Title
Novel Applications of Machine Learning and Statistics for Genome-resolved Metagenomic Data

Author
Rahman, Sumayah

Publication Date
2019

Peer reviewed|Thesis/dissertation

Novel Applications of Machine Learning and Statistics for Genome-resolved Metagenomic Data

by

Sumayah F Rahman

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Microbiology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jillian Banfield, Chair
Professor Rodrigo Almeida
Professor Lexin Li

Spring 2019

Abstract

Novel Applications of Machine Learning and Statistics
for Genome-resolved Metagenomic Data

by

Sumayah F Rahman

Doctor of Philosophy in Microbiology

Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Jillian Banfield, Chair

By sequencing environmental DNA and reconstructing microbial genomes, we can obtain insight into the previously hidden microbial world. This approach, known as genome-resolved metagenomics, has been utilized to study microorganisms in a variety of environments. Small sample sizes were common in genome-resolved metagenomics studies of the past, and thus few statistical methods of analysis were applied to the data resulting from these small-*n* studies. Instead, the analyses were focused on other aspects that did not require statistical methods, such as the identification of metabolic pathways possessed by the genomes and the phylogenetic relationships between organisms. However, in recent years, decreased sequencing costs and greater availability of computational resources have enabled scientists to sequence and process hundreds of samples for a single study. This dissertation demonstrates the application of several statistical and machine learning methods for the interpretation and strategic analysis of data from high-throughput genome-resolved metagenomic studies. Through the combination of new methods with previously existing methods, this work illustrates potential benefits that quantitative methods of analysis can offer to the field of genome-resolved metagenomics.

The first chapter of this dissertation serves as an example of a traditional genome-resolved metagenomics study, using primarily manual methods of analysis after the main steps of the data processing pipeline (including assembly, binning, and annotation) are complete. The manual methods of analysis applied in this small-scale study enable us to understand what microbes are present in a particular bioreactor community, and what metabolic functions these microbes are capable of. This contrasts with the much more data-intensive studies in the latter chapters, in which manual analyses would not be an efficient use of the data.

The second and third chapters, which are both focused on very large-scale data from the premature infant gut microbiome, illustrate the use of statistical methods for deciphering relationships in complex systems. This includes machine learning techniques applied to metagenome-associated genomes to make predictions that may potentially be useful in determining optimal care for a patient, as well as more basic statistical methods that allow us to better understand the gut microbiome and how it is influenced by external factors.

The fourth chapter is focused on the development of a new method that takes the hierarchical structure of genome-resolved metagenomic data into account. With genes in pathways, pathways in genomes, and genomes in communities of microorganisms, traditional ways of comparing samples fail to fully elucidate the biological systems because not all levels of the hierarchy are accounted for. To address this problem, the new concept described here allows for the inclusion of both functional and phylogenetic data to best utilize the wide breadth of information available in genome-resolved metagenomic data. The combination of quantitative approaches with genome-resolved metagenomics may lead to a more robust understanding of microbial communities.

Table of Contents

Acknowledgements

First and foremost, I would like to thank my advisor, Professor Jill Banfield. It has been an honor to be a part of her lab and do research under her guidance. I also deeply enjoyed the opportunity to teach her metagenomics lab class and microbial ecology class as a graduate student instructor. Jill's support in my various pursuits has been vital to my career thus far as a graduate student and will surely have an enduring effect on my future. For that, I am very grateful. I would also like to thank Professor Rodrigo Almeida and Professor Lexin Li for serving on my dissertation committee.

The members of the Banfield lab, both current and alumni, gave me so much help and advice as I dove into the world of metagenomics. I am especially grateful to Rose Kantor, who mentored me during my rotation. I also thank Matt Olm, for everything that he did for the human microbiome component of the Banfield lab, and Alex Thomas for time spent working together on new methods to analyze metagenomic data.

My projects have collaborators around the world. At the University of Cape Town in South Africa, I would like to thank Robert Huddy, Andries Wynand van Zyl, and Susan Harrison. At the University of Pittsburgh, so much thanks goes to Michael Morowitz, Robyn Baker, and Brian Firek. I would also like to thank the infants and the parents who enrolled their infants in our studies of the human microbiome, for their generous contributions to science.

Lastly, I thank my family and friends for the never-ending love and support.

Chapter 1

Genome-resolved metagenomics of a bioremediation system for degradation of thiocyanate in mine water containing suspended solid tailings

Sumayah F. Rahman[1], Rose S. Kantor[1], Robert Huddy[2], Brian C. Thomas[3], Andries Wynand van Zyl[2], Susan T.L. Harrison[2] and Jillian F. Banfield[3,4]

[1]Department of Plant and Microbial Biology, University of California, Berkeley, California, USA
[2]Center for Bioprocess Engineering Research, Department of Chemical Engineering, University of Cape Town, Cape Town, South Africa
[3]Department of Earth and Planetary Sciences, University of California, Berkeley, California, USA
[4]Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, USA

Abstract

Thiocyanate (SCN^-) is a toxic compound that forms when cyanide (CN^-), used to recover gold, reacts with sulfur species. SCN^--degrading microbial communities have been studied using bioreactors fed synthetic wastewater. The inclusion of suspended solids in the form of mineral tailings, during the development of the acclimatized microbial consortium, led to the selection of an active planktonic microbial community. Preliminary analysis of the community composition revealed reduced microbial diversity relative to the laboratory-based reactors operated without suspended solids. Despite minor upsets during the acclimation period, the SCN^- degradation performance was largely unchanged under stable operating conditions. Here we characterized the microbial community in the SCN^- degrading bioreactor that included solid particulate tailings and determined how it differed from the biofilm-based communities in solids-free reactor systems inoculated from the same source. Genome-based analysis revealed that the presence of solids decreased microbial diversity, selected for different strains, suppressed growth of thiobacilli inferred to be primarily responsible for SCN^- degradation, and promoted growth of *Trupera*, an organism not detected in the reactors without solids. In the solids reactor community, heterotrophy and aerobic respiration represent the dominant metabolisms. Many organisms have genes for denitrification and sulfur oxidation, but only one *Thiobacillus* sp. in the solids reactor has SCN^- degradation genes. The presence of the solids prevented floc and biofilm formation, leading to the observed reduced microbial diversity. Collectively the presence of the solids and lack of biofilm community may result in a process with reduced resilience to process perturbations, including fluctuations in the influent composition and pH. The results from this investigation have provided novel insights into the community composition of this industrially-relevant community, giving potential for improved process control and operation through ongoing process monitoring.

Introduction

Cyanide (CN^-) is used globally in the gold mining industry as a lixiviant to dissolve and remove gold from ore. During gold extraction by cyanidation, CN^- can react with reduced sulfur species, forming thiocyanate (SCN^-) in the gold mining effluents. Although SCN^- is not as toxic as CN^-, it is known to be harmful to humans and aquatic organisms (Boening & Chew, 1999; Erdogan, 2003; Shifrin, Beck, Gauthier, Chapnick, & Goodman, 1996), requiring the use of chemical or biological methods for its removal. The use of microbes for biological remediation of

SCN^- from contaminated wastewater has been successful at both the laboratory scale (Boucabeille, Bories, Ollivier, & Michel, 1994; Du Plessis, Barnard, Muhlbauer, & Naldrett, 2001; Van Zyl, Huddy, Harrison, & Van Hille, 2014; Zyl, Harrison, & Hille, 2011) and in commercial operations (van Buuren, Makhotla, & Olivier, 2011). Engineers have developed and commercialized a SCN^- biodegradation process known as Activated Sludge Tailings Effluent Remediation (ASTER™) that involves continuous feeding of SCN^--containing solutions into aerated bioreactors to promote microbial degradation (van Buuren et al., 2011).

Two metabolic pathways have been proposed for the biological degradation of SCN^-. In one pathway, thiocyanate hydrolase converts SCN^- to sulfide and cyanate (OCN^-). OCN^- is further hydrolyzed to carbon dioxide and ammonium, while sulfide is oxidized to sulfate. In the other degradation pathway, SCN^- is hydrolyzed into carbonyl sulfide (OCS) and ammonium. OCS can be broken down into carbon monoxide and sulfide, which is then oxidized to sulfate (Katayama et al., 1998, 1992).

To identify the microorganisms responsible for SCN^- degradation, microbial communities in experimental reactors have been characterized by molecular fingerprinting (Felföldi et al., 2010; Huddy, Van Zyl, Van Hille, & Harrison, 2015; Quan et al., 2006) and genome-resolved metagenomic analysis (Kantor et al., 2017, 2015). Analysis of the 16S and 18S rRNA in a reactor established with an ASTER™ consortium revealed that the microbial community was much more diverse than previously expected (Huddy et al., 2015). Metagenomic analysis of the same system predicted the metabolic potential of the key organisms (e.g., *Thiobacillus* spp.) and described the potential flow of carbon, sulfur, and nitrogen through the community (Kantor et al., 2015).

In the laboratory-based SCN^--degrading system described by previous studies, SCN^--containing synthetic wastewater was fed to the laboratory reactors and, where the SCN^- feed concentration was sufficiently high, thick biofilms formed on all reactor surfaces. Biofilm improves SCN^- degradation rates, in part by ensuring biomass retention during continuous flow mode and by enhancing process robustness for dynamic waste streams (Huddy et al., 2015). Typically, the ASTER™ process is not performed in the presence of particulate tailings (i.e., mineral particles left behind after separating the gold from ore concentrate). However, at some mining sites, the removal of solid tailings from the effluent is not achieved fully due to site topography, particle size, density of the tailings and other factors (Van Zyl et al., 2014). In a bioreactor inoculated with the microbial consortium of the SCN^- stock reactor (Kantor et al., 2015), van Zyl et al. (2014) acclimatized the microbial community to an incrementally increasing loading of solids of density 2.7 g/l to a final concentration of 5.5% mass/volume, and showed that, following acclimatization, SCN^- degradation still occurred. However, biofilm did not form on the submerged surfaces of the reactor. Following an extended period of continuous operation, this solids-containing reactor was operated in 'draw and fill' mode, meaning that fluid was removed periodically and the volume replaced with untreated fluid.

This study was motivated by the use of the acclimatized microbial culture, as developed by van Zyl et al. (2014), as the inoculum for an ASTER™ process to treat the effluent from a bioleaching operation exploiting a refractory gold deposit in the Philippines. The aim of the research was to resolve the microbial community associated with an active ASTER™ solids reactor system and to compare that with previously resolved (Kantor et al., 2017, 2015) ASTER™ microbial communities. In this study, we used genome-resolved metagenomics to elucidate the microbial community composition and metabolic potential of the solids-containing SCN^- degradation bioreactor. We hypothesized that due to the lack of biofilm in the solids reactor (Van Zyl et al., 2014), there would be differences in community membership compared to the reactors without solids. Moreover, we hypothesized that given the lower SCN^- loading in this system, key SCN^- degrading organisms may be at lower relative abundances in this reactor compared to solids-

free reactors at higher loading rates. Here we report the composition and metabolic potential of the solids reactor microbial community.

Materials and Methods

Mineral solids

The mineral solids were generated by SGS (Johannesburg) and provided by Gold Fields, as described by van Zyl et al. (2014). The fine grained particulates had a D_{50} of 6.122 µm (D_{10} of 0.939 µm and D_{90} of 38.026 µm) and a density of 2.677 g/ml.

The ASTER™ culture

The mixed microbial consortium used to inoculate the reactors was derived from the stock ASTER™ culture, with prior characterization reported by Huddy et al., (2015) and Kantor et al. (2015). It was acclimatized to cultivation in the presence of suspended solids as described by van Zyl et al. (2014).

Reactor system

The work was conducted using a stirred tank reactor, with an operating volume of 1 L, as described by van Zyl et al. (2014). The microbial culture, acclimatized during the investigation by van Zyl et al. (2014), was maintained in a "draw-and-fill" culture with a 10% volume replacement by a feed solution, containing the solids (5.5% m/v), SCN^- (450 mg/L, as KSCN), molasses (150 mg/L) and phosphate (27 mg/L, as KH_2PO_4) on a weekly basis. The molasses was provided to support heterotrophic growth. The pH of the feed was initially adjusted using potassium hydroxide to maintain the reactor at approximately pH 7.0.

DNA extraction and sequencing

Two separate samples of approximately 15 mL were drawn from the well-mixed suspended solids reactor operated under the same conditions at an interval of 25 days. These samples were processed independently. The biomass was harvested by centrifugation (14,000 rpm for 10 min at 22°C). Total DNA was extracted using a NucleoSpin® soil genomic DNA extraction kit (Machery-Nagel, Germany) with the inclusion of a repeated extraction step, according to the manufacturer's instructions. Paired end library preparation and sequencing were performed with Illumina HiSeq 2500 run at the rapid mode at the Joint Genome Institute (Walnut Creek, CA). An insert size of 500 bp was used to yield 251 bp reads.

Read processing, assembly and initial functional annotation

For both datasets, reads were hard trimmed to 150 bp and processed by BBtools to remove Illumina adapters and trace contaminants. The reads were then trimmed for quality using Sickle with default settings (https://github.com/najoshi/sickle). The datasets were assembled independently using idba_ud with the pre-correction option, for normalization of highly represented kmers (Peng, Leung, Yiu, & Chin, 2012). Genes on scaffolds ≥1000 bp were predicted using Prodigal with the metagenome option (Hyatt, Locascio, Hauser, & Uberbacher, 2012). For annotation, similarity searches were performed using USEARCH, which compares sequences against the KEGG, UniRef100, and UniProt databases. KEGG and UniRef100 were searched in the forward and reverse direction to identify reciprocal best hits, while only forward searches were done for UniProt. The phylogenetic affiliation to the lowest possible taxonomic level was determined based on the best hit against the UniRef100 database. 16S rRNA genes were predicted

based on the ssu-align-0p1.1.cm database, and transfer RNA genes were predicted using tRNAscanSE (Lowe & Eddy, 1997).

Genome binning and dereplication

Genome bins were assigned based on coverage, GC content, and the phylogenetic best-hit profile of scaffolds ≥1000 bp using ggkbase binning tools (ggkbase.berkeley.edu). Emergent-self organizing maps (ESOMs) based on di- and tri-nucleotide frequencies and differential coverage across the samples were created for each of the two datasets (Dick et al., 2009), and the tentative bin information was superimposed onto the ESOMs as class files using the Databionic ESOM Tool, esomana (Ultsch and Moerchen, 2005).The bins were checked manually, and any mis-binned scaffolds were transferred to the correct bin. The bacterial genomes were curated to resolve assembly errors, extend scaffolds, and join scaffolds. Genome completeness for the bacterial bins was assessed based on the presence or absence of 51 bacterial single copy genes that are widely conserved. The genome bins from the solids 1 and solids 2 samples were aligned, and bins with >98% nucleotide identity across 50% of the genome were classified as the same genome. The winning genome was chosen based on genome completeness and included in the dereplicated solids dataset. To determine which organisms from the solids bioreactor have been found previously in thiocyanate bioreactors, genomes were clustered at >98% average nucleotide identity using the MinHash technique (Ondov et al., 2016). Read mapping for coverage calculation was performed using Bowtie2 with default settings (Langmead & Salzberg, 2012). If an organism occurred with coverage >1x, it was considered to be present in that sample.

Phylogenetic analysis based on ribosomal protein sequences

The genes for 16 ribosomal proteins (L2, L3, L4, L5, L6, L14, L16, L18, L22, L24, S3, S8, S10, S17 and S19) were collected from the solids 1 and solids 2 datasets, as well as the SCN-only (SCN^- loading rate of 1.9 mM h^{-1}; 12 h residence time; Kantor et al., 2015), CN-SCN (SCN^- and CN^- loading rate of 0.9 and 0.14 mM h^{-1} respectively at a 14 h residence time; Kantor et al., 2015), and SCN^- two-reactor time series datasets (SCN^- loading rate of 0.07 – 1.4 mM h^{-1} at a 12 h residence time and SCN^- feed concentrations from 50 to 1000 mg/l; Kantor et al., 2017), excluding those from bins labeled as eukaryotes, viruses, phage, plasmids, or mitochondria. These 16 genes, along with the 16 ribosomal protein genes from a custom reference set, were aligned independently with MUSCLE (Edgar, 2004). The alignments were trimmed to remove ambiguously aligned termini and columns composed of more than 95% gaps. The alignments were then concatenated to form an alignment with 2,454 columns, and taxa that had less than 50% of the alignment were removed. Due to incomplete sequences, some organisms from the datasets did not get incorporated into the final concatenated alignment. This alignment was used to construct a maximum likelihood phylogeny with RAxML using the PROTGAMMALG model (Stamatakis, 2014).

Metabolic analysis

Genome-specific metabolic potential was determined by (1) searching all predicted ORFs in a genome with Pfam30, TIGRfam31, Panther32 and custom HMM profiles of marker genes for specific pathways using hmmscan33 (Anantharaman et al., 2016) (2) assessment of metabolic pathways using annotations on ggKbase (ggkbase.berkeley.edu) (3) searching particular proteins of interest using BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990). For generation of custom HMM profiles, references for each marker gene were aligned using MUSCLE and the start and ends of the alignment were manually trimmed. The alignment was converted into Stockholm format and databases were built using hmmscan33. For RuBisCO and hydrogenases34, different hmm databases were constructed for each distinct group. Individual cutoffs for all HMMs were

determined by manual inspection. To compare genomes found in thiocyanate bioreactors with published genomes, the genomes of interest were downloaded from NCBI, and reciprocal BLASTs were utilized to identify shared and unique genes.

Results and Discussion

Genome recovery and community structure

Metagenomic sequencing for two samples from the SCN^- bioreactor with solids was obtained: the "solids 1" sample (4.6 Gbp of sequence) was taken one month prior to "solids 2" (5.1 Gbp of sequence). Raw read data for solids 1 and solids 2 can be accessed at NCBI with accession numbers SAMN05509838 and SAMN05509839. *De novo* assembly of the metagenomes resulted in a 178.9 Mbp assembly for solids 1 (with 51% of the assembly in contigs ≥ 5 kb) and a 213.2 Mbp assembly for solids 2 (with 77% in contigs ≥ 5 kb). Genome binning based on GC content, coverage, di-and tri-nucleotide frequencies, and differential coverage across the solids 1 and solids 2 samples yielded 34 bacterial genomes from solids 1 and 25 bacterial genomes from solids 2. The taxonomic compositions of the two samples were similar. However, in solids 1, we also reconstructed draft mitochondrial genomes for two protozoa and a partial genome of a yeast belonging to the Saccharomycetales.

Based on coverage data, Sphingobacteriales_2 was consistently the dominant organism and *Thiobacillus* spp. were present at only moderate abundance (Figure 1.1), unlike in SCN^- bioreactors without solids where thiobacilli were the most abundant community members (Kantor et al., 2015). Phylogenetic analysis revealed that the two thiobacilli in the solids reactor are closely related to, but distinct from, the strains reported from solids-free reactors with the same inoculum (Figure 1.2).

To identify overlapping genomes in solids 1 and solids 2, genomes with >98% nucleotide identity were clustered. This resulted in a dereplicated dataset of 40 bacterial genomes, available at http://ggkbase.berkeley.edu/scnpilot_solids_dereplicated/organisms. Out of the 34 distinct bacterial genomes in solids 1 that were abundant enough for at least partial genome-based analysis (>0.25% of the community), 19 were also sufficiently abundant for genome-based analysis in solids 2 (Figure 1.1). The relative abundances of some of these organisms did not change between time points (e.g., Truepera_1) whereas others decreased (e.g., Sphingobacteriales_2) or increased dramatically (e.g., Rhodanobacter_2) (Figure 1.1). However, based on an analysis involving stringent read mapping that enabled detection of organisms at relative abundance levels of ~0.06% of the community, 39 out of the 40 unique bacterial genomes were detected in both samples (Table 1.1).

Dominance of heterotrophy and aerobic metabolism

Metabolic analyses revealed that no genomes from the solids bioreactor possess the genes for the Wood–Ljungdahl pathway or the reverse TCA cycle. Four genomes carry genes for form I and/or form II RuBisCO (*rbc*), the key enzyme of the Calvin–Benson–Bassham cycle (Table 1.1). The lack of carbon fixation genes in 90% of genomes from this dataset indicates that the community was mainly composed of heterotrophs. In fact, only one of the five most abundant (and genomically well-defined) organisms in the solids reactor is an autotroph (Figure 1.1). Heterotrophs likely consume the molasses in the media as well as biomass and/or organic exudates or lysis products from autotrophs. In contrast, three of the five most abundant organisms in the solids-free bioreactor are autotrophs (Kantor et al., 2015).

To determine the oxygen requirements for organisms in the solids bioreactor, we searched the genome bins for the presence of cytochrome oxidase genes. The vast majority of genomes

contain at least one cytochrome oxidase, indicating the ability to use oxygen as a terminal electron acceptor. Just one near-complete genome, the predicted endosymbiont Cytophagia_1, lacks cytochrome oxidase (Table 1.1). The presence of most of the glycolysis pathway, in addition to pyruvate dehydrogenase, suggests that Cytophagia_1 may ferment pyruvate, possibly producing acetate as a metabolic byproduct. The dominance of aerobic organisms is not surprising, given that the solids reactor is well aerated and does not develop biofilm (Van Zyl et al., 2014), which would provide anaerobic and microaerobic environments (Falsetta, McEwan, Jennings, & Apicella, 2010; Fox et al., 2014).

Thiocyanate, nitrogen and sulfur compound metabolic pathways

Thiobacillus_2, the more abundant of the two identified *Thiobacillus* strains (Figure 1.1), possesses a thiocyanate hydrolase (*scnABC*) (Table 1.1), the enzyme involved in the degradation of SCN^- in *Thiobacillus thioparus* THI115 (Arakawa et al., 2007; Kataoka et al., 2006), and the genes for this enzyme are located in a conserved operon as previously described (Kantor et al., 2015). Both Thiobacillus_1 and Thiobacillus_2 possess genes involved in sulfur oxidation and denitrification (Table 1.1). Thus, it is clear that *Thiobacillus* spp. have important roles in the solids reactor, although they are not the dominant organisms (Figure 1.1) as they were in the SCN^- stock reactor (Kantor et al., 2015). The decrease in the proportion of *Thiobacillus* in the solids bioreactor relative to the reactors operated without solids and the reduced number of species with SCN^- degradation ability may explain the increased sensitivity of the SCN^- degradation to process perturbation and stress as reported by van Zyl et al. (2014).

