-75	1.80E-74
hmp	species with count > 5	combined_ref	gtdb_ref	106	106	0.819366415	0.873990843
hmp	species with count > 5	combined_ref	scg_ref	106	4	5.14E-84	1.08E-82
hmp	species with count > 5	gtdb_ref	scg_ref	106	4	6.77E-84	1.08E-82
humanT1D	total hashes assigned	combined_ref	gtdb_ref	104829.5	104246.5	0.792522646	0.873990843
humanT1D	total hashes assigned	combined_ref	scg_ref	104829.5	4996.5	4.52E-21	1.61E-20

humanT1D	total hashes assigned	gtdb_ref	scg_ref	104246.5	4996.5	4.52E-21	1.61E-20
humanT1D	species with count > 5	combined_ref	gtdb_ref	121	123	0.786757619	0.873990843
humanT1D	species with count > 5	combined_ref	scg_ref	121	4	1.26E-14	2.37E-14
humanT1D	species with count > 5	gtdb_ref	scg_ref	123	4	1.26E-14	2.37E-14
lab_mouse	total hashes assigned	combined_ref	gtdb_ref	34587	25295	1.00E-17	2.67E-17
lab_mouse	total hashes assigned	combined_ref	scg_ref	34587	11820	1.34E-46	6.14E-46
lab_mouse	total hashes assigned	gtdb_ref	scg_ref	25295	11820	2.32E-18	6.75E-18
lab_mouse	species with count > 5	combined_ref	gtdb_ref	23	25	0.102407583	0.131081706
lab_mouse	species with count > 5	combined_ref	scg_ref	23	4	7.31E-56	3.90E-55
lab_mouse	species with count > 5	gtdb_ref	scg_ref	25	4	1.75E-57	1.12E-56
tara	total hashes assigned	combined_ref	gtdb_ref	33375	33380	0.956837144	0.987702858
tara	species with count > 5	combined_ref	gtdb_ref	79.5	79.5	1	1
wild_mouse	total hashes assigned	combined_ref	gtdb_ref	28864	18836.5	2.17E-05	3.30E-05
wild_mouse	total hashes assigned	combined_ref	scg_ref	28864	11664.5	1.08E-05	1.82E-05
wild_mouse	total hashes assigned	gtdb_ref	scg_ref	18836.5	11664.5	2.17E-05	3.30E-05
wild_mouse	species with count > 5	combined_ref	gtdb_ref	19.5	20.5	0.568708623	0.699949074
wild_mouse	species with count > 5	combined_ref	scg_ref	19.5	4	1.63E-04	2.28E-04
wild_mouse	species with count > 5	gtdb_ref	scg_ref	20.5	4	1.64E-04	2.28E-04

Table 7. Results of two-sided Mann-Whitney tests comparing the change in performance of Sourmash-LCA when using different reference databases. Yellow shaded rows mark tests where the B-H adjusted p-value was less than 0.1. Green shaded rows also fulfill that condition, but additionally signify that the comparison is not trivial, i.e. it involves the SCG-only or combined database outperforming the default GTDB database.

