

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Strain-resolved metagenomic analysis of the premature infant microbiome and other natural microbial communities

Permalink

<https://escholarship.org/uc/item/2v5870fc>

Author

Olm, Matthew Raymond

Publication Date

2019

Peer reviewed|Thesis/dissertation

Strain-resolved metagenomic analysis of the premature infant microbiome and other natural
microbial communities

By

Matthew Raymond Olm

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Microbiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jillian F. Banfield, Chair

Professor Steven E. Lindow

Professor Britt Koskella

Spring 2019

Abstract

Strain-resolved metagenomic analysis of the premature infant microbiome and other natural microbial communities

By

Matthew Raymond Olm

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor Jillian F. Banfield, Chair

Microorganisms are critical to immune system development and physiology, yet the factors that drive initial colonization and human microbiome assembly are largely unknown. Early molecular approaches to study the microbiology of human body sites relied upon the direct amplification and sequencing of 16S rRNA genes, but this only produces a coarse catalog of the organisms present. In contrast, genome-resolved metagenomics involves recovery of genomes directly from whole community DNA, enabling prediction of biosynthetic capacities and providing the ability to differentiate the capabilities of closely related strains. However, genome-resolved metagenomics is computationally challenging and validated methods for many types of analyses are lacking. In this thesis, custom genome-resolved metagenomic methods were developed to determine the structure of microbial communities, monitor the activities of bacteria *in situ*, and track their evolution. The research focused primarily on the colonization and development of microbial communities that live in, and on, premature infants. Discovered patterns of fine-scale bacterial diversity, evolution, and functional potential shed light on early microbiome assembly, and highlight factors that contribute to necrotizing enterocolitis, one of the most common diseases of premature infants. Transmission pathways and reservoirs of bacteria and microbial eukaryotes that can cause nosocomial infections were identified.

Due to the sensitivity of metagenomic methods, foreign DNA sequences can be detected in metagenomic datasets. A case-study involving the identification of the source of introduced sequences in metagenomes was conducted and the specific physical source of the contaminant identified. Though very detailed genome sequence comparisons it was possible to measure the *in situ* evolution rate of the reagent contaminant over a three year period, enabling the near correct estimation of the introduction time of the contaminant into the reagent production facility. The research established the methods for genome-resolved metagenomics-based microbial forensics and strain tracking and showed that application is possible, even in extremely complex environments like soil.

High-resolution analyses are needed to determine if microbes in different environments are the result of strain transfer events. The necessary genome-wide comparisons cannot be performed with 16S rRNA sequencing, the most common method for study of the human microbiome, leaving basic questions related to the strain-level diversity and body-site specificity unanswered. To

address these questions, strain-tracking analyses were applied to metagenomes derived from samples of the mouth, skin, and gut microbiomes of premature infants. The results highlight the extreme lack of body-site diversity during very early colonization of premature infants in the neonatal intensive care unit. Surprisingly, identical bacterial genomes for organisms such as *Escherichia coli*, *Klebsiella pneumoniae* and *Citrobacter koseri* were found in the mouth, skin, and gut microbiomes. Differential genome coverage was used to measure their bacterial population replication rates *in situ*. In all cases, the replication rates for same bacterial populations in different body sites were faster in mouth and skin compared to the gut, despite the fact that these bacteria are traditionally considered gut colonists. Finally, strain-level analysis of polymorphic sites across the *C. koseri* genome were used to define 10 subpopulations, implying initial colonization of premature infants by multiple individual cells with distinct genotypes.

Methods for rapid, accurate and reproducible “de-replication”, the process of grouping recovered genomes together based on similarity and choosing the best representative genome from each group, are needed for genome-resolved metagenomic analyses. This task requires a number of steps, including mass pairwise genome comparison, evaluation of completeness and contamination, and generation of explanatory figures to visualize and validate the dereplication process. An open-source program, “dRep”, was developed and validated. The method achieves very similar results as naive pairwise clustering algorithms, but with an order of magnitude speed increase due to use of a biphasic algorithm. Importantly, individual samples in sample sets can now be assembled independently and the genomes effectively de-replicated, reducing the incidence of chimeric sequences and improving genome recovery over results for co-assemblies.

Delineation of bacteria as belonging to the same versus different species is a longstanding problem in microbiology. Fundamental questions related to the existence of species and how species should be differentiated remain, and the answers have both practical and evolutionary implications. A large public dataset comprised of >5,000 genomes acquired directly from metagenomes was analyzed. In conjunction, genome-based metrics that could be used to define bacterial species boundaries were evaluated. A distinct gap in the distribution of average nucleotide identity (ANI) values at 95% ANI exists, supporting the existence of discrete species. ANI was compared with metrics of selection for non-synonymous versus synonymous substitutions and for homologous recombination to identify processes that could lead to species clusters. The 95% ANI value corresponds approximately with the genetic distance beyond which homologous recombination drops to near zero. The findings implicate sequence divergence-based breakdown in homologous recombination as the evolutionary force responsible for bacterial speciation. 50 genes were evaluated to provide a practical means to define species content when genomes are not recovered from metagenomes for most community members. Although 16S rRNA gene sequences cannot be used for this purpose, the nucleotide sequences of several ribosomal proteins were found to be reasonable proxies for the relevant genome ANI value.

Microbial eukaryotes are particularly understudied in the human microbiome, yet they are considered to be emerging health threats. Genomes from microbial eukaryotes can be reconstructed from human microbiome metagenomic datasets. The results are greatly improved through use of a recently developed machine learning algorithm, EukRep, which can identify eukaryotic DNA based on the sequences alone. EukRep was used to scan thousands of metagenomes from the premature infant gut and hospital room environments and fourteen novel eukaryotic genomes were reconstructed. Two of these, for a Diptera (fly) and a Rhabdida (worm),

were novel at the class level. Importantly from the perspective of tracking nosocomial agents, genomes from the same eukaryotic species were recovered from both infants and hospital room environments. Population heterogeneity and zygoty of genomes were lower in genomes recovered from the hospital room as compared to those recovered from premature infant samples, which could reflect years of inbreeding or strong selection imposed by room conditions. Together this work indicates that the hospital room, especially the sink, may be a reservoir of infant-colonizing fungal strains.

Necrotizing enterocolitis (NEC) involves extreme bowel inflammation and necrosis and has a mortality rate of around 30%. Various lines of evidence point to the human microbiome as a central factor in disease development, yet no consistent microbial signal for NEC onset has been identified. We performed large-scale genome-resolved metagenomic analyses of thousands of prospectively collected premature infant fecal samples and used a machine learning classifier to identify signals that predict imminent development of NEC. Significant associations were found related to the abundance of *Klebsiella*, bacteria encoding fimbriae, and specific types of secondary metabolite clusters, and the *in situ* growth rate of bacteria overall. NEC development could be promoted by metabolic imbalances related to the rampant growth of particular bacterial strains and/or the stimulation of human TLR4 receptors by *Klebsiella* and fimbriae. Together, the results identify potential biomarkers for early detection of NEC and possible targets for microbiome-based therapeutics and probiotics

Table of Contents

Acknowledgements	iii
1 The Source and Evolutionary History of a Microbial Contaminant Identified Through Soil Metagenomic Analysis	1
1.1 Introduction	1
1.2 Materials and methods	2
1.3 Results	5
1.4 Discussion	8
1.5 Figures	10
2 Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates	14
2.1 Introduction	14
2.2 Materials and methods	16
2.3 Results	19
2.4 Discussion	23
2.5 Figures	25
3 dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication	32
3.1 Introduction	32
3.2 Results	33
3.3 Conclusions	34
3.4 Figures	35
4 Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms	37
4.1 Introduction	37
4.2 Materials and Methods	39
4.3 Results	44
4.4 Discussion	48
4.5 Figures	51
5 Necrotizing enterocolitis is preceded by increased gut bacterial replication, <i>Klebsiella</i>, and fimbriae-encoding bacteria	58
5.1 Introduction	58
5.2 Materials and Methods	59
5.3 Results	65
5.4 Discussion	69
5.5 Figures	71

6	Consistent metagenome-derived metrics verify and define bacterial species boundaries	77
6.1	Introduction	77
6.2	Materials and Methods	78
6.3	Results	80
6.4	Discussion	82
6.5	Figures	83
	Concluding remarks and future work	94
	References	97
	Appendix 1: Select co-authored publications	119

Acknowledgements

I would first like to acknowledge Jill Banfield for her mentorship, scientific guidance, and support. I am extremely grateful for the countless hours she dedicated to fostering my growth as a scientist, especially in teaching me to write succinctly and persuasively, to approach questions with curiosity and enthusiasm, and to conduct research with integrity and to hold others to the same standard. I could not ask for a better scientific role model.

Thank you to Steven Lindow and Britt Koskella for their scientific and career advice, and to Dr. Michael Morowitz and his lab at the University of Pittsburgh, especially Robyn Baker and Brian Firek, for their invaluable medical and technical expertise. Thank you to all members of the Banfield lab for their support and bioinformatics advice, including Karthik Anantharaman, Evan Starr, Keith Gouma-Gregson, Alex Crits-Christoph, and Patrick West. Special thanks to Chris Brown and Bubba Brooks for their extensive mentorship in the fields of bioinformatics and the human microbiome.

I would also like to acknowledge those that led me to attend graduate school in the first place. Deborah Jacobs-Sera provided an unparalleled environment to first conduct academic research, and the mentorship of Debbie, Lauren Oldfield, and Graham Hatfull, combined with the camaraderie of Alex Cathcart, Enoch Tse, Emilee Shine, Victoria Schneider, and Terrence Parker, led me to fall in love with academic science. Finally, I would like to thank my grade school teachers Mr. Mariner for sparking my interest in science and Mr. Parvin for encouraging me to enjoy intellectual challenges.

Many others deserve thanks for their friendship and emotional support, especially John and Noelle Rabiah, Jack Kaufmann, Matthew Schroeder, Rodger Dilla, Joey Neubert, my family, and Kitty. The curiosity that I picked up from by Dad and rational skepticism from my Mom were critical to my scientific development, as were grammar lessons from my Grandma Marilyn.

Finally, I offer my deepest thanks to my wife Rylee Kercher Olm. Rylee's continual encouragement since high school has been instrumental in leading me to pursue higher education, and I am extremely grateful for her support in my pursuit of an academic career. Kayaking, backpacking, and hiking around the Bay Area with Rylee made my time here a joy, and I look forward to seeing where we end up next.

1 The Source and Evolutionary History of a Microbial Contaminant Identified Through Soil Metagenomic Analysis

Olm, Matthew R; Butterfield, Cristina N; Copeland, Alex; Boles, T Christian; Thomas, Brian C; Banfield, Jillian F

Published in *mBio*, February 2017

In this study, strain-resolved metagenomics was used to solve a mystery. A 6.4-Mbp complete closed genome was recovered from a soil metagenome and found to be astonishingly similar to that of *Delftia acidovorans* SPH-1, which was isolated in Germany a decade ago. It was suspected that this organism was not native to the soil sample because it lacked the diversity that is characteristic of other soil organisms; this suspicion was confirmed when PCR testing failed to detect the bacterium in the original soil samples. *D. acidovorans* was also identified in 16 previously published metagenomes from multiple environments, but detailed-scale single nucleotide polymorphism analysis grouped these into five distinct clades. All of the strains indicated as contaminants fell into one clade. Fragment length anomalies were identified in paired reads mapping to the contaminant clade genotypes only. This finding was used to establish that the DNA was present in specific size selection reagents used during sequencing. Ultimately, the source of the contaminant was identified as bacterial biofilms growing in tubing. On the basis of direct measurement of the rate of fixation of mutations across the period of time in which contamination was occurring, we estimated the time of separation of the contaminant strain from the genomically sequenced ancestral population within a factor of 2. This research serves as a case study of high-resolution microbial forensics and strain tracking accomplished through metagenomics-based comparative genomics. The specific case reported here is unusual in that the study was conducted in the background of a soil metagenome and the conclusions were confirmed by independent methods.

1.1 Introduction

Microbial strains of the same species can have very different traits, including virulence and drug resistance (Greenblum et al., 2015; Luo et al., 2015; Schloissnig et al., 2012), thus tracking of specific strain populations is important in a number of different contexts. Microbial source tracking (MST) via quantitative PCR is routinely used to determine the source of fecal bacteria in environmental waters (Hagedorn et al., 2011), and in some cases can discriminate between fecal profiles of different types of animals (Harwood et al., 2014). Tracing pathogenic strains within hospitals via sequencing of isolated strains can uncover vectors of nosocomial infections (Snitkin et al., 2012) and larger scale studies have improved understanding of the intercontinental spread of pathogens (Harris et al., 2010). The forensic investigation launched following the 2001 *B.*

anthracis bioterrorism attack has been called “one of the largest and most complex in the history of law enforcement” (“Amerithrax or Anthrax Investigation,” n.d.). Significant effort since has been invested to develop new methods, including CRISPR-Cas analysis (McGhee and Sundin, 2012), to deploy in the case of future bioterrorism events (Budowle et al., 2014).

A key component of microbial forensics is strain typing. Strain resolution is essential to trace the spread of a population, and more sensitive strain-typing methods can provide higher resolution transmission maps. The most common methods identify unique (but small) markers of the microbial population’s genomic DNA sequence and take advantage of the fact that random mutations develop in all growing populations. Examples of methods to identify these mutations include restriction endonuclease analysis (REA), pulse-field gel electrophoresis (PFGE), ribotyping, and multi-locus sequencing typing (MLST) (Chan et al., 2001; Olive and Bean, 1999). Decreasing sequencing costs have also allowed an increasing number of studies to take advantage of genome sequencing, the “gold standard” of microbial-typing (Snitkin et al., 2012). This approach can discriminate between microbial populations that differ by even a single nucleotide, but typically this requires culturing of the organism before DNA extraction, and this is not feasible in all cases. Metagenomics, on the other hand, has the potential to trace and characterize virulent strains without cultivation, and could be used to detect strains of interest in environmental samples (Gilchrist et al., 2015).

The National Academy of Sciences stated that metagenomics “will bring about a transformation in biology, medicine, ecology, and biotechnology that may be as profound as that initiated by the invention of the microscope” (Handelsman et al., 2007). In metagenomics, shotgun sequencing of DNA extracted directly from environmental samples allows characterization of microbes without the need for cultivation. Assembly and binning of short metagenomic reads can yield hundreds of genomes from metagenomic samples (Anantharaman et al., 2016; Brown et al., 2015; Dombrowski et al., 2016; Lee et al., 2016). However, sequencing projects can be contaminated with exogenous DNA (Salter et al., 2014). Significant efforts have been made to determine where these contaminants originate (Knights et al., 2011), but precise sources of contaminant sequences are seldom identified. The mystery of determining the source of contaminant DNA in the background of a complex metagenomic sample represents a useful test case for genomics-based microbial forensics.

1.2 Materials and methods

1.2.1 Sample collection

Soil collection and DNA extraction was performed as reported previously (Butterfield et al., In Press) and briefly described here. Soil samples were collected from the Angelo Coast Range Reserve (with permission under APP # 27790) meadow (39°44'21.4"N 123°37'51.0"W) and from a nearby ridge within the Eel Critical Zone Observatory. Approximately 1 kg of soil was removed for each depth using sterilized stainless steel hand trowels. Samples were immediately flash frozen in a mixture of dry ice and ethanol and placed on dry ice for transport to the lab. For each depth, DNA was extracted using MoBio Laboratories PowerMax Soil DNA Isolation kits from 10 g of soil. We optimized the protocol for our samples, to maximize DNA yield while minimizing shearing: each sample was only vortexed for 1 minute, followed by a 30-minute heat step at 65 °C,

inverting every 10 minutes. We performed two elution steps of 5 mL each, and precipitated the DNA using sodium acetate and glycogen, resuspending in 100 μ L of 10 mM Tris buffer.