SCN^- degradation results in the production of ammonium that could be converted to nitrite and removed by denitrification. Only one genome bin, Nitrosospira_1, contains genes for ammonium oxidation, *amo* and *hao* (Table 1.1), suggesting that this organism is critical for nitrite production in the system. We detected no genes for anaerobic ammonium oxidation in the dataset, as was the case in studies of solids-free reactors (Kantor et al., 2017, 2015). Six organisms in the solids reactor contain a full denitrification pathway (including *nar*, *nir*, *nor*, and *nos* genes) for the complete reduction of nitrate to N_2 (Table 1.1). Other genomes were missing one or more genes in the denitrification pathway, although this may be due to incomplete genome recovery. The dominant organism, Sphingobacteriales_2, is likely the main contributor to denitrification in the system (Figure 1.1).

An important step in the SCN^- breakdown pathway is sulfur oxidation. For the oxidation of sulfide, either SoxCD or rDsrAB is required. The gene for SoxC, which forms a complex with SoxD and works in conjunction with the other Sox enzymes, is present in four genomes (Table 1.1). Thiobacillus_1 contains *dsrAB,* which may function in the reverse dissimilatory sulfite reductase pathway that can oxidize sulfur to sulfite. We identified genes for APS reductase (*apr*) in Thiobacillus_2 and ATP sulfurylase (*atpS*) in Xanthomonadales_1; these may complete the oxidation by converting sulfite to sulfate. Other genes known to be involved in the oxidation of sulfur compounds, such as *fcc* and *sqr*, were found in several genomes in this dataset (Table 1.1). Overall, we conclude that based on its high abundance, Burkholderiales_1 is the most important organism involved in sulfur compound oxidation, although Rhizobiales_1 and Afipia_1 likely also contribute to these reactions (Figure 1.1).

An organism in the solids reactor not found in the solids-free reactors

A bacterium of the phylum Deinococcus-Thermus occurred in both the solids 1 and solids 2 samples (Figure 1.1). To our knowledge, this is the first reporting of a Deinococcus-Thermus in bioreactors inoculated with the ASTER^{TM} microbial consortium (Du Plessis et al., 2001; Huddy et al., 2015; Kantor et al., 2017, 2015; van Buuren et al., 2011; Van Zyl et al., 2014). We

reconstructed a draft Truepera_1 genome that is 1.22 Mbp in length with 90% of expected single copy genes (Table 1.1). In comparison, the published complete genome of *Truepera radiovictrix*, the only genome available from the *Truepera* genus, is 3.23 Mbp in length (Ivanova et al., 2011). The 16S rRNA gene of *T. radiovictrix* shares only 89% identity with the sequence from Truepera_1, so it is possible that the two organisms do not belong to the same Genus; however, *T. radiovictrix* is the nearest sequenced relative. Members of Deinococcus-Thermus are known to be highly resistant to environmental hazards; specifically, *T. radiovictrix* is resistant to ionizing radiation and can grow under extreme conditions such as high alkalinity (Albuquerque et al., 2005). Given that *T. radiovictrix* is an alkaliphile, it was surprising that Truepera_1 was not also detected in the solids-free reactor, which has a higher pH than the solids reactor (Huddy et al., 2015; Van Zyl et al., 2014).

We compared the newly reconstructed Truepera_1 genome to that of *T. radiovictrix*, as it is the closest reference available. Like the published *T. radiovictrix* strain RQ-24^T, Truepera_1 is predicted to be an aerobic heterotroph. Unlike the reference sequence, Truepera_1 has genes for the export of heavy metals. There were also several genes present in the published *Truepera* genome that are not in Truepera_1, although this may be due to the fact that the Truepera_1 genome is incomplete. These included genes for L-lactate dehydrogenase, which *T. radiovictrix* strain RQ-24^T utilizes when it switches to homolactic fermentation, and manganese catalase, an antioxidant defense metalloenzyme that may be involved in strain RQ-24^T's resistance to ionizing radiation.

Truepera_1 thrives in the well-aerated solids bioreactor as the fifth most abundant organism (Figure 1.1), where it most likely consumes the molasses in the reactor feed. As the genome harbors a copper-containing nitrite reductase NirK, Truepera_1 may play a role in the denitrification process within the bioreactor (Table 1.1). The differing conditions in the solids reactor, including the SCN^- loading rate, likely resulted in the enrichment of low-abundance organisms that were not detected previously, such as Truepera_1. A notable feature of the solids bioreactor is that the high agitation of solid tailings causes shear stress that prevents biofilm formation (Illing & Harrison, 1999; Van Zyl et al., 2014). Mechanisms for resistance to shear stress have been identified in other members of the Deinococcus-Thermus; e.g., the SlpA protein in *Deinococcus radiodurans* R1 maintains cell envelope integrity (Rothfuss, Lara, Schmid, & Lidstrom, 2006). One S-layer protein gene was found in the Truepera_1 genome, and its best hit in the NCBI Protein database is the S-layer protein of *D. radiodurans* R1. If Truepera_1 has capabilities similar to *D. radiodurans* that allow it to resist the shear stress brought about by the agitated solids, this may contribute to its proliferation in this bioreactor.

Viruses and eukaryotes

Viruses and phage were abundant in the solids reactor. Two eukaryotic viruses and twenty phage were binned from the dataset, with five of the phage occurring in both the solids 1 and solids 2 samples. Some phage genomes were found within the genome bins of specific bacteria based on co-abundance patterns, suggesting possible affiliations. These bacteria include Rhizobiales_1, Rhizobiales_2, Xanthomonadales_1, Burkholderiales_1, Afipia_1, Chryseobacterium_1, and Rhodanobacter_2. A virus was found in the eukaryotic genome bin Saccharomycetales_1. These findings may indicate that viruses and phage play important roles in carbon turnover in the bioreactor. Metagenomic analysis of the bioreactors without solids also suggested that predation by eukaryotes and phage affects community dynamics (Kantor et al., 2015).

The genome for the yeast Saccharomycetales_1 was 8.52 Mbp in length and appears to be around half-complete, given 690 complete single-copy Benchmarking Universal Single-Copy Orthologs (BUSCOs) out of 1438 total BUSCO groups searched (Simão, Waterhouse, Ioannidis,

Kriventseva, & Zdobnov, 2015). Other organisms belonging to the order Saccharomycetales have been previously found in this system. An analysis of 18S rRNA genes from the solids bioreactor revealed the presence of a yeast that is a close relative of *Candida palmioleophila* (Van Zyl et al., 2014), and the ASTER™ microbial consortium has been found to include *Candida humulis* (van Buuren et al., 2011). In a study that utilized light microscopy, yeast-like cells and other eukaryotes such as filamentous fungi were present in the biofilm of the SCN^- reactor without solids (Huddy et al., 2015). The presence of Saccharomycetales_1 in the solids bioreactor, which does not contain biofilm, indicates that yeasts can also occur in the liquid portion of the bioreactor.

Mitochondrial genomes for two protozoa were identified in the solids 1 dataset. Mitochondria_Protozoa_1 was classified as *Acanthamoeba castellanii*, a unicellular amoeba that frequently captures prey by phagocytosis and harbors bacterial endosymbionts (Khan, 2001). In the solids reactor, Protozoa_1 likely carried the bacterial symbiont Cytophagia_1 based on co-abundance patterns. Mitochondria_Protozoa_2, which was also observed in one of the solids-free bioreactors (Figure 1.2), was classified as a Schizopyrenida. The majority of the contigs within the Mitochondria_Protozoa_2 genome bin corresponded to *Naegleria*, which are organisms known for their ability to transform from an amoeba to a flagellate (Marciano-Cabral, 1988).

An interesting phenomenon is the lack of rotifers in the solids reactor, which have been identified and observed in bioreactors without solids (Kantor et al., 2017). The rotifers prefer the planktonic portion of these reactors and feed on the edges of the biofilm. The shear stress caused by the highly agitated tailings material in the solids bioreactor may hinder the survival of these pseudocoelomate animals. Additionally, the slightly acidic conditions in the solids reactor (Van Zyl et al., 2014) may not be ideal for these planktonic rotifers (Berziņš & Pejler, 1987).

Conclusions

Bacteria from seven phyla were detected in the solids bioreactor (present at > 0.06% of the community) (Table 1.1). In contrast, a sample from the SCN^- stock reactor, sequenced to approximately the same depth, contained bacteria from nine different bacterial phyla (Kantor et al., 2015). The draw and fill mode of operation of the solids reactor at the time of analysis should have favored retention of slow-growing cells relative to the continuous flow mode of operation of the solids-free reactor, which could have led to increased diversity in the solids reactor. However, this effect would have been countered by biofilm formation in the reactor without solids, which likely prevented washout of slow-growing species, potentially increasing diversity in the solids-free reactor. The bacteria detected represent only a subset of all organisms present in the SCN^- stock reactor, given that organisms from 17 bacterial phyla have been detected across three experiments that were inoculated from that source (SCN^- stock reactor + the CN-SCN^- reactor + SCN^- two-reactor time series; Kantor et al., 2015; Kantor et al., 2017). The finding of lower bacterial diversity in the solids reactor compared with the solids-free SCN-stock reactor expands on results of a previous study that used a 16S rRNA gene clone library (30 sequences) to suggest lowering of diversity in reactors operated with solids and in continuous culture (Van Zyl et al., 2014).

Differing conditions in the solids bioreactor, including the mode of operation, SCN^- loading rate, absence of biofilm, increased shear stress, and lower pH (Van Zyl et al., 2014), likely affected the community composition. Truepera_1, an organism not previously detected in other reactors, was relatively abundant in the solids reactor, and may have been selected for due to its ability to withstand shear stress. The other bacteria enriched in the solids reactor were different species or different strains of species present in the other reactors derived from the same inoculum. The solids reactor exhibited lower diversity than any solids-free reactor at the strain as well as

phylum level. Performance of the solids reactor over an extended duration achieved a sustained SCN^- degradation rate of 56 mg/l/h and was similar to the biofilm-based communities in the solids-free reactors, in terms of SCN^- degradation rates achieved relative to SCN^- loading. The solids reactor also exhibited short periods of compromised degradation in response to perturbation of preferred operating conditions over the experimental period (Van Zyl et al., 2014). This may have been a consequence of decreased species diversity in the solids reactor relative to the solids-free reactors and the presence of only one organism capable of SCN^- degradation.

Organisms that can respire aerobically dominated the solids reactor community (Table 1.1), as was expected since the reactor was well aerated and biofilm was not present (Van Zyl et al., 2014). A moderately abundant *Thiobacillus* in the solids reactor possessed the genes for SCN^- degradation (Figure 1.1), whereas in the solids-free reactors, SCN^--degrading thiobacilli were the dominant organisms (Kantor et al., 2015; Kantor et al., 2017). The comparatively lower relative abundance of *Thiobacillus* explains the reduced resilience of the solids reactor system to perturbation in terms of SCN^- degradation reported by van Zyl et al. (2014). Despite the differences between this reactor and the solids-free reactors, several organisms in the solids bioreactor harbored genes for denitrification and sulfur oxidation (Table 1.1), key steps in the remediation of thiocyanate from wastewater.

We reconstructed genomes for 40 bacteria present in the solids reactor but only six of these were genomically sampled from bioreactors operated without solids (Figure 1.2). Thus this genome-resolved metagenomic analysis of the solids reactor expanded knowledge regarding organisms present in ASTER™ microbial consortium and increased available information about their metabolic potential.

Acknowledgements

Funding was provided by the NSF Sustainable Chemistry grant (1349278) and by the Department of Science and Technology (DST) and National Research Foundation of South Africa through the SARChI Chair in Bioprocess Engineering (UID 64778). The Joint Genome Institute's Emerging Technologies Opportunity Program (ETOP) grant supported sequencing, and we would particularly like to thank Susannah Tringe from the Joint Genome Institute. We also gratefully acknowledge Karthik Anantharaman, David Burstein, Christopher Brown, Alexander Probst, and Patrick West for their assistance.

Figure 1.1. Metabolic potential of the 19 organisms present at a high enough abundance in both the solids 1 and solids 2 samples to allow for genome-based analysis. The number of raw reads for each sample was used to normalize the coverage data, in order to accurately compare the two samples.

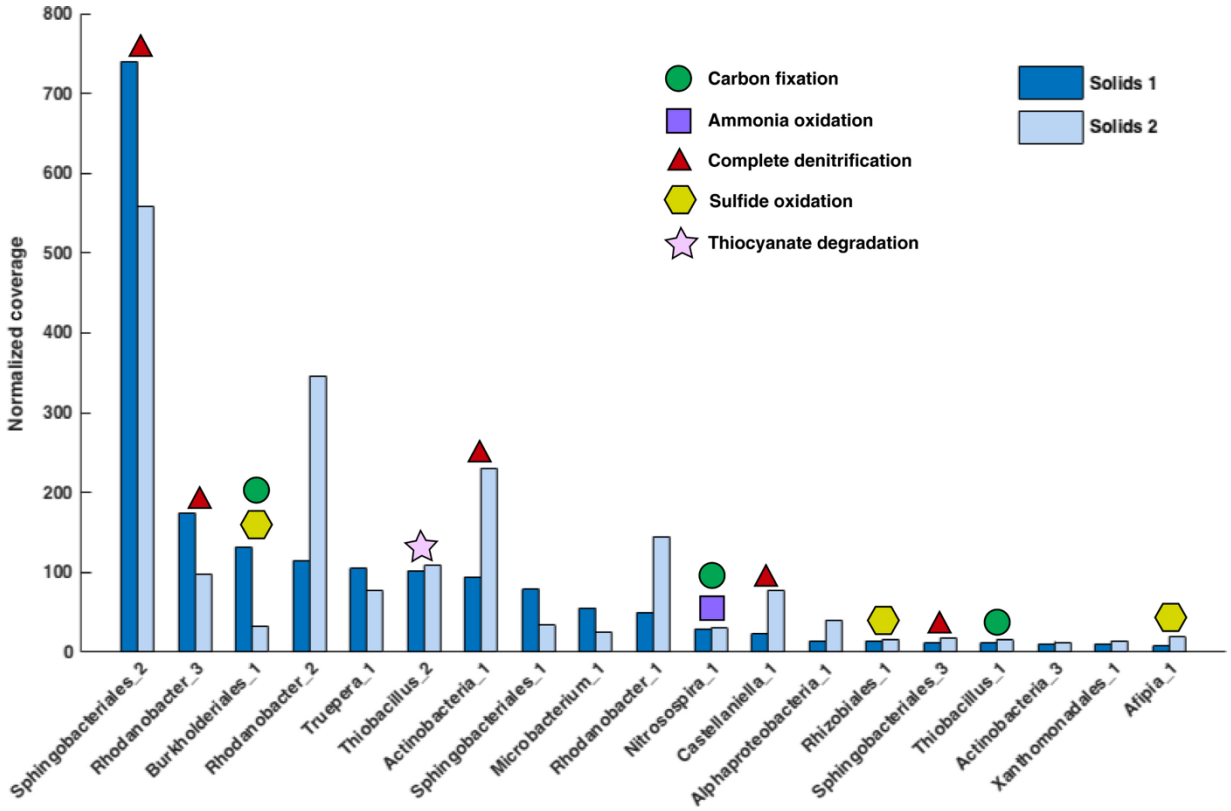

Figure 1.2. Illustration of the overlaps among reactor communities. To identify overlapping genomes, representative parts of the genome bins were aligned and clustered based on >98% average nucleotide identity.

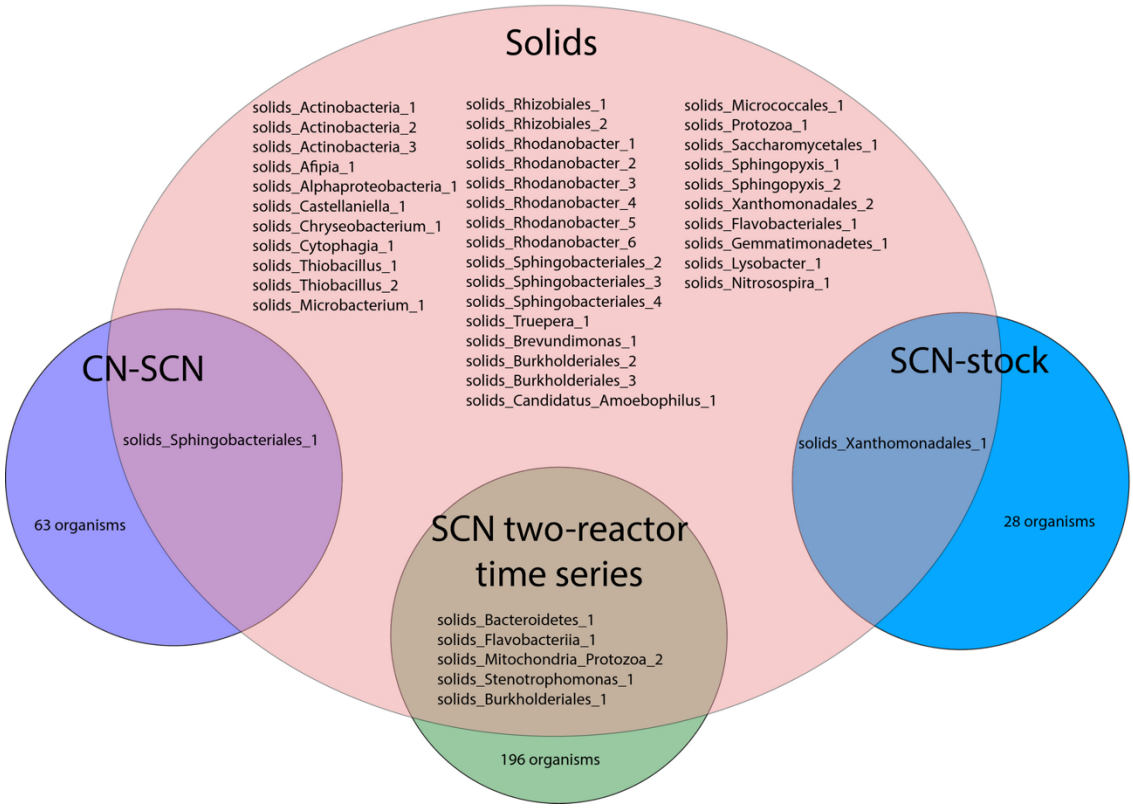

Table 1.1. Genome statistics and metabolic potential for the 40 bacterial genome bins in the solids bioreactor dataset (http://ggkbase.berkeley.edu/scnpilot_solids_dereplicated/organisms), ordered by normalized coverage in the solids 1 sample. The numbers represent the count of each indicated gene.