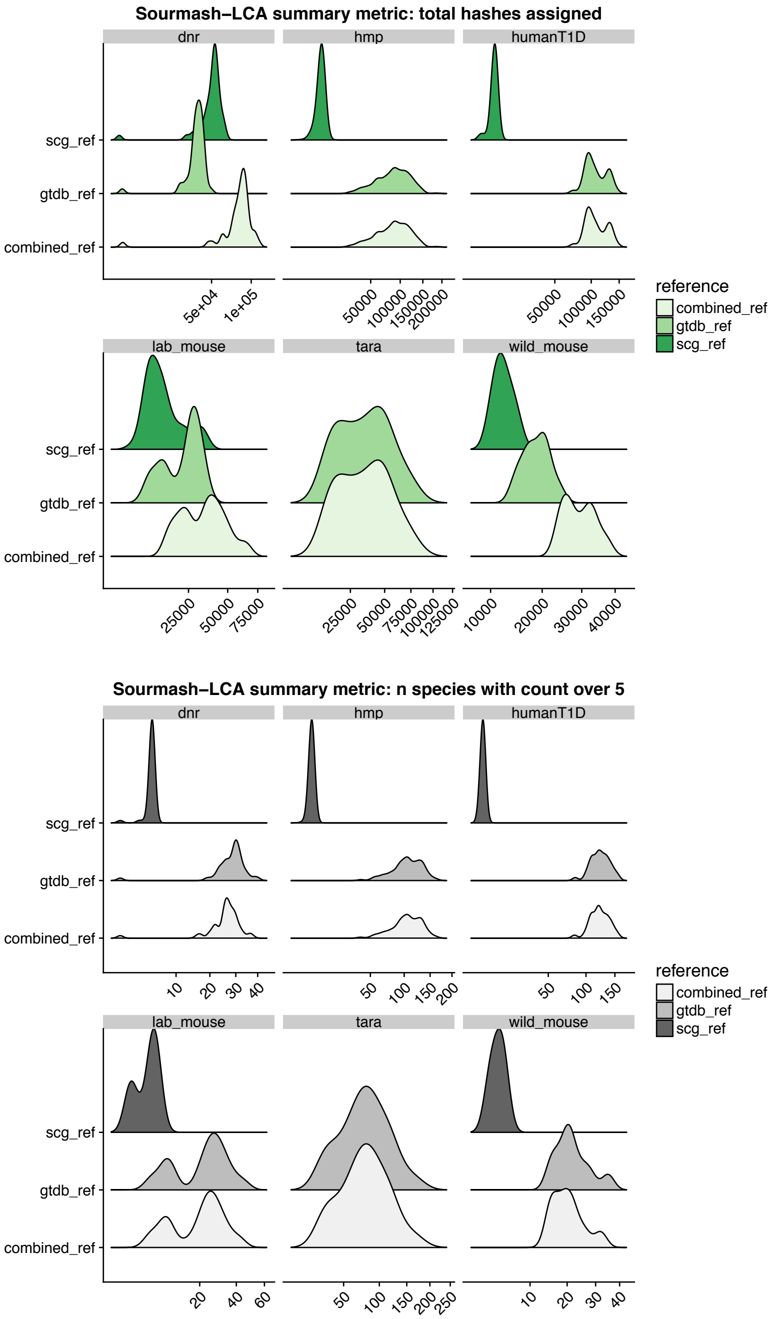

Figure 13. Ridgeline plots of 2 Sourmash-LCA performance metrics, plotted with each reference type per line, facetted by test dataset. Ridgelines are missing for the Tara dataset tested with the single cell only reference because no hashes were assigned in those runs.

3.5 Conclusions from examining the genome features annotated in single cell genomes

CRISPR typing/subtyping continues to be an evolving line of computational research, and the assignments we have obtained will likely change as the field moves forward. Even with the coarse classification approach that is available now, we can already see an interesting separation that occurs between phylogenetically disparate bacteria in this modestly sized set of genomes. It appears that the Type VI CRISPR-Cas system is favored by bacteria of the newly named "*Candidatus* Homeothermaceae" family (previously Bacteroidales family S24-7), which is also the family with one of the highest percentages of phylogenetic gain. This taxon has frequently appeared in 16S studies of various diseases as an OTU of interest[106][107][108]. Despite being a common member of the warm-blooded animal microbiota, it has not been studied as thoroughly as other clades. A recent effort to survey this family more closely has been carried out by Ormerod et al[109], who examined 30 metagenome assembled genomes from new and previously sequenced stool samples.

We can also see that that although the most popular AntiSMASH categories are the ubiquitous saccharide biosynthesis and to a lesser degree fatty acid biosynthesis, there are a few rare hits that can be of potential interest for a deeper dive. For example, resorcinol, which seems to be concentrated in single cell genomes assigned to the Bacteroidaceae family. The signature gene for this cluster, DarB, is present in the KEGG database as K00648, which is part of the fatty acid biosynthesis pathway. Closer investigation of the literature regarding the putative products of this BGC reveals that bacterial dialkylresorcinols have a wide variety of effects, exhibiting antibiotic, antiproliferative, and anti-inflammatory activities[110]. Intriguingly, the resorcinol BGCs in our genomes are not evenly spread between the two biological samples – 36 are from the DNR sample while only 6 are from the WT. This suggests an interesting future line of inquiry into the relationship between this colitis-attenuating compound[111] and the pro-inflammatory host genotype.