1.2.2 Metagenomic sequencing, assembly, and binning

Metagenomic DNA was sequenced at the Joint Genome Institute (JGI). DNA was sheared to 800 bp using the Covaris LE220 (Covaris) and size selected using the Pippin Prep (Sage Science). The fragments were treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Inc) using the KAPA-Illumina library creation kit (KAPA biosystems). 250 bp paired end reads were generated on an Illumina HiSeq2500, with sequencing depth enumerated in **Table S1.1**. Reads were trimmed with Sickle (Joshi and Fass, 2011), and assembled using IDBA-UD (Peng et al., 2012). Resulting scaffolds greater than 1 kb in length were annotated using Prodigal (Hyatt et al., 2010) to predict open reading frames using default metagenomic settings. Annotated protein sequences were searched against KEGG (Kanehisa et al., 2014), UniRef100 (Suzek et al., 2007), and Uniprot databases using USEARCH (Edgar, 2010). All matches with bit scores greater than 60 were saved, and reciprocal best hits with a bit score greater than 300 were also cataloged. We identified rRNA sequences using Infernal (Nawrocki and Eddy, 2013) by searching against databases from the SSU-Align package (Nawrocki, 2009), and tRNAs using tRNAscan_SE (Lowe and Eddy, 1997). Genome binning was carried out using the online interface within ggKbase as described previously (Raveh-Sadka et al., 2015) (<http://ggkbase.berkeley.edu/>). This method takes into account phylogenetic profile, GC content, and coverage information. The completeness of bacterial bins was evaluated based on the presence or absence of single copy genes. Shannon diversity calculations were performed as described previously (Spellerberg and Fedor, 2003) using assembled ribosomal protein S3 (rps3) genes.

1.2.3 Genome Curation

Once *Delftia acidovorans* scaffolds were binned they were manually curated in order to close gaps between scaffolds. This was done within the Geneious software package version 8.1.5 (Kearse et al., 2012). Project reads were first mapped back to the contigs using Bowtie 2 (Langmead and Salzberg, 2012), and Geneious was used to extend the contigs. Contigs were next ordered and oriented by mapping them to the *D. acidovorans* SPH-1 reference genome using ABACAS (Assefa et al., 2009). Areas of overlap between adjacent contigs were used to curate two contigs into one. In order to circularize the genome, reads were mapped directly to the reference genome. The reference genome was edited to be in agreement with the reads, and regions between contigs were filled in with corresponding pieces of the edited reference genome. Reads were finally mapped back to the circularized genome, and every base was manually inspected to ensure sequence read-support. In order to catalogue the differences between genomes, Mauve (Rissman et al., 2009) was used to align the genomes, and the alignment was manually inspected for mutations.

1.2.4 Phylogenetic tree

An alignment was generated using all rpS3 genes in the Angelo metagenomes, as well as previously published rpS3 sequences identified as similar to Angelo rpS3 sequences by BLAST. All rpS3 amino acid sequences longer than 180 amino acids were aligned using MUSCLE (Edgar, 2004). The full alignments were stripped of columns containing 95% or more gaps. A maximum-likelihood phylogeny was inferred using RAXML (Stamatakis, 2006) run using the PROTGAMMLG model of evolution. The RAXML interface included calculation of 100 bootstrap iterations (MRE-based Bootstrapping criterion).

1.2.5 PCR testing

Primers amplifying randomly selected portions of the *D. acidovorans* were synthesized from IDT to generate a 500 bp insert [F: 5'GGGTTGCACCATTGGTATT), R: (5'GTCAGCGCCTTCTTTTCAA]. Primers which amplify a 150 bp product of the 16S gene were also synthesized (F: 5'GTGSTGCAYGGYTGTCGTCA, R: 5'ACGTCRTCCMCACCTTCCTC) (Horz et al., 2005). Pure *D. acidovorans* DNA was purchased from the DSMZ culture collection (DSMZ No. 14801). Reactions were performed using the 5 PRIME MasterMix 50 uL reactions, with 1uL of each primer at 10 mM, run in the thermocycler for 35 cycles with a T_m of 50 and an extension time of 1:30. Both sets of primers and a no primer control were run on 1) 0.05 ng of *D. acidovorans* DNA, 2) 4.95 ng of original extracted soil DNA (sample 13_1_20cm_4), and 3) both of the above combined into a single reaction.

1.2.6 Paired read insert length profiling

Reads were mapped from each analyzed project to the *D. acidovorans* SPH-1 genome using Bowtie 2 (Langmead and Salzberg, 2012). For comparative purposes, reads from project 13_1_20cm_4 were also mapped to the second-most abundant bacterial genome, *Gemmatimonadetes*. The resulting .sam file was converted to a sorted .bam file using samtools (Li et al., 2009), and the insert size was profiled using Picard ("Picard Tools - By Broad Institute," n.d.) (command: `java -jar picard_tools/CollectInsertSizeMetrics.jar MINIMUM_PCT=0.4`).

1.2.7 Comparison of *D. acidovorans* strains

To determine if soil and sediment metagenomes contained the *D. acidovorans* strain ANG1 genome, PileupProfile.py (source code available at <https://github.com/banfieldlab/mattolm-public-scripts>) was used to calculate the average nucleotide identity of reads mapping to the ANG1 genome. ANI was at least 99.9% in all cases where the median coverage was ≥ 2 (**Table S1.1**). To compare strains present in 46 public metagenomic projects sequenced at JGI, which had been flagged as possibly containing *D. acidovorans* contamination, reads were mapped to the *D. acidovorans* SPH-1 genome using Bowtie 2 (Langmead and Salzberg, 2012). Projects with at least 20% of bases having 5x coverage were said to have significant *D. acidovorans* present and were analyzed in more detail. To compare the strains in each metagenome, the custom script ReadComparer.py was used (source code available at <https://github.com/banfieldlab/mattolm-public-scripts>). Briefly, the program first aligns reads from all project to the same genome, and compares the mutational patterns between projects to determine relatedness of strains. VarScan (Koboldt et al., 2009) was used to create the input files (command: `java -jar VarScan.v2.3.8.jar pileup2cns --min-coverage 3`), and the script was run using the command: `ReadComparer.py --min_breadth 0.2 --matrix --dend --smart_ignore`. The similarity matrix was then plotted into a dendrogram using R, and clusters were determined and colored based on manual inspection of the resulting dendrogram.

1.2.8 Mutation identification and profiling

A number of methods were used to identify differences between *D. acidovorans* strains SPH-1 and ANG1. Mutations were identified using Breseq (Deatherage and Barrick, 2014) mapping reads from project 13_1_20cm_4 to the *D. acidovorans* SPH-1 genome. To identify larger indels, *D. acidovorans* genomes were aligned using Mauve (Rissman et al., 2009) and the alignment was manually inspected for differences. Finally, VarScan (Koboldt et al., 2009) was used on reads

mapping from 13_1_20cm_4 to the *D. acidovorans* ANG1 genome. All called mutations were manually verified by inspection of the region using the software package Geneious (Kearse et al., 2012).

To track the frequency of mutations among all projects of the contaminant clade, `polymorpher2.py` (available at <https://github.com/banfieldlab/mattolm-public-scripts>) was used to determine the frequency of each base at all positions identified. Variants that were closer than 315 bp apart (a 1% chance of occurring assuming a random distribution of variants) were excluded from mutation rate calculations based on the assumption that they were likely acquired in a recombination event. Positions were required to have some coverage in all projects, which excluded 4.4% of variants from analysis. Polymorphisms with a frequency of at least 50% at the first time-point and became near-fixed at the last time-point ($\leq 0.1\%$) were said to be evolving and used in rate calculations (including a correction to account for variants without sufficient coverage for analysis). Full calculation details are available in the **Supplemental Jupyter Notebook**.

ORFs predicted by Prodigal were used to determine the total number of synonymous and non-synonymous sites in the genome, as well as to classify each variant as synonymous, non-synonymous, or intergenic. Genome wide pN/pS ratios were calculated using the formula: $[(\# \text{ non-synonymous substitutions} / \# \text{ non-synonymous sites}) / (\# \text{ synonymous substitutions} / \# \text{ synonymous sites})]$.

1.3 Results

1.3.1 Recovery of a complete *Delftia acidovorans* genome from soil

Soil samples were collected from two locations in the Angelo Coast Range Reserve in northern California. A meadow within the reserve was sampled at two soil depths (20 cm and 40 cm) over a six-week period, and a ridge nearby was sampled at 6 depths (15 cm – 115 cm, including both soil and underlying weathered shale) during one sampling event. The sites are located within the Eel Critical Zone Observatory (CZO). Metagenomic DNA was extracted from all samples, and between 15.7 and 44.0 Gbp of Illumina shotgun paired-end sequencing was generated for each sample. In total, 0.4 Tbp of sequence data was generated (**Table S1.1**). Shannon diversity was calculated for all samples using ribosomal protein S3 (rpS3) genes. The mean alpha diversity of soil collected in this study was 4.65 (SD: 0.28), similar to that of many previously studied soils (Fierer et al., 2013; Howe et al., 2014; Williamson et al., 2005) (**Table S1.1**).

An initial binning analysis revealed genome fragments that were profiled as deriving from *D. acidovorans* populations in six samples (other genomes are reported in Butterfield et al. (Butterfield et al., In Press)). We reconstructed a 6.41 Mbp high-quality draft genome of *D. acidovorans*, consisting of 53 scaffolds ranging in length from 6 Kbp to 500 Kbp. The genome was further curated by read mapping to fill scaffolding gaps and extend and join the contigs. These resulting 16 contigs could be ordered and oriented to the previously sequenced *D. acidovorans* SPH-1 genome (Genbank Accession: NC_010002.1). The contigs spanned 97.95% of the SPH-1 genome, with an average nucleotide identity (ANI) of 99.99% (Goris et al., 2007). Due to the remarkably high nucleotide similarity between our recovered genome and strain SPH-1, we further curated the 16 contigs into one circular genome by using the SPH-1 genome to identify reads in our metagenomic dataset that filled the gaps. When visualized using the software package

Geneious (Kearse et al., 2012) single nucleotide polymorphisms (SNPs) and insertions and deletions relative to the isolate genome sequence were easily identified (**Figure S1.1**). The circular closed genome represents the bacterial strain *D. acidovorans* ANG1.

1.3.2 Identification of *D. acidovorans* as a contaminant

Based on metagenomic read mapping, *D. acidovorans* ANG1 was detected in all fourteen samples from two sites. No pattern between sampling location and abundance of *D. acidovorans* could be detected (**Figure 1.1A**). Based on mapping of reads to the ANG1 genome, it was determined that the *D. acidovorans* reads in all samples derived from a single genotype (**Table S1.1**). This result contrasts with our general findings from soil metagenomics, which typically indicate the presence of multiple closely related strains. For example, a phylogenetic tree constructed using *rpS3* gene sequences shows dandelion-like strain diversity patterns in Betaproteobacteria from the same samples (**Figure 1.1B**). The lack of strain diversity in *D. acidovorans*, in combination with the high similarity of the *D. acidovorans* genome to that of the SPH-1 strain isolated years earlier, raised the possibility that *D. acidovorans* ANG1 was not a native member of the soil community. A PCR test using primers designed to target random regions of *D. acidovorans* genome failed to detect DNA from this bacterium in the original DNA extracted from the samples (**Figure S1.2**). This indicated that the DNA was probably introduced into our samples during sequencing. Moreover, the detection of the same genotype in samples sequenced at different times suggested that *D. acidovorans* was a persistent contaminant at this facility.

1.3.3 Source tracking of contaminant *D. acidovorans*

To further investigate the possibility that DNA from *D. acidovorans* ANG1 was introduced in the sequencing facility, we screened 43 publicly available metagenomes sequenced at this facility between June 2012 and January 2015. Seventeen of these projects had $\geq 20\%$ of the *D. acidovorans* genome present, with coverage $> 5x$ (**Table S1.2**). Multiple strains were present in these projects, based on analysis of the patterns of single nucleotide polymorphisms (SNPs) relative to the *D. acidovorans* ANG1 genome (**Figure 1.2A**) (**Table S1.3**). Six projects, including three plant genomes and our soil samples, contained sequences that clustered with the *D. acidovorans* ANG1 genome. We refer to this as the contaminant clade, and hypothesized that some reagent in the sequencing pipeline may have been the source of the *D. acidovorans* DNA in these six projects. Records provided by the sequencing facility revealed that each project containing the *D. acidovorans* ANG1 contaminant clade used a Pippin Prep size selection cassette. All other analyzed projects did not use these cassettes (**Figure 1.2A**).

To test the hypothesis that the Pippin Prep size selection cassettes were the source of *D. acidovorans* ANG1 contamination, we analyzed the insert sizes of reads from all projects mapped to the ANG1 genome. If the contaminant DNA was introduced from the library preparation cassette, it should have a more random fragment size profile, and thus insert size profile, than the tight size profile generated during library preparation. A histogram of reads mapping to the *D. acidovorans* ANG1 genome from a sample exhibiting substantial contamination is shown in **Figure 1.2B**, along with a histogram of reads mapping to the genome of a *Gemmatimonadetes* known to be native to the same sample. The fragment size of paired reads mapped to the *Gemmatimonadetes* genome was ~ 800 bp. In contrast, read mapping to *D. acidovorans* indicates that many of the sequencing reads were generated from fragments of around 250 bp (in these cases, the 250 bp *D. acidovorans* reads overlap completely). However, the *D. acidovorans* peak is

strongly skewed, and some of the fragments were > 1000 bp in length. These observations indicate that the contaminant DNA was present in the gel and/or buffer used for size selection. All projects that contained *D. acidovorans* from the contaminant clade had similar insert size histograms, whereas projects with other *D. acidovorans* clades had normal insert sizes.

Sage Science, the producer of the Pippin size selection cassettes, was contacted regarding our observations. They explained that bacterial biofilms were present in tubing that delivered buffer to the cassettes, and stated that the problem was corrected in 2013. All 6 projects that contain the contaminant clade used cassettes made prior to the correction, and libraries made with cassettes produced after Sage Science revised their manufacturing process to keep buffer tubing bacteria-free did not reveal *D. acidovorans* ANG1 contamination (data not shown). The sequencing facility continued to detect *D. acidovorans* in samples sequenced after Sage Science corrected the problem, and so concluded that sequences were not from the Pippin cassettes. We show here that the newly detected sequences were from the other clades, either strains actually present in the samples or contaminants from a different source. For example, clade 1 is associated with 6 metagenomes of the upper troposphere and clade 2 with three metagenomes of thiocyanate bioreactor communities (**Table S1.2**). In the case of clade 2, we conclude that the *D. acidovorans* was native to the sample because the population was growing rapidly, based on differential coverage at the origin compared to terminus of replication (peak-to-trough ratios, PTR) (full range 1.5 – 1.9) (Brown et al., 2016; Korem et al., 2015) (**Table S1.4**). In contrast, bacteria of the contaminant clade (clade 5) had consistently low growth rates, based on PTRs between 1.4 and 1.1 (**Figure 1.3**).

1.3.4 *In situ evolution of a bacterial contaminant*

Fifteen large (>100 bp) insertions/deletions distinguished the SPH-1 and ANG1 *D. acidovorans* genomes (**Table S1.5**). Four of these were insertions and 11 were deletions (**Figure 1.4A**). Specifically, two prophages were inserted into the ANG1 relative to the SPH-1 genome, but six prophages and five transposons were lost. Another difference involved three adjacent CRISPR repeat-spacer sequences. Finally, we identified a 195 bp insertion that added a pair of transmembrane helices that converted a predicted major facilitator superfamily protein into a predicted transporter that confers drug resistance.

We documented SNP-based genomic variation over the 1.5-year period to determine whether an evolutionary rate could be measured. A total of 203 single nucleotide polymorphisms (SNPs) distinguished the ANG1 genome and the *D. acidovorans* SPH-1 isolate genome (**Table S1.5**). 37 SNPs were very closely spaced on the genome (**Figure 1.4A**) and are statistically unlikely to have formed by individual random mutation events (see methods). Specifically, 11 SNPs and 4 single bp indels occurred in a 64 bp intergenic region upstream of an integrase and 24 SNPs occurred in a 447 bp region within a YD repeat-containing protein possibly involved in carbohydrate binding. We infer that these regions may have been acquired via homologous recombination with a very distinct genotype (<95% nucleotide identity), thus we did not include them in analysis of *in situ* genomic change via SNP formation.