									Carbon fixation		Aerobic respiration										SCN hydrolase	Ammonia oxidation		NO_3 reduction		NO_2 reduction		NO reduction		N_2O reduction		S^{2-} oxidation			H_2S oxidation			SO_3^{2-} oxidation	
Genome Bin	**Phylum/Class**	**Solids 1 coverage**	**Solids 2 coverage**	**Bin length (Mbp)**	**GC%**	**Number of scaffolds**	**Longest scaffold (bp)**	**Single copy genes (total: 51)**	*rbc* - Form I	*rbc* - Form II	*coxA*	*coxB*	*ccoN*	*ccoO*	*ccoP*	*cyoA*	*cyoD*	*cyoE*	*cydA*	*cydB*	*scnABC*	*amoC*	*hao*	*narG*	*narH*	*nirK*	*nirS*	*norB*	*norC*	*nosD*	*nosZ*	*soxC*	*dsrA*	*dsrB*	*fcc*	*sqr*	*apr*	*atpS*	
solids_Sphingobacteriales_2	Bacteroidetes	739	559	3.560973	40.11	254	71409	49					1	1				1						1	1	2		1	2	1	1								
solids_Rhodanobacter_3	Gammaproteobacteria	174	98	3.459585	68.74	123	139761	48			1	1	1		1			1	1	1				1	1	2		4	1	1					1				
solids_Burkholderiales_1	Betaproteobacteria	132	33	3.556883	65.94	27	790171	48	1		1	1	2	1	1			1					1			1	1	1			2			3				
solids_Rhodanobacter_2	Gammaproteobacteria	115	346	2.95749	68	332	48323	27			1	1	3	1	1			1	1	1						2		4	1									
solids_Truepera_1	Deinococcus-Thermus	105	77	1.220355	66.58	327	45361	46			1	1						1								1												
solids_Thiobacillus_2	Betaproteobacteria	101	108	0.671397	62.87	176	25668	39													3			1						1	1				1	2	1	
solids_Actinobacteria_1	Actinobacteria	93	230	3.443405	68.93	71	208764	50			2							1						1	1											1		
solids_Burkholderiales_2	Betaproteobacteria	82	4.9	2.272635	65.43	417	102603	32	1		1		1		1			3						2	1		1			1	1	1			1			
solids_Sphingobacteriales_1	Bacteroidetes	80	35	3.293488	39.22	18	591082	50					2	1	1			1						1	1	1			1	1	1							
solids_Microbacterium_1	Actinobacteria	56	25	3.028969	69.03	17	468500	51			1	1						1	1	1				1	1	1												
solids_Rhodanobacter_1	Gammaproteobacteria	49	144	2.762428	68.02	322	59675	30				1	3	1	1			1	2	1						2		2		1								
solids_Rhodanobacter_6	Gammaproteobacteria	47	28	0.749597	68.15	209	44573	9																				2	1									
solids_Sphingopyxis_1	Alphaproteobacteria	39	1.9	0.670628	65.34	306	13764	38			1	1	1		1	1		1	1	1																		
solids_Rhodanobacter_5	Gammaproteobacteria	36	13	3.602337	66.95	175	120909	37			1	1	1		1			1	1	1								1							1			
solids_Rhodanobacter_4	Gammaproteobacteria	30	29	3.432851	66.33	89	285871	23			1	1	2	1	1			1	1	1						2		1	1						1			
solids_Nitrosospira_1	Betaproteobacteria	28	30	3.080869	54.42	44	307813	51	1		1	1			1			1				1	1			1		1	1						2			
solids_Sphingobacteriales_4	Bacteroidetes	25	7.3	4.668254	38.37	70	293464	51					1		1			1						1	1	1				1	1							
solids_Brevundimonas_1	Alphaproteobacteria	23	1.6	1.485446	67.83	670	26323	45			1	1						1	1																1			
solids_Castellaniella_1	Betaproteobacteria	23	77	3.708385	66.07	51	574638	51			1	1	3	2	1	1	1	1	1					1	1	2		2	1	1	1				1			
solids_Stenotrophomonas_1	Gammaproteobacteria	21	1.2	3.463406	69.59	89	203045	50					1	1		1	1	1	1	1				1	1	1		1										
solids_Burkholderiales_3	Betaproteobacteria	16	4.6	1.404795	65.94	575	69432	18					1			1			1	1																		
solids_Alphaproteobacteria_1	Alphaproteobacteria	14	39	3.200314	61.99	66	253970	49			1	1	2	1	1			1	1	1								1										
solids_Rhizobiales_1	Alphaproteobacteria	14	16	4.689115	63.33	69	485647	51			2	1	1		1			1	1	2				1	1	3		1	1			1						
solids_Sphingobacteriales_3	Bacteroidetes	12	18	5.329888	42.32	116	223840	51					2	1	1			1						1	1	2		1		1	1							
solids_Thiobacillus_1	Betaproteobacteria	11	16	1.935166	64.27	229	53321	44	1	1	1	1	2		3			1	1	1				1				1		1	1		2	1	1			
solids_Actinobacteria_3	Actinobacteria	9.7	12	3.028281	70.55	224	59440	50			2	1							1					1	1													
solids_Xanthomonadales_1	Gammaproteobacteria	9.5	13	3.327705	65.7	24	411702	45			1	1	2	2	1	1	1	1	1					1	1			1										1
solids_Xanthomonadales_2	Gammaproteobacteria	8.4	3.7	1.825173	68.58	191	34686	18				1	1	1	1			1									1	1										
solids_Lysobacter_1	Gammaproteobacteria	7.9	3.4	0.432826	68.46	156	16506	26																1	1	1												
solids_Afipia_1	Alphaproteobacteria	7.5	19	2.809209	61.18	27	315568	51						1	1									1	1	2						1						
solids_Sphingopyxis_2	Alphaproteobacteria	7.3	2.4	1.436326	66.44	641	11619	30									1											1										
solids_Gemmatimonadetes_1	Gemmatimonadetes	6.3		3.096685	65.71	623	29115	47			1	1	2	1	1									1		1		1		1	1							
solids_Cytophagia_1	Bacteroidetes	5.2	8.5	1.403249	34.1	138	53453	49																														
solids_Micrococcales_1	Actinobacteria	5	16	1.843805	68.98	1001	8486	24							1									1														
solids_Bacteroidetes_1	Bacteroidetes	4.6	12	3.159377	37.08	83	248971	46					1	1	2			1								1				1	1							
solids_Flavobacteriales_1	Bacteroidetes	4.2	2.7	1.307045	39.03	814	8954	11						1																								
solids_Rhizobiales_2	Alphaproteobacteria	2.9	9.1	2.693058	66.69	167	56037	19			1	1	2	1	1			1	1	1						1												
solids_Flavobacteriia_1	Bacteroidetes	2.1	53	3.160681	39.49	20	531725	49					2					1																				
solids_Actinobacteria_2	Actinobacteria	1.9	14	5.135461	69.54	86	293326	45				1			1			1	1	1				1	1													
solids_Chryseobacterium_1	Bacteroidetes	1.6	40	2.280943	39.38	23	311201	50					2	2	1											1			1	1	1							

Chapter 2

Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome

Sumayah F. Rahman[1], Matthew R. Olm[1], Michael J. Morowitz[2] and Jillian F. Banfield[3*]

[1] Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA
[2] Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA
[3] Department of Earth and Planetary Sciences, and Environmental Science, Policy and Management, University of California, Berkeley, CA 94720, USA

Abstract

Antibiotic resistance in pathogens is extensively studied, yet little is known about how antibiotic resistance genes of typical gut bacteria influence microbiome dynamics. Here, we leverage genomes from metagenomes to investigate how genes of the premature infant gut resistome correspond to the ability of bacteria to survive under certain environmental and clinical conditions. We find that formula feeding impacts the resistome. Random forest models corroborated by statistical tests revealed that the gut resistome of formula-fed infants is enriched in class D beta-lactamase genes. Interestingly, *Clostridium difficile* strains harboring this gene are at higher abundance in formula-fed infants compared to *C. difficile* lacking this gene. Organisms with genes for major facilitator superfamily drug efflux pumps have faster replication rates under all conditions, even in the absence of antibiotic therapy. Using a machine learning approach, we identified genes that are predictive of an organism's direction of change in relative abundance after administration of vancomycin and cephalosporin antibiotics. The most accurate results were obtained by reducing annotated genomic data into five principal components classified by boosted decision trees. Among the genes involved in predicting if an organism increased in relative abundance after treatment are those that encode for subclass B2 beta-lactamases and transcriptional regulators of vancomycin resistance. This demonstrates that machine learning applied to genome-resolved metagenomics data can identify key genes for survival after antibiotics and predict how organisms in the gut microbiome will respond to antibiotic administration.

Introduction

Antibiotic use has been steadily increasing over the past several decades and is correlated with the prevalence of antibiotic resistance in bacteria (Goossens, Ferech, Vander Stichele, & Elseviers, 2005). Widespread antibiotic resistance, in combination with the decline in development of new antibiotics, presents a major threat to human health (Spellberg et al., 2008). The gut microbiome is a reservoir for antibiotic resistance genes (Penders, Stobberingh, Savelkoul, & Wolffs, 2013) and may be involved in the spread of resistance genes to pathogens (Simonsen, Lvseth, Dahl, & Kruse, 1998; Teuber, Meile, & Schwarz, 1999; Van Braak et al., 1998). Additionally, antibiotics are often prescribed to treat infections without considering how the drug will affect the gut microbial community, which can lead to negative consequences for the human host (Langdon, Crook, & Dantas, 2016). It is therefore important to study how the antibiotic resistance genes harbored by organisms in the gut microbiome impact community dynamics.

The preterm infant gut resistome is considered a research priority because premature infants are almost universally administered antibiotics during the first week of life (Clark, Bloom,

Spitzer, & Gerstmann, 2014). Early life is a critically important time for community establishment (J. J. Faith et al., 2013), and neonatal antibiotic therapies have both transient and persistent effects on the gut microbial community. Included among the many ways that antibiotics have been shown to affect the microbiome are lower bacterial diversity (Greenwood et al., 2014), enrichment of *Enterobacteriaceae* (Arboleya et al., 2015; Greenwood et al., 2014), reduction of *Bifidobacterium* spp. (Tanaka et al., 2009), and enrichment of antibiotic resistance genes (Jernberg, Löfmark, Edlund, & Jansson, 2007), including those that have no known activity against the particular antibiotic administered (Gibson et al., 2016). Previous studies have shown that the community composition of the infant microbiome is affected by diet, with artificial formula selecting for *Escherichia coli* and *Clostridium difficile* (Penders et al., 2005), and breast milk selecting for certain strains of *Bifidobacterium* (Costello, Stagaman, Dethlefsen, Bohannan, & Relman, 2012). The effect of birth mode on the microbiome is contested, with most studies finding that it has an effect on the gut microbiome (Penders et al., 2006; Wampach et al., 2017; Yassour et al., 2016) although some show no effect (Chu et al., 2017; Stewart et al., 2017). Gender (Cong, Xu, Janton, Henderson, & Matson, 2016) and maternal antibiotics before or during birth (Fouhy et al., 2012; Keski-nisula et al., 2013; Mshvildadze et al., 2010) also influence microbiome assembly.

Here we use genome-resolved metagenomics coupled with statistical and machine learning approaches to investigate the gut resistome of 107 longitudinally sampled premature infants. We show that certain antibiotic resistance genes in particular genomes affect how clinical factors influence the gut microbiome and, in turn, how the antibiotic resistance capabilities of a gut organism influence its growth and relative abundance.

Materials and Methods

Sample collection, sequencing, assembly, and gene prediction

Fecal samples were collected from 107 infants that resided in the Neonatal Intensive Care Unit (NICU) at the Magee Women's hospital in Pittsburgh, Pennsylvania during the sampling period. Briefly, DNA was extracted using the PowerSoil DNA isolation kit (MoBio Laboratories, Carlsbad, CA, USA) and sequenced using the Illumina HiSeq platform. Details on sample recovery, extraction, library preparation, and sequencing have been previously reported (Brooks et al., 2017; Raveh-Sadka et al., 2016, 2015). Using default parameters for all the programs, the reads were trimmed with Sickle (https://github.com/najoshi/sickle), cleared of human contamination following mapping to the human genome with Bowtie2 (Langmead & Salzberg, 2012), and assembled with idba_ud (Peng et al., 2012). Additionally, idba_ud was used to generate co-assemblies for each infant by simultaneously assembling all the samples belonging to the infant. Prodigal (Hyatt et al., 2010) run in the metagenomic mode was used for gene prediction.

Genome recovery and relative abundance calculation

For each infant, reads from all samples from that infant were mapped to all individual assemblies from that infant as well as the infant's co-assembly using SNAP (Zaharia et al., 2011). Coverage of scaffolds was calculated and used to run concoct (Alneberg et al., 2014) with default parameters on all individual assemblies and co-assemblies. To remove redundant bins, all bins recovered from each infant were de-replicated using dRep (Olm, Brown, Brooks, & Banfield, 2017) v0.4.0 with the command: *dRep dereplicate_wf --S_algorithm gANI -comp 50 -con 25 -str 25 -l 50000 -pa .9 -nc .1*.

Using Bowtie2 (Langmead & Salzberg, 2012), the reads from each sample were mapped to the set of genomes that were recovered from that particular infant. The read mapping output

files were used to calculate the average coverage of each genome in each sample, and the coverage values were converted to relative abundance by utilizing the read length, total number of reads in the sample, and genome length.

iRep calculation

For each sample, a set of representative genomes was first chosen from the complete collection of de-replicated genomes. First, all genomes were clustered at 98% ANI using dRep (Olm et al., 2017). A pangenome was then generated for each of these clusters using PanSeq (Laing et al., 2010), creating a list of fragments representing the entire sequence-space of each cluster. All pangenomes of all clusters were merged, and reads from all samples were mapped to the resulting pangenome set using SNAP (Zaharia et al., 2011). By analyzing the coverage of all fragments in the pangenome set, the breadth of each genome in each sample was calculated (number of genome fragments > 1x coverage / total genome fragments). Genomes with less than 85% breadth were removed from analysis. For all remaining genomes, the genome from each cluster with the highest breadth was added to that sample's representative genome list.

Next, reads from each sample were mapped to its representative genome list using bowtie2 (Langmead & Salzberg, 2012) default parameters. iRep (Brown, Olm, Thomas, & Banfield, 2016) was run on the resulting mapping files using default parameters and without GC correction. Only values that passed iRep's default filtering and were < 3 were considered for analysis.

Annotation

The amino acid sequences of genes predicted by the metaProdigal gene finding algorithm (Hyatt et al., 2010) were searched against Resfams (Gibson, Forsberg, & Dantas, 2014), an antibiotic resistance gene specific profile hidden Markov model (HMM) database using the *hmmscan* function of HMMER v 3.1b2 (Finn, Clements, & Eddy, 2011). The *--cut_ga* option was used to set the reporting and inclusion limits as the profile-specific gathering threshold, which have been manually optimized on a profile-by-profile basis to ensure Resfams prediction accuracy (Gibson et al., 2014). The Resfams annotation output and the coverage of each scaffold that had a hit to a Resfams profile were used to generate sample resistance gene summaries. Each sample resistance gene summary, which represents the antibiotic resistance potential of a particular infant gut microbiome at a particular point in time, displays the counts per million reads (CPM) for each of the 170 antibiotic resistance gene families in the Resfams database. Additionally, genome resistance gene profiles that indicated the count of each resistance gene were developed for each genome. Information about the database, including descriptions of the antibiotic resistance genes represented by each accession code, is available at http://www.dantaslab.org/resfams/.

To gather general metabolism data, all binned sequences were searched against Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000) HMMs and the results were parsed for genome profiling. This resulted in a KEGG metabolism profile for each organism that displayed the fraction of each KEGG module encoded by that genome.

Statistical and computational analysis

To evaluate the effect of feeding regimen, delivery mode, gender, maternal antibiotics, and the infant's current antibiotic therapy, three cross-sectional PERMANOVA (McArdle & Anderson, 2013) tests for weeks two, four, and six were performed using the *adonis2* function of the *vegan* package in R (Dixon, 2003). For each infant, the first sample of each week was identified and the resistance gene summary of that sample was included in the PERMANOVA. If antibiotics were being administered on the day of sampling (which also indicates a current disease diagnosis), the infant was labeled as currently receiving antibiotics. Infants that were exclusively fed breast

milk and infants that were given breast milk at any point were both labeled as receiving breast milk. The Bray-Curtis dissimilarity metric was used and 9,999 permutations were performed to assess the marginal effects of the terms. The factor revealed to have a significant difference in antibiotic resistance gene content ($p < 0.05$) was selected for continued analysis. To identify antibiotic resistance genes associated with either formula feeding or breast milk during the weeks indicated by the PERMANOVA results, the infant's diet was used to classify sample resistance gene summaries using random forest models (Pedregosa et al., 2012). Mann-Whitney U tests were performed on Resfams that had feature importance scores above 0.07 in the random forest models, as calculated by the Gini importance metric. Genomes containing resistance genes significantly associated with a particular feeding type, along with genomes of the same species lacking these genes, were further investigated. The ribosomal protein S3 (RPS3) genes for each genome were identified by rp16.py (https://github.com/christophertbrown/bioscripts/blob/master/bin/rp16.py). The RPS3 nucleotide sequences were aligned with MUSCLE (Edgar, 2004) using default parameters, and a maximum-likelihood phylogenetic tree was built with RAxML (Stamatakis, 2014). Pairwise Pearson correlations of Resfams with KEGG modules within these genomes were calculated.

The Pearson correlation of mean replication index (iRep) for a sample and the sample's total resistance gene content was determined for samples collected within five days following antibiotic treatment. The replication rates of organisms harboring antibiotic resistance genes were compared to those lacking resistance genes of the same category, All p-values were Bonferonni corrected for multiple testing.

Infants for which there was a sample taken both before and after post-week antibiotic treatment were identified and the before and after samples were selected (no samples were available prior to the empiric antibiotics administered during the first week). Genomes from the selected samples were identified and labeled as either increasing or decreasing in relative abundance from the pre-antibiotic sample to the post-antibiotic sample. Using scikit-learn (Pedregosa et al., 2012), development of a machine learning model to predict the direction of change in relative abundance for each genome based on its Resfams and KEGG metabolism was attempted; yet an adequate model could not be developed, presumably due to variation in the ways that organisms respond to different antibiotic combinations. Therefore, the dataset was narrowed to include the six infants that received either cefotaxime or cefazolin (both cephalosporin antibiotics) in conjunction with vancomycin. 70% of the genomes obtained from these infant samples were used for training, 15% was used as a validation set for model improvement, and 15% was held out as a final test set. Several attempts to improve model performance through algorithm choice, feature engineering, and parameter tuning were applied, and the model that exhibited the best results with regard to precision and recall was selected. This model was then used to make predictions on the final test set. Each feature constructed for the model was a principal component of the Resfams and KEGG metabolic data, and the genes/modules contributing most strongly to each of these principal components were identified. The tendency for each of the genes and modules to occur in the increase class was calculated by adding -1 to the gene's mean value in the increase class divided by its mean value in the decrease class.

Data availability

The dataset used is comprised of 597 previously reported samples (Brooks et al., 2017; Raveh-Sadka et al., 2016, 2015), as well as 305 new samples. These samples are available at NCBI under accession number SRP114966 (https://www.ncbi.nlm.nih.gov/sra/?term=SRP114966). The code for the analysis, along with all the data and metadata used in the analysis, is hosted at https://github.com/SumayahR/antibiotic-resistance.

Results and Discussion

Antibiotic resistance of the premature infant microbiome

107 premature infants were studied during the first three months of life. The median birthweight was 1228 g (IQR = 902 - 1462), with 35% of the infants extremely low birthweight (< 1000 g) and 65% of infants with birthweight > 1000 g. Birthweight is closely linked to gestational age, which is divided into the following categories: late preterm (34 to < 37 weeks gestation), moderate preterm (32 to < 34 weeks gestation), very preterm (28 to < 32 weeks gestation) and extremely preterm (< 28 weeks gestation) (Glass et al., 2016). 30% of infants in this study were extremely preterm; these infants tend to have significant health problems, including higher rates of necrotizing enterocolitis and extreme dysbiosis of the microbiota (Underwood & Sohn, 2017). The majority of infants in our study (60%) were classified as very preterm, just 10% of our infants were classified as moderate preterm, and no infants were late preterm. Because the infants in this study were mostly very or extremely preterm, it should be noted that the biological characteristics reported here are highly divergent from that of typical full-term infants (Schwiertz et al., 2003).

Longitudinal sampling of each infant resulted in a total of 902 samples that were sequenced and analyzed. All 107 infants received gentamicin and ampicillin during the first week of life, and 36 of those infants received additional antibiotics in the later weeks due to disease (Table 2.1). All samples were subject to Illumina short-read shotgun sequencing and the sequence data assembled using idba-ud (see methods for details). Binning resulted in a de-replicated set of 1483 genomes. The taxonomic composition of these samples is typical for the premature infant gut (Figure 2.1A, Figure 2.1B). Resfams (Gibson et al., 2014) annotations of predicted amino acid sequences from the resulting scaffolds revealed that the most abundant resistance mechanisms were resistance-nodulation-division (RND) efflux pumps and ATP-binding-cassette (ABC) transporters (Figure 2.1C, Figure 2.1D). It is important to note that in addition to their ability to contribute to antibiotic resistance, efflux pumps and transporters have been associated with stress response (Nagayama, Fujita, Takashima, Ardin, & Ooshima, 2014; Poole, 2008, 2014) and may reflect a rapidly changing environment during the first few months of life.

For infants that did not receive additional antibiotics (Figure 2.1C), a decreasing trend in total antibiotic resistance potential is observed over time ($p < 0.005$). During the first week of life, empiric antibiotic therapy perturbs the microbiome by preferentially enriching for antibiotic resistant organisms. This is consistent with prior results showing temporarily elevated resistance gene levels after administration of antibiotics (Yassour et al., 2016). Microbial community recovery begins following this period. For infants that received antibiotics after the first week of life (Figure 2.1D), there was no consistent trend of decreasing resistance potential. This suggests that administration of antibiotics to premature infants after the first week of life can prolong the enrichment of the resistome.

Approximately 20% of resistance genes annotated by Resfams were not assignable to specific organisms in the microbiome. This is partly due to some genes being carried on plasmids, which were excluded from the genomic analysis.

Formula feeding influences the gut resistome through strain-level selection

Permutational multivariate analysis of variance (PERMANOVA) tests, which discern and isolate the effects of factors through partitioning of variance (Anderson, 2006), were performed to investigate the effect of feeding regimen, delivery mode, gender, maternal antibiotics, and the infant's current antibiotic therapy on the resistome. Tests were performed on the resistomes of samples taken at weeks two, four, and six to avoid the bias of repeated measures in longitudinal

sampling. At week two, formula-fed infants did not have a significantly different distribution of antibiotic resistance genes compared to infants that received breast milk. However, a difference was detected at weeks four and six ($p < 0.05$), accompanied by an increase in effect size as assessed by PERMANOVA F-statistic (Table 2.2). This signals divergence of the resistomes of formula-fed and breast-fed infants over time. The PERMANOVA tests were not sensitive enough to detect any effects on the resistome resulting from delivery mode, gender, or antibiotics, which may be because the test displays conservatism when variances are positively related to group sample size(Anderson & Walsh, 2013). Because these factors have been shown to alter the microbiota(Cong et al., 2016; Fouhy et al., 2012; Keski-nisula et al., 2013; Penders et al., 2006), it is unlikely that the resistome was truly unchanged. Since feeding type was the only factor that produced a detectable response, we further investigated its effects.

Random forest models were used to classify resistomes as either belonging to a formula-fed baby or a breast-fed baby, and we used the trained model's feature importance scores to select resistance genes for further study (Table 2.3). One out of the four selected resistance genes was significantly associated with a feeding group: Class D beta-lactamase was enriched in formula-fed infants ($p < 0.05$) (Figure 2.2A). Genome-resolved analysis revealed that Class D beta-lactamase genes are most frequently carried by *Clostridium difficile*. Of the 67 *C. difficile* genomes in the de-replicated dataset, 38 of these organisms harbor a Class D beta-lactamase gene. Phylogenetic analysis reveals that these 38 organisms are very closely related (Figure 2.2B). To ascertain if this *C. difficile* strain is involved in the enrichment of Class D beta-lactamase in the formula-fed infant gut resistome, the relative abundance of *C. difficile* with and without a Class D beta-lactamase gene in the gut microbiome of breast-fed and formula-fed infants was assessed. In infants that only receive formula, *C. difficile* with Class D beta-lactamase is consistently more abundant than *C. difficile* lacking this gene; while in infants that receive breast milk, both types of *C. difficile* are low in relative abundance (Figure 2.2C). Even with the lower relative abundance of some *C. difficile*, there was no significant difference in genome completeness and N50 between the two groups, assuring us that there was no methodological issue that reduced ability to detect beta-lactamase. Prior studies have reported an increased abundance of *C. difficile* in the gut microbiomes of formula-fed infants (Penders et al., 2005), but here we reveal that formula feeding enriches for a particular *C. difficile* strain.