3.6 Conclusions from evaluating the dataset as a reference for metagenomics studies

In line with my expectations, the draft genomes improved performance for metagenomic read classification in similar environments, but were not useful when tested in a dissimilar complex microbiome. The improvements in classifying reads from wild mice with the kmer-based tool Sourmash and the combined reference are particularly encouraging. Performance was expected to drop off for that dataset due to the increased diversity of wild mouse microbiomes[103]. The lack of significant improvement in more of the comparisons involving human samples can potentially be explained by heterogeneity in those samples, since the distributions in Figure 12 show bimodality in some of the human dataset panels. The findings in this section echo similar investigations into expanding reference databases in general[112], which have noted that taxonomic classifier performance is as much a function of the reference database as it is of the algorithm used.

4 Conclusion

The work I have presented in this dissertation highlights the importance of gathering the right kind of data and examining it with appropriate statistical methods. In Chapter 2, I detailed the evolution of my approach for studying the temporal changes that occur in a mouse model of IBD. This study yielded multiple promising biological pathways that could serve as candidate biomarkers, as well as a handful of species with a marked response to the inflamed environment. Changes that occurred early on in the disease trajectory were particularly useful, as they could serve as hypotheses for intervention experiments aimed at curtailing IBD early on. In Chapter 3, I showcase the surprising effectiveness of a modest sequencing effort for the purposes of creating a custom reference database and investigating an understudied set of genomes. The genomes that our lab has generated provide a new look into the characteristic features of a mouse gut symbiont. They also expand the representation of a clade that previously had only 16S markers available, and had only recently been more thoroughly investigated with longer sequences reconstructed from metagenomes. The improvements achieved in supporting metagenomic classification in mice should indirectly lead to improvements in human studies as well, as we learn what features are specific to the most widely used model animal, versus what findings are truly generalizable.

References

1. Ursell LK, Metcalf JL, Parfrey LW, Knight R. Defining the human microbiome. Nutr Rev. 2012;70 Suppl 1 Suppl 1:S38-44. doi:10.1111/j.1753-4887.2012.00493.x.

2. Kamada N, Chen GY, Inohara N, Núñez G. Control of pathogens and pathobionts by the gut microbiota. Nat Immunol. 2013;14:685–90. doi:10.1038/ni.2608.

3. LeBlanc JG, Milani C, de Giori GS, Sesma F, van Sinderen D, Ventura M. Bacteria as vitamin suppliers to their host: A gut microbiota perspective. Current Opinion in Biotechnology. 2013;24:160–8.

4. Tremaroli V, Bäckhed F. Functional interactions between the gut microbiota and host metabolism. Nature. 2012;489:242–9.

5. Round JL, Mazmanian SK. The gut microbiota shapes intestinal immune responses during health and disease. Nature Reviews Immunology. 2009;9:313–23.

6. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, et al. The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes. Cell Host Microbe. 2015;17:260–73. doi:10.1016/j.chom.2015.01.001.

7. Kappelman MD, Moore KR, Allen JK, Cook SF. Recent trends in the prevalence of Crohn’s disease and ulcerative colitis in a commercially insured US population. Dig Dis Sci. 2013;58:519–25.

8. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology. 2012;142:46–54. doi:10.1053/j.gastro.2011.10.001.

9. Ng SC, Tang W, Leong RW, Chen M, Ko Y, Studd C, et al. Environmental risk factors in inflammatory bowel disease: A population-based case-control study in Asia-Pacific. Gut. 2015;64:1063–71.

10. Rook GAW. Hygiene hypothesis and autoimmune diseases. Clin Rev Allergy Immunol. 2012;42:5–15. doi:10.1007/s12016-011-8285-8.

11. Ng SC, Bernstein CN, Vatn MH, Lakatos PL, Loftus E V., Tysk C, et al. Geographical variability and environmental risk factors in inflammatory bowel disease. Gut. 2013;62:630–49.

12. Hold GL. Western lifestyle: a "master" manipulator of the intestinal microbiota? Gut. 2014;63:5–6. doi:10.1136/gutjnl-2013-304969.

13. Basson A, Trotter A, Rodriguez-Palacios A, Cominelli F. Mucosal Interactions between Genetics, Diet, and Microbiome in Inflammatory Bowel Disease. Front Immunol. 2016;7:290. doi:10.3389/fimmu.2016.00290.