We identified 15 SNPs that distinguished the genotypes present in the first compared to last metagenome (a separation time of ~1.5 years) and tracked their frequencies over time. Notably, the SNP frequencies cluster into three patterns, consistent with linkage and thus the existence of three sub-populations. Frequencies do not show a simple trend towards fixation, but all SNPs are

fixed by the last time point (**Figure 1.4A**) (**Table S1.6**). Cohort three is defined by a single non-synonymous mutation (Gly to Ser) that becomes fixed in a gene encoding a rod-shape determining protein.

By correcting for missing information due to lack of coverage at SNP sites in the ANG1 genome, we estimate that ~16 SNPs were fixed over the 1.5-year period. Thus, the rate of fixation is estimated as 10.1 SNPs/year. The remaining 150 SNPs that distinguish the SPH-1 genome sequenced in 2006 (Kjelleberg, written communication) and ANG1 genomes likely arose between the time the SPH-1 genome was sequenced and the first metagenome time point (~6.5 years).

Given information about the SNP accumulation rate, we estimated how long ago the strain was separated from the original source culture. Assuming that the measured value approximates the rate of SNP formation over longer time periods, we calculated that the ANG1 genotype present in the soil was separated from the SPH-1 population 16.5 years ago.

We classified all 166 fixed SNPs that distinguish the SPH-1 and ANG1 genome as synonymous, non-synonymous, or intergenic (**Figure 1.4B**). Overall, the pN/pS ratio is 0.98..

1.4 Discussion

We reconstructed a complete *D. acidovorans* genome from a soil metagenome. Recovery of a complete genome from soil is an extremely unusual achievement given that assembly becomes more difficult as sample complexity increases. We infer that this occurred in the current study because the genome was relatively abundant (~1% of total DNA) and the population was near-clonal, avoiding assembly problems that arise due to strain-variation. The lack of micro-heterogeneity is atypical of soil populations. This, and the uncanny similarity to an isolate genome raised the possibility that the genome derived from a contaminant, a hypothesis that was confirmed by PCR-based testing. The extremely high similarity between the contaminant *D. acidovorans* populations and *D. acidovorans* SPH-1 genome gives us high confidence that one derived from the other. In this specific case, we could determine the source of the contaminant DNA. Consequently, the research serves as a case study for microbial metagenomics-based forensics. Importantly, we showed that a contaminant clade could be identified through base pair-by-base pair analysis of variable sites in population genomic datasets collected over time. This made it possible to distinguish strains that were native to the samples in which they occurred from those that were the contaminant. This approach should be generally applicable in strain tracking investigations so long as high-quality genomes can be recovered.

Metagenomics-derived genomes have only very rarely been used to identify and track strains. The closest example to our work was a study by Loman et al. (Loman et al., 2013) that targeted pathogenic *E.coli* in fecal samples and associated it with disease. However, their methods were specifically designed for the clinical setting and generated a draft genome that was too incomplete and fragmented for use in accurate strain tracking and *in situ* evolutionary analysis. Long stretches of contiguous sequence are needed to identify indels and regions associated with horizontal gene transfer, and higher genome completeness leads to more accurate estimates of evolutionary distance.

During analysis of the Anthrax attack, the crux of the analysis related to determining the similarity between the weaponized strain and the Ames laboratory strain (Hoffmaster et al., 2002; Rasko et al., 2011). As we show here, high-resolution determination of strain relatedness is possible given comprehensive comparative genomic information for the bacterium of interest. Given a reasonable estimate of mutation rate and information about the genomic similarity between the weaponized strain and the Ames laboratory strain, it would be possible to estimate how long ago the cultures were separated.

Most short-term evolution experiments used to determine mutation rates are carried out under laboratory conditions. These rates may differ significantly from rates that are applicable under conditions experienced by populations growing in natural systems. Because we had access to datasets archived over a 1.5-year period, we could estimate the rate of evolution of the contaminant population in the relevant environment (tubing) and use this calibration to approximate the length of time separating it from the previously sequenced isolate. Our calculated value (10.1 substitutions / year) is very close to previously reported values of evolving pathogen populations (9.6 substitutions / year) (Harris et al., 2010), and orders of magnitude faster than estimates of natural *E. coli* populations (Ochman et al., 1999). Using our calculated value, we estimate separation of ANG1 from the source population 16 years ago. SPH-1 was sequenced in 2006 and ANG1 was sequenced in 2014, hence the longest time that the contaminant ANG1 population could have been separated from the SPH-1 isolate is nine years. This result is within a factor of 2 of the expected value, despite the possibility of large errors in mutation accumulation rates due to variations in growth rates and stress (Bjedov et al., 2003). Thus, we conclude that the comparative genomics approach used here can constrain the time of separation of two populations. From a technical perspective, it is worthwhile to note that this detailed analysis was successful even in the background of soil.

We defined three distinct SNP cohorts at time points intermediate between the first and last metagenomic sequencing events, suggesting the existence of three genotypic variants. The fact that all 15 SNPs are fixed in the final ANG1 population is unexpected, and inconsistent with separate coevolving populations. Further, the frequencies of SNPs in cohorts 1 and 3 undergo dramatic fluctuations in frequency (in some cases dropping below our detection level) rather than exhibiting a simple trajectory towards fixation. We attribute both observations to extensive redistribution of SNPs through homologous recombination, a common process among closely related bacteria (Rosen et al., 2015; Smillie et al., 2011; Tenailon et al., 2010). The homogenization of population variation is consistent with recombination acting as a cohesive force, countering the diversifying SNP formation processes that otherwise could lead to speciation.

The original SPH-1 strain was isolated from municipal sewage sludge, an environment very different from the biofilm tubing environment in which the ANG1 strain was growing (Schleheck et al., 2004). The pN/pS (the population-based equivalent to dN/dS) value of ~ 1 , determined based on comparison of the isolate and ANG1 population, is inconsistent with stabilizing selection. It could reflect minimal selective pressure or the combination of positive selection and negative selection. Most populations studied previously have values consistent with stabilizing selection (dN/dS of ~ 0.1) because they are shaped by overall negative selection with some genes under positive selection (Friedman, 2004; Schloissnig et al., 2012). Given this, we consider it more likely that the *D. acidovorans* population was experiencing a mixture of positive and negative selection rather than no selection. Positive selection is not surprising, given a population that is evolving

and adapting to an environment different from where it was isolated. An alternative explanation is that the non-synonymous mutations have not yet had enough time to be selected against (Rocha et al., 2006). The larger number of deletions (particularly of phage and transposon sequences) compared to insertions in the ANG1 relative to the SPH-1 genome is indicative of genome-streamlining, consistent with its adaptation to a defined laboratory environment. Again, the observation of streamlining rather than genomic expansion is informative regarding the recent population history.

In conclusion, we show that detailed strain-resolved metagenomic studies can detect a specific organism of interest in very complex samples, provide evidence that the strain is not native to the environment from which the sample was collected, and constrain its recent history. We demonstrate this using the example of a contaminant that was introduced during the laboratory handling of metagenomic samples, but the approach is far more broadly applicable. We used statistical analysis to link the contaminant genotypic group to its source and comparative genomics to uncover aspects of its recent evolutionary history. Because there was confirmation of many conclusions by independent methods, this work serves as a case study for strain-resolved forensic metagenomics.

1.5 Figures

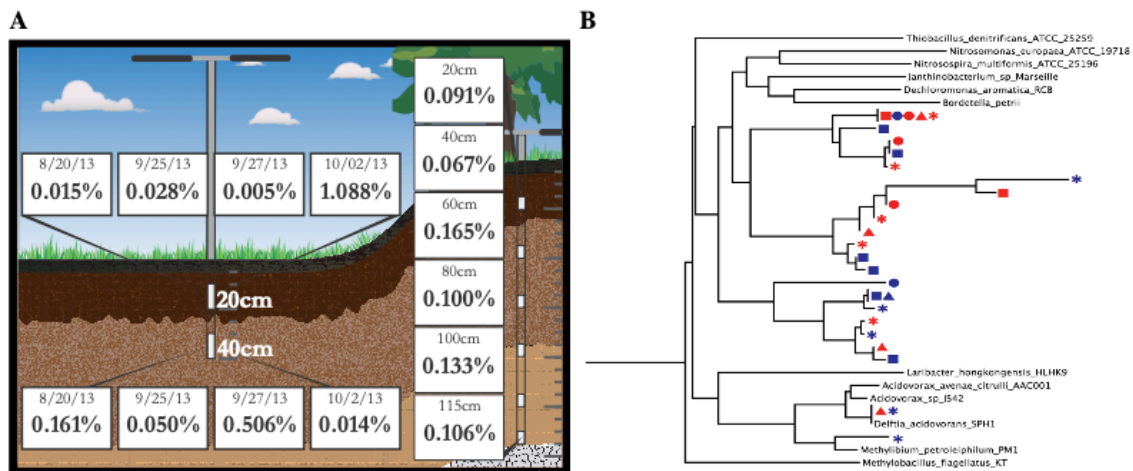


Figure 1.1 *Delftia acidovorans* is present in all sequenced soil and sediment samples, but the population structure is distinct from that of other bacteria. (A) The percentage of all reads from each sample that map to the *Delftia acidovorans* ANG1 genome. (B) A *rps3* phylogenetic tree for a typical bacterial group (Betaproteobacteria) shows a dandelion-like pattern of strain diversity that is dissimilar to that for *Delftia acidovorans* (only the full-length assembled sequences are shown). Branches ending with a taxonomic identification are reference sequences and the soil sequences are indicated by colored shapes representing their soil sample depth and

time of origin around the first rainfall in August – October 2013 (10 – 20 cm, blue and 30 – 40 cm, red) before (squares) and after the rain events (4 days after: circles, 6 days after: triangles, and 2 days after the second rain: asterisks).

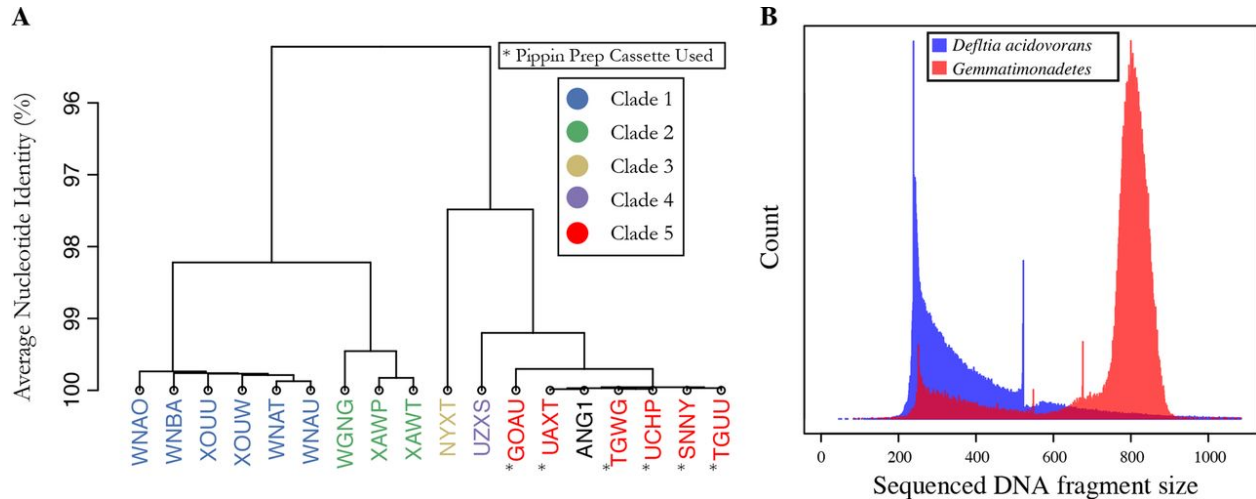


Figure 1.2 *Delftia acidovorans* contamination originates from library size selection cassettes. **(A)** Hierarchical clustering of *D. acidovorans* strains present in shotgun sequencing projects sequenced at the same facility as soil metagenomics in this study, based on shared SNP frequencies. Samples prepared with Pippin Size Selection Cassettes (Sage Science) are marked with an asterisk (*). Sequences of *D. acidovorans* in those samples form a monophyletic group that includes the ANG1 genome. **(B)** Histogram of sequenced DNA fragment sizes of metagenomic reads mapping to the recovered *D. acidovorans* genome and the second most abundant recovered genome in the sample, from a *Gemmatimonadetes* bacterium. The placements of reads mapping to the *Gemmatimonadetes* genome indicate the expected 800 bp fragment length, yet reads mapping to *D. acidovorans* show a different and skewed profile, consistent with introduction of DNA after size selection.

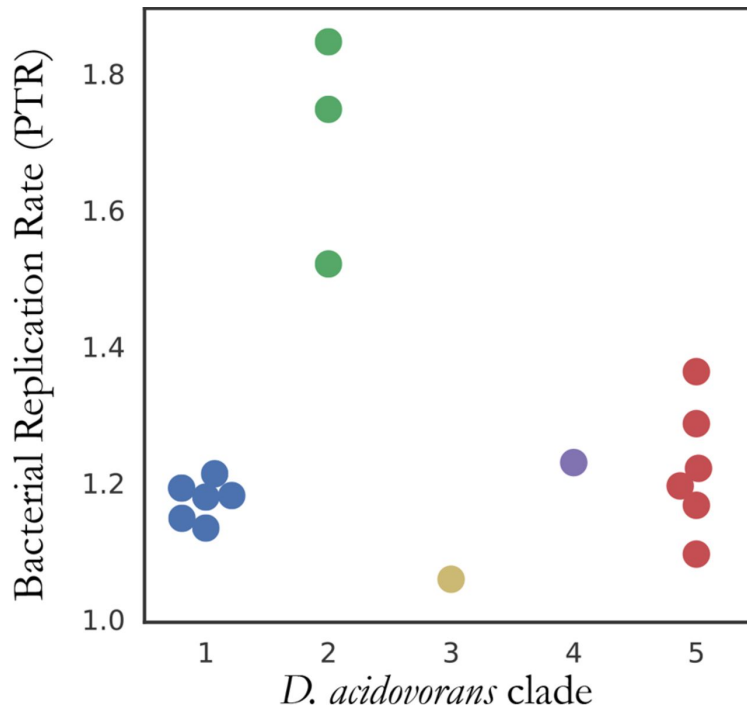


Figure 1.3 *D. acidovorans* strains native to a bioreactor have distinct replication rates. Clades of similar *D. acidovorans* genomes were clustered based on shared SNP frequencies (Figure 2A). The peak to trough ratio values are significantly higher ($p = 0.005$) for the strains growing in a thiocyanate bioreactor (clade 2) than the other environments. As the other clades are likely made up of contaminants (see main text), this observation supports the conclusion that the bioreactor strains are active community members and not contaminants.