Class D beta-lactamase hydrolyzes beta-lactam antibiotics (Szarecka, Lesnock, Ramirez-Mondragon, Nicholas, & Wymore, 2011), and there is no known connection between host diet and its antibiotic resistance function. It is thus unlikely that Class D beta-lactamase offers a selective advantage to organisms in the gut of formula-fed infants, but this gene may be linked to other genes that confer an advantage. Pairwise correlations of the Resfams and KEGG metabolism modules in *C. difficile* genomes revealed that one KEGG module, the cytidine 5'-monophospho-3-deoxy-d-*manno*-2-octulosonic acid (CMP-KDO) biosynthesis module, was perfectly correlated with the presence of the Class D beta-lactamase gene. CMP-KDO catalyzes a key reaction in lipopolysaccharide biosynthesis (Wang & Quinn, 2010). Further inspection of the KEGG annotations revealed that only one gene from this module was present in *C. difficile*: arabinose-5-phosphate isomerase. This gene typically occurs in Gram-negative bacteria, where it plays a role in synthesis of lipopolysaccharide for the outer membrane (Meredith, Aggarwal, Mamat, Lindner, & Woodard, 2006), yet a recent study identified arabinose-5-phosphate isomerase in a Gram-positive organism, *Clostridium tetani* (Cech, Markin, & Ronald, 2017). Although the function of this gene in Gram-positive bacteria is unknown, it is hypothesized to be a regulator and may modulate carbohydrate transport and metabolism (Cech et al., 2017). If so, *C. difficile* (Gram-positive) strains with arabinose-5-phosphate isomerase may have a competitive advantage because they are able to rapidly respond to availability of the carbohydrates that are abundant in formula.

It is also possible that other, potentially unknown, genes are responsible for the observed effect; and these genes may not necessarily relate to metabolism of compounds in formula. Breast-fed babies have increased abundance of *Bifidobacterium* (Costello et al., 2012), so the ways that different strains of *C. difficile* interact and compete with *Bifidobacterium* may contribute to the observed trend.

Major facilitator superfamily pumps are associated with increased replication

A previous analysis revealed that antibiotic administration is associated with elevated bacterial replication rates (iRep values), which was hypothesized to be due to high resource availability after elimination of antibiotic susceptible strains (Brown et al., 2016). Extending upon this result, we show here that a sample's mean replication index in the days following antibiotic treatment is positively correlated with total resistance gene content ($p < 0.05$) (Figure 2.3A). To be present in the period following antibiotic administration, all organisms must be antibiotic resistant; it is thus unclear why a larger inventory of resistance genes should lead to faster growth rates.

To characterize the effect of antibiotic resistance genes on iRep values in isolation from the confounding effects of antibiotics, we studied infants that did not receive any antibiotics after the first week of life. In these infants, organisms carrying genes for major facilitator superfamily (MFS) transporters have significantly higher iRep values than those that do not have MFS genes ($p < 5 \times 10^{-5}$) (Figure 2.3B). As there are known differences in median iRep values among phyla (Brown et al., 2016), the comparison was repeated within each phylum that contained members with and without MFS genes. The trend of higher iRep values for organisms with MFS was most apparent in Firmicutes ($p < 5 \times 10^{-4}$) (Figure 2.3B). The genomes lacking MFS show comparatively high completeness scores, suggesting that this finding is not due to missed detection of the MFS genes. Therefore, the presence of these antibiotic resistance genes appears to inherently increase replication, even when no antibiotics are being administered. This could be due to protection from antibiotics being produced at a low level by other gut organisms (Modi, Collins, & Relman, 2014) or a result of MFS pumps' naturally beneficial physiological functions (Piddock, 2006). We also acknowledge that this finding may simply reflect high incidence of organisms with MFS genes during periods of fast replication without a causal link.

A model that predicts an organism's response to vancomycin and cephalosporins

We modeled the relationship between gene content of a gut organism and its direction of change in relative abundance (increase vs. decrease) after a premature infant is administered a combination of glycopeptide (vancomycin) and beta-lactam (cephalosporin, either cefotaxime or cefazolin) antibiotics. Principal component analysis was performed on Resfams (Gibson et al., 2014) and KEGG (Kanehisa & Goto, 2000) annotations to generate a low-dimensional representation of each organism's metabolic potential and resistance potential. The first five principal components (PCs) cumulatively explained 48% of the variation in the dataset. Using these PCs as input, the AdaBoost-SAMME algorithm (Zhu, Zou, Rosset, & Hastie, 2009) was applied, with decision tree classifiers as base estimators. The model, trained on 70% of the data, performed extremely well on the validation set, with a precision of 1.0 and recall of 1.0, indicating that every genome was correctly classified. Because the validation set was utilized for testing during the preliminary stages of model development, the model was also evaluated with a final test set, on which it achieved 0.9 precision and 0.7 recall.

Of the features that acted as the strongest contributors to each of the PCs, five genes with a tendency to occur in microbes that increase in relative abundance after antibiotic treatment were identified (Figure 2.4). One of these is subclass B2 beta-lactamase, which is carried by several

organisms that persisted after antibiotics including *Enterococcus faecalis*, *Clostridium baratii,* and *Bradyrhizobium sp.* Subclass B2 beta-lactamase has been shown to hydrolyze carbapenems and displays much lower levels of resistance to cephalosporins (Valladares et al., 1997). Considering its substrate specificity for carbapenems, this beta-lactamase may not be directly contributing to an organism's ability to persist after treatment with cephalosporins; rather, it may be linked to other, potentially unknown, genes. However, the substrate specificity of an antibiotic resistance gene can depend on the organismal context of that gene (Hansen, Jensen, Sørensen, & Sørensen, 2007), and a single base substitution in a beta-lactamase gene can alter substrate specificity (Jacoby & Medeiros, 1991), so the possibility that beta-lactamases falling into the B2 subclass may confer some gut organisms with resistance to cephalosporins should not be discounted.

Furthermore, our model shows that a gene linked to vancomycin resistance, vanR, is among the genes predictive of an organism's propensity to increase in relative abundance after antibiotic treatment (Figure 2.4). VanR is the transcriptional activator of an operon encoding genes involved in peptidoglycan modification (VanH, VanA, and VanX), which prevents vancomycin from binding to its target (Hughes, 2003). This gene cluster usually resides on plasmids (Boyce, 1997; Périchon & Courvalin, 2009). VanR, an essential gene for the initiation of the vancomycin resistance operon promoter (Arthur & Quintiliani, 2001), was chromosomally encoded in several genomes of organisms that increased after antibiotics, such as *Enterococcus faecalis* and *Clostridium perfringens*. Because our genomic analysis precluded the assignment of genes on plasmids, VanR was the best indicator of resistance.

In addition to genes specifically encoding for resistance to beta-lactams or glycopeptides, efflux pumps and transporters were also strong contributors to the PCs used as input to the model. Mex genes (of the resistance nodulation cell division family of drug efflux pumps) and ATP-binding cassette (ABC) transporter genes were associated with microbes that increase in relative abundance after antibiotics (Figure 2.4). Multidrug efflux pumps are essential for the intrinsic drug resistance of many bacteria, and overexpression of the genes for these pumps leads to elevated resistance levels (Li & Nikaido, 2009). *Bacteroides ovatus* and *Bacteroides helcogenes* carried multiple copies of Mex efflux pumps, while *Enterococcus faecalis* and *Clostridium baratii* harbored several ABC transporter genes. Although the genomes of these organisms also encoded target-specific resistance genes such as the subclass B2 beta-lactamase, the more general pumps and transporters likely enhanced their ability to flourish after antibiotic treatment.

Previous studies have utilized data from 16S rRNA gene amplicon sequencing or read-based metagenomics of the human microbiome to predict life events and disease states of the human host using machine learning or other modeling techniques (DiGiulio et al., 2015; Yazdani et al., 2016). However, read-based metagenomics lacks resolution at the genomic level, and due to strain-level differences in antibiotic resistance (Kumar et al., 2011), taxonomy data from marker gene studies cannot be used to predict how particular organisms in a community will respond to antibiotics. Here, for the first time, we utilize the data obtained by reconstructing genomes from metagenomes to make predictions about the future states of individual gut microbes. This has tremendous potential for application in the fields of medicine and microbial ecology. For example, such a model can be used before administering drugs to a patient to verify that a particular combination of antibiotics will not lead to overgrowth of an undesirable microbe. Our study serves as a proof of concept for this application of machine learning used in conjunction with genome-resolved metagenomics to derive biological insight.

Acknowledgements

Funding was provided through the National Institutes of Health (NIH) under grant RAI092531A and the Alfred P. Sloan Foundation under grant APSF-2012-10-05. This work used the Vincent J. Coates Genomics Sequencing Laboratory, supported by NIH S10 OD018174 Instrumentation Grant. The study was approved by the University of Pittsburgh Institutional Review Board (IRB) (Protocol PRO12100487).

We acknowledge Robyn Baker for recruiting infants, and Brian Firek for performing DNA extractions. We would also like to thank David Burstein for the KEGG HMM annotation pipeline, and Christopher Brown for scripts to identify ribosomal proteins and to calculate genome coverage.

Figure 2.1. Microbiome and resistome of the premature infant gut microbial community. The numbers of samples included in each week's average are as follows: for the infants that did not receive antibiotics after the first week, week 2 n = 197, week 3 n = 188, week 4 n = 110, week 5 n = 16, week 6 n = 20, week 7 n = 7, and week 8 n = 8; for the infants that received antibiotics after the first week, week 2 n = 72, week 3 n = 73, week 4 n = 53, week 5 n = 24, week 6 n = 16, week 7 n = 8, and week 8 n = 13. (A) The genus-level taxonomic composition of the gut community for the infants that did not receive antibiotics after the first week of life. (B) The genus-level taxonomic composition of the gut community for the infants that received antibiotics beyond the first week of life. (C) For the infants that do not receive antibiotics after the first week, the total resistance content of the premature infant gut microbiome has a slight negative correlation with age ($p = 0.003$). (D) The resistance gene levels of infant microbiomes that were exposed to additional antibiotics did not display a significant trend ($p = 0.265$).

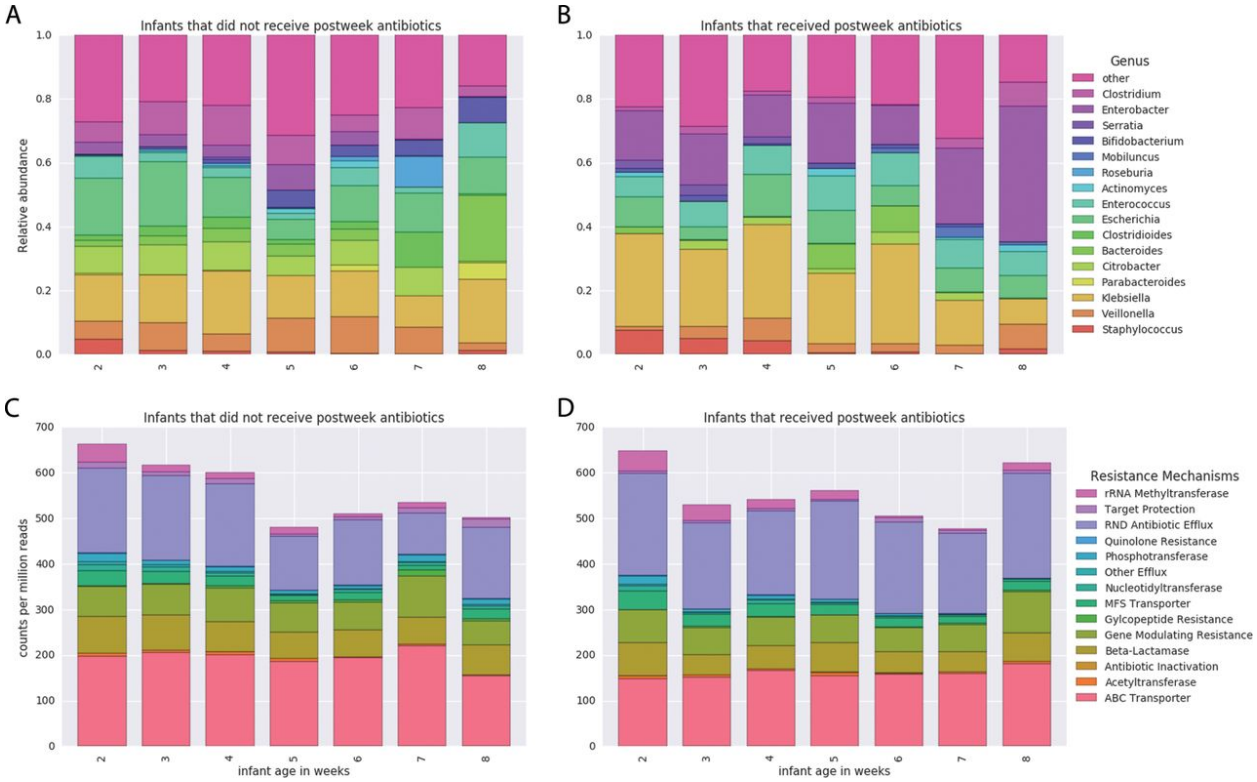

Figure 2.2. Formula feeding affects the resistome. (A) Class D beta-lactamase is enriched in formula-fed infants at 4 weeks of age (Mann-Whitney U = 66, Bonferroni-corrected p = 0.031). (B) Phylogenetic tree of Clostridium difficile genomes based on the ribosomal protein S3 gene. Names of genomes harboring a class D beta-lactamase are colored green and labeled with an asterisk. (C) The relative abundances of C. difficile genomes with class D beta-lactamase in formula-fed and breast-fed infants (top) (n = 38) and the relative abundances of C. difficile genomes lacking class-D betalactamase in formula-fed and breast-fed infants (bottom) (n = 29). Only the infants harboring C. difficile were included in calculations of average relative abundances.

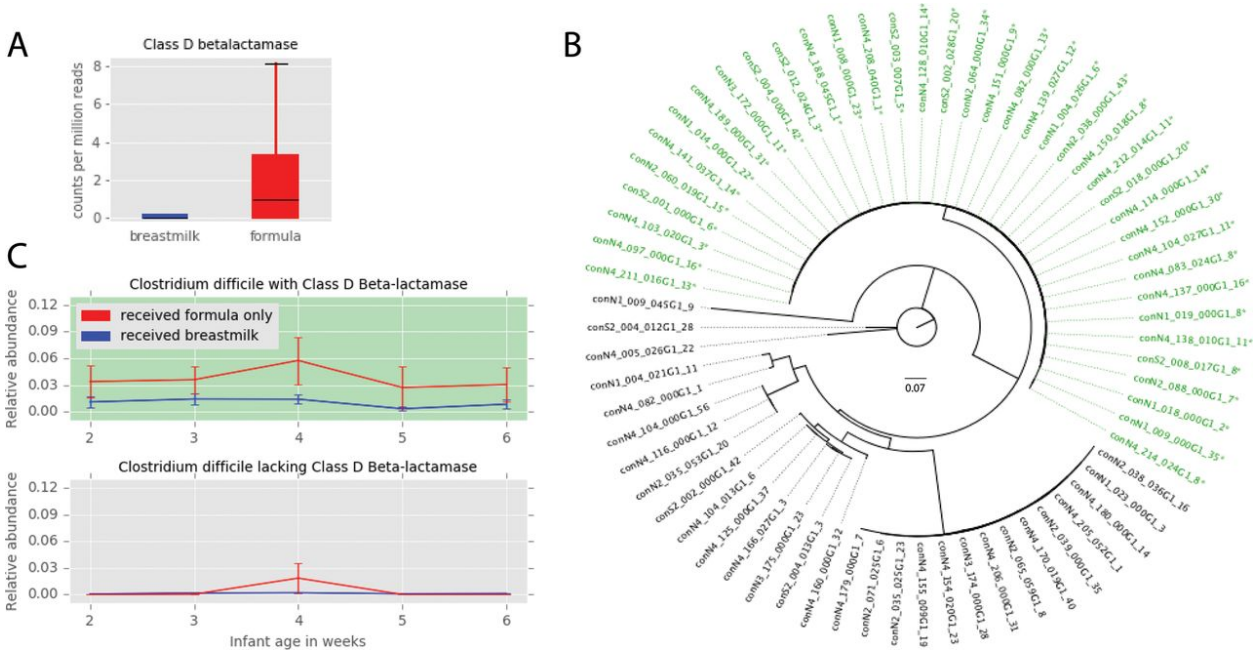

Figure 2.3. Antibiotic resistance and replication. (A) Among the samples taken within 5 days after antibiotic treatment, the antibiotic resistance potential of each sample is correlated with its mean replication index value (Pearson's $r = 0.39$, $p = 0.03$). (B) In infants that did not receive antibiotics after the first week of life, bacteria harboring at least one major facilitator superfamily (MFS) transporter gene had significantly higher iRep values (Mann-Whitney $U = 827{,}176.0$, $p = 1.55 \times 10^{-5}$), and this pattern is apparent within the members of the *Firmicutes* phylum (Mann-Whitney $U = 136{,}756.0$, $p = 0.0002$).

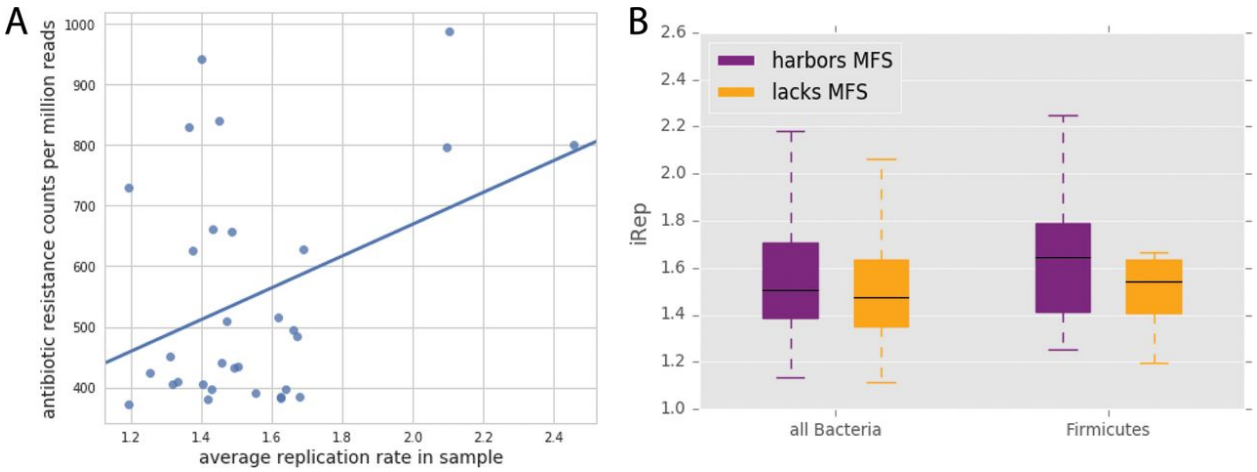

Figure 2.4. The tendency of genes to occur in the class of genomes that increased in relative abundance after antibiotics. Genes and modules strongly contributing to the principal components used in the machine learning model were identified, and class tendency was calculated using the ratio of the gene's prevalence in the increased-abundance group to its prevalence in the decreased-abundance group. Genes associated with the increased-abundance class of genomes are colored red.

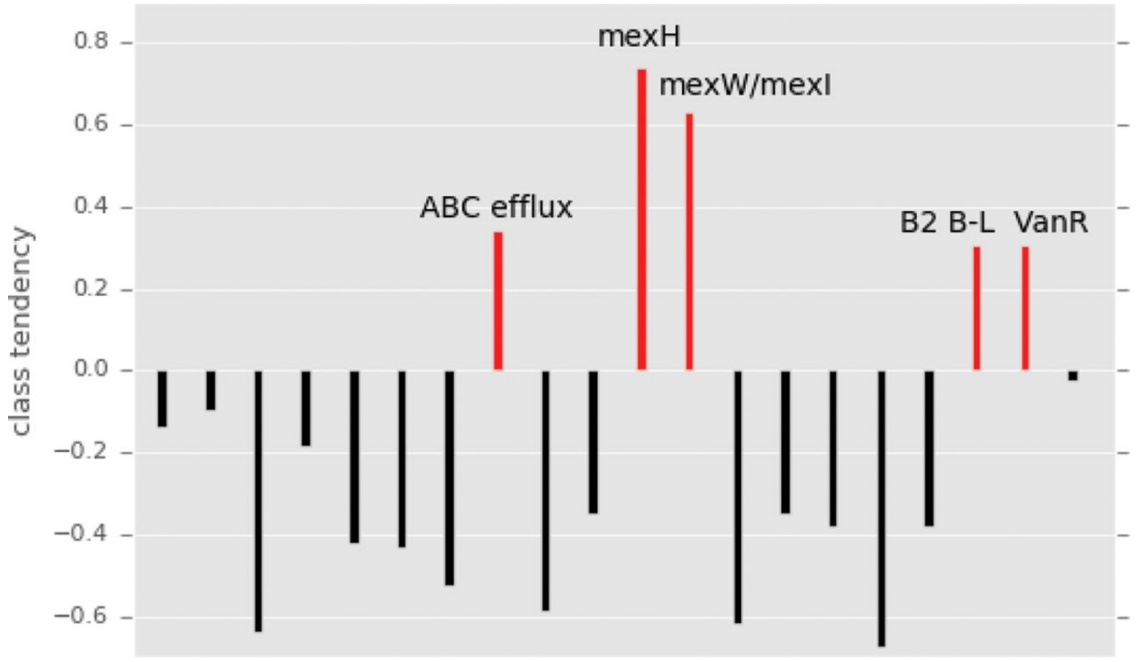

Table 2.1. Infant characteristics.

Characteristic	Value for infants who:	
	Received no antibiotics after the first week	**Received antibiotics after the first week[a]**
No. of samples	604	298
Total no. of infants[b]	71	36
No. of infants who received breast milk	52	32
No. of infants who were delivered by C-section	54	22
No. of infants of male sex	34	17
No. of infants with maternal antibiotics	24	20

[a] The infants represented in the column corresponding to those who received antibiotics after the first week (right) were administered antibiotics while in the NICU beyond the first week of life due to late-onset sepsis, necrotizing enterocolitis, or another disease.
[b] All 107 premature infants were in the neonatal intensive care unit (NICU) of the Magee-Women's Hospital in Pittsburgh, PA.