14. Binder V. Epidemiology of IBD during the twentieth century: An integrated view. Best Practice and Research: Clinical Gastroenterology. 2004;18:463–79.

15. Halme L, Paavola-Sakki P, Turunen U, Lappalainen M, Farkkila M, Kontula K. Family and twin studies in inflammatory bowel disease. World J Gastroenterol. 2006;12:3668–72.

doi:10.3748/WJG.V12.I23.3668.

16. Gordon H, Trier Moller F, Andersen V, Harbord M. Heritability in inflammatory bowel disease: From the first twin study to genome-wide association studies. Inflammatory Bowel Diseases. 2015;21:1428–34.

17. Ananthakrishnan AN. Epidemiology and risk factors for IBD. Nat Rev Gastroenterol Hepatol. 2015;12:205–17. doi:10.1038/nrgastro.2015.34.

18. Integrative HMP (iHMP) Research Network Consortium TIH (iHMP) RN. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe. 2014;16:276–89. doi:10.1016/j.chom.2014.08.014.

19. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol. 2015;15:66. doi:10.1186/s12866-015-0351-6.

20. Martiny JBH, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: A phylogenetic perspective. Science (80-). 2015;350:aac9323. doi:10.1126/science.aac9323.

21. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014;505:559–63. doi:10.1038/nature12820.

22. Maurice CF, Haiser HJ, Turnbaugh PJ. Xenobiotics Shape the Physiology and Gene Expression of the Active Human Gut Microbiome. Cell. 2013;152:39–50.

doi:10.1016/J.CELL.2012.10.052.

23. Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature. 2018.

24. Wu H, Esteve E, Tremaroli V, Khan MT, Caesar R, Mannerås-Holm L, et al. Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. Nat Med. 2017;23:850–8. doi:10.1038/nm.4345.

25. Kelly JR, Kennedy PJ, Cryan JF, Dinan TG, Clarke G, Hyland NP. Breaking down the barriers: the gut microbiome, intestinal permeability and stress-related psychiatric disorders. Front Cell Neurosci. 2015;9. doi:10.3389/fncel.2015.00392.

26. Thaiss CA, Zeevi D, Levy M, Zilberman-Schapira G, Suez J, Tengeler AC, et al. Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. Cell. 2014;159:514–29.

27. Tamburini S, Shen N, Wu HC, Clemente JC. The microbiome in early life: Implications for health outcomes. Nature Medicine. 2016;22:713–22.

28. Bokulich NA, Chung J, Battaglia T, Henderson N, Jay M, Li H, et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. Sci Transl Med. 2016;8:343ra82. doi:10.1126/scitranslmed.aad7121.

29. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science (80-). 2016;352:565–9.

30. Engen PA, Green SJ, Voigt RM, Forsyth CB, Keshavarzian A. The Gastrointestinal Microbiome: Alcohol Effects on the Composition of Intestinal Microbiota. Alcohol Res. 2015;37:223–36. http://www.ncbi.nlm.nih.gov/pubmed/26695747. Accessed 19 May 2018.

31. Sanjabi S, Flavell RA. Overcoming the hurdles in using mouse genetic models that block TGF-B signaling. J Immunol Methods. 2010;353:111–4. doi:10.1016/j.jim.2009.12.008.

32. Kiesler P, Fuss IJ, Strober W. Experimental Models of Inflammatory Bowel Diseases. Cell Mol Gastroenterol Hepatol. 2015;1:154–70. doi:10.1016/J.JCMGH.2015.01.006.

33. Monteleone G, Kumberova A, Croft NM, McKenzie C, Steer HW, MacDonald TT. Blocking Smad7 restores TGF-beta1 signaling in chronic inflammatory bowel disease. J Clin Invest. 2001;108:601–9.

34. Monteleone G, Neurath MF, Ardizzone S, Di Sabatino A, Fantini MC, Castiglione F, et al. Mongersen, an oral SMAD7 antisense oligonucleotide, and Crohn's disease. N Engl J Med. 2015;372:1104–13. doi:10.1056/NEJMoa1407250.

35. Sharpton T, Lyalina S, Luong J, Pham J, Deal EM, Armour C, et al. Development of inflammatory bowel disease is linked to a longitudinal restructuring of the gut metagenome in mice. mSystems. 2017;2.