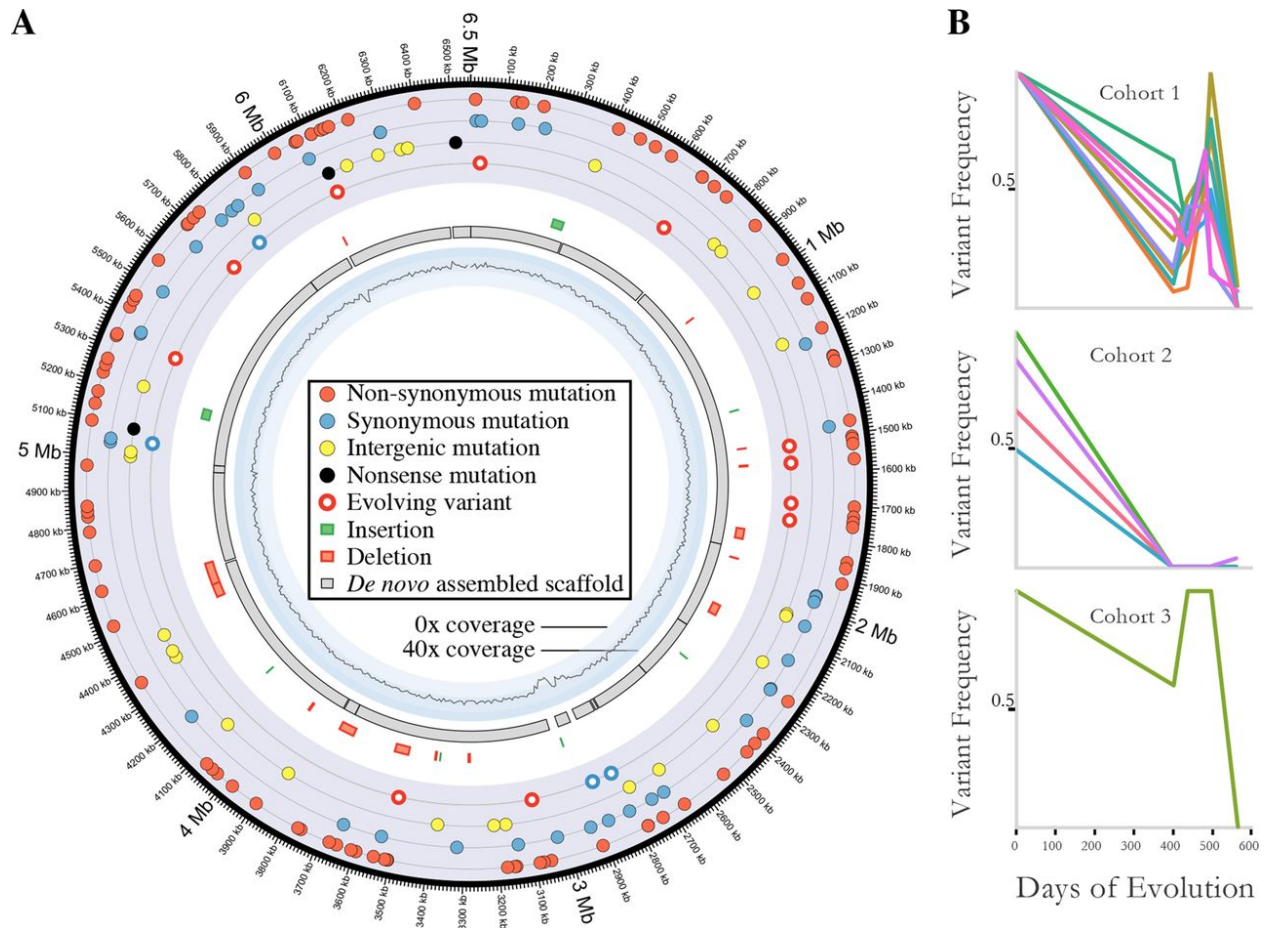


Figure 1.4 Detailed analysis of *D. acidovorans* sequence variation indicates the presence of subpopulations that are evolving in situ (A) The frequencies of specific SNPs in samples sequenced at different time points exhibit three distinct patterns suggestive of three subpopulations (cohorts). Abundances fluctuate, but there is an overall progression towards fixation. (B) Shown are locations of all differences between the contaminant *D. acidovorans* ANG1 genome recovered in this study and the *D. acidovorans* SPH-1 genome, as well as the alignment of the assembled contigs and coverage of the genome.

For supplemental figures, tables, and information for Chapter 1, see <https://doi.org/10.1128/mBio.01969-16>

2 Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates

Olm, Matthew R; Brown, Christopher T; Brooks, Brandon; Firek, Brian; Baker, Robyn; Burstein, David; Soenjoyo, Karina; Thomas, Brian C; Morowitz, Michael; Banfield, Jillian F

Published in *Genome Research*, April 2017

The initial microbiome impacts the health and future development of premature infants. Methodological limitations have led to gaps in our understanding of the habitat range and subpopulation complexity of founding strains, as well as how different body sites support microbial growth. Here, we used metagenomics to reconstruct genomes of strains that colonized the skin, mouth, and gut of two hospitalized premature infants during the first month of life. Seven bacterial populations, considered to be identical given whole-genome average nucleotide identity of >99.9%, colonized multiple body sites, yet none were shared between infants. Gut-associated *Citrobacter koseri* genomes harbored 47 polymorphic sites that we used to define 10 subpopulations, one of which appeared in the gut after 1 wk but did not spread to other body sites. Differential genome coverage was used to measure bacterial population replication rates in situ. In all cases where the same bacterial population was detected in multiple body sites, replication rates were faster in mouth and skin compared to the gut. The ability of identical strains to colonize multiple body sites underscores the habit flexibility of initial colonists, whereas differences in microbial replication rates between body sites suggest differences in host control and/or resource availability. Population genomic analyses revealed microdiversity within bacterial populations, implying initial inoculation by multiple individual cells with distinct genotypes. Overall, however, the overlap of strains across body sites implies that the premature infant microbiome can exhibit very low microbial diversity.

2.1 Introduction

Infants are born near sterile and continually acquire microbial colonists until reaching an adult-like state at around 2–3 yr of age (Cilieborg et al. 2012; Faith et al. 2015). The microbiota during the first 100 d of life is especially important, as dysbiosis during this “critical window” has been linked to a number of problems later in life, especially relating to the developing immune system (Costello et al. 2012; Cahenzli et al. 2013; Sim et al. 2013; Arrieta et al. 2015). The nature of dysbiosis during the critical window has yet to be clearly defined, but a number of studies have implicated low-diversity as a marker (Cahenzli et al. 2013; Arrieta et al. 2015). Initial colonists are acquired maternally and from the immediate environment, but early life clinical factors (such as birth by cesarean section and neonatal antibiotic administration) can disrupt the normal acquisition process (Ding and Schloss 2014; Bäckhed et al. 2015; Mueller et al. 2015). Among premature infants, who generally harbor microbial communities of limited diversity and instability (Costello et al. 2013; Sim et al. 2013; Ward et al. 2016), this disruption can lead to colonization by resident microbes of the neonatal intensive care unit (NICU) (Brooks et al. 2014; Shin et al. 2015).

Operational taxonomic units (OTUs) identified from 16S rRNA hypervariable region surveys have approximately genus-level resolution (Tu et al. 2014; Jovel et al. 2016). Using this methodology, it has been suggested that, within 24 h after birth, the microbiomes of the mouth, skin, and gut are undifferentiated (Dominguez-Bello et al. 2010) and that site-specific communities develop over the first weeks of life (Costello et al. 2013; Dominguez-Bello et al. 2016). This could imply a common inoculum to all body sites followed by body site-specific selection and immigration. However, even organisms with identical 16S rRNA sequences have been shown to have different genomic and functional profiles (Prosser et al. 2007; Achtman and Wagner 2008; Luo et al. 2015). Such differences could imply different inoculum sources and processes as well as differences in antibiotic susceptibility and strain complexity. Further, if bacterial populations occupy multiple sites, a strain eliminated from one body site could be replaced by dispersal of the same strain from another site (Costello et al. 2012). This could contribute to both pathogen persistence and retention of founding “keystone species” (Dominguez-Bello et al. 2016). Clearly, more sensitive methods like whole-genome sequencing (considered the gold standard of strain typing) (Snitkin et al. 2012) are needed to determine if strains are the “same” or “different.”

Genome-resolved methods have yet to be widely applied to the human microbiome, and thus the level of microdiversity present within human-associated microbial populations is largely unknown. Strain-level diversity is common across other ecosystems and is hypothesized to contribute to the stability of populations of related organisms in the face of phage predation and changing environmental conditions (Jaspers and Overmann 2004; Erkus et al. 2013; Sharon et al. 2013). Methods based on assembly-free metagenomics have attempted to document strain diversity, but reliance on reference genome sequences limits identification of strains to those that have already been analyzed. Other methods have been proposed to document deviations from reference strain sequences, including ConStrains (Luo et al. 2015) and PanPhlan (Ward et al. 2016), but these methods only consider portions of the genome (specific marker genes and coding regions, respectively) and thus fall short of the resolving power needed to account for small-scale differences (Snitkin et al. 2012). While requiring more computational time and manual curation, genome-resolved metagenomics has been used to successfully investigate strain-level differences in the infant gut microbiota several times (Morowitz et al. 2011; Sharon et al. 2013; Raveh-Sadka et al. 2015) and, in combination with previously developed methods (Lang et al. 2013; Luo et al. 2015), has the potential to identify subpopulations of microbes that differ by even a single nucleotide.

A recent study by Browne et al. (2016) found that over half of microbes in the human gut can enter nonvegetative states. This is an important fact to consider when interpreting the studies referenced above, as while the same microbes may be present in different environments, their activity levels in distinct body sites have yet to be investigated. A number of laboratory methods have been developed to discriminate between live and dead cells, including the use of propidium monoazide (Nocker et al. 2006), redox sensing probes (Rodriguez et al. 1992), and the incorporation of radioactive substrates (Karl 1979). However, these methods have limited ability to discriminate between levels of activity, require extensive testing for use with different organisms, and would be difficult to perform in the context of human hosts. An attractive solution is the utilization of differential genome coverage, a method recently described by Korem et al. (2015), that measures the fraction of the bacterial population currently undergoing active DNA replication. However,

assembly errors in reference genome databases (Salzberg and Yorke 2005) and divergence among microbial genomes (Tenailon et al. 2010; Luo et al. 2011; Rosen et al. 2015; King et al. 2016) cast doubt on methods that map directly to genomes in reference collections. Using draft genomes recovered from the samples themselves would solve this problem, but as circular genomes are only rarely recovered, a method to determine the order of the contigs before calculation of growth rates is needed.

2.2 Materials and methods

2.2.1 Patient recruitment and sample collection

Fecal samples from two preterm infants hospitalized in the NICU in Magee-Womens Hospital of UPMC (Pittsburgh, PA) were collected as available over the first month of life. Both infants were of low gestational age (<30 wk), and Infant 2 was of extremely low birth weight (<1000 g). See Supplemental Table S2.1 for additional clinical information.

Fecal samples were spontaneously expelled and collected from diapers or acquired directly using an established perineal stimulation procedure (Morowitz et al. 2011). Skin and oral samples were obtained by a member of the study team using a BD BBL Culture Swab EZ. Oral swabs were collected by rolling the swab head 5–10 times over the dorsal surface of the tongue. If intubated, the sample was collected by swabbing any exposed surface of the tongue. Skin swabs were collected by first dipping the swab into a 0.5 mL aliquot of a sterile solution of 0.15 M NaCl and 0.1% Tween 20. The swab head was then rolled 5–10 times over the left anterior upper chest wall. Stool samples were placed promptly into -20°C storage and transferred to a -80°C freezer for long-term storage as soon as possible. Swab samples were placed promptly in a -80°C freezer for storage.

DNA was extracted using either the MO BIO PowerSoil DNA Isolation kit (single tube extractions) or PowerSoil-htp 96-Well DNA Isolation kit. DNA extracted from stool using the single tube format followed the protocol as previously described (Raveh-Sadka et al. 2016). For DNA extracted from feces with the 96-well kit, fecal samples were added to individual wells of the bead plate and stored overnight at -80°C . The next day, the Bead Solution and Solution C1 were added, and the plates were incubated at 65°C for 10 min. The plates were shaken on a Retsch Oscillating Mill MM400 with 96-well plate adaptors for 10 min at speed 20. The plates were rotated 180° and shaken again for 10 min at speed 20. All remaining steps followed the manufacturer's centrifugation protocol. For swab samples, the swab head was cut off directly into the wells of the bead plate and stored overnight at -80°C . The next day, the Bead Solution and Solution C1 were added, and the plates were incubated at 65°C for 10 min. The plates were shaken on a Retsch Oscillating Mill MM400 with 96-well plate adaptors for 5 min at speed 20. The plates were rotated 180° and shaken again for 5 min at speed 20. The Solution C2 and C3 steps were combined (200 μL of each added) to improve DNA yield. All remaining steps followed the manufacturer's centrifugation protocol.

2.2.2 Metagenomic sequencing and assembly

Sample preparation and sequencing of skin and oral samples were performed at the University of Illinois at Urbana-Champaign sequencing facility, and sample preparation and sequencing of fecal samples were performed at the University of California at Berkeley Vincent J. Coates Genomics

Sequencing Laboratory. Paired end reads of 160 bp with a combination of 1000- and 600-bp library insert sizes were sequenced using an Illumina HiSeq 2500 (Supplemental Table S2.2). Reads were trimmed with Sickle (<https://github.com/najoshi/sickle>). Reads that mapped to the human genome with Bowtie 2 (Langmead and Salzberg 2012) under default settings were discarded. An additional step of mapping with BMap (Bushnell 2014) was performed on all projects with at least 10% of reads removed with Bowtie 2 mapping. See Supplemental Table S2.2 for depth of sequencing and levels of human contamination in each sample.

Reads were assembled using *idba_ud* (Peng et al. 2012) under default settings. Resulting scaffolds >1 kb in length were annotated using Prodigal (Hyatt et al. 2010) to predict open reading frames using default metagenomic settings. Annotated protein sequences were searched against KEGG (Kanehisa et al. 2014), UniReff100 (Suzek et al. 2007), and UniProt databases using USEARCH (Edgar 2010). All matches with bit scores greater than 60 were saved, and reciprocal best hits with a bit score greater than 300 were also cataloged. We identified rRNA sequences using Infernal (Nawrocki and Eddy 2013) by searching against databases from the SSU-Align package (Nawrocki 2009) and tRNAs using tRNAscan_SE (Lowe and Eddy 1997).

Genome binning was carried out using the online interface within ggKbase as described previously (Raveh-Sadka et al. 2015; <http://ggkbase.berkeley.edu/>). This method takes into account phylogenetic profile, GC content, and coverage information. Bins were refined based on differential coverage implemented using time-series emergent self-organizing maps as described previously (Sharon et al. 2013). The completeness of bacterial bins was evaluated based on the presence or absence of single-copy genes (Raes et al. 2007; Raveh-Sadka et al. 2015). Phage sequences were identified based on the presence of typical phage genes such as capsid, terminase, and tail-fiber and as distinct clusters in ESOMs; phage-host relationships were inferred based on phylogeny of annotated proteins and abundance patterns.

2.2.3 Genome recovery

All genome bins from all samples were pooled based on (1) the infant the genome was recovered from, and (2) the genome's identity as either of phage/plasmid origin or of bacterial origin. Genomes within each pool were then compared in a pairwise fashion based on ANIm (Richter and Rosselló-Móra 2009). For all clusters of genomes with high ANI values among members, a representative genome was chosen based on the highest total bin length, lowest scaffold fragmentation, and most complete complement of single-copy genes. Ambiguous genome clusters were visualized using Mauve alignments in Geneious (Kearse et al. 2012) to decide whether genomes could be included in the cluster.

Next, for each infant, ANIm was determined for all representative bacterial bins and all representative phage/plasmid bins together. The resulting ANI matrix was manually curated to resolve cases of overlap between the two genome sets. Most cases of overlap were of bacterial genomes containing prophage that were also represented in the phage list. These were resolved by removal of the prophage scaffold from the bacterial bin. The final genome list for each infant was verified by mapping reads from each project to the genomes to confirm strong coherence between the reads and the genomes, as well as to verify the completeness of the list based on total percentage of mapped project reads. Read-mapping data are available in Supplemental Table S2..3, and the final genome list is available in Supplemental Table S2..6.

2.2.4 *Sample profiling*

Reads from all samples were mapped to the corresponding infant's genome list generated above. SAMtools (Li et al. 2009) was used to convert mapping files (.sam) to mpilup format, and `calculate_breadth.py` and `pileup_profile.py` were used to determine the depth of coverage, breadth of coverage, and average nucleotide identify of each genome in each project. As SAMtools has an implicit coverage limit of 8,000 \times , coverage values from `calculate_breadth.py` were used. The results of both scripts were manually combined and are available in Supplemental Table S2.7.

A strain was considered a “colonist” of a body site if at least 1% of the reads from at least one sample from at least one body site mapped to the recovered genome. We chose to define colonization in this way to be consistent with previous studies of infant colonization (Ward et al. 2016) and because using a coverage-based threshold would have biased against samples with less sequencing reads (Supplemental Fig. S2.4). The same analysis was performed on all phage and plasmid sequences, using 99% breadth to define carriage (Supplemental Fig. S2.2). To identify when strains below the 1% threshold first appeared in body sites, we normalized to account for different sampling depths by using a read percentage cutoff which corresponds to 0.1 \times coverage of the most shallowly sampled data set (see Jupyter notebook, `CallingColonisits`, for details).