Table 2.2. Results of marginal PERMANOVAs with 9,999 random permutations for weeks 2, 4, and 6 performed on the antibiotic resistance gene content of infant samples as annotated by Resfams. The Bray-Curtis distance metric was used in the PERMANOVA and Bonferonni corrections were applied on the p-values to correct for multiple testing.

WEEK 2	Degrees of freedom	Sum of Squares	F-statistic	p value	corrected p value
received_breastmilk	1	0.3163	2.2328	0.0506	0.1518
infant_antibiotics	1	0.3958	2.7941	0.023	0.069
birth_mode	1	0.1875	1.3237	0.2146	0.6438
gender	1	0.0968	0.6837	0.6444	1
maternal_antibiotics	1	0.2846	2.0089	0.0728	0.2184
Residual	83	11.7574			

WEEK 4	Degrees of freedom	Sum of Squares	F-statistic	p value	corrected p value
received_breastmilk	**1**	**0.5348**	**4.2009**	**0.0026**	**0.0078**
infant_antibiotics	1	0.0824	0.647	0.6049	1
birth_mode	1	0.2427	1.9069	0.0791	0.2373
Gender	1	0.1169	0.9182	0.4462	1
maternal_antibiotics	1	0.2132	1.6747	0.1152	0.3456
Residual	64	8.1471			

WEEK 6	Degrees of freedom	Sum of Squares	F-statistic	p value	corrected p value
received_breastmilk	**1**	**0.4619**	**5.005**	**0.0024**	**0.0072**
infant_antibiotics	1	0.0994	1.0765	0.3168	0.9504
birth_mode	1	0.0869	0.9411	0.4227	1
gender	1	0.1601	1.7351	0.1285	0.3855
maternal_antibiotics	1	0.05	0.5417	0.7404	1
Residual	25	2.307			

Table 2.3. Features selected using the random forest Gini importance metric after training on resistomes of formula-fed infants and breast-fed infants.

Resfams category	**Feature importance score**	**Mann-Whitney U value**	***P* value**	**Corrected *P* value[a]**
ANT6	0.071	17	0.327	1
Class D beta-lactamase	0.089	66	0.008	0.031
mexX	0.098	11	0.106	0.426
soxR mutant	0.071	10	0.081	0.324

[a] Bonferroni corrections were applied to the *P* values obtained from Mann-Whitney *U* tests to adjust for multiple testing.

Chapter 3

Functional potential of bacterial strains in the premature infant gut microbiome is associated with gestational age

Sumayah F. Rahman[1], Matthew R. Olm[1], Michael J. Morowitz[2] and Jillian F. Banfield[3*]

[1]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA
[2]Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA
[3]Department of Earth and Planetary Sciences, and Environmental Science, Policy and Management, University of California, Berkeley, CA 94720, USA

Abstract

The gut microbiota of premature and full-term infants have many known differences, but the extent to which the degree of prematurity influences the structure and functional potential of the microbiome has not been deeply explored. Here, we used genome-resolved metagenomics to address how gestational age impacts the premature infant gut microbiome. Our analyses leveraged a genome-resolved metagenomic dataset derived from 106 infants, utilizing multiple linear regression and other data mining techniques. We found that gestational age is associated with species richness, with more premature infants having lower species richness; this effect lasts until the fourth week of life. Novel *Clostridium* species and strains related to *Streptococcus salivarius* and *Enterococcus faecalis* colonize infants of different gestational ages, and the metabolic potential of these organisms can be distinguished. Thus, we conclude that the extent of prematurity, or directly linked factors, can be an important influence on the microbiome and its functions.

Introduction

The human gut microbiome plays many important roles, including the extraction of nutrients from food, metabolizing toxins, immunomodulation, and protection from pathogens (Jandhyala et al., 2015). Infants, near-sterile when born, obtain microbes from their mother and their environment (Brooks et al., 2014; Koenig et al., 2011; Makino et al., 2013). The gut microbiome of infants is known for its simplicity and low complexity compared to the gut microbiome of older children and adults (Yatsunenko et al., 2012). Premature infants, born before they have reached 37 weeks *in utero*, harbor gut microbial communities of even lower complexity than full-term infants, as they are colonized by tenfold fewer species (Gibson et al., 2016). Premature infant gut microbiomes display abrupt shifts in composition (Costello, Carlisle, Bik, Morowitz, & Relman, 2013), and may have a different taxonomic makeup than microbiomes of full-term infants (Gibson et al., 2016; Rodríguez et al., 2015; Sim et al., 2013). Studies on premature infants suggest that the extent of the infant's prematurity influences their gut microbiome. A recent study, utilizing 16S rRNA gene sequencing, revealed that bacterial alpha diversity varies based on the infants' gestational age, with more premature infants having less diverse microbiomes. This study also showed that infants born at a later gestational age had greater abundance of *Bifidobacterium* and *Streptococcus* (Chernikova et al., 2018). However, there were no analyses of differences in metabolic potential of microorganisms colonizing infants of different gestational age, due to the low resolution of the 16S method. A metaproteomics study revealed that gut bacteria of extremely preterm infants (< 28 weeks gestational age) produced more

translation and membrane transport proteins, while the microbiomes of infants with a gestational age of 30 weeks produced more energy metabolism proteins (Zwittink et al., 2017).

Genome-resolved metagenomics involves sequencing all the DNA extracted from a sample and then reconstructing genomes for the relatively abundant microorganisms present. Previous genome-resolved metagenomics studies found that the infant gut microbiome is influenced by factors such as formula feeding, the hospital room environment, and antibiotic exposure (Brooks et al., 2014; Brown et al., 2016; S.F. Rahman, Olm, Morowitz, & Banfield, 2018). Here, we utilize genome-resolved metagenomics to analyze the effects of gestational age on the composition and metabolic potential of the premature infant gut microbiome. It is well-established that prematurity is associated with increased disease and infant mortality (Kramer et al., 2000), and this is partially due to factors involving the microbiome (Morrow et al., 2013). Understanding the effect that gestational age, i.e. the extent of prematurity, has on the microbiome may improve understanding of disease in premature infants. We found that certain bacteria occurring in infants of different gestational ages carry distinct sets of metabolic genes, thus utilizing genome-centric metagenomics to resolve how the gut microbiome is influenced by extent of prematurity.

Methods

Sample collection and metagenomic data processing for these samples were previously described (Sumayah F. Rahman, Olm, Morowitz, & Banfield, 2017). Briefly, fecal samples were collected from premature infants residing in the neonatal intensive care units (NICU) of the Magee-Women's Hospital in Pittsburgh, PA, and a PowerSoil DNA isolation kit (Mo Bio Laboratories, Carlsbad, CA) was used to extract the DNA, which was then sequenced on an Illumina platform (further details available in: (Brooks et al., 2017; Raveh-Sadka et al., 2016, 2015)). The samples analyzed in this study have been previously reported, and the reads are publicly available at NCBI as described in: (Brooks et al., 2017; S.F. Rahman et al., 2018; Raveh-Sadka et al., 2016, 2015). The reads were then trimmed using Sickle (https://github.com/najoshi/sickle) and cleared of human contamination through read mapping with Bowtie2 (Langmead & Salzberg, 2012). IDBA-UD (Peng et al., 2012) was used to assemble the reads of each sample and was also used to generate co-assemblies by combining the reads of all the samples from a particular infant. The genes on the scaffolds were predicted using Prodigal (Hyatt et al., 2010). The scaffolds were grouped into genome bins using concoct (Alneberg et al., 2014) and redundant bins were dereplicated using dRep (Olm et al., 2017) v0.4.0. The sequences were searched against the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000) using profile hidden Markov models, and the results were used to generate a KEGG metabolism profile for each organism that displayed the fraction of each KEGG module encoded by that genome. For specialized identification of biosynthetic gene clusters, antiSMASH (Medema et al., 2011) was used to annotate genes from particular scaffolds of interest. Organisms were considered to be present in a particular sample if the genome bin showed full breadth of coverage in the sample.

When calculating correlations, Pearson's product-moment correlation was used for two continuous variables and the point biserial correlation was used for pairs that contained at least one categorical variable. To model the effects of infant characteristics and clinical treatments on gut species richness levels, linear regression from the scikit-learn (Pedregosa et al., 2012) package was utilized. Mann-Whitney U tests were performed for comparison of relative abundance values and for the comparison of richness values at each week, and p-values were corrected using the False Discovery Rate (FDR) method. Spearman correlations were performed to identify trends over time. When carrying out the strain-focused portion of the study, separate analyses were

performed for each week of life and for each species, to remove effects of these factors. Only one sample per infant was included in each analysis to avoid bias due to repeated samples. When a genome was considered to be lacking a particular gene, the entire dRep set (all genomes belonging to the same secondary cluster in dRep) was checked to ensure that another genome in that set did not harbor the gene of interest. Associations were considered spurious and removed if the validation step did not confirm the findings.

Results

We analyzed 900 previously reported samples from 106 premature infants with gestational ages of 24 to 32 weeks at birth. Samples were collected over the first two to three months of life. Among these infants, just 10% were classified as moderate preterm (defined as 32 to < 34-week gestation), 60% of the infants were very preterm (28 to < 32-week gestation), and 30% were extremely preterm (< 28-week gestation) (Figure 3.1A). A correlation analysis revealed that gestational age is closely associated with birth weight ($r = 0.84$) (Figure 3.1B), but gestational age does not display a correlation with any of the other variables in the infant metadata (Figure 3.1C). Reconstructing genome bins from the 900 samples sequenced resulted in a dereplicated set of 1,483 genomes with an average completeness of 92% as evaluated based on the presence of bacterial single copy genes. There was no significant difference in genome completeness or sequencing depth of infants of different gestational age.

A linear regression model was applied to evaluate the effect of the infant's characteristics as well as environmental factors on the species richness of the gut microbial community. As expected, administration of antibiotics due to a disease diagnosis after the first week of life caused a significant decrease in richness ($p < 1 \times 10^{-6}$) (Table 3.1). The model also revealed that gestational age has a significant effect on species richness ($p < 1 \times 10^{-6}$) (Table 3.1). To understand how gestational age's effect on species richness changes over the course of the first few months of life, the richness of microbiomes of infants with gestational age < 28 weeks (extremely premature) was compared to that of infants with gestational age ≥ 28 weeks, at each week of life. In the first few weeks of life, microbiomes of extremely premature infants have significantly lower richness levels (Figure 3.2). This effect is no longer present at the fifth week of life and onward.

We compared the average taxonomic composition of microbiomes of infants with gestational age < 28 weeks and infants with gestational age ≥ 28 weeks (Figure 3.3). The figure shows small fluctuations in composition over time, but the microbiomes of individual infants can display more drastic shifts in composition. In infants with gestational age < 28 weeks, *Klebsiella* was consistently the most abundant taxa, except for during the seventh week of life where *Escherichia* was most abundant (Figure 3.3A). In infants with gestational age ≥ 28 weeks, *Escherichia* and *Klebsiella* were initially abundant but the relative abundance of these taxa appeared to decline over time (Figure 3.3B); however, this decline was not statistically significant, as the interindividual variation was substantial. When making direct comparisons of relative abundance between the two infant cohorts in the same week of life, the infants with gestational age ≥ 28 weeks had significantly higher populations of *Veillonella*, *Clostridioides* and *Clostridium* throughout the first month of life ($p < 0.001$). When making comparisons based on corrected age, which is calculated from the time of conception to adjust for prematurity, we found no significant difference between the relative abundance values of *Veillonella* and *Clostridioides* in the two infant cohorts. The lack of difference when matching samples based on corrected age supports the hypothesis that the previously mentioned finding is due to prematurity—as the extremely premature infants reach the second month of life, they become more developmentally similar to

the less premature infants, and the (likely prematurity-induced) effects that were present in early life are no longer detectable.

We investigated the differences in metabolic potential of bacterial strains colonizing infants of varying gestational ages. Separate analyses were performed for each week of life and for each species or species group, in the case of *Clostridium*. Within each week we considered only one sample per infant to reduce bias due to resampling of the same strain in subsequent samples, allowing us to test for patterns that were consistent across infants. We uncovered trends related to the extent of prematurity, in which particular organisms exclusively colonized infants of a certain gestational age. In the second week of life, species of a novel group in *Clostridium,* all of which have genes for vitamin B biosynthesis, were only present in infants of gestational age greater than 30 weeks (Figure 3.4A). *Streptococcus salivarius*-related strains containing genes for L-Cystine transport only occurs in infants of gestational age ≤ 30 weeks (Figure 3.4B). In the third and fourth week of life, very-closely related *Enterococcus faecalis* strains with genes for the RaxABRaxC type I secretion system are present exclusively in infants of gestational age ≥ 28 weeks (Figure 3.4C). Most of the *E. faecalis* carrying genes for this secretion system also harbor biosynthetic gene clusters for bacteriocin and lantibiotic, while none of the *E. faecalis* lacking the secretory genes were found to harbor these biosynthetic gene clusters.

Discussion

Our study involved analysis of the microbiomes of 107 premature infants for which a variety of metadata was collected, including each infant's gestational age, birthweight, disease incidence, antibiotic exposure, feeding method, gender, and birth mode. Agreeing with population-based references built from historical data (Talge, Mudd, Sikorskii, & Basso, 2014), the gestational age of infants in this study was closely correlated with birthweight (Figure 3.1B). However, the other collected metadata did not display a correlation with gestational age (Figure 3.1C), indicating that the findings of this microbiome study can be attributed to either (1) the gestational age, directly, or (2) differing clinical treatment among babies of varying gestational age that was not recorded in the collected metadata. This differing clinical treatment could be a particular feeding regimen (e.g., more specific than whether the infant received breastmilk, formula, or a combination), usage of a mechanical ventilator, length and timing of attachment to intravenous nutrition lines, or another factor.

Regardless of whether it is a direct or indirect effect, gestational age was strongly associated with species richness (Table 3.1), and this effect is only present during the first month of life (Figure 3.2). This indicates that while gestational age influences the microbes present in the few weeks immediately following birth, it does not have a persistent impact on microbiome complexity over the study period. We found that no other factors besides gestational age and post-week infant antibiotic exposure had a significant impact on richness of the gut microbiome (Table 3.1). This contrasts with findings of previous studies that associated formula feeding with increased diversity (Mueller, Bakacs, Combellick, Grigoryan, & Maria, 2015) and intrapartum maternal antibiotic use with decreased diversity (Mshvildadze, M., and Neu, 2010). It is important to note, however, that in the current study, organisms can be grouped at the strain or species level, whereas prior studies mostly relied on 16S rRNA gene fragment profiling, which typically has genus-level resolution.

Perhaps the most surprising result was that the day of life (infant's age in days) did not have a significant influence on species richness (Table 3.1). Previous studies have shown that the microbiome of full term infants gains species over time and displays a clear increase in diversity (Bäckhed et al., 2015). The difference between the patterns reported here and prior studies may be

largely accounted for by prematurity and, in some cases the administration of antibiotics that cause a sharp decline in microbiome diversity (Relman, 2012). Because these infants were in the neonatal intensive care unit throughout our study, the consortia available to colonize them likely had lower diversity than would have been encountered in the home environment and includes bacteria considered to be hospital-associated pathogens (Brooks et al., 2017).

We found that certain organisms only colonized infants of a particular gestational age range. One such case is *Clostridium*, which are anaerobes that have been previously found in the infant gut (Ferraris et al., 2012). We identified a group of *Clostridium* that was not closely related to previously sequenced organisms, and found that some organisms in this group harbor genes for biosynthesis of pantothenate, also called vitamin B_5, a water-soluble vitamin and an essential nutrient typically supplied by intestinal bacteria (Said, 2011). These novel *Clostridium* species with genes for pantothenate production were only found in less premature infants (Figure 3.4A). It should be noted that two of the infants harboring *Clostridium* with the pantothenate biosynthesis genes are twins, born at a gestational age of 32 weeks. Genetic relatedness, exposure to the same mother's microbiota, or other factors may have contributed to colonization by similar bacteria. However, the other infants showing gestational age-dependent colonization were unrelated. The observation that very premature infants may have comparatively lower access to vitamin B_5 than less premature infants due to strain colonization may be important because lack of pantothenic acid can adversely affect the immune system, producing a pro-inflammatory state (Gominak, 2016). It has been shown previously that the production of pantothenate in the gut is negatively impacted by low availability of vitamin D, which is often the case with very premature infants (Gurmeet, 2017). Thus, selection against *Clostridium* strains with the capacity for pantothenate production may be explained by increased prematurity.

Streptococcus salivarius-related bacteria with genes for transport of L-cystine, an amino acid essential for infants, are present only in infants of less than 31 weeks gestational age (Figure 3.4B). The infants in this study received cysteine, which forms the cystine dimer, as part of an amino acid mixture included in the parenteral nutrition. A study evaluating plasma amino acid concentrations in infants given parenteral nutrition found that infants of lower birthweight have less of an ability to use cystine/cysteine compared to infants of higher birthweight (W.C. et al., 1988). Since the more premature infants have lower birthweight (Figure 3.1B), the inability of the human cells to uptake cystine may lead to higher concentrations of cystine in the gut, indicating why cystine-transporting bacteria are selected for in these infants.

The trends described with *Clostridium* and *S. salivarius* both occur in the second week of life. However, *Enterococcus faecalis* displays interesting pattern in the third and fourth week of life: strains with a RaxABRaxC type I secretion system occur only in infants of greater gestational age, while strains lacking this secretion system occur exclusively in extremely premature infants (Figure 3.4C). The RaxABRaxC type I system is involved in the secretion of double-glycine-type leader peptides (da Silva et al., 2004), which occur in bacteriocins and lantibiotics (Aymerich et al., 1996). All the *E. faecalis* with the RaxABRaxC type I secretion system have gene clusters for production of lantipeptides, bacteriocin, or both. In contrast, most of the *E. faecalis* lacking the secretion system (i.e., the *E. faecalis* occurring in the infants of lowest gestational age) did not have these biosynthetic gene clusters. Since bacteriocins are toxins that inhibit the growth of closely related strains, the gestational age of an infant could indirectly influence the contribution of *E. faecalis* to bacteriocin production and thus influence microbiome composition. As *E. faecalis* is a common and often abundant member of the gut microbiomes of premature infants (Moles et al., 2015), it is possible that bacteriocin production is less common in infants of very low gestational ages.

The findings discussed above offer a strain-level perspective to what is known about how the taxonomic and functional characteristics of the gut microbial community change depending on the infant's gestational age (Chernikova et al., 2018; Zwittink et al., 2017). By analyzing the metabolic potential of each genome, we found evidence that the extent of prematurity, either directly or indirectly, can affect the gut microbiome. Given evidence that a lower gestational age may limit bacteriocin and vitamin production, which are factors that can impact community structure and lead to inflammation (Hibberd et al., 2017; Umu et al., 2016), these findings may inform our understanding of diseases associated with dysbiosis of the microbiome, especially in very premature infants.

Acknowledgements

We acknowledge Robyn Baker for recruiting infants, Brian Firek for performing DNA extractions, Christopher Brown for scripts to calculate genome coverage, and David Burstein for the KEGG HMM annotation pipeline.

Figure 3.1. Gestational age and related infant metadata. (A) The distribution of gestational age in weeks. (B) Gestational age is correlated with birthweight ($r = 0.84$, $p < 1 \times 10^{-26}$). (C) Gestational age is not strongly correlated with any variable other than birthweight.

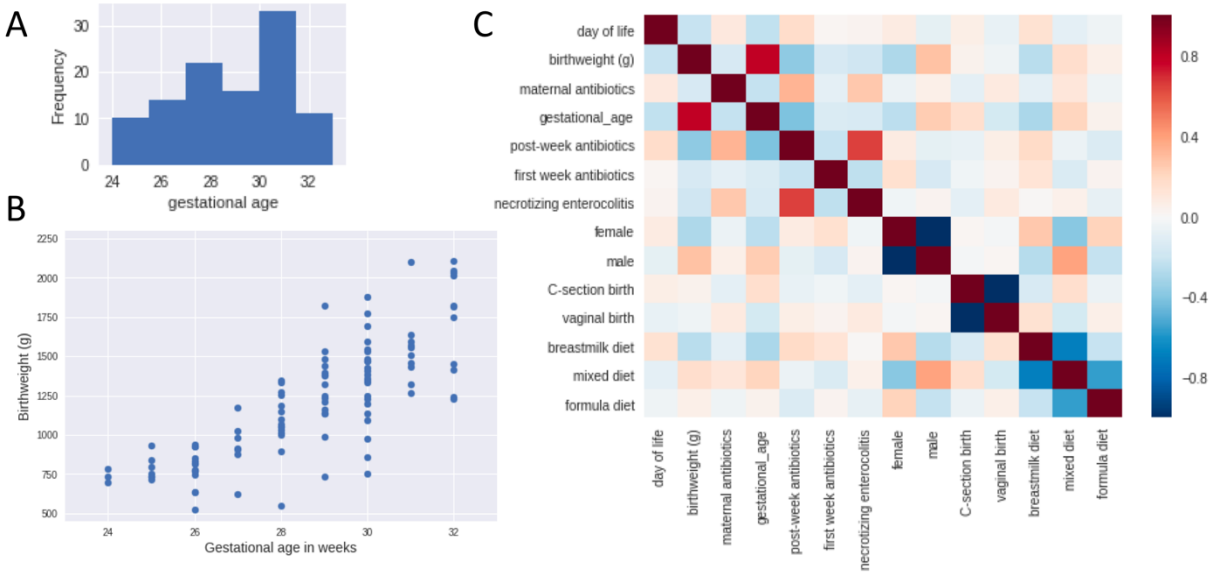

Figure 3.2. Mann-Whitney U tests were performed to compare species richness in infants with < 28 week gestational age and infants >= 28 week gestational age, at each week of life. Two asterisks indicate statistical significance at $p < 0.005$, while 3 asterisks indicate statistical significance at $p < 0.0005$. Error bars represent standard error of the mean. The number of samples in each week for infants < 28 weeks are as follows: week 2 $n = 79$, week 3 $n = 91$, week 4 $n = 79$, week 5 $n = 13$, week 6 $n = 14$, week 7 $n = 5$, week 8 $n = 12$. The number of samples in each week for infants ≥ 28 weeks are as follows: week 2 $n = 190$, week 3 $n = 170$, week 4 $n = 85$, week 5 $n = 29$, week 6 $n = 22$, week 7 $n = 10$, week 8 $n = 10$.