36. Nayfach S, Bradley PH, Wyman SK, Laurent TJ, Williams A, Eisen JA, et al. Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. PLoS

Comput Biol. 2015;11.

37. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 2016;26:1612–25. doi:10.1101/gr.201863.115.

38. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008;36 SUPPL. 1.

39. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol. 2009;24:127–35. doi:10.1016/j.tree.2008.10.008.

40. Nayfach S, Pollard KS. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biol. 2015;16:51. doi:10.1186/s13059-015-0611-7.

41. Bates D. Linear mixed model implementation in lme4. 2012.

42. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J. 2017;9:378–400.

43. Bolker B, Skaug H, Magnusson A, Nielsen A. Getting started with the glmmADMB package. R Packag ver 20–8. 2012;:12.

44. R Foundation for Statistical Computing. R: A Language and Environment for

Statistical Computing. In: R Foundation for Statistical Computing. 2016. doi:10.1007/978-3-540-74686-7.

45. El-Shaarawi AH, Zhu R, Joe H. Modelling species abundance using the Poisson-Tweedie family. Environmetrics. 2011;22:152–64.

46. Zhang Y. Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models. Stat Comput. 2013;23:743–57.

47. Vuong QH. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica. 1989;57:307. doi:10.2307/1912557.

48. Hejblum BP, Skinner J, Thiébaut R. Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. PLoS Comput Biol. 2015;11:e1004310. doi:10.1371/journal.pcbi.1004310.

49. Wu S, Wu H. More powerful significant testing for time course gene expression data using functional principal component analysis approaches. BMC Bioinformatics. 2013;14:6. doi:10.1186/1471-2105-14-6.

50. Xiao L, Zipunnikov V, Ruppert D, Crainiceanu C. Fast covariance estimation for high-dimensional functional data. Stat Comput. 2016;26:409–21.

51. Rizzo ML, Székely GJ. DISCO analysis: A nonparametric extension of analysis of variance. Ann Appl Stat. 2010;4:1034–55.

52. Ye Y, Doak TG. A Parsimony Approach to Biological Pathway

Reconstruction/Inference for Metagenomes. In: Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches. 2011. p. 453–60.

53. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995;57:289–300. doi:10.2307/2346101.

54. Wilson WG, Lundberg P, Vázquez DP, Shurin JB, Smith MD, Langford W, et al. Biodiversity and species interactions: Extending Lotka-Volterra community theory. Ecol Lett. 2003;6:944–52.

55. Samuelson PA. Generalized Predator-Prey Oscillations in Ecological and Economic Equilibrium. Proc Natl Acad Sci. 1971;68:980–3. doi:10.1073/pnas.68.5.980.

56. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, et al. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLoS Comput Biol. 2013;9:e1003388. doi:10.1371/journal.pcbi.1003388.

57. Buffie CG, Bucci V, Stein RR, McKenney PT, Ling L, Gobourne A, et al. Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. Nature. 2014. doi:10.1038/nature13828.

58. Chung M, Krueger J, Pop M. Robust Parameter Estimation for Biological Systems: A Study on the Dynamics of Microbial Communities. ArXiv. 2015;:1–33. http://arxiv.org/abs/1509.06926. Accessed 5 Jul 2016.

59. Honaker J, King G, Blackwell M. Amelia II: A Program for Missing Data. J Stat Softw.

2011;45:1–47. doi:10.18637/jss.v045.i07.

60. Soetaert K, Petzoldt T. Inverse Modelling , Sensitivity and Monte Carlo Analysis in R Using Package FME. J Stat Softw. 2010;33:1–28.

61. Martyn A, Best N, Cowles K, Vines K, Bates D, Almond R, et al. Package R ' coda ' correlation. R news. 2016;6:7–11. https://cran.r-project.org/web/packages/coda/coda.pdf.

62. Momeni B, Xie L, Shou W. Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions. Elife. 2017;6:e25051. doi:10.7554/eLife.25051.

63. Faber F, Bäumler AJ. The impact of intestinal inflammation on the nutritional environment of the gut microbiota. Immunol Lett. 2014;162:48–53. doi:10.1016/J.IMLET.2014.04.014.