2.2.5 *ANI calculation from metagenomic reads*

Metagenomic reads from each sample were mapped to the genome list described previously, and nucleotide variants between reads and genomes were determined using `pileup_profile.py`. See the data availability section for full source code. Briefly, the script calculates ANI by masking regions of DNA near the ends of scaffolds, in conserved regions (tRNAs and rRNAs), or of insufficient coverage, locating all base pairs along the unmasked genome in which at least 80% of reads conflict with the reference genome, and calculating consensus ANI as $(1 - [\# \text{ variant positions} / \text{unmasked genome length}])$. As shown in Supplemental Table S2.3, the number of SNPs found using this method was extremely low (average 34.2), with an average consensus ANI of 99.998%. We attempted to reduce the number of erroneously called SNPs and found that some appear to represent errors made during the process of metagenomic assembly (generation of the reference sequence) or unmasked regions of high sequence conservation (which recruit reads from other genomes) rather than real biological differences. Given this, and the extremely high reference ANI between strains on different body sites, we defined the strains as identical if they met the criterion of >99.9% consensus ANI.

2.2.6 *Growth rate determination*

To attain growth rates for the incomplete genomes recovered in this study, genome fragments were ordered and oriented to previously isolated reference genomes and the peak-to-trough coverage ratio was determined using `bPTR.py` based on the method previously described (Korem et al. 2015; Brown et al. 2016) (Supplemental Table S2.8). Circular reference genomes of the same species as draft genomes from this study were downloaded from NCBI GenBank. The expected form of the cumulative GC skew of genomes (Grigoriev 1998) was manually verified using the program `gc_skew.py`, and genomes with aberrant patterns were discarded. The ANI of each draft genome to all reference genomes was determined using the previously described ANIm method, and the reference genome with the highest ANI was chosen. Draft genome fragments were aligned to the

reference genome using BLAST (Altschul et al. 1990), and any fragment with <20% alignment coverage was discarded. The remaining draft sequence fragments were then aligned to the reference genome using progressive Mauve (Rissman et al. 2009) (`java -Xmx500m -cp Mauve.jar org.gel.mauve.contigs.ContigOrderer`), resulting in an ordered and oriented draft “core genome.” All core genomes were verified by manual inspection of the cumulative GC skew and genome coverage plots generated by the script (Supplemental Figs. S2.5, S2.6). Projects with aberrant plots or coverage below 5× were excluded from analysis.

2.2.7 *Microdiversity of C. koseri*

To identify differences between the reads of specific data sets relative to reconstructed genomes (described above), VarScan (Koboldt et al. 2009) was run on all .pileup files using the `pileup2cns` command with the flag `-min-coverage = 3`. The frequency of each single nucleotide variant was tabulated for all samples using the script `polymorpher2.py` (Supplemental Table S2.9). Variants were then filtered based on a number of criteria and clustered based on changing relative frequency. Briefly, variants were required to have a minimum of 10× coverage in all samples, over 0.2 frequency in at least two samples, not be defined by polymorphisms present in conserved regions (rRNAs, tRNAs) (Supplemental Table S2.9), and pass an auto-correlation threshold. Clustering of variants was done using the Scipy hierarchical clustering package, with the cutoff threshold of 0.275 decided based on manual inspection of the resulting clusters. Full source code of analysis performed is available in the Jupyter notebook, `CitroK_microdiversity` (Supplemental Material).

2.2.8 *CRISPR analysis*

CRISPR arrays were identified in bacterial draft genomes using the program CRISPRFinder (Grissa et al. 2007). CRISPR spacer targets (protospacers) were identified by searching spacers for full-length BLAST hits in a sequence database, followed by filtering our results that also included full-length matches to CRISPR repeats (to remove instances of the CRISPR array itself). In addition to the assemblies of this study, we also searched for protospacers in previously published studies from the same NICU (Raveh-Sadka et al. 2015, 2016) and the NCBI nt database (accessed January 2016). DNA fragments with identifiable CRISPR arrays were excluded from the protospacer search. The alignments of *E. faecalis* CRISPR arrays were manually curated within Geneious (Kearse et al. 2012). The ratio of variant arrays was determined by comparing the number of reads that mapped (using Bowtie 2 default settings) to unique regions of both arrays. Mutations elsewhere in the genome that varied in frequency with CRISPR variants were identified based on a significant Pearson correlation. Source code is available in the Jupyter notebook, `E. faecalis microdiversity` (Supplemental Material).

2.3 Results

2.3.1 *Community profile*

Two premature infants were recruited for this study with parental consent at Magee-Womens Hospital of the University of Pittsburgh Medical Center. Clinical information for these infants is summarized in Supplemental Table S2.1. Both infants were born via vaginal delivery to women with pregnancies complicated by chorioamnionitis. Each infant was treated immediately after birth with 7 d of initial antibiotic treatment (ampicillin and gentamycin). We collected 17 and 20 fecal

samples during the first month of life for the two infants (referred to as #1 and #2). Additionally, two skin swabs and two tongue swabs were collected for each infant (see Fig. 2.1 for detailed information about timing of skin and oral swabs). For the skin and mouth, sample choice was based on the availability of sufficient DNA and, where more than two samples were available, to span the longest time period. In total, 183 gigabase-pairs of Illumina shotgun DNA sequencing were generated for 45 samples (Supplemental Table S2.2).

Historically, assembly-based metagenomics of the skin and mouth has been hampered by human DNA contamination (Liu et al. 2012; Tsai et al. 2016). While all but one skin or mouth sample consisted of >50% human DNA (range 34.2%–93.8%), the deep sequencing effort ensured that all samples had enough reads from microbial genomes for successful de novo genome reconstruction. Overall, an average of 98.7% (Infant 1) and 95.0% (Infant 2) of nonhuman reads could be assigned to assembled genomes (in some cases, by mapping to genomes reconstructed from another sample) (Supplemental Table S2.3).

Using ggKbase (Raveh-Sadka et al. 2015), we manually binned assembled DNA sequences to genome bins based on G+C content, coverage, and phylogenetic profile and subsequently used sequencing coverage patterns for binned scaffolds to verify the bins (Raveh-Sadka et al. 2015). For the bin refinement step, the clustering of fragments was analyzed using emergent self-organizing maps (ESOMs). Since identical genomes were assembled from different samples, the genomes were de-replicated based on similar average nucleotide identity (ANI), and the best genome for each strain was selected for downstream analyses. Across all three body sites for Infant 1, we recovered nine near-complete bacterial genomes and three partial genomes (all partial genomes were from the skin data sets). From Infant 2, we recovered 11 near-complete bacterial genomes and three partial genomes (all partial genomes were from the mouth). The community composition of all samples from all body sites and both infants are presented in Figure 2.1. It is interesting that the first two gut samples for Infant 2 that were collected during the antibiotic treatment course were dominated by *Streptococcus agalactiae*. The placenta showed heavy growth of *S. agalactiae* (group B streptococcus), yet, as noted above, the infant's blood cultures were negative.

The overlap in community composition between body sites is shown in Figure 2.2. Strains were considered identical if they had over 99.9% whole-genome ANI and were considered colonizers of a body site if they accounted for >1% of reads in any sample from the site (Supplemental Table S2.4). Despite obvious differences in habitat characteristics, we identified some identical strain populations in all three sampled body sites for both infants. Infant 1 body sites were heavily colonized by *Citrobacter koseri*, which comprised over 60% of all three communities. Six strains were colonists of more than one body site of Infant 2: *Escherichia coli*, *Pseudomonas aeruginosa*, *Klebsiella pneumoniae*, and *Serratia marcescens* colonized all three body sites, *Enterococcus faecalis* colonized the mouth and skin, and *Staphylococcus epidermidis* colonized the mouth and gut. Interestingly *E. coli*, which is traditionally thought of as a gut colonist (Tenailon et al. 2010), accounted for the highest portion of Infant 2 reads at all three body sites. The infants were housed in different NICU rooms ~3 mo apart, and no bacterial strains were shared between infants.

We also tested for the presence of organisms in multiple body sites at low abundance (<1% of the community) and, to the extent possible, evaluated the time periods in which specific strains

appeared at these sites (for details, see Methods). For Infant 1, *Enterococcus faecalis* had appeared in the skin and gut by the time of collection of the first samples, persisted in the gut, and was present in both habitats at the later time point when samples from both body sites were collected (day of life [DOL] 23). A *Haemophilus parainfluenzae* population was not detected in the first collected samples from all body sites but was present in the gut on DOL 21 and had appeared in the mouth and skin 2 d later. For Infant 2, *Staphylococcus epidermidis* and *E. coli* were present in all body sites at most time points. *Pseudomonas aeruginosa* and *Klebsiella pneumoniae* were undetectable in gut samples collected during antibiotic administration (Fig. 2.1; Supplemental Table S2.3) but appeared in all three body sites the day after cessation of antibiotics (DOL 7). *Staphylococcus sp.* M0480 was present by the time the first skin (DOL 12) and mouth (DOL 14) samples were collected but was undetectable in the mouth at the second time point. *E. faecalis* was present in the first-sampled gut and mouth communities (and persisted there), was absent in the skin on DOL 12, but had appeared there by DOL 22. *Serratia marcescens* was a relatively late colonist, appearing first in the gut on DOL 19 and was present in all three body sites at later time points (Supplemental Table S2.3). In general, strains became detectable at all three body sites at around the same day of life, with the exception of *E. faecalis* in Infant 2, which persisted in the gut and mouth for over a week before being detected in the skin.

2.3.2 Growth rates are different across body sites

Recently, it was shown that accurate growth rates of microbial strains in their natural environment can be determined by measuring the ratio of the coverage of DNA at the origin and terminus of replication (Korem et al. 2015). However, this method requires complete closed circular genomes (which are rarely acquired from metagenomic studies) in order to locate the origin and terminus. We were able to circumvent that requirement by orienting each strain assembly (median number of contigs 67.5, range 12–1460) to a representative isolate genome in order to determine the order and orientation of the contigs. When available, we used multiple isolate genomes to confirm the best assembly, as some genomes in the RefSeq database were found to be incorrectly assembled around rRNA operons (Fig. 2.3A,B). The peak-to-trough ratio (origin to terminus coverage ratio) was then determined by mapping reads to the oriented assembly (see Methods).

Surprisingly, in all cases for which we could determine growth rates for the same strain from multiple body sites, growth was faster in the skin and mouth than in the gut (Fig. 2.3). When all growth rates were analyzed, microbes in the skin and mouth had significantly higher growth rates than microbes in the gut ($P < 0.00001$), whereas strains in the skin and mouth did not differ significantly in their growth rates ($P = 0.12$; Mann-Whitney U test). The strains exhibiting the fastest growth in the skin, mouth, and gut were *P. aeruginosa*, *Streptococcus mitis*, and *Clostridium perfringens*, respectively.

In general, growth rates measured for strains in the gut increased with increasing infant age, consistent with population recovery after early antibiotic treatment (Spearman rank correlation, $R = 0.30$, $P = 0.0005$, $n = 135$) (Fig. 2.4). In Infant 2, *S. agalactiae*, also known as group B Streptococcus, accounted for over 80% of the microbial community during antibiotic treatment in Infant 2 for a presumed group B Streptococcus infection. Despite being abundant, the organism exhibited a low and decreasing growth rate during the treatment period. Several microbial strains exhibited sharp changes in growth rates over the first month of life (*Clostridium difficile*, *C. perfringens*, *P. aeruginosa*). However, these events did not coincide with medical

events indicated by the clinical metadata and were probably short lived, as they did not always lead to changes in relative abundance of these strains in the next-collected sample.

2.3.3 *Microdiversity*

We classified strains colonizing multiple body sites as the same based on >99.9% genome-wide ANI, but this analysis is insensitive to very small-scale differences that could be used to document subpopulation dynamics and constrain inoculum diversity. Thus, we performed high-resolution analyses of *Citrobacter koseri*, the strain with the highest coverage in all body sites of Infant 1. The 4.66-Mbp genome was initially assembled de novo into 27 scaffolds. By reference to an isolate genome, we confirmed potential joins supported by sequence overlaps and identified gap-filling reads so that these scaffolds could be reconstructed into a complete circular genome.

Single nucleotide polymorphisms were identified by mapping reads from each body site to the circularized *C. koseri* genome. No fixed mutations distinguished populations colonizing the skin, mouth, and gut. However, 47 polymorphic sites with a minimum frequency of 0.2 were identified along the genome. Of these, 21 occurred in intergenic regions and five within the coding region of *yadA*, a gene encoding adhesins with known pathogenic alleles (El Tahir and Skurnik 2001). Using hierarchical clustering, 10 cohorts of polymorphisms with similar variations in frequency over the sample series were identified, five of which had only one member (referred to as “singletons”) (Fig. 2.5). Each cohort is inferred to represent a strain subpopulation. Particularly interesting is cohort 1, which underwent a dramatic purge event around DOL 23, and singleton 3, which rapidly rose above detection level on DOL 17.

To evaluate differences in the populations across body sites, we focused on DOL 23, where gut, skin, and mouth samples are all available for Infant 1. Three variant positions had significantly different fractions of polymorphisms between the gut and both the skin and mouth (Fisher's exact test with Bonferroni correction, $\alpha = 0.01$). These three positions make up the entirety of cohort 4 and singleton 3. Singleton 3 rises in abundance to comprise ~40% of the gut population on DOL 17 but is only ever detected at ~2% of the mouth and skin populations (Fig. 2.5B). There were no genomic positions with significantly different polymorphism levels between the mouth and skin on either DOL 6 or DOL 23. The detection of differences in subpopulation frequencies between body sites shows that there is limited gene flow between body sites and may indicate the start of in situ diversification.

2.3.4 *CRISPR and phage*

In addition to the bacterial genomes referenced above, our assembly-based metagenomic pipeline resulted in the recovery of 21 bacteriophage genomes and 18 plasmid genomes. An average of 2.1% of reads from all samples mapped to bacteriophage genomes, and the most significant bacteriophage bloom occurred in the gut of Infant 2 on DOL 7 (immediately following cessation of antibiotic administration) (Fig. 2.1). Three bacteriophage genomes alone accounted for ~50% of the DNA sequenced during this bloom: *K. pneumoniae* phage A (7.9% of community; 12,700× coverage), *K. pneumoniae* phage B (39.1% of community; 19,400× coverage), and *E. coli* phage A (2.7% of community; 4100× coverage). A second bloom of *K. pneumoniae* phage B and *E. coli* phage A also occurred in the same environment on DOL 24. Interestingly, although the bacteriophages' abundance in the second bloom was only a fraction of their abundance in the first bloom, both blooms dramatically shifted the microbial community composition (Supplemental

Fig. S2.1). Overall, however, most phage, plasmid, and host population abundance patterns were highly correlated (partly due to integration). Consequently, bacterial distribution patterns across the three body sites generally predicted patterns for the associated phage and plasmids (Supplemental Fig. S2.2). For circularly recovered phage and plasmids (inferred to be nonintegrated), we also found GC skew and coverage patterns that could be indicative of DNA replication style (Supplemental Fig. S2.3). Unfortunately, no phage or plasmids with these coverage patterns were detected at multiple time-points, so the consistency of this pattern could not be evaluated. However, such analyses could possibly be used to elucidate plasmid and phage replication regulation and growth rate in future studies.

CRISPR-Cas loci confer bacterial phage resistance (Horvath and Barrangou 2010). Because CRISPR spacers are added uni-directionally, the loci provide a record of population history (Sun et al. 2016). We identified CRISPR arrays in 50% of bacterial genomes in Infant 1 and 43% of bacterial genomes in Infant 2, with a total of 111 and 182 unique spacer sequences, respectively (Supplemental Table S2.5). However, we found only four spacer targets in the samples from the infant from which the spacers were recovered, even using relaxed (1 mismatch allowed) search parameters. Thus, we broadened our search to include previously published gut metagenomes from the same NICU as this study (Raveh-Sadka et al. 2015, 2016) and the current NCBI database. This revealed 26 and 16 spacers/protospacer matches, respectively.