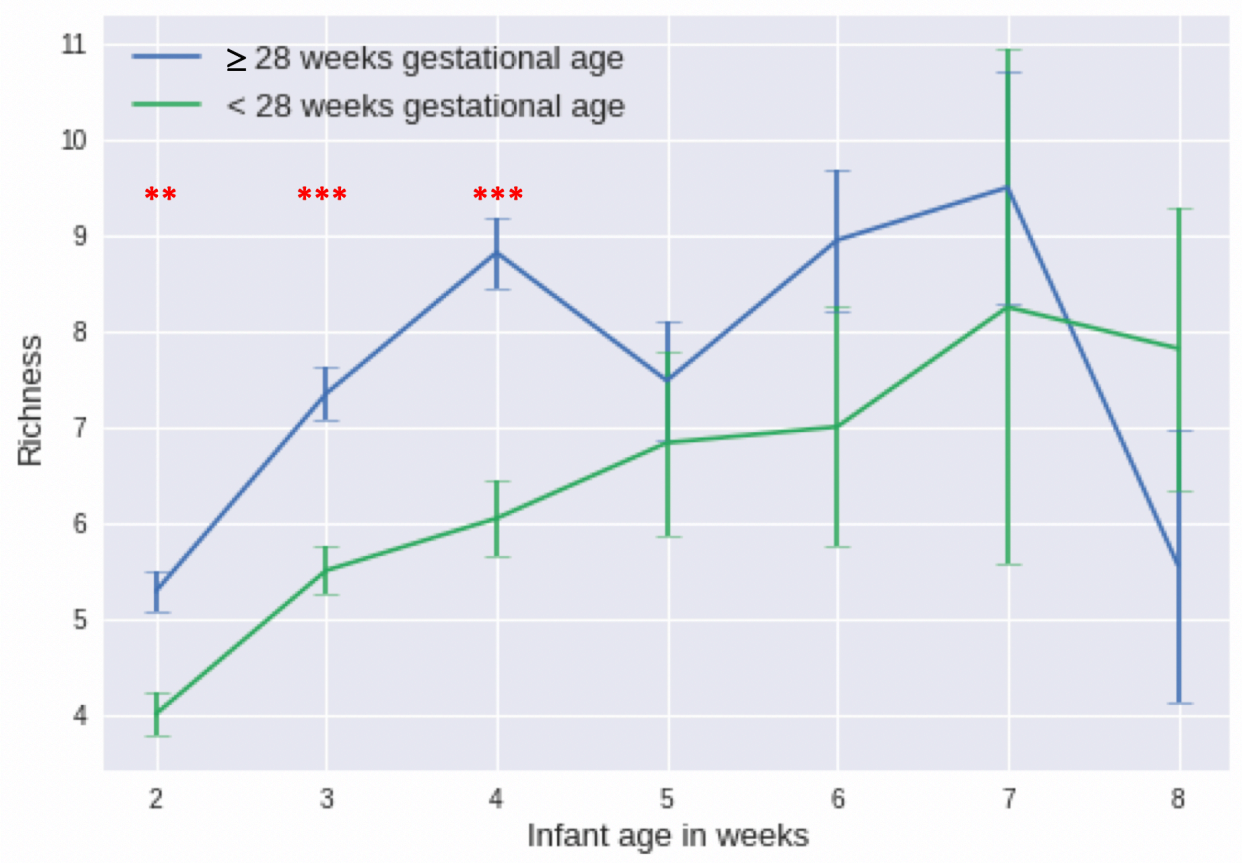

Figure 3.3. (A) The genus-level taxonomic composition of the gut community for the infants with < 28 week gestational age. The number of samples in each week for infants < 28 weeks are as follows: week 2 *n* = 79, week 3 *n* = 91, week 4 *n* = 79, week 5 *n* = 13, week 6 *n* = 14, week 7 *n* = 5, week 8 *n* = 12. (B) The genus-level taxonomic composition of the gut community for the infants with >= 28 week gestational age. The number of samples in each week for infants ≥ 28 weeks are as follows: week 2 *n* = 190, week 3 *n* = 170, week 4 *n* = 85, week 5 *n* = 29, week 6 *n* = 22, week 7 *n* = 10, week 8 *n* = 10.

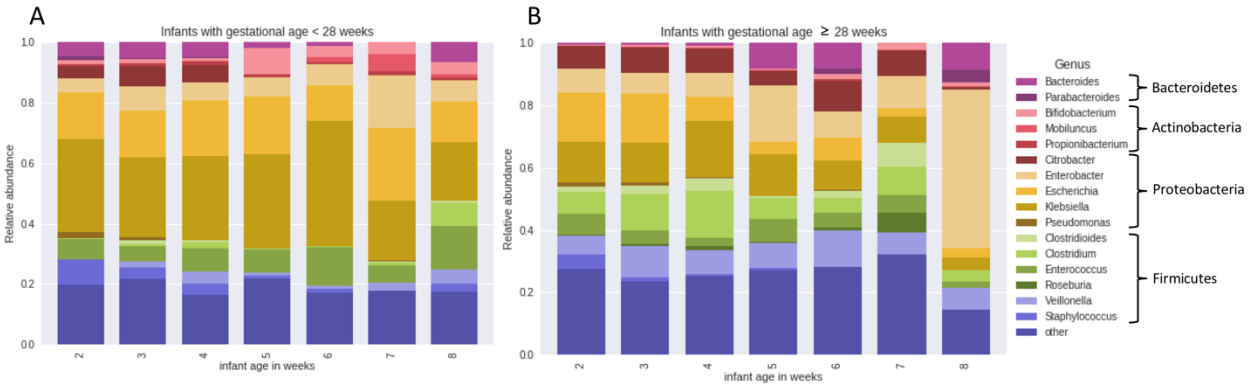

Figure 3.4. Trends in strain functional potential related to extent of prematurity. Each dot represents one genome, and the location of the dot on the y-axis indicates the gestational age of the infant in which the organism was found. Green dots indicate that the genome has the particular metabolic function listed in the title of the plot, and blue dots indicate that the genome lacks this function. (A) Genomes part of a novel group in *Clostridium* labeled as having or lacking genes for pantothenate biosynthesis. (B) *Streptococcus salivarius*-related strains labeled as having or lacking genes for an L-cystine transport system. (C) Strains of *Enterococcus faecalis* labeled as having or lacking genes for RaxAB-RaxC type 1 secretion system.

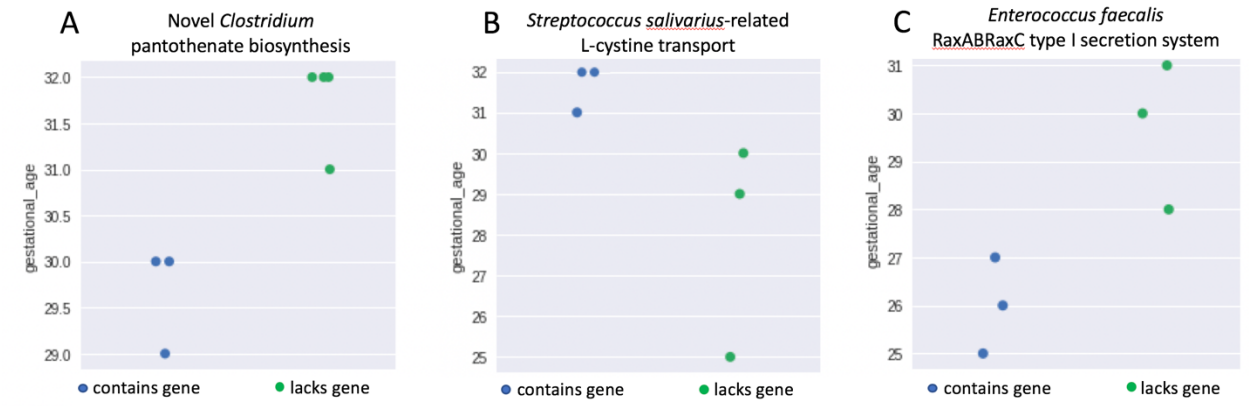

Table 3.1. Linear regression using infant metadata variables to predict species richness.

	Coefficient Estimate	**p-value**
intercept	-4.819817	0.999999
infant age (days)	0.096755	0.558479
gestational age (weeks)	0.385138	0.000000
birthweight (g)	-0.001211	0.979418
maternal antibiotics	-0.321744	0.215535
antibiotic exposure during first week	0.005125	0.915728
post-week antibiotic exposure	-1.427714	0.000000
male gender	0.098894	1
vaginal birth	0.141521	1
breastmilk	0.566750	1
formula	-1.014005	1
combination of breastmilk and formula	0.447255	1

Chapter 4

A new concept for the usage of genome functional potential in the quantification of community similarity

Introduction

Over the course of the past century, community ecology emerged as a prominent area of research. Although the focus was on macro-biological communities rather than microbes, the tools and methods developed during this time are still relevant as we are uncovering and understanding the systems invisible to the naked eye. Several techniques were developed to detect changes or differences in community structure (Anderson, 2006; Clarke, 1993), and methods were developed to determine exactly how similar or different two communities of organisms are. For many years, the Bray-Curtis dissimilarity index, which is based on counts of species or operational taxonomic units (OTUs) at each site, has been considered the standard in community ecology (D. P. Faith et al., 2010). More recently, new similarity measures have been introduced that take the phylogenetic relatedness of OTUs into account (Lozupone & Knight, 2005).

By sequencing all the environmental DNA of a sample taken at a particular site, we can also obtain information about the functional capabilities of microbes, which adds a new dimension to analysis of the community. Read-based metagenomics studies have used dissimilarity measures to evaluate community similarity based on the communities' metabolic potential (Forsberg et al., 2014). Yet, read-based metagenomics studies are lacking information regarding the organization of the sets of genes that individual organisms in the community are carrying. Genome-resolved metagenomics, on the other hand, reveals this information through the reconstruction of each genome in the metagenome. However, for genome-resolved communities, there is no established method of evaluating community similarity. In previous genome-resolved metagenomics studies, either functional or taxonomic data was used for evaluating similarity because there are no existing measures that combine these aspects.

In order to best represent a microbial community, the metabolic potential of each organism and the structure of the community should be included in the representation. By considering only phylogenetic information, the functional capabilities are not accounted for, because even organisms of the same species can have very different functional potential (Israel et al., 2001; Rasko et al., 2008). Similarly, by studying the genes in the environment without assigning them to genomes, two communities that appear to have similar functions may in reality be surprisingly different. For example, if an entire biosynthetic pathway composed of ten genes is carried by one organism, there is a higher likelihood that the compound is actually being synthesized than if each of the ten genes were in ten separate organisms. Read-based metagenomics may not be able to differentiate between the two situations described. This chapter describes development and initial testing of a new concept that aims to utilize genome-resolved metagenomic data to more accurately calculate community similarity.

Methods and Results

The usage of both phylogenetic and metabolic information for quantitative comparison of microbial communities was explored. In order to create a standard at which the proposed method and initial results could be evaluated, a test dataset using human infant gut samples was developed. A flowchart summarizing the test dataset development process is shown in Figure 4.1. The

sequencing and metagenomic data processing pipeline for these samples, including assembly, read mapping, and genome binning was described in chapters two and three of this dissertation. Centrifuge (Kim, Song, Breitwieser, & Salzberg, 2016) was used for taxonomy assignment. In previous studies, the metabolic potential of each genome was measured by using profile hidden Markov models (HMMs) to compare sequences against the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000). For this study, an additional sample module completeness profile using the KEGG annotation data was calculated; this sample profile indicates the fraction of a particular KEGG module that is carried by a sample as whole. If a genome displayed 100% coverage breadth, it was considered to be present in a sample, and its metabolism was thus considered as a component of the whole sample metabolism.

The sample module completeness profiles were clustered using density-based spatial clustering of applications with noise (DBSCAN) (Ester, Hans-Peter, Jorg, & Xiaowei, 2010). Various clustering parameters were tested and indicator species analysis (ISA) (Dufrene & Legendre, 1997) was performed on each of the clustering schemes to determine which parameters resulted in the strongest sample indicators. For this application of ISA, the indicators were KEGG modules rather than species. After evaluation and subsequent selection of DBSCAN parameters, the minimum number of samples for a cluster was 16 and eps value (which refers to the maximum allowed distance for two samples in the same cluster) was 0.11. This resulted in two clusters of samples and some samples classified as noise, i.e. not belonging to either cluster (Figure 4.2A). To understand what factors most strongly influenced the clustering, the point-biserial correlation between the continuous metadata variables (such as infant age in days, gestational age at birth, days of antibiotic exposure, etc.) and the cluster that the sample fell into was calculated. There were no significant correlations. A chi-square contingency test between the categorical metadata variables and the cluster of a sample was performed, and it revealed that (1) maternal disease/antibiotic exposure, (2) infant disease/post-week antibiotic exposure, and (3) the specific infant that provided the sample had a significant influence on the clustering; the birth mode (vaginal vs. C-section) had an effect that approached significance; and the gender of the infant and feeding type did not have an influence (Table 4.1, Figure 4.2B).

ISA revealed which modules best represent a shared sample function for the group of samples in the same cluster. This is based on two measures: exclusivity (it is exclusively present in that cluster) and fidelity (it occurs in all samples within that cluster). These measures are combined into one statistic that represents the extent to which a particular module should be considered an indicator module of that cluster. To identify the samples within a cluster that are most similar (and are thus the best representatives of that cluster), the indicator modules in the top fifth percentile of that cluster were selected and considered as the top indicators for the cluster, and then samples with the following properties were selected: they have all the top indicators of their own cluster and none of the top indicators of the other cluster. This resulted in a test dataset in which samples in one cluster (hereby referred to as Cluster 1) are very similar to each other and are distinctly different than the samples in Cluster 2, and vice versa.

This test dataset was utilized to investigate how a microbial community could be accurately represented and how changes or differences between communities can be measured. Many avenues were explored and the process is still ongoing. One of these avenues, the results of which are described here, relates to the calculation of community similarity using both functional and phylogenetic information. For each sample in the test dataset, a matrix of species and modules was generated. This matrix represents which species are carried in a particular sample, and if a species is present, what functional modules are harbored by the genome. Traditional distance or dissimilarity measures (e.g. Euclidean or Bray-Curtis) cannot be applied to data in 2-D matrix form, as 1-D vectors are required for these methods. Therefore, alternatives were investigated.

Procrustes analysis is a type of geometric analysis typically used for the comparison of shapes. In this method, shapes are optimally superimposed. Generating species-module matrices for each of the samples as in the method described above creates representations that can be perfectly overlaid and then compared by matching corresponding points on the matrices. As in Procrustes analysis, least-squares orthogonal mapping can be performed to determine the difference between the datasets. The disparity at each point in the dataset can be calculated and then summed to produce a dissimilarity score between 0 and 1. With a greater number of points in the input matrix, an increased dissimilarity score is expected. Because our sample-module matrices have 97,012 points each and the sparseness of the matrices creates issues for Procrustes analysis, truncated Singular Value Decomposition (SVD) was used to create a reduced rank approximation of each matrix, and these approximations were used in the calculations. Combinations for all possible pairs of samples within clusters and between clusters were generated, and twenty difference subsets of these pairs were randomly selected to use as replicates. Procrustes analysis was performed on each pair of samples using the low-rank approximations of the matrices, and this was done for all subsets. The mean dissimilarity score for samples in different clusters was 0.79, and the mean dissimilarity score for samples in the same cluster was 0.38 for Cluster 1 and 0.31 for Cluster 2 (Figure 4.3A). The Bray-Curtis distance between samples based on their metabolism using the KEGG sample module completeness profiles was also calculated; using this method, the mean dissimilarity score for samples in different clusters was 0.30, and the mean dissimilarity score for samples in the same cluster was 0.15 for both Cluster 1 and Cluster 2 (Figure 4.3B). Finally, the Bray-Curtis distance between samples based on the counts of the species that were present or absent in a particular sample was calculated; using this method, the mean dissimilarity score for samples in different clusters was 0.88, and the mean dissimilarity score for samples in the same cluster was 0.62 for Cluster 1 and 0.78 for Cluster 2 (Figure 4.3C).

Discussion

This report describes initial steps of exploration of a new concept in ecology that utilizes the unique nature of genome-resolved data to quantitatively compare microbial communities based on both phylogenetic structure and metabolic potential. To investigate this concept, a test set was developed using unsupervised learning and repurposing of a traditional ecological approach (ISA), illustrating the value of statistical learning methods in microbial ecology research. Samples from the same infant frequently fell into the same cluster group in the test set, which was expected. Other than the infants themselves, the strongest factors determining the grouping were intrapartum maternal antibiotic exposure and infant antibiotic exposure due to occurrence of disease after the first week of life (Figure 4.2B). This suggests that antibiotic administration may induce major changes in the metabolic potential of a microbial community that are even more expansive than changes in just antibiotic resistance levels as has been shown previously (Jernberg et al., 2007).

The genome-resolved microbial communities were represented as matrices, and these matrices were compared using least-squares orthogonal mapping. One method for creating lower-dimensional representations of these matrices (SVD) was utilized, but the many possible alternatives (e.g. autoencoders) also have potential as useful techniques for improving the matrix representation. The results indicated that Procrustes analysis applied to a combined approximation of metabolism and taxonomy data was better able to discern samples in different clusters (representing highly "different" communities) than Bray-Curtis distance of sample metabolisms, as evidenced by the higher dissimilarity scores resulting from the former (Figure 4.3A and Figure 4.3B). This suggests that microbial communities can be better differentiated from each other when community structure is accounted for in addition to metabolism. In both the aforementioned

methods, the between-cluster dissimilarity score was significantly higher than the within-cluster dissimilarity scores (Figure 4.3A and Figure 4.3B), confirming the validity of the test. However, the Bray-Curtis distance calculations based on the counts of the species that were present or absent in samples led to fairly high dissimilarity scores even within clusters (Figure 4.3C), illustrating the method's inability to recognize similar metabolisms that can arise in communities of dissimilar taxonomic makeups. Overall, it appears that the new method using both metabolism and taxonomy may be a middle ground that recognizes major differences between communities while still acknowledging function-based similarity. Because the clusters were significantly influenced by maternal and infant antibiotic exposure (Figure 4.2B), the results from the application of this method strongly suggest that the administration of antibiotics influences the microbial community both taxonomically and functionally.

Nevertheless, this new method has clear weaknesses. Although species were used as a grouping in the matrices so that the sample datasets could be consistently overlaid, there is substantial reason to consider the species an inappropriate partitioning of communities (Fraser, Alm, Polz, Spratt, & Hanage, 2009; Staley, 2006). Alternatives that use naturally hierarchical rank-free representations have been proposed (Tikhonov, 2015), and these may be more appropriate for usage in similarity calculation methods. Moreover, the matrices were generated based on the presence or absence of organisms, rather than their relative abundance. By using coverage data obtained from read mapping of genomes across metagenomes, relative abundance information can be calculated and then applied as a multiplicative factor to the metabolisms carried by an organism, which would result in a more accurate representation of the community than one based on the presence/absence of organisms.

Similarity calculation just scratches the surface of the host of possibilities for quantitative analysis of genome-resolved communities. Methods that identify the particular genomes carrying functions that are characteristic of a set of samples have substantial utility in research studies of microbial systems. Genome-resolved metagenomics creates a new form of biological data, but there are few established methods for analysis of this hierarchical data structure. This report makes a first attempt at combining a gene-centric approach with a phylogenetic approach to better analyze genome-resolved metagenomic data.

Acknowledgements

This work involved several important contributions from Alex Thomas, who was integral to this project. Additionally, thanks goes to Matt Olm for sequence data processing and David Burstein for HMM annotation pipelines.

Figure 4.1. Process for development of a test set of similar and distinct microbial communities.

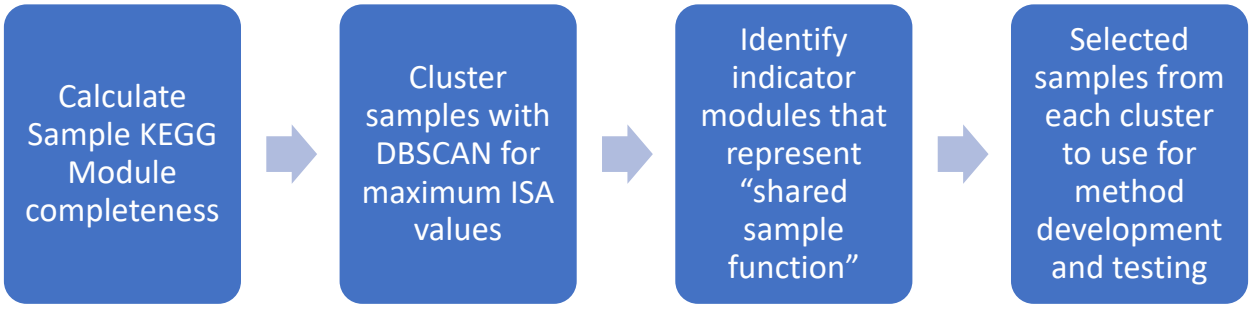

Figure 4.2. (A) Principal component analysis was used to visualize how the samples were clustered by DBSCAN. (B) False discovery rate corrected p-values resulting from chi-square contingency tests of metadata variables and sample clusters reveal that clusters were influenced by maternal antibiotic exposure, infant antibiotic exposure, and the specific infant that provided the sample. The orange dashed line represents the $p = 0.05$ significance threshold.