64. Winter SE, Winter MG, Xavier MN, Thiennimitr P, Poon V, Keestra AM, et al. Host-derived nitrate boosts growth of E. coli in the inflamed gut. Science. 2013;339:708–11. doi:10.1126/science.1232467.

65. Galan JE, Wolf-watz H. Protein delivery into eukaryotic cells by type III secretion machines. Nature. 2006;444:567–73. doi:10.1038/nature05272.

66. Salomon D, Orth K. Type VI secretion system. Curr Biol. 2015;25:R265-6. doi:10.1016/j.cub.2015.02.031.

67. Hachani A, Wood TE, Filloux A. Type VI secretion and anti-host effectors. Current

Opinion in Microbiology. 2016;29:81–93.

68. Cummings JH, Macfarlane GT, Hang HC, Bertozzi CR, Green HC, Fisher JC, et al. Mucin glycan foraging in the human gut microbiome. Mol Cell Proteomics . 2017;82:141–8. doi:10.1016/0003-9861(76)90329-5.

69. McGuckin MA, Lindén SK, Sutton P, Florin TH. Mucin dynamics and enteric pathogens. Nat Rev Microbiol. 2011;9:265–78.

70. Stojiljkovic I, Perkins-Balding D. Processing of Heme and Heme-Containing Proteins by Bacteria. DNA Cell Biol. 2002;21:281–95. doi:10.1089/104454902753759708.

71. Nobles CL, Maresso AW. The theft of host heme by Gram-positive pathogenic bacteria. Metallomics. 2011;3:788–96. doi:10.1039/c1mt00047k.

72. Rouault TA. Pathogenic bacteria prefer heme. Science. 2004;305:1577–8.

73. Anzaldi LL, Skaar EP. Overcoming the heme paradox: heme toxicity and tolerance in bacterial pathogens. Infect Immun. 2010;78:4977–89. doi:10.1128/IAI.00613-10.

74. Radogna F, Ghibelli L. Melatonin: A pleiotropic molecule regulating inflammation. Biochem Pharmacol. 2010;80:1844–52.

75. Carrillo-Vico A, Lardone PJ, Alvarez-Sánchez N, Rodríguez-Rodríguez A, Guerrero JM. Melatonin: buffering the immune system. Int J Mol Sci. 2013;14:8638–83. doi:10.3390/ijms14048638.

76. Vital M, Howe A, Tiedje J. Revealing the Bacterial Synthesis Pathways by Analyzing

(Meta) Genomic Data. MBio. 2014;5:1–11. doi:10.1128/mBio.00889-14.Editor.

77. Hamer HM, Jonkers D, Venema K, Vanhoutvin S, Troost FJ, Brummer RJ. Review article: The role of butyrate on colonic function. Alimentary Pharmacology and Therapeutics. 2008;27:104–19.

78. Alexander C, Rietschel ET. Invited review: Bacterial lipopolysaccharides and innate immunity. J Endotoxin Res. 2001;7:167–202. doi:10.1177/09680519010070030101.

79. Vázquez M, Santana O, Quinto C. The NodI and NodJ proteins from Rhizobium and Bradyrhizobium strains are similar to capsular polysaccharide secretion proteins from Gram-negative bacteria. Mol Microbiol. 1993;8:369–77.

80. Bloom SM, Bijanki VN, Nava GM, Sun L, Malvin NP, Donermeyer DL, et al. Commensal Bacteroides species induce colitis in host-genotype-specific fashion in a mouse model of inflammatory bowel disease. Cell Host Microbe. 2011;9:390–403.

81. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High-throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. bioRxiv. 2017;:225342. doi:https://doi.org/10.1101/225342.

82. Hartmann K, Steel M, Faith D. Maximizing Phylogenetic Diversity in Biodiversity Conservation: Greedy Solutions to the Noah's Ark Problem. Syst Biol. 2006;55:644–51. doi:10.1080/10635150600873876.

83. Moulton V, Spillner A. Phylogenetic diversity and the maximum coverage problem. Appl Math Lett. 2009;22:1496–9. doi:10.1016/J.AML.2009.03.017.

84. Roy J., Mitchell SA, Duquesne C-M. PuLP. https://github.com/coin-or/pulp.

85. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012;19:455–77. doi:10.1089/cmb.2012.0021.

86. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from. Cold Spring Harb Lab Press Method. 2015;1:1–31. doi:10.1101/gr.186072.114.

87. Chaumeil P-A, Hugenholtz P, Parks DH. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. 2018. https://github.com/Ecogenomics/GtdbTk.

88. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A proposal for a standardized bacterial taxonomy based on genome phylogeny. bioRxiv. 2018;:256800. doi:10.1101/256800.

89. Parks DH. GenomeTreeTk. 2018. https://github.com/dparks1134/GenomeTreeTk.

90. Foster ZSL, Sharpton TJ, Grünwald NJ. Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. PLoS Comput Biol. 2017;13.

91. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11.

92. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-

generation sequencing data. Bioinformatics. 2012;28:3150–2.

93. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44:D286–93.

94. Galperin MY, Makarova KS, Wolf YI, Koonin E V. Expanded Microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015;43:D261–9.

95. Qin J, Li R, Raes J, Arumugam M, Burgdorf S, Manichanh C, et al. A human gut microbial gene catalog established by metagenomic sequencing. Nature. 2010;464:59–65.

96. Xiao L, Feng Q, Liang S, Sonne SB, Xia Z, Qiu X, et al. A catalog of the mouse gut metagenome. Nat Biotechnol. 2015;33:1103–8. doi:10.1038/nbt.3353.

97. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science (80-). 2015;348.

98. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45:W36–41.

99. Zhang Q, Ye Y. Not all predicted CRISPR-Cas systems are equal: Isolated cas genes and classes of CRISPR like elements. BMC Bioinformatics. 2017;18:92. doi:10.1186/s12859-017-1512-4.

100. Crawley AB, Henriksen JR, Barrangou R. CRISPRdisco: An Automated Pipeline for the Discovery and Analysis of CRISPR-Cas Systems. Cris J. 2018;1:171–81. doi:10.1089/crispr.2017.0022.

101. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8:28–36.

102. Titus Brown C, Irber L. sourmash: a library for MinHash sketching of DNA. J Open Source Softw. 2016;1:27. doi:10.21105/joss.00027.

103. Rosshart SP, Vassallo BG, Angeletti D, Hutchinson DS, Morgan AP, Takeda K, et al. Wild Mouse Gut Microbiota Promotes Host Fitness and Improves Disease Resistance. Cell. 2017;171:1015–1028.e13. doi:10.1016/j.cell.2017.09.016.

104. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016;2.

105. Human T, Project M. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14. doi:10.1038/nature11234.

106. Starke RM, McCarthy DJ, Komotar RJ, Connolly ES. Gut Microbiome and Endothelial TLR4 Activation Provoke Cerebral Cavernous Malformations. Neurosurgery. 2017;81:N44–6. doi:10.1093/neuros/nyx450.

107. Krych Ł, Nielsen D, Hansen A, Hansen C. Gut microbial markers are associated with

diabetes onset, regulatory imbalance, and IFN-γ level in NOD Mice. Gut Microbes. 2015;6:101–9. doi:10.1080/19490976.2015.1011876.

108. Harach T, Marungruang N, Duthilleul N, Cheatham V, Mc Coy KD, Frisoni G, et al. Reduction of Abeta amyloid pathology in APPPS1 transgenic mice in the absence of gut microbiota. Sci Rep. 2017;7:41802. doi:10.1038/srep41802.

109. Ormerod KL, Wood DLA, Lachner N, Gellatly SL, Daly JN, Parsons JD, et al. Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. Microbiome. 2016;4:36. doi:10.1186/s40168-016-0181-2.

110. Schöner TA, Kresovic D, Bode HB. Biosynthesis and function of bacterial dialkylresorcinol compounds. Applied Microbiology and Biotechnology. 2015;99:8323–8.

111. Forbes E, Murase T, Yang M, Matthaei KI, Lee JJ, Lee NA, et al. Immunopathogenesis of experimental ulcerative colitis is mediated by eosinophil peroxidase. J Immunol. 2004;172:5664–75. doi:10.4049/JIMMUNOL.172.9.5664.

112. Nasko DJ, Koren S, Phillippy AM, Treangen TJ. RefSeq database growth influences the accuracy of k-mer-based species identification. bioRxiv. 2018;:304972. doi:10.1101/304972.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature ____________ Date 06/04/18