No changes in CRISPR spacer inventories over the study period of 28 d were found, but two coexisting *E. faecalis* populations were identified in Infant 2 based on distinct CRISPR spacer inventories. The variant A locus contained 13 spacers and the variant B locus contained 11 unique spacers, with one spacer appearing in the array twice (Fig. 2.6). The ratio of the two variants fluctuated over the study period, with variant B rising from 0% of the population on DOL 5 to over 60% during the DOL 12–24 period and declining to ~5% by DOL 28. As the spacers that differentiate variants A and B had no protospacer matches, we sought other genomic features for which selection might explain the abundance changes. One polymorphism was significantly correlated with the variant A locus and another significantly correlated with the variant B locus (Pearson correlation with Bonferroni correction, $\alpha = 0.01$) (Fig. 2.7). Both polymorphisms were single-nucleotide substitutions within the coding regions of separate hypothetical proteins. Bacteria with both CRISPR locus variants and the associated gene variants were detected in the skin on DOL 22 (the only sample from another body site with sufficient coverage for detection). CRISPR arrays reconstructed in this study were also compared to a number of previously isolated *E. faecalis* strains (Fig. 2.6; Palmer and Gilmore 2010). Surprisingly, we found remarkable similarity in spacer content (with no polymorphisms in the spacer sequences) for *E. faecalis* strains isolated as many as 82 yr apart. Additionally, we found that isolates from the 1970s still retain CRISPR that match the sequences of phage still found in the NICU today.

2.4 Discussion

Twenty-six bacterial genomes, as well as 39 phage/plasmid genomes, were recovered from two premature infants. Given that 98.7% of all reads mapped to 12 reconstructed bacterial and 18 phage/plasmid genomes from Infant 1, and 95.0% to 14 bacterial and 21 phage/plasmid genomes from Infant 2, we conclude that the majority of the community was accounted for. This result confirms the overall low diversity of the early community when compared to full-term infants (Costello et al. 2013; Gibson et al. 2016; Ward et al. 2016). Overlap of strains across body sites

contributes to the low total diversity of these premature infant microbiomes compared to those of other infants.

For *C. koseri*, we confirmed the complete absence of any fixed mutations that would distinguish populations in the mouth, skin, and gut, allowing us to conclude that identical populations colonized all three body sites. However, colonizing populations are typically not clonal (Luo et al. 2015). Analysis of subpopulation microdiversity, needed, for example, to constrain inoculum diversity, is complicated because recruitment of reads from different taxa to homologous regions can cause miscalculations in commonly used variant-detecting programs (Wilm et al. 2012; Deatherage and Barrick 2014). In this study, *C. koseri* had sufficient coverage in all Infant 1 samples to reliably detect variant positions (average coverage 493, full range 73–769), and no other similar taxa colonized the infant concurrently (preventing erroneous read recruitment). This allowed us to identify seven early colonizing *C. koseri* subpopulations (present in at least 20% of the reads) defined by between one and 29 polymorphisms, likely reflecting inoculation by at least seven distinct cell genotypes. Microdiversity has been linked with taxon stability in other environments (Jaspers and Overmann 2004; Rodriguez-Brito et al. 2010; Erkus et al. 2013), so seemingly low microdiversity in infant-associated populations may contribute to the observed low community stability as body habitats change along with infant development (Costello et al. 2013).

Our approach can constrain the timing and directionality of colonization events. For both infants, the presence of the same strains in multiple body sites at the first sampling event may indicate an early widespread inoculation event from the same source. The presence of the same population in multiple body sites suggests the ability of body sites to act as strain reservoirs for one another in early life. For example, colonization of the gut and mouth of Infant 2 by *E. faecalis* was followed by dispersal to the skin. However, we identified a strain of *C. koseri* that appeared later in the gut colonization process, but this strain did not spread to other body sites. This result may indicate increasing body site specificity as infant age increases and would imply functional significance of single mutations. This deduction may be supported by the dramatic shifts in abundances of genotypically near-identical *E. faecalis*, which were likely due to single nucleotide polymorphisms, given that CRISPR spacers that otherwise distinguish the variants did not have targets in the same samples.

The lack of coexisting CRISPR spacer targets (yet presence of targets in other samples) is likely due to the phage immunity conferred by the CRISPR spacers. Conservation of *E. faecalis* CRISPR spacers over many decades without mutation (Fig. 2.6) suggests that they target mutation-resistant phage genome regions. Slow CRISPR evolution contrasts with the dynamic seen in other systems (Tyson and Banfield 2007; Pride et al. 2011) but is consistent with observations of conserved CRISPR arrays in other common enteric organisms (Touchon and Rocha 2010; Touchon et al. 2011).

To our knowledge, this study represents the first comparison of in situ bacterial growth rates of multiple body sites, and the comparison is especially powerful as the measurements were for identical strains in different environments. In all cases when growth was measured for the same strain in multiple body sites, growth was slowest in the gut (Fig. 2.3C). Several mechanisms could explain the difference in growth between body sites, including differences in (1) higher resource availability (including oxygen) on the skin and in the mouth compared to the gut, (2) higher levels

of competition among microbes in the gut, or (3) host control through the innate and/or adaptive immune system (Donaldson et al. 2015).

Previous studies have described the gut microbiome of premature infants as relatively simple and prone to rapid changes in composition (Costello et al. 2013; Gibson et al. 2016; Ward et al. 2016). To our knowledge, this is the first study to investigate the body habitat range of individual genotypes and to compare microbial activity of the same populations across body sites. Colonization of the three studied body sites by the same populations may be due to overall low inoculum diversity in the highly cleaned NICU and limited human contact. Given the rapid measured growth rates, the premature infant skin and mouth appear to be desirable microbial habitats (Fig. 2.3). It remains to be seen whether similar observations hold true for full-term infants, how long features of the founding communities persist, and whether differences in community composition arising from prematurity have long-term health consequences.

2.5 Figures

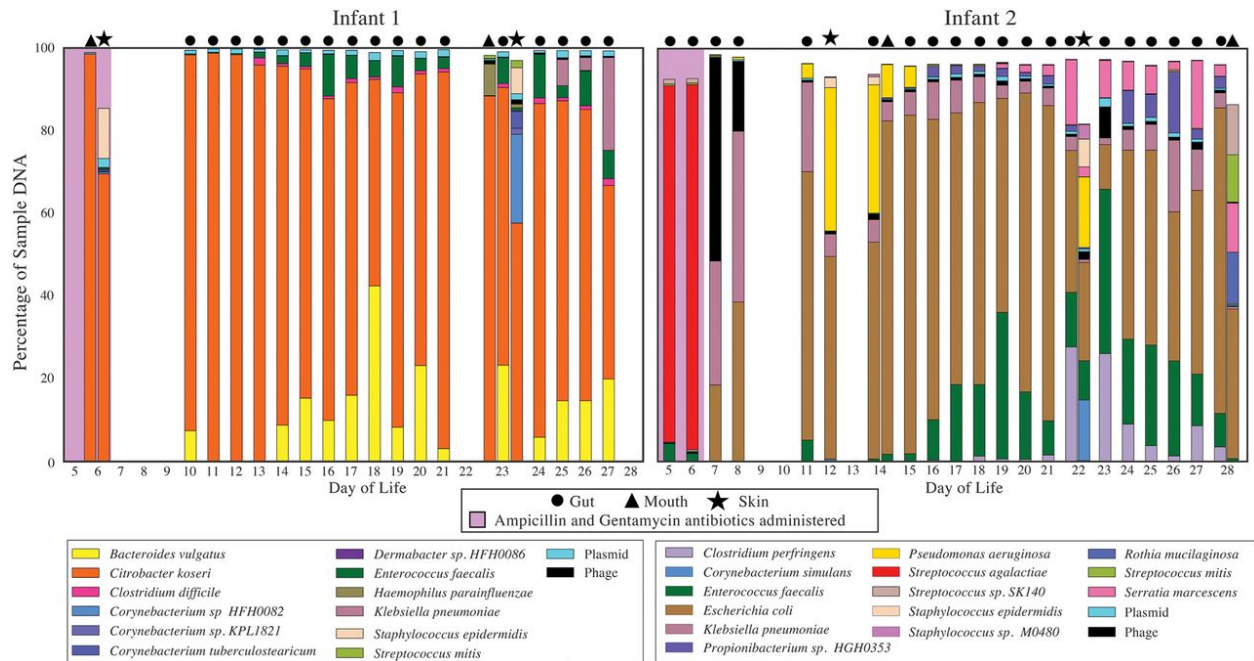


Figure 2.1 Compositional profile of microbial communities colonizing the mouth, skin, and gut of two premature infants. Each colored box represents the percentage of nonhuman reads mapping to an assembled genome, and the stacked boxes for each sample show the fraction of the reads in that data set accounted for by the genomes from that sample.

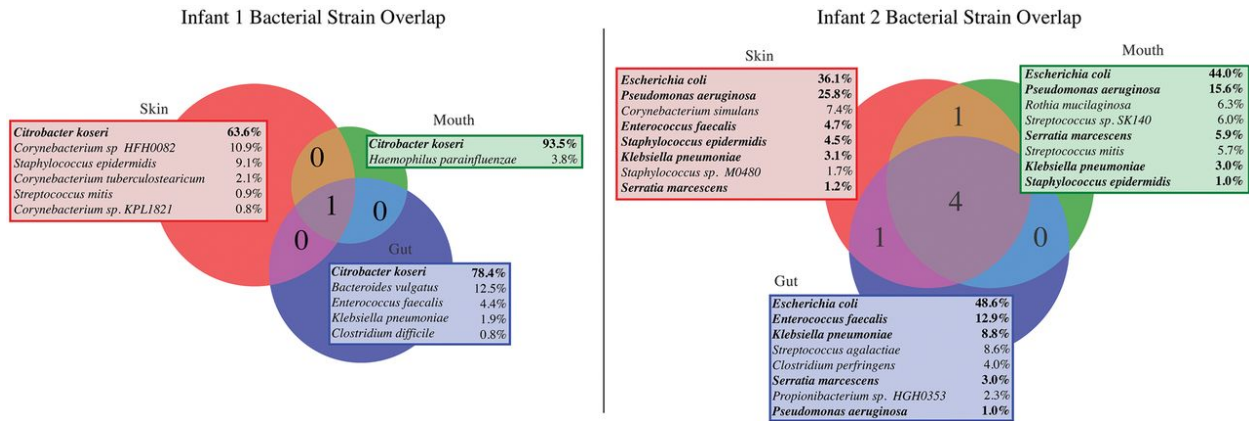


Figure 2.2 Identical bacterial strains colonize multiple body sites of premature infants. Microbes were considered colonists of a body site if they make up >1% of a community. All colonists of each site are shown, along with the total percentage of the community they make up across all sampling events. Colonists of multiple sites are shown in bold.

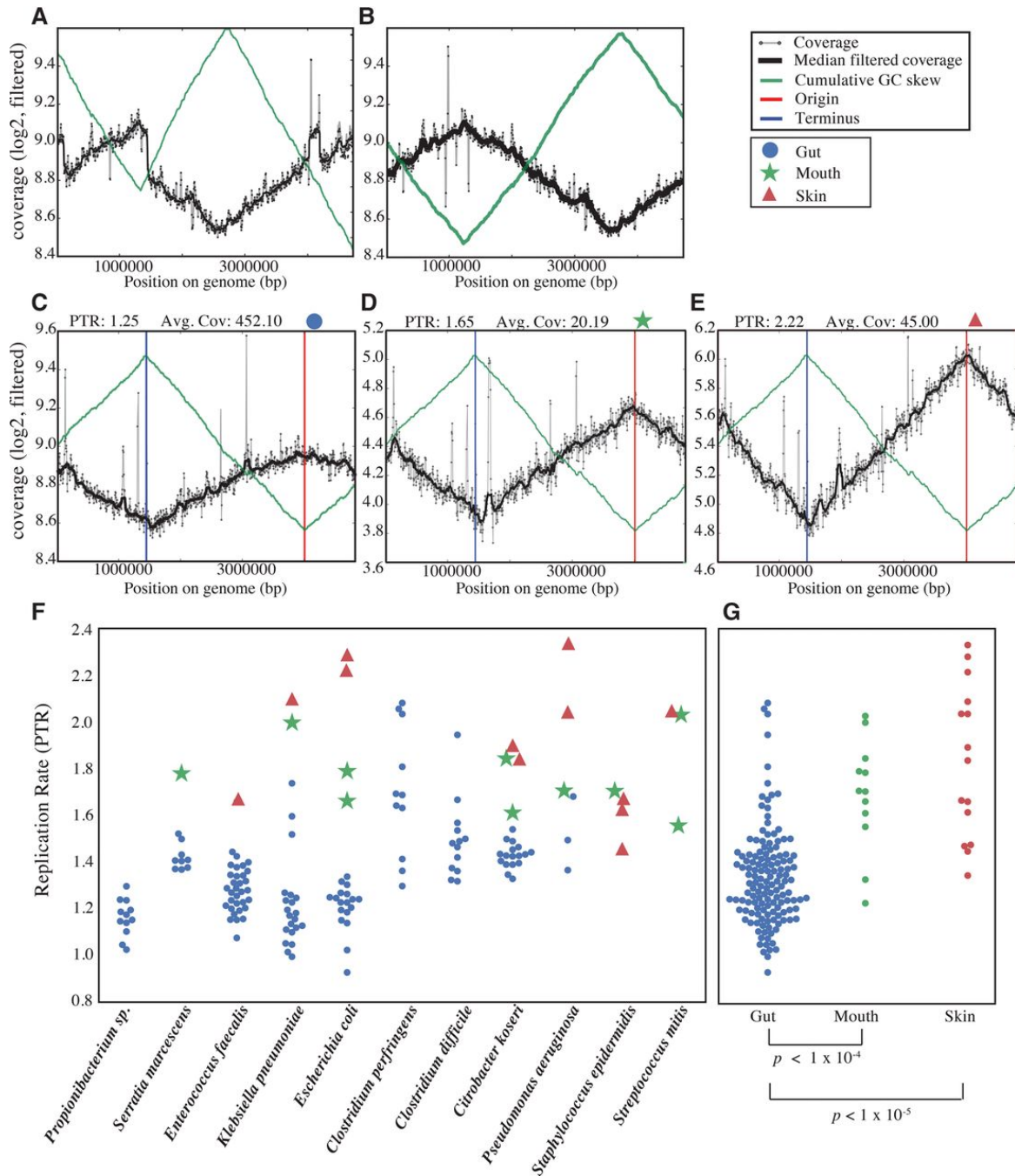


Figure 2.3 In situ bacterial growth rates are faster in the mouth and skin than the gut. (A,B) Cumulative GC-skew (green line) and coverage (black line) for our reconstructed *Citrobacter koseri* genome aligned to the RefSeq genome *C. koseri* strain ATCC BAA-895 (A) and another reference genome (*C. koseri* strain FDAARGOS_86) (B). Based on the irregularity of trends in the *C. koseri* strain ATCC BAA-895 plot, we conclude that this genome was improperly assembled at the rRNA operons. Thus, we used *C. koseri* strain FDAARGOS_86 for ordering and orienting our genome fragments. The ability to uncover assembly errors by inspection of PTR plots underlines the value of these displays. (C–E) Cumulative GC-skew and coverage of our ordered and oriented *Escherichia coli* genome from Infant 2 using reads mapped from a gut

sample (C), mouth sample (D), and skin sample (E). Inspection of the PTR plots ensures that the origin and terminus are determined properly. (F) Aggregate of all peak-to-trough ratio (PTR) measurements for each bacterial species for which at least three measurements were available. In all cases where measurements are available for the same strain growing in multiple body sites, growth is slowest in the gut. (G) Direct comparison of all growth rate measurements for each body site. P-values for Mann-Whitney U test between body sites are shown below.

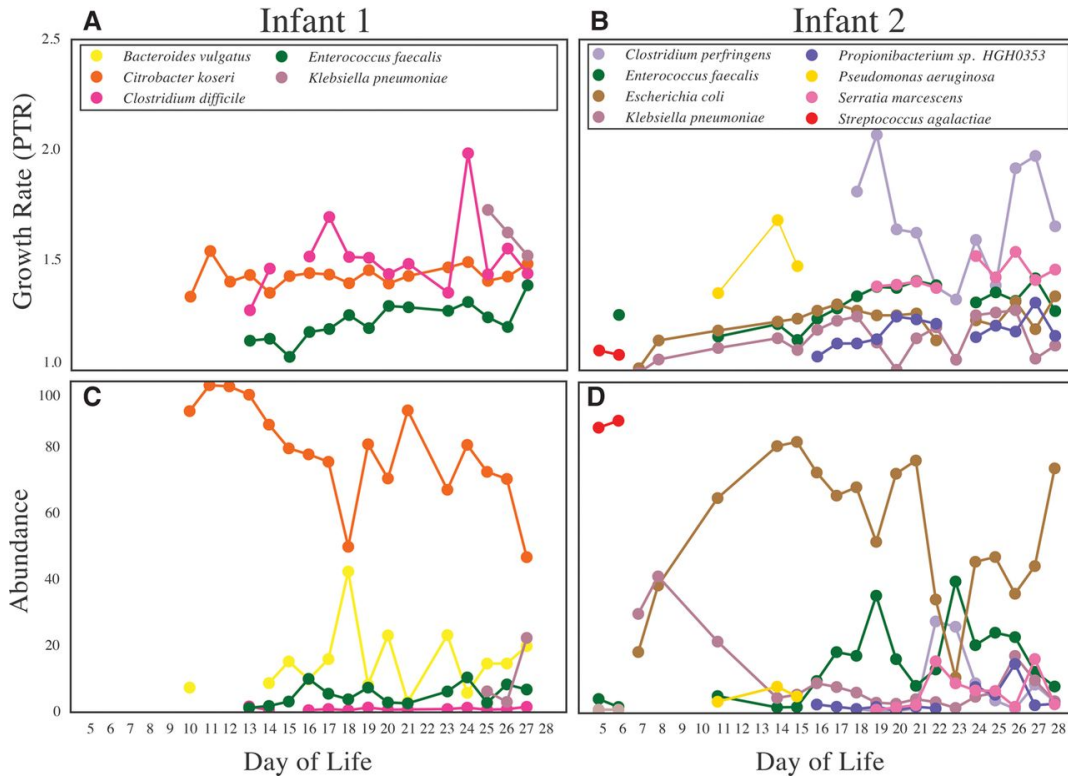


Figure 2.4 Growth rates (determined by PTR) often do not predict changes in relative abundance of a population in subsequent samples. Growth rate measurements for all gut colonists are shown in A and B. Corresponding relative abundance information is shown in C and D. The lack of correspondence between increased PTR in one sample and increased relative abundance in the next sample could be due either to the transient nature of growth spurts or to the fact that cell death is not accounted for in this analysis.

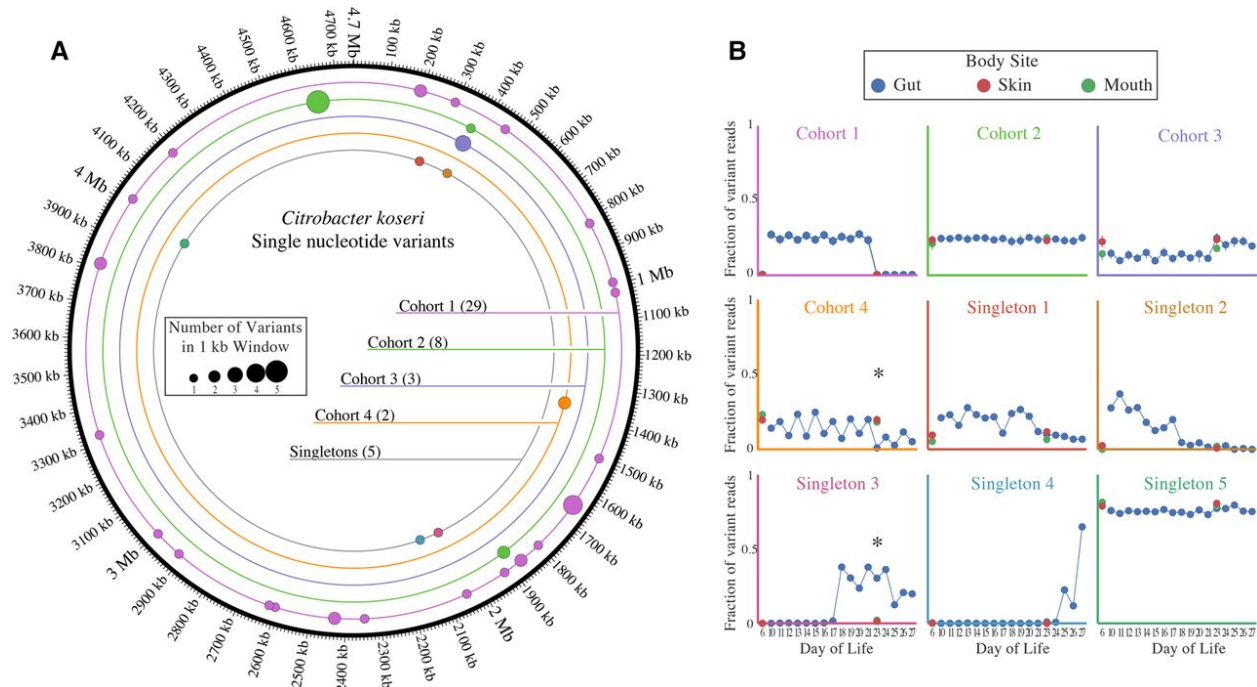


Figure 2.5 Subpopulations exist within colonizing *C. koseri* populations. (A) Single nucleotide variants were identified by mapping reads from all Infant 1 samples to the draft genome of *C. koseri* recovered from Infant 1. The total number of variants in each cohort is listed in parentheses. (B) The frequency of each variant in each sample. Cohorts are plotted as the average of all variant frequencies, with error bars representing standard deviation of the mean. Asterisks represent cases where the frequency of variants is statistically different between body sites (Fisher's exact *t*-test with Bonferroni correction, $P < 0.01$).

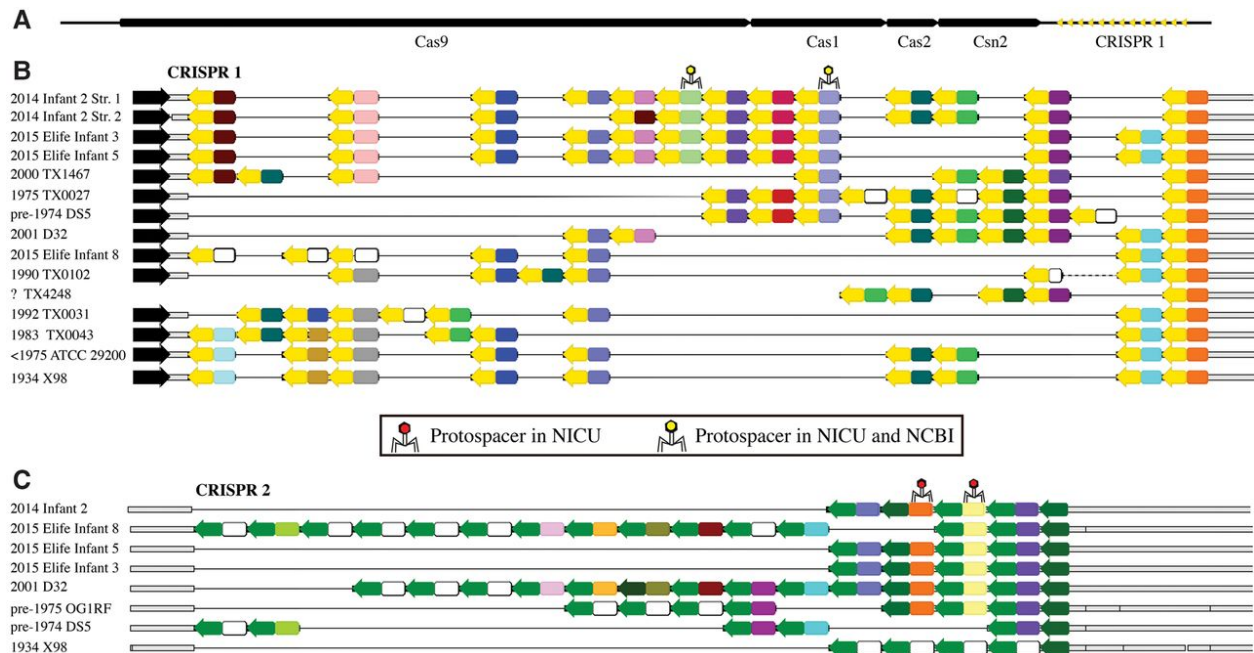


Figure 2.6 CRISPR spacers are maintained over decades in *Enterococcus faecalis*. (A) Genomic organization of CRISPR-Cas array #1. (B) Alignments of array #1 and (C) array #2 from *E. faecalis* from Infant 2 of this study compared to arrays reconstructed from publicly available genomes for isolates. The year of isolation of all *E. faecalis* isolates is provided to the extent possible. Infants marked “Elife” are those from a previous publication from the same NICU (Raveh-Sadka et al. 2015). Arrows represent repeats and colors represent spacers; identical colors symbolize identical spacers, whereas white spacers are unique. Phage symbols represent spacers with a protospacer match (max 1 mismatch) in a sequence assembled from infants in the same NICU as this study (red), and spacers with matches in both the same NICU and a separate genome in NCBI (yellow).

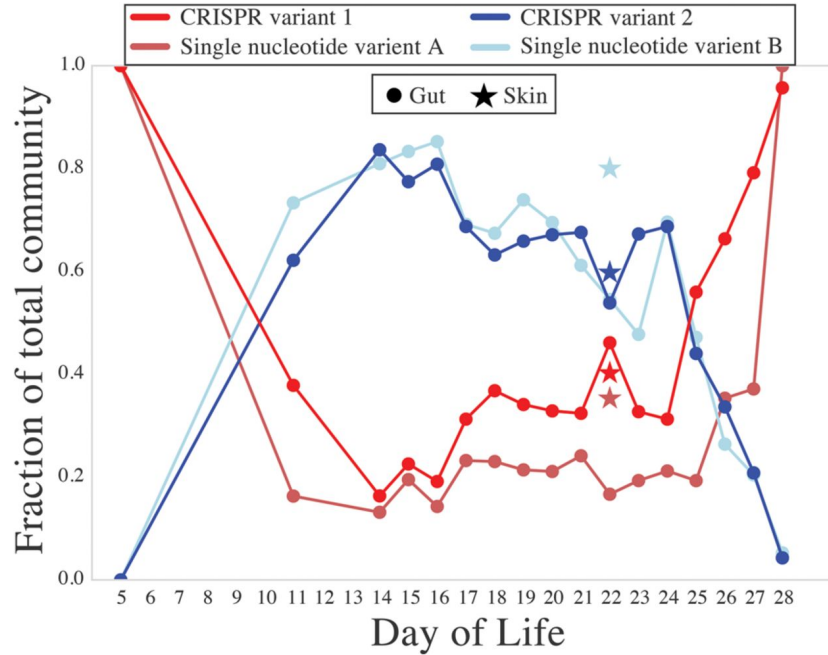


Figure 2.7 *E. faecalis* CRISPR variants change frequency within the *E. faecalis* population over the colonization period for Infant 2. The relative abundances of each CRISPR variant (diagrammed in Fig. 6) are shown. Additionally, two single nucleotide variants within *E. faecalis* correlated significantly with CRISPR variant frequencies (Pearson correlation with Bonferroni correction, $P < 0.01$). Both variants are located in the coding regions of hypothetical proteins.

For supplemental figures, tables, and information for Chapter 2, see <https://doi.org/10.1101/gr.213256.116>

3 dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication

Olm, Matthew R; Brown, Christopher T; Brooks, Brandon; Banfield, Jillian F

Published in *The ISME journal*, December 2017

The number of microbial genomes sequenced each year is expanding rapidly, in part due to genome-resolved metagenomic studies that routinely recover hundreds of draft-quality genomes. Rapid algorithms have been developed to comprehensively compare large genome sets, but they are not accurate with draft-quality genomes. Here we present dRep, a program that reduces the computational time for pairwise genome comparisons by sequentially applying a fast, inaccurate estimation of genome distance, and a slow, accurate measure of average nucleotide identity. dRep achieves a $28 \times$ increase in speed with perfect recall and precision when benchmarked against previously developed algorithms. We demonstrate the use of dRep for genome recovery from time-series datasets. Each metagenome was assembled separately, and dRep was used to identify groups of essentially identical genomes and select the best genome from each replicate set. This resulted in recovery of significantly more and higher-quality genomes compared to the set recovered using co-assembly.

3.1 Introduction

Genome-resolved metagenomics involves the recovery of genomes directly from environmental shotgun DNA sequence datasets (Tyson et al., 2004). This is generally performed by assembling short-read sequences into longer scaffolds, followed by binning together scaffolds belonging to the same genomes. Metagenomic analysis of related samples from the same ecosystem is often employed to investigate compositional stability and spatial or temporal variation. The approach can also reveal microbial co-occurrence patterns and identify factors or processes that control organism abundances. Analysis of sample series data is also important technically, as different abundance patterns across the sample series for different organisms provide valuable constraints for binning (Sharon et al., 2013). In this process, reads from individual samples are mapped back to a collection of genomes that is often obtained by combining the reads from all samples and assembling them together (co-assembly; Bendall et al., 2016; Lee et al., 2017; Vineis et al., 2016). However, co-assembly dramatically increases the data set size and complexity, especially when multiple different strains of the same species are present across the sample series, and can result in fragmented assemblies (Sczyrba et al., 2017).

Independent assembly should generate more and higher-quality genomes than the co-assembly based approach because the complexity of individual samples is lower than that of the combination of samples (Supplementary Figure S3.1). The challenge that arises from independent assembly is that de-replication of the resulting genome set is required (Raveh-Sadka et al., 2015; Probst et al., 2016; Olm et al., 2017). De-replication involves identifying genomes that are the ‘same’ from a

larger set, as well as determining the highest quality genome in each replicate set. This is important to maximize the accuracy of metabolic predictions and other downstream analyses.

De-replication requires pairwise genome comparisons, and the number of comparisons required scales quadratically with an increasing number of genomes. Hundreds of thousands of CPU hours may be needed to de-replicate larger genome sets with robust algorithms (gANI; Varghese et al., 2015). Mash, a recently developed algorithm that utilizes MinHash distance to estimate similarity between genomes, is an attractive alternative due to its incredibly fast speed (Ondov et al., 2016). However, we found that the accuracy of MASH decreases as the completeness of the compared genome bins decreases (Figure 3.1a). Thus, it cannot be used to de-replicate collections of partial genomes.

Here we present dRep, a program that utilizes both gANI and Mash in a bi-phasic approach to dramatically reduce the computational time required for genome de-replication, while ensuring high accuracy. The genome set is first divided into primary clusters using Mash, and then each primary cluster is compared in a pairwise manner using gANI, forming secondary clusters of near-identical genomes that can be de-replicated. Using published information about time required for genome comparisons, we performed an *in silico* simulation of de-replication time for Mash, gANI, and dRep (Figure 3.1b). The results indicate that dRep affords a multiple orders of magnitude increase in computation efficiency compared to naïve gANI.

3.2 Results

To verify this prediction and to test dRep's accuracy, we ran dRep on 1,125 genomes assembled from 195 fecal metagenomes collected from 21 premature infants during the first months of life (Raveh-Sadka et al., 2016). Genomes were 50–100% complete and contained between 0% and 24% contamination according to checkM (Parks et al., 2015). dRep clustered genomes in an identical manner to naïve gANI using default parameters, and showed near-perfect precision and recall using a variety of other parameters (Supplementary Table S1). Mash, on the other hand, resulted in a recall of 51.3% and a precision of 99.9% when compared to gANI, consistent with underestimation of similarity between incomplete genomes (Figure 3.1a). The actual run times were also very close to those predicted by our simulation: 92 versus 93 CPU hours for dRep, 2673 vs 2784 CPU hours for naïve gANI (actual vs predicted run times), and <1 CPU hour in both cases for Mash. As the run-time of dRep depends on the diversity of the genome set, and pre-term infant gut communities are especially non-diverse (Gibson et al., 2016; Ward et al., 2016), even greater increases in computational efficiency are expected from most other environments than predicted by our simulation.

We analyzed the same 195 metagenomes to test the prediction that, for each infant, individual assembly and de-replication would generate more and higher-quality genomes than co-assembly of the read datasets. We de-replicated genomes obtained from assemblies generated from each sample individually as well as from a co-assembly (to recover low-abundance genomes), and recovered a de-replicated genome set with 39% more bins (>75% complete, <5% contaminated) than were obtained from co-assembly alone (270 vs 194 genomes; Figure 3.1c). 76 bins were recovered only from individual assemblies, 35 only from co-assemblies, and 159 from both methods.

We next compared genomes recovered using both methods. A Wilcoxon signed-ranks test indicated that scaffold length, as measured by N50, was significantly higher in genomes from dRep (median=34 046 bp) than genomes assembled from co-assemblies (median=26 103 bp), $P=4.0e-11$. Completeness was also significantly higher in genomes from dRep than co-assembly overall ($P=6.0e-8$), and although the median value was the same for genomes from both sets (median=98.3%), the 5% quantiles were different (91.6% vs 84.9%, respectively; Supplementary Table S2). Visualizations of the similarity between groups of genomes were also generated using dRep (Figures 3.1d and e; Supplementary Figure S3.2). These may be particularly valuable for comparing the population structures of groups of genomes. Taken together, dRep enabled recovery of more and better genomes than co-assembly alone, and is an effective tool for exploring the similarity among large set of genomes.

We used a published fecal metagenome data set with known strain heterogeneity to explore the effect of within-population variation on assembly and genome recovery (Sharon et al., 2013). Samples from the single infant were either co-assembled or samples were assembled individually and then de-replicated using dRep. In the case of *Staphylococcus hominis*, co-assembly generated a contaminated bin (that is, many duplicate and triplicate single copy genes; Figure 3.2a). In contrast, a near-complete, uncontaminated genome was recovered from several individual time-points. Previous work on the same data set (Eren et al., 2015) has shown manual bin curation of the co-assembled bin with *anvi'o* can increase the *S. hominis* bin quality (73% complete; 6.6% redundant), but still not to the level of the un-curated bin from the individual assembly (98% complete; 0% redundant).

3.3 Conclusions

It is both logical, based on the well-known effects of sample complexity, and clear from the analysis of human microbiome samples presented here, that assembly of data from individual samples followed by de-replication has major advantages over co-assembly (especially as co-assembled genomes can be included in the de-replication process). Because it relies on Mash, dRep can only be used if the genomes in the comparison set are >50% complete. dRep combines checkM for completeness-based genome filtering (Parks et al., 2015), Mash (Ondov et al., 2016) for fast grouping of similar genomes, gANI (Varghese et al., 2015) or ANIm (Richter and Rosselló-Móra, 2009) for accurate genomic comparisons, and Scipy (Jones et al., 2001) for hierarchical clustering. In the case of viruses and plasmids, dRep requires use of an independent method to estimate genome completeness because there are currently no established metrics for this in checkM.

dRep is easy to use, highly customizable, and parallelizable. If desired, dRep can perform rapid pairwise genomic comparisons (without de-replication) to enable visualization of the degree of similarity among groups of similar genomes (Figure 3.1; Supplementary Figure 3.2). Version 0.5.5 of dRep is available in the Supplementary Information section (Supplemental Data 1 and 2), and for up-to-date source-code, installation instructions, and the manual, see <https://github.com/MrOlm/drep>.

3.4 Figures

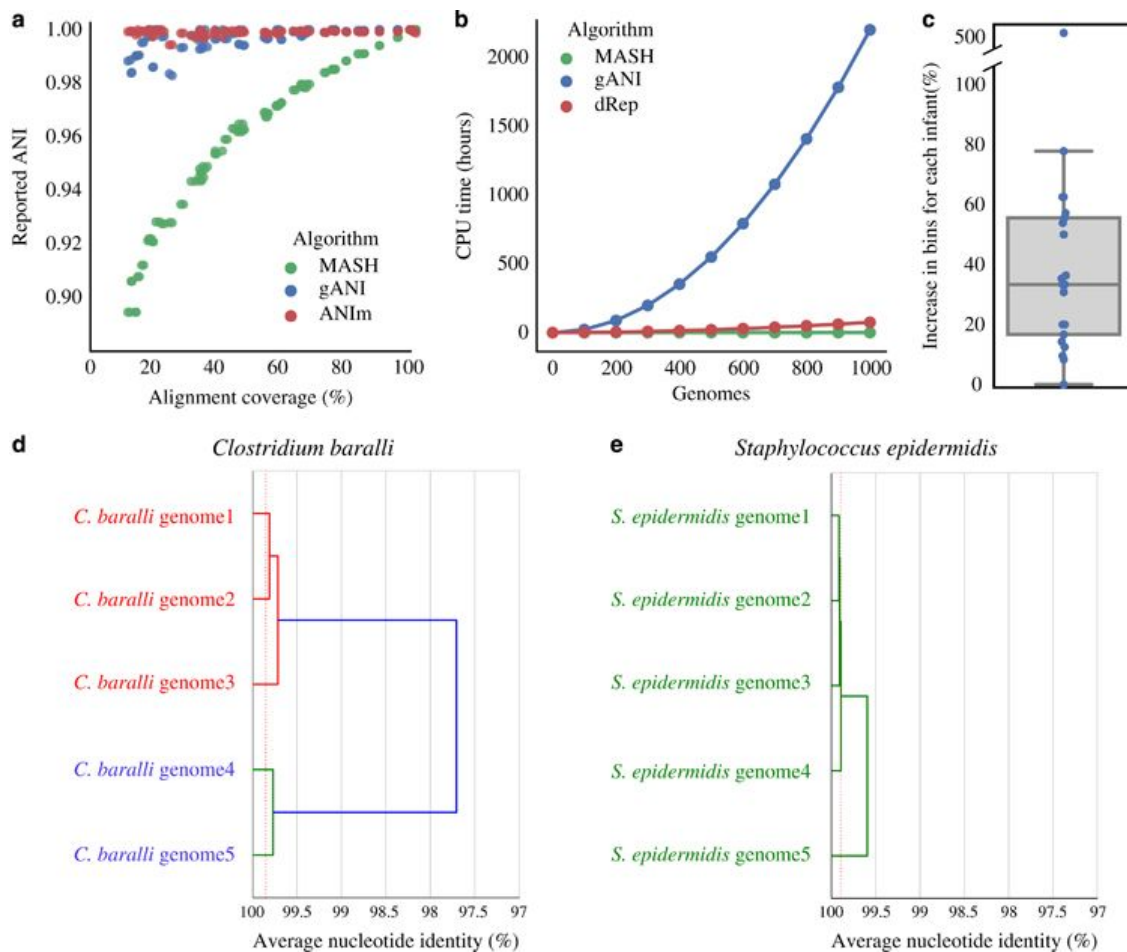


Figure 3.1 Assembly and de-replication with dRep results in more and higher-quality genome bins as compared to co-assembly. **(a)** A complete *Escherichia coli* genome was subset 10 times in increments of 10% (10%, 20%, 30% etc.). Subsets were compared to each other in a pairwise manner (100 total comparisons) using three algorithms- ANIm, MASH and gANI. For each pair of subsets, the alignment coverage between the two genomes as determined by MUMmer is shown on the *x* axis (aligned length / average genome length), and the ANI reported from each algorithm is shown on the *y* axis. ANIm and gANI are accurate when genomes are incomplete, but MASH is only accurate when genomes are essentially complete. **(b)** Using previously reported algorithm runtimes, we estimated the time required to de-replicate genome sets of various sizes. gANI exhibits a sharp exponential climb, limiting its use on larger genome sets; MASH and dRep do not. **(c)** De-replication of bins from individual assemblies and co-assembly (dRep assembly method) resulted in more bins (>75% complete, <5% contaminated) than co-assembly alone. **(d and e)** Examples of genome relatedness figures generated by dRep. The red dotted line is the value of the lowest ANI resulting from a self-vs-self alignment of each genome in the cluster.

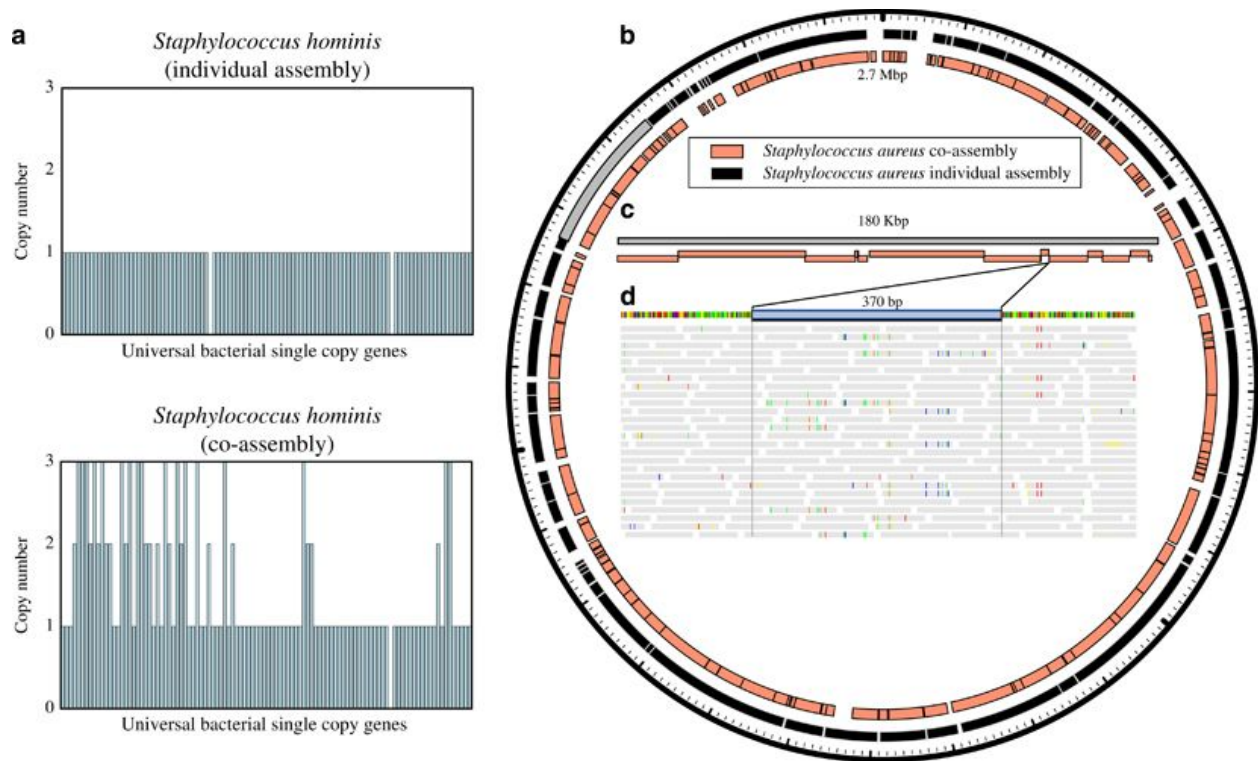


Figure 3.2 Strain heterogeneity reduces genome assembly quality and causes fragmentation in areas of extensive population-level variation. **(a)** Compared to individual assembly, co-assembly resulted in many duplicate and triplicate single copy genes. **(b–d)** The *Staphylococcus aureus* bin obtained from co-assembly is more fragmented than that from an individual assembly. **(b)** Scaffolds from both bins are aligned to a complete reference genome (2.7 Mbp). **(c)** Scaffolds from the co-assembly are aligned to a single scaffold (shown in gray in **b**) from the individual assembly. **(d)** Reads from all samples aligned to a gap in the alignment in **c**. Reads mapped to the area where co-assembly failed to recover a genome sequence (highlighted in blue) show signs of population-level strain variation. Gray boxes represent reads, and colored lines represent discrepancies between reads and reference sequence.

For supplemental figures, tables, and information for Chapter 3, see <https://doi.org/10.1038/ismej.2017.126>

4 Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms

Olm, Matthew R; West, Patrick T; Brooks, Brandon; Firek, Brian A; Baker, Robyn; Morowitz, Michael J; Banfield, Jillian F

Published in *Microbiome*, February 2019

Fungal infections are a significant cause of mortality and morbidity in hospitalized preterm infants, yet little is known about eukaryotic colonization of infants and of the neonatal intensive care unit as a possible source of colonizing strains. This is partly because microbiome studies often utilize bacterial 16S rRNA marker gene sequencing, a technique that is blind to eukaryotic organisms. Knowledge gaps exist regarding the phylogeny and microdiversity of eukaryotes that colonize hospitalized infants, as well potential reservoirs of eukaryotes in the hospital room built environment. Genome-resolved analysis of 1,174 time-series fecal metagenomes from 161 premature infants revealed fungal colonization of 10 infants. Relative abundance levels reached as high as 97%, and were significantly higher in the first weeks of life ($p = 0.004$). When fungal colonization occurred, multiple species were present more often than expected by random chance ($p = 0.008$). 24 metagenomic samples were analyzed from hospital rooms of six different infants. Compared to floor and surface samples, hospital sinks hosted diverse and highly variable communities containing novel species, including from Diptera (fly) and Rhabdida (worm) for which genomes were assembled. With the exception of Diptera and two other organisms, zygosity of the newly-assembled diploid eukaryote genomes was low. Interestingly, similar *Malassezia* and *Candida* species were present in both room and infant gut samples. Increased levels of fungal co-colonization may reflect synergistic interactions or differences in infant susceptibility to fungal colonization. Discovery of eukaryotic organisms that have not been sequenced previously highlights the benefit of genome-resolved analyses, and low zygosity of assembled genomes could reflect inbreeding or strong selection imposed by room conditions. Together with the detection of the same eukaryotic species in the infant gut and hospital room, these observations indicate the possible role of rooms as reservoirs of infant-colonizing fungal strains.

4.1 Introduction

Eukaryotes are common members of the human microbiome (Baley et al., 1986; Schulze and Sonnenborn, 2009; Tamburini et al., 2016). The colonization density and diversity of eukaryotes are lower than their bacterial counterparts (Ott et al., 2008; Parfrey et al., 2011; Schulze and Sonnenborn, 2009), but they can have substantial health consequences. The yeast *Saccharomyces boulardii* can significantly reduce rates of antibiotic-associated diarrhea (Surawicz et al., 1989), protozoa limit *Salmonella* populations through predation (Wildschutte et al., 2004), and high abundances of *Candida* and *Rhodotorula* are associated with asthma development in neonates (Fujimura et al., 2016). Fungal disease is most prevalent in immunocompromised patients,

including premature infants (Fridkin and Jarvis, 1996; Manzoni et al., 2015), although their incidence has declined in recent decades (Aliaga et al., 2014).

While infant fungal disease is an active area of study, little is known about routine infant colonization by fungi or other eukaryotes. Studies have reported 0%, 26%, 50%, and 63% of premature infants being colonized by fungi (Baley et al., 1986; LaTuga et al., 2011; Stewart et al., 2013, 2012), with variation in methodological sensitivity probably at the heart of these differences. Methods used to analyze the mycobiome, including culturing, DGGE, and ITS sequencing, also separate the fungal fraction from the community at large. This has left basic knowledge gaps about the relative abundance of fungi in early life, an important point as fungi-infant interactions in early life are known to affect allergy development (Bush and Portnoy, 2001; Fujimura et al., 2016, 2010). In fact, recent review articles have referred to eukaryotes as a “Missing Link in Gut Microbiome Studies” (Laforest-Lapointe and Arrieta, 2018), and stated that “Studies addressing how the infant mycobiome develops and shapes the host immune system will be required for a more comprehensive understanding of the early-life microbiome.” (Tamburini et al., 2016). Particular highlighted knowledge gaps relate to the ecological roles, growth dynamics, and source of eukaryotes in the human and hospital microbiomes (Huffnagle and Noverr, 2013; Laforest-Lapointe and Arrieta, 2018).

The hospital is a known source for bacterial infant colonists (Brooks et al., 2017). Fungal outbreak strains have been traced to the built environment (Mesquita-Rocha et al., 2013), yet the eukaryotic built environment microbiome remains understudied. This is because the vast majority of high-throughput studies of the hospital microbiome and the human gut microbiome use bacteria-specific 16S rRNA marker gene sequencing, and thus are blind to eukaryotes. Of five recent studies of the hospital microbiome, only one included primers to target the internal transcribed spacer (ITS) sequences to detect eukaryotes as well as bacteria (Bokulich et al., 2013; Hewitt et al., 2013; Lax et al., 2017; Oberauner et al., 2013; Shin et al., 2015). It remains to be seen if eukaryotes in the room have the genetic potential to colonize infants, and if so, where in the room these eukaryotes are located.

An alternative approach to microbiome characterization involves shotgun metagenomics. In this method, all DNA from a sample is