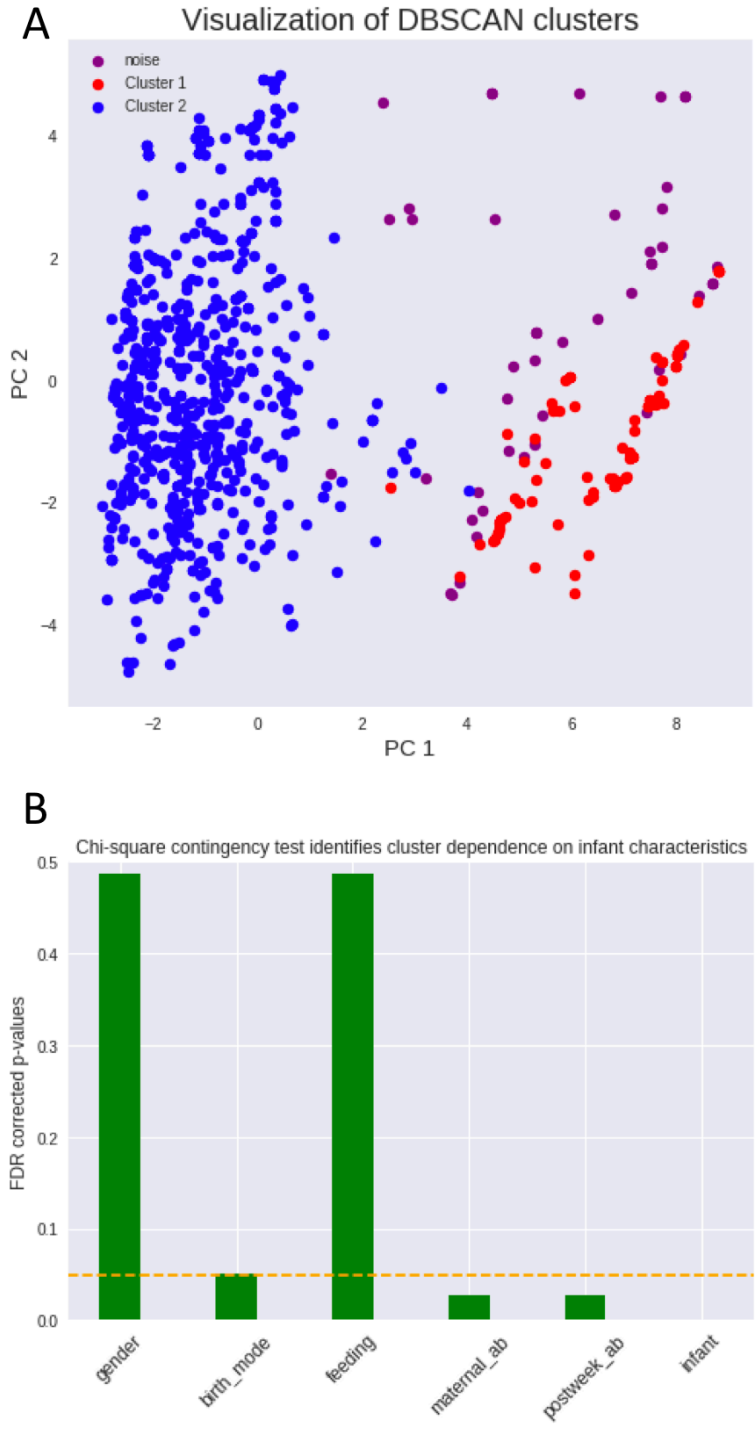

Figure 4.3. Calculated sample dissimilarity with (A) the Procrustes method using metabolism and taxonomy, (B) the Bray-Curtis method using metabolism, and (C) the Bray-Curtis method using taxonomy.

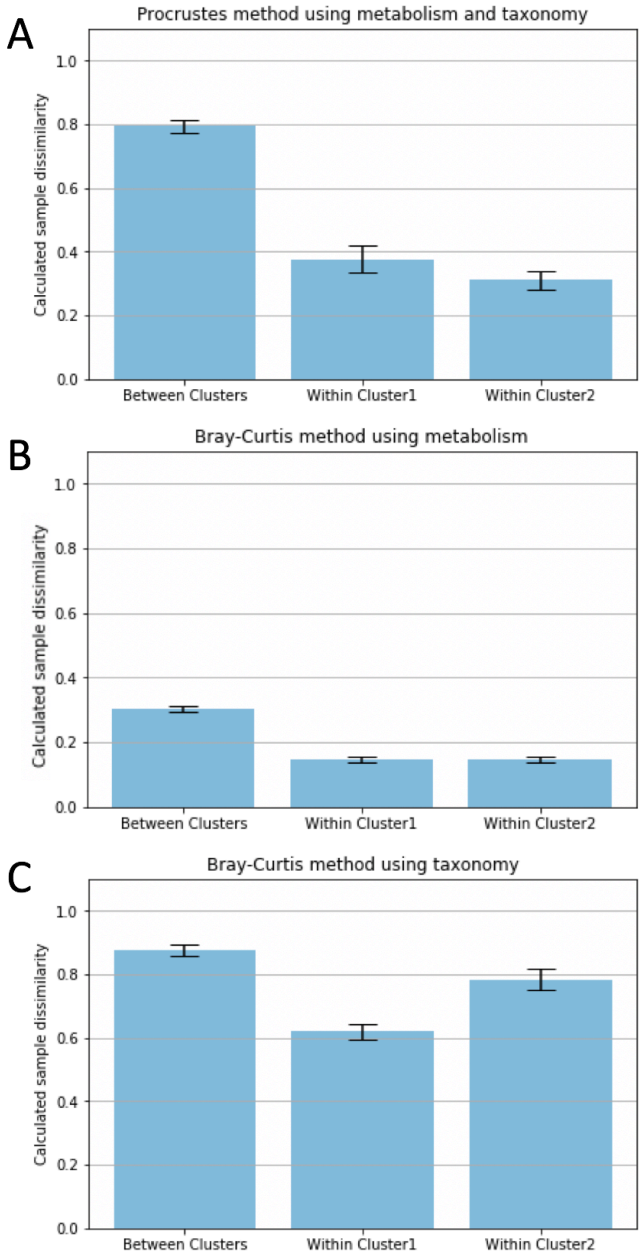

Table 4.1. Results of chi-square contingency tests for categorical metadata variables and sample clusters. P-values are FDR corrected.

	chi_square_statistic	p-values
gender	0.481339	4.878163e-01
birth_mode	4.511499	5.050160e-02
feeding	1.516498	4.878163e-01
maternal_ab	6.089120	2.720336e-02
postweek_ab	6.261633	2.720336e-02
infant	395.846830	1.198883e-36

Conclusion

This dissertation, focused on the development and application of quantitative methods for the analysis of genome-resolved metagenomic data, covers new ground on both the biological and methodological side. The biological conclusions resulting from this research have industrial and clinical applications. The study on the microbial community of a bioreactor used for bioremediation of thiocyanate that contains solid particulate tailings suggests that the presence of the solids prevents biofilm community formation, which may result in a process with reduced resilience to perturbations. Although meaningful conclusions can be obtained from this "small data" study, the transition to "big data" requires a change from manual analysis of metagenome-associated genomes to computational and statistical analysis that can glean patterns of interest. The analysis of approximately one thousand samples of the premature infant gut (>4 terrabases of sequence data) reveals that formula feeding selects for antibiotic resistant bacterial strains and that gestational age at birth is a strong predictor of gut community diversity, among other findings.

On the methodological side, this dissertation makes advances in the area of applied statistical learning and illustrates the first instance that a machine learning method was used to predict the future state of gut microbes. This chapter was recommended in F1000 prime as being of special significance in its field and was selected as an *mSystems* "Editor's Pick." In the aforementioned study, the metagenomes were collected in a time series design, which is a requirement for this application. The ability to predict how a particular organism will respond to antibiotics, or any external factor, has significant value if samples are sequenced on a clinically relevant timescale, which may be the case in the near future. In a more general sense, the unique form of genome-resolved metagenomic data allows for greater information gain than other methods of community analysis, such as the taxonomic identification of organisms present in a community (e.g. as resulting from 16S rRNA gene sequencing) or functional potential of the community (e.g. as revealed by read-based metagenomics). This dissertation introduces a new concept to best utilize genome-resolved metagenomic data: quantitatively evaluating community dissimilarity based on both functional and phylogenetic information. This represents a shift in the way that ecological measures are typically performed and has the potential to start a new paradigm in community ecology.

In conclusion, the combination of genome-resolved metagenomics with statistical learning and other quantitative techniques can lead to a better understanding of biological systems and potentially useful applications. The convergence of metagenomics and quantitative analysis of biological systems is still a relatively new area of research and is likely to be an area in which major advances are made in the coming years.

References

Albuquerque, L., Simoes, C., Nobre, M. F., Pino, N. M., Battista, J. R., Silva, M. T., … Da Costa, M. S. (2005). Truepera radiovictrix gen. nov., sp. nov., a new radiation resistant species and the proposal of Trueperaceae fam. nov. *FEMS Microbiology Letters*, *247*(2), 161–169. https://doi.org/10.1016/j.femsle.2005.05.002

Alneberg, J., Bjarnason, B. S., Bruijn, I. De, Schirmer, M., Quick, J., Ijaz, U. Z., … Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), 1144–1154. https://doi.org/10.1038/nmeth.3103

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., … Banfield, J. F. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, *7*, 1–11. https://doi.org/10.1038/ncomms13219

Anderson, M. J. (2006). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*, 32–46.

Anderson, M. J., & Walsh, D. C. I. (2013). PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs*, *83*(4), 557–574. https://doi.org/10.1890/12-2010.1

Arakawa, T., Kawano, Y., Kataoka, S., Katayama, Y., Kamiya, N., Yohda, M., & Odaka, M. (2007). Structure of Thiocyanate Hydrolase: A New Nitrile Hydratase Family Protein with a Novel Five-coordinate Cobalt(III) Center. *Journal of Molecular Biology*, *366*(5), 1497–1509. https://doi.org/10.1016/j.jmb.2006.12.011

Arboleya, S., Borja, S., Milani, C., Duranti, S., Solis, G., Fernandez, N., … Gueimonde, M. (2015). Intestinal Microbiota Development in Preterm Neonates and Effect of Perinatal Antibiotics. *Journal of Pediatrics*, *166*(3), 538–544. https://doi.org/10.1016/j.jpeds.2014.09.041

Arthur, M., & Quintiliani, R. (2001). Regulation of VanA- and VanB-Type Glycopeptide Resistance in Enterococci. *Antimicrobial Agents and Chemotherapy*, *45*(2), 375–381. https://doi.org/10.1128/AAC.45.2.375

Aymerich, T., Holo, H., Håvarstein, L. S., Hugas, M., Garriga, M., & Nes, I. F. (1996). Biochemical and genetic characterization of enterocin A from Enterococcus faecium, a new antilisterial bacteriocin in the pediocin family of bacteriocins. *Applied and Environmental Microbiology*, *62*(5), 1676–1682. https://doi.org/

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., … Wang, J. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life (Cell Host and Microbe (2015) 17(5) (690–703)). *Cell Host and Microbe*, *17*(6), 852. https://doi.org/10.1016/j.chom.2015.05.012

Berziņš, B., & Pejler, B. (1987). Rotifer occurrence in relation to pH. In L. May, R. Wallace, & A. Herzig (Eds.), *Hydrobiologia* (Vol. 147, pp. 107–116). Dordrecht: Springer Netherlands. https://doi.org/10.1007/BF00025733

Boening, D. W., & Chew, C. M. (1999). A critical review: General toxicity and environmental fate of three aqueous cyanide ions and associated ligands. *Water, Air, and Soil Pollution*, *109*(1–4), 67–79. https://doi.org/10.1023/A:1005005117439

Boucabeille, C., Bories, a., Ollivier, P., & Michel, G. (1994). Microbial degradation of metal complexed cyanides and thiocyanate from mining wastewaters. *Environmental Pollution*,

84(1), 59–67. https://doi.org/10.1016/0269-7491(94)90071-X
Boyce, J. M. (1997). VANCOMYCIN-RESISTANT ENTEROCOCCUS - Detection, epidemiology and control measures. *Infect Dis Clin North Am*, *11*(2), 367–384. https://doi.org/10.1016/S0891-5520(05)70361-5
Brooks, B., Firek, B. A., Miller, C. S., Sharon, I., Thomas, B. C., Baker, R., … Banfield, J. F. (2014). Microbes in the neonatal intensive care unit resemble those found in the gut of premature infants. *Microbiome*, *2*(1), 1. https://doi.org/10.1186/2049-2618-2-1
Brooks, B., Olm, M. R., Firek, B. A., Baker, R., Thomas, B. C., Morowitz, M. J., & Banfield, J. F. (2017). Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nature Communications*, *8*(1), 1–7. https://doi.org/10.1038/s41467-017-02018-w
Brown, C. T., Olm, M. R., Thomas, B. C., & Banfield, J. F. (2016). Measurement of bacterial replication rates in microbial communities. *Nature Biotechnology*, *34*(12), 1256–1263. https://doi.org/10.1101/057992
Cech, D., Markin, K., & Ronald, W. (2017). Identification of a D-arabinose-5-phosphate isomerase in the Gram-positive Clostridium tetani. *Journal of Bacteriology*. https://doi.org/10.1128/JB.00246-17
Chernikova, D. A., Madan, J. C., Housman, M. L., Zain-ul-abideen, M., Lundgren, S. N., Morrison, H. G., … Hoen, A. G. (2018). The premature infant gut microbiome during the first 6 weeks of life differs based on gestational maturity at birth. *Pediatric Research*, *84*(1), 71–79. https://doi.org/10.1038/s41390-018-0022-z
Chu, D. M., Ma, J., Prince, A. L., Antony, K. M., Seferovic, M. D., & Aagaard, K. M. (2017). Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nature Medicine*, *23*(3), 314–326. https://doi.org/10.1038/nm.4272
Clark, R., Bloom, B., Spitzer, A. R., & Gerstmann, D. R. (2014). Reported Medication Use in the Neonatal Intensive Care Unit: Data From a Large National Data Set. *Pediatrics*, *29*(1), 18–20. https://doi.org/10.1016/j.ijpharm.2014.03.004
Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, *18*(117–143).
Cong, X., Xu, W., Janton, S., Henderson, W. A., & Matson, A. (2016). Gut Microbiome Developmental Patterns in Early Life of Preterm Infants : Impacts of Feeding and Gender. *PLoS ONE*, *11*(4), e0152751. https://doi.org/10.1371/journal.pone.0152751
Costello, E. K., Carlisle, E. M., Bik, E. M., Morowitz, M. J., & Relman, D. A. (2013). Microbiome Assembly across Multiple Body Sites in Low-Birthweight Infants. *MBio*, *4*(6). Retrieved from http://mbio.asm.org/content/4/6/e00782-13.abstract
Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M., & Relman, D. A. (2012). The Application of Ecological Theory Toward an Understanding of the Human Microbiome. *Science*, *336*(June), 1255–1263.
da Silva, F. G., Shen, Y., Dardick, C., Burdman, S., Yadav, R. C., de Leon, A. L., & Ronald, P. C. (2004). Bacterial Genes Involved in Type I Secretion and Sulfation Are Required to Elicit the Rice *Xa21* -Mediated Innate Immune Response. *Molecular Plant-Microbe Interactions*, *17*(6), 593–601. https://doi.org/10.1094/MPMI.2004.17.6.593
Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., & Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biology*, *10*(8), R85. https://doi.org/10.1186/gb-2009-10-8-r85
DiGiulio, D. B., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., … Relman, D. A. (2015). Temporal and spatial variation of the human microbiota during

pregnancy. *Proceedings of the National Academy of Sciences*, *112*(35), 11060–11065. https://doi.org/10.1073/pnas.1502875112
Dixon, P. (2003). Computer program review VEGAN , a package of R functions for community ecology. *Journal of Vegetation Science*, *14*(6), 927–930.
Du Plessis, C. A., Barnard, P., Muhlbauer, R. M., & Naldrett, K. (2001). Empirical model for the autotrophic biodegradation of thiocyanate in an activated sludge reactor. *Letters in Applied Microbiology*, *32*(2), 103–107. https://doi.org/10.1046/j.1472-765X.2001.00859.x
Dufrene, M., & Legendre, P. (1997). Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical Approach. *Ecological Monographs*, *67*(3), 345–366. https://doi.org/10.2307/2963459
Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340
Erdogan, M. F. (2003). Thiocyanate overload and thyroid disease. *BioFactors (Oxford, England)*, *19*(3–4), 107–111. https://doi.org/10.1002/biof.5520190302
Ester, M., Hans-Peter, K., Jorg, S., & Xiaowei, X. (2010). Density-Based Clustering Algorithms for Discovering Clusters. *Comprehensive Chemometrics*, *2*, 635–654. https://doi.org/10.1016/B978-044452701-1.00067-3
Faith, D. P., Minchin, P. R., Belbin, L., Faith, P., Minchin, P. R., & Box, G. P. O. (2010). Compositional Dissimilarity as a Robust Measure of Ecological Distance Stable URL : http://www.jstor.org/stable/20038103 Compositional dissimilarity as a robust measure of ecological distance, *69*(1), 57–68.
Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., … Gordon, J. I. (2013). The Long-Term Stability of the Human Gut Microbiota. *Science*, *341*(6141), 1237439. https://doi.org/10.1126/science.1237439
Falsetta, M. L., McEwan, A. G., Jennings, M. P., & Apicella, M. A. (2010). Anaerobic metabolism occurs in the substratum of gonococcal biofilms and may be sustained in part by nitric oxide. *Infection and Immunity*, *78*(5), 2320–2328. https://doi.org/10.1128/IAI.01312-09
Felföldi, T., Székely, A. J., Gorál, R., Barkács, K., Scheirich, G., András, J., … Márialigeti, K. (2010). Polyphasic bacterial community analysis of an aerobic activated sludge removing phenols and thiocyanate from coke plant effluent. *Bioresource Technology*, *101*(10), 3406–3414. https://doi.org/10.1016/j.biortech.2009.12.053
Ferraris, L., Butel, M. J., Campeotto, F., Vodovar, M., Rozé, J. C., & Aires, J. (2012). Clostridia in premature neonates' gut: Incidence, antibiotic susceptibility, and perinatal determinants influencing colonization. *PLoS ONE*, *7*(1). https://doi.org/10.1371/journal.pone.0030594
Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server : interactive sequence similarity searching. *Nucleic Acids Research*, *39*(May), 29–37. https://doi.org/10.1093/nar/gkr367
Forsberg, K. J., Patel, S., Gibson, M. K., Lauber, C. L., Knight, R., Fierer, N., & Dantas, G. (2014). Bacterial phylogeny structures soil resistomes across habitats. *Nature*, *509*(7502), 612–616. https://doi.org/10.1038/nature13377
Fouhy, F., Guinane, C. M., Hussey, S., Wall, R., Ryan, C. A., Dempsey, E. M., … Cotter, P. D. (2012). High-throughput sequencing reveals the incomplete, short-term recovery of infant gut microbiota following parenteral antibiotic treatment with ampicillin and gentamicin. *Antimicrobial Agents and Chemotherapy*, *56*(11), 5811–5820. https://doi.org/10.1128/AAC.00789-12
Fox, E. P., Cowley, E. S., Nobile, C. J., Hartooni, N., Newman, D. K., & Johnson, A. D. (2014). Anaerobic bacteria grow within candida albicans biofilms and induce biofilm formation in

suspension cultures. *Current Biology*, *24*(20), 2411–2416. https://doi.org/10.1016/j.cub.2014.08.057

Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G., & Hanage, W. P. (2009). The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity. *Science*, *323*(5915), 741 LP-746. https://doi.org/10.1126/science.1159388

Gibson, M. K., Forsberg, K. J., & Dantas, G. (2014). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME Journal*, *9*(1), 207–216. https://doi.org/10.1038/ismej.2014.106

Gibson, M. K., Wang, B., Ahmadi, S., Burnham, C.-A. D., Tarr, P. I., Warner, B. B., & Dantas, G. (2016). Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nature Microbiology*, *1*(March), 16024. https://doi.org/10.1038/nmicrobiol.2016.24

Glass, H. C., Andrew, C. T., Stayer, S. A., Brett, C., Cladis, F., & Davis, P. J. (2016). Outcomes for Extremely Premature Infants. *Anesth Analg*, *120*(6), 1337–1351. https://doi.org/10.1213/ANE.0000000000000705.Outcomes

Gominak, S. C. (2016). Vitamin D deficiency changes the intestinal microbiome reducing B vitamin production in the gut. The resulting lack of pantothenic acid adversely affects the immune system, producing a "pro-inflammatory" state associated with atherosclerosis and autoimmun. *Medical Hypotheses*, *94*, 103–107. https://doi.org/10.1016/j.mehy.2016.07.007

Goossens, H., Ferech, M., Vander Stichele, R., & Elseviers, M. (2005). Outpatient antibiotic use in Europe and assaciation with resistance: a cross-national database study. *The Lancet*, *365*(9459), 579–587.

Greenwood, C., Morrow, A. L., Lagomarcino, A. J., Altaye, M., Taft, D. H., Yu, Z., … Schibler, K. R. (2014). Early empiric antibiotic use in preterm infants is associated with lower bacterial diversity and higher relative abundance of enterobacter. *Journal of Pediatrics*, *165*(1), 23–29. https://doi.org/10.1016/j.jpeds.2014.01.010

Gurmeet, S. (2017). Vitamin D levels in preterm and term neonates at birth. *International Journal of Contemporary Pediatrics*, *4*(1), 48–52. Retrieved from http://www.ijpediatrics.com/index.php/ijcp/article/view/80

Hansen, L. H., Jensen, L. B., Sørensen, H. I., & Sørensen, S. J. (2007). Substrate specificity of the OqxAB multidrug resistance pump in Escherichia coli and selected enteric bacteria. *Journal of Antimicrobial Chemotherapy*, *60*(May), 145–147. https://doi.org/10.1093/jac/dkm167

Hibberd, M. C., Wu, M., Rodionov, D. A., Li, X., Cheng, J., Griffin, N. W., … Gordon, J. I. (2017). The effects of micronutrient deficiencies on bacterial species from the human gut microbiota. *Science Translational Medicine*, *9*(390). https://doi.org/10.1126/scitranslmed.aal4069

Huddy, R. J., Van Zyl, A. W., Van Hille, R. P., & Harrison, S. T. L. (2015). Characterisation of the complex microbial community associated with the ASTER^{TM} thiocyanate biodegradation system. *Minerals Engineering*, *76*, 65–71. https://doi.org/10.1016/j.mineng.2014.12.011

Hughes, D. (2003). Exploiting genomics, genetics and chemistry to combat antibiotic resistance. *Nature Reviews Genetics*, *4*(June), 432–441. https://doi.org/10.1038/nrg1084

Hyatt, D., Chen, G., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal : prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119.

Hyatt, D., Locascio, P. F., Hauser, L. J., & Uberbacher, E. C. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, *28*(17), 2223–2230.

https://doi.org/10.1093/bioinformatics/bts429
Illing, S., & Harrison, S. T. L. (1999). The kinetics and mechanism of Corynebacterium glutamicum aggregate breakup in bioreactors. *Chemical Engineering Science*, *54*(4), 441–454. https://doi.org/10.1016/S0009-2509(98)00253-X
Israel, D. A., Salama, N., Arnold, C. N., Moss, S. F., Ando, T., Wirth, H. P., … Peek, R. M. (2001). Helicobacter pylori strain-specific differences in genetic content, identified by microarray influence host inflammatory responses. *Journal of Clinical Investigation*, *107*(5), 611–620. https://doi.org/10.1172/JCI11450
Ivanova, N., Rohde, C., Munk, C., Nolan, M., Lucas, S., Del Rio, T. G., … Lapidus, A. (2011). Complete genome sequence of Truepera radiovictrix type strain (RQ-24). *Standards in Genomic Sciences*, *4*(1), 91–99. https://doi.org/10.4056/sigs.1563919
Jacoby, G. A., & Medeiros, A. A. (1991). More Extended-Spectrum Beta-Lactamases. *Antimicrobial Agents and Chemotherapy*, *35*(9), 1697–1704.
Jandhyala, S. M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M., & Reddy, D. N. (2015). Role of the normal gut microbiota. *World Journal of Gastroenterology*, *21*(29), 8836–8847. https://doi.org/10.3748/wjg.v21.i29.8787
Jernberg, C., Löfmark, S., Edlund, C., & Jansson, J. K. (2007). Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *The ISME Journal*, *1*, 56–66. https://doi.org/10.1038/ismej.2007.3
Kanehisa, M., & Goto, S. (2000). KEGG : Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30.
Kantor, R. S., Huddy, R. J., Iyer, R., Thomas, B. C., Brown, C. T., Anantharaman, K., … Banfield, J. F. (2017). Genome-Resolved Meta-Omics Ties Microbial Dynamics to Process Performance in Biotechnology for Thiocyanate Degradation. *Environmental Science and Technology*, *51*(5), 2944–2953. https://doi.org/10.1021/acs.est.6b04477
Kantor, R. S., van Zyl, a. W., van Hille, R. P., Thomas, B. C., Harrison, S. T. L., & Banfield, J. F. (2015). Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unraveled with genome-resolved metagenomics. *Environmental Microbiology*, *17*(12), 4929–4941. https://doi.org/10.1111/1462-2920.12936
Kataoka, S., Arakawa, T., Hori, S., Katayama, Y., Hara, Y., Matsushita, Y., … Odaka, M. (2006). Functional expression of thiocyanate hydrolase is promoted by its activator protein, P15K. *FEBS Letters*, *580*(19), 4667–4672. https://doi.org/10.1016/j.febslet.2006.07.051
Katayama, Y., Matsushita, Y., Kaneko, M., Kondo, M., Mizuno, T., & Nyunoya, H. (1998). Cloning of genes coding for the three subunits of thiocyanate hydrolase of Thiobacillus thioparus THI 115 and their evolutionary relationships to nitrile hydratase. *Journal of Bacteriology*, *180*(10), 2583–2589.
Katayama, Y., Narahara, Y., Inoue, Y., Amano, F., Kanagawa, T., & Kuraishi, H. (1992). A thiocyanate hydrolase of thiobacillus thioparus: A novel enzyme catalyzing the formation of carbonyl sulfide from thiocyanate. *Journal of Biological Chemistry*, *267*(13), 9170–9175.
Keski-nisula, L., Kyynäräinen, H.-R., Kärkkäinen, U., Karhukorpi, J., Heinonen, S., & Pekkanen, J. (2013). Maternal intrapartum antibiotics and decreased vertical transmission of Lactobacillus to neonates during birth. *Acta Pædiatrica*, *102*, 480–485. https://doi.org/10.1111/apa.12186
Khan, N. A. (2001). Acanthamoeba castellanii Cell Culture. In *eLS*. John Wiley & Sons, Ltd. https://doi.org/10.1038/npg.els.0002577
Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729. https://doi.org/10.1101/gr.210641.116

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., … Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, *108*(Supplement_1), 4578–4585. https://doi.org/10.1073/pnas.1000081107

Kramer, M. S., Demissie, K., Yang, H., Platt, R. W., Sauvé, R., & Liston, R. (2000). The contribution of mild and moderate preterm birth to infant mortality. Fetal and Infant Health Study Group of the Canadian Perinatal Surveillance System. *Jama*, *284*(7), 843–849. https://doi.org/joc00258 [pii]

Kumar, V., Sun, P., Vamathevan, J., Li, Y., Ingraham, K., Palmer, L., … Brown, J. R. (2011). Comparative genomics of Klebsiella pneumoniae strains with different antibiotic resistance profiles. *Antimicrobial Agents and Chemotherapy*, *55*(9), 4267–4276. https://doi.org/10.1128/AAC.00052-11

Laing, C., Buchanan, C., Taboada, E. N., Zhang, Y., Kropinski, A., Villegas, A., … Gannon, V. P. J. (2010). Pan-genome sequence analysis using Panseq : an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics*, *11*, 461.

Langdon, A., Crook, N., & Dantas, G. (2016). The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Medicine*, *8*(1), 39. https://doi.org/10.1186/s13073-016-0294-z

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

Li, X.-Z., & Nikaido, H. (2009). Efflux-Mediated Drug Resistance in Bacteria: an Update. *Drugs*, *69*(12), 1555–1623. https://doi.org/10.2165/11317030-000000000-00000.Efflux-Mediated

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*(5), 955–964. https://doi.org/10.1093/nar/25.5.955

Lozupone, C., & Knight, R. (2005). UniFrac : a New Phylogenetic Method for Comparing Microbial Communities, *71*(12), 8228–8235. https://doi.org/10.1128/AEM.71.12.8228

Makino, H., Kushiro, A., Ishikawa, E., Kubota, H., Gawad, A., Sakai, T., … Tanaka, R. (2013). Mother-to-infant transmission of intestinal bifidobacterial strains has an impact on the early development of vaginally delivered infant's microbiota. *PLoS ONE*, *8*(11). https://doi.org/10.1371/journal.pone.0078331

Marciano-Cabral, F. (1988). Biology of Naegleria spp. *Microbiological Reviews*, *52*(1), 114–133. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC372708/

McArdle, B., & Anderson, M. (2013). Fitting Multivariate Models to Community Data : A Comment on Distance-Based Redundancy Analysis. *Ecology*, *82*(1), 290–297.

Medema, M. H., Blin, K., Cimermancic, P., De Jager, V., Zakrzewski, P., Fischbach, M. A., … Breitling, R. (2011). AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*, *39*(SUPPL. 2), 339–346. https://doi.org/10.1093/nar/gkr466

Meredith, T. C., Aggarwal, P., Mamat, U., Lindner, B., & Woodard, R. W. (2006). Redefining the Requisite Lipopolysaccharide. *ACS Chemical Biology*, *1*(1), 33–42. https://doi.org/10.1021/cb0500015

Modi, S. R., Collins, J. J., & Relman, D. A. (2014). Antibiotics and the gut microbiota. *Journal of Clinical Investigation*, *124*(10), 4212–4218. https://doi.org/10.1172/JCI72333

Moles, L., Gómez, M., Jiménez, E., Fernández, L., Bustos, G., Chaves, F., … del Campo, R. (2015). Preterm infant gut colonization in the neonatal ICU and complete restoration 2 years later. *Clinical Microbiology and Infection*, *21*(10), 936.e1-936.e10.

https://doi.org/10.1016/j.cmi.2015.06.003
Morrow, A. L., Lagomarcino, A. J., Schibler, K. R., Taft, D. H., Yu, Z., Wang, B., … Newburg, D. S. (2013). Early microbial and metabolomic signatures predict later onset of necrotizing enterocolitis in preterm infants. *Microbiome*, *1*(1), 13. https://doi.org/10.1186/2049-2618-1-13
Mshvildadze, M., and Neu, J. (2010). The Infant Intestinal Microbiome: Friend or Foe? *Early Human Development*, *86*(Suppl 1), 67–71. https://doi.org/10.1016/j.earlhumdev.2010.01.018
Mshvildadze, M., Neu, J., Shuster, J., Theriaque, D., Li, N., & Mai, V. (2010). Intestinal Microbial Ecology in Premature Infants Assessed with Non-Culture-Based Techniques. *The Journal of Pediatrics*, *156*(1), 20–25. https://doi.org/10.1016/j.jpeds.2009.06.063
Mueller, N. T., Bakacs, E., Combellick, J., Grigoryan, Z., & Maria, G. (2015). The infant microbiome development: mom matters. *Trends Mol Med*, *21*(2), 109–117. https://doi.org/10.1016/j.molmed.2014.12.002.The
Nagayama, K., Fujita, K., Takashima, Y., Ardin, A. C., & Ooshima, T. (2014). Role of ABC Transporter Proteins in Stress Responses of Streptococcus mutans. *Oral Health Dent Manag*, *13*(2), 359–365.
Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. (2017). dRep: A tool for fast and accurate genome de-replication that enables tracking of microbial genotypes and improved genome recovery from metagenomes. *The ISME Journal*. Retrieved from http://biorxiv.org/content/early/2017/02/13/108142.abstract
Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/s13059-016-0997-x
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2
Penders, J., Stobberingh, E. E., Savelkoul, P. H. M., & Wolffs, P. F. G. (2013). The human microbiome as a reservoir of antimicrobial resistance. *Frontiers in Microbiology*, *4*(APR), 1–7. https://doi.org/10.3389/fmicb.2013.00087
Penders, J., Thijs, C., Vink, C., Stelma, F. F., Snijders, B., & Kummeling, I. (2006). Factors Influencing the Composition of the Intestinal Microbiota in Early Infancy. *Pediatrics*, *118*(2), 511–521. https://doi.org/10.1542/peds.2005-2824
Penders, J., Vink, C., Driessen, C., London, N., Thijs, C., & Stobberingh, E. E. (2005). Quantification of Bifidobacterium spp., Escherichia coli and Clostridium difficile in faecal samples of breast-fed and formula-fed infants by real-time PCR. *FEMS Microbiology Letters*, *243*, 141–147. https://doi.org/10.1016/j.femsle.2004.11.052
Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174
Périchon, B., & Courvalin, P. (2009). VanA-type vancomycin-resistant Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, *53*(11), 4580–4587. https://doi.org/10.1128/AAC.00346-09
Piddock, L. J. V. (2006). Multidrug-resistance efflux pumps — not just for resistance. *Nature Reviews Microbiology*, *4*(August), 629–636.
Poole, K. (2008). Bacterial Multidrug Efflux Pumps Serve Other Functions. *Microbe-American Society for Microbiology*, *3*(4), 179–185. https://doi.org/10.1111/j.1364-3703.2009.00558.x
Poole, K. (2014). Stress responses as determinants of antimicrobial resistance in Pseudomonas

aeruginosa: multidrug efflux and more. *Canadian Journal of Microbiology*, *60*(12), 783–791. https://doi.org/10.1139/cjm-2014-0666
Quan, Z. X., Rhee, S. K., Bae, J. W., Baek, J. H., Park, Y. H., & Lee, S. T. (2006). Bacterial community structure in activated sludge reactors treating free or metal-complexed cyanides. *Journal of Microbiology and Biotechnology*, *16*(2), 232–239.
Rahman, S. F., Olm, M. R., Morowitz, M. J., & Banfield, J. F. (2017). Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *BioRxiv*.
Rahman, S. F., Olm, M. R., Morowitz, M. J., & Banfield, J. F. (2018). Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *MSystems*, *3*(1). https://doi.org/10.1128/mSystems.00123-17
Rasko, D. A., Rosovitz, M. J., Myers, G. S. A., Mongodin, E. F., Fricke, W. F., Gajer, P., … Ravel, J. (2008). The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of Bacteriology*, *190*(20), 6881–6893. https://doi.org/10.1128/JB.00619-08
Raveh-Sadka, T., Firek, B., Sharon, I., Baker, R., Brown, C. T., Thomas, B. C., … Banfield, J. F. (2016). Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants. *The ISME Journal*, *10*(12), 2817–2830. https://doi.org/10.1038/ismej.2016.83
Raveh-Sadka, T., Thomas, B. C., Singh, A., Firek, B., Brooks, B., Castelle, C. J., … Banfield, J. F. (2015). Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. *ELife*, *2015*(4), e05477. https://doi.org/10.7554/eLife.05477
Relman, D. A. (2012). The human microbiome: ecosystem resilience and health. *Nutr Rev.*, *70*(Suppl 1), 1–12. https://doi.org/10.1111/j.1753-4887.2012.00489.x.The
Rodríguez, J. M., Murphy, K., Stanton, C., Ross, R. P., Kober, O. I., Juge, N., … Collado, M. C. (2015). The composition of the gut microbiota throughout life, with an emphasis on early life. *Microbial Ecology in Health & Disease*, *26*(0), 1–17. https://doi.org/10.3402/mehd.v26.26050
Rothfuss, H., Lara, J. C., Schmid, A. K., & Lidstrom, M. E. (2006). Involvement of the S-layer proteins Hpi and SlpA in the maintenance of cell envelope integrity in Deinococcus radiodurans R1. *Microbiology*, *152*(9), 2779–2787. https://doi.org/10.1099/mic.0.28971-0
Said, H. M. (2011). Intestinal absorption of water-soluble vitamins in health and disease. *Biochemical Journal*, *437*(3), 357 LP-372. Retrieved from http://www.biochemj.org/content/437/3/357.abstract
Schwiertz, A., Gruhl, B., Löbnitz, M., Michel, P., Radke, M., & Blaut, M. (2003). Development of the intestinal bacterial composition in hospitalized preterm infants in comparison with breast-fed, full-term infants. *Pediatric Research*, *54*(3), 393–399. https://doi.org/10.1203/01.PDR.0000078274.74607.7A
Shifrin, N. S., Beck, B. D., Gauthier, T. D., Chapnick, S. D., & Goodman, G. (1996). Chemistry, toxicology, and human health risk of cyanide compounds in soils at former manufactured gas plant sites. *Regulatory Toxicology and Pharmacology : RTP*, *23*(2), 106–116. https://doi.org/10.1006/rtph.1996.0032
Sim, K., Powell, E., Shaw, A. G., McClure, Z., Bangham, M., & Kroll, J. S. (2013). The neonatal gastrointestinal microbiota: the foundation of future health? *Archives of Disease in Childhood - Fetal and Neonatal Edition*, *98*(4), F362 LP-F364. Retrieved from http://fn.bmj.com/content/98/4/F362.abstract
Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015).

BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simonsen, G. S., Lvseth, A., Dahl, K. H., & Kruse, H. (1998). Enterococci and vanA Resistance Elements between Chicken. *Microbial Drug Resistance*, *4*(4), 313–318.

Spellberg, B., Guidos, R., Gilbert, D., Bradley, J., Boucher, H. W., Scheld, W. M., … Edwards, J. (2008). The Epidemic of Antibiotic-Resistant Infections: A Call to Action for the Medical Community from the Infectious Diseases Society of America. *Clinical Infectious Diseases*, *46*(2), 155–164. https://doi.org/10.1086/524891

Staley, J. T. (2006). The bacterial species dilemma and the genomic-phylogenetic species concept. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *361*(1475), 1899–1909. https://doi.org/10.1098/rstb.2006.1914

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

Stewart, C. J., Embleton, N. D., Clements, E., Luna, P. N., Smith, D. P., Fofanova, T. Y., … Cummings, S. P. (2017). Cesarean or Vaginal Birth Does Not Impact the Longitudinal Development of the Gut Microbiome in a Cohort of Exclusively Preterm Infants. *Frontiers in Microbiology*, *8*(June), 1–9. https://doi.org/10.3389/fmicb.2017.01008

Szarecka, A., Lesnock, K. R., Ramirez-Mondragon, C. A., Nicholas, H. B., & Wymore, T. (2011). The Class D beta-lactamase family: residues governing the maintenance and diversity of function. *Protein Engineering, Design & Selection*, *24*(10), 801–809. https://doi.org/10.1093/protein/gzr041

Talge, N. M., Mudd, L. M., Sikorskii, A., & Basso, O. (2014). United States Birth Weight Reference Corrected For Implausible Gestational Age Estimates. *Pediatrics*, *133*(5), 844–853. https://doi.org/10.1542/peds.2013-3285

Tanaka, S., Kobayashi, T., Songjinda, P., Tateyama, A., Tsubouchi, M., Kiyohara, C., … Nakayama, J. (2009). Influence of antibiotic exposure in the early postnatal period on the development of intestinal microbiota. *FEMS Immunol Med Microbiol*, *56*, 80–87. https://doi.org/10.1111/j.1574-695X.2009.00553.x

Teuber, M., Meile, L., & Schwarz, F. (1999). Acquired antibiotic resistance in lactic acid bacteria from food. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, *76*(1–4), 115–137. https://doi.org/10.1023/A:1002035622988

Tikhonov, M. (2015). Theoretical ecology without species. *Arxiv*, (APRIL 2015), 28–31. Retrieved from http://arxiv.org/abs/1504.0255

Ultsch, A., Moerchen, F. (2005). *ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM*.

Umu, Ö. C. O., Bäuerl, C., Oostindjer, M., Pope, P. B., Hernández, P. E., Pérez-Martínez, G., & Diep, D. B. (2016). The potential of class II bacteriocins to modify gut microbiota to improve host health. *PLoS ONE*, *11*(10), 1–22. https://doi.org/10.1371/journal.pone.0164036

Underwood, M. A., & Sohn, K. (2017). The Microbiota of the Extremely Preterm Infant. *Clinics in Perinatology*, *44*(2), 407–427. https://doi.org/10.1016/j.clp.2017.01.005

Valladares, M. H., Felici, A., Weber, G., Adolph, H. W., Zeppezauer, M., Rossolini, G. M., … Galleni, M. (1997). Zn(II) Dependence of the Aeromonas hydrophila AE036 Metallo-beta-lactamase Activity and Stability. *Biochemistry*, *36*(38), 11534–11541.

Van Braak, N. Den, Van Belkum, A., Van Keulen, M., Vliegenthart, J., Verbrugh, H. A., & Endtz, H. P. (1998). Molecular characterization of vancomycin-resistant enterococci from hospitalized patients and poultry products in the Netherlands. *Journal of Clinical*

Microbiology, *36*(7), 1927–1932.
van Buuren, C., Makhotla, N., & Olivier, J. W. (2011). The Aster Process: Technology Development through to Piloting, Demonstration and Commercialisation. *ALTA 2011 Nickel-Cobalt-Copper, Uranium and Gold Conference*, 236–253.
Van Zyl, A. W., Huddy, R., Harrison, S. T. L., & Van Hille, R. P. (2014). Evaluation of the ASTERTM process in the presence of suspended solids. *Minerals Engineering*, *76*, 72–80. https://doi.org/10.1016/j.mineng.2014.11.007
W.C., H., W., H., R.A., H., M.C., S., S., K., & R.B., D. (1988). Pediatric parenteral amino acid mixture in low birth weight infants. *Pediatrics*, *81*(1), 41–50. https://doi.org/10.1038/sj.bdj.2015.151
Wampach, L., Heintz-Buschart, A., Hogan, A., Muller, E. E. L., Narayanasamy, S., Laczny, C. C., … Wilmes, P. (2017). Colonization and succession within the human gut microbiome by archaea, bacteria and microeukaryotes during the first year of life. *Frontiers in Microbiology*, *8*, 738. https://doi.org/10.3389/FMICB.2017.00738
Wang, X., & Quinn, P. J. (2010). Progress in Lipid Research Lipopolysaccharide: Biosynthetic pathway and structure modification. *Progress in Lipid Research*, *49*(2), 97–107. https://doi.org/10.1016/j.plipres.2009.06.002
Yassour, M., Vatanen, T., Siljander, H., Hämäläinen, A., Härkönen, T., Ryhänen, S. J., … Vlamakis, H. (2016). Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science Translational Medicine*, *8*(343), 343ra81. https://doi.org/10.1126/scitranslmed.aad0917
Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., … Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*, 222–228. https://doi.org/10.1038/nature11053
Yazdani, M., Taylor, B. C., Debelius, J. W., Li, W., Knight, R., & Smarr, L. (2016). Using Machine Learning to Identify Major Shifts in Human Gut Microbiome Protein Family Abundance in Disease. In *IEEE International Conference on Big Data* (pp. 1273–1280).
Zaharia, M., Bolosky, W. J., Curtis, K., Fox, A., Patterson, D., Shenker, S., … Berkeley, U. C. (2011). Faster and More Accurate Sequence Alignment with SNAP. *ArXiv*. Retrieved from https://arxiv.org/abs/1111.5572
Zhu, J., Zou, H., Rosset, S., & Hastie, T. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, *2*, 349–360.
Zwittink, R. D., van Zoeren-Grobben, D., Martin, R., van Lingen, R. A., Groot Jebbink, L. J., Boeren, S., … Knol, J. (2017). Metaproteomics reveals functional differences in intestinal microbiota development of preterm infants. *Molecular & Cellular Proteomics*, *16*(9), 1610–1620. https://doi.org/10.1074/mcp.RA117.000102
Zyl, A. W. Van, Harrison, S. T. L., & Hille, R. P. Van. (2011). Biodegradation of thiocyanate by a mixed microbial population. *Imwa*, 119–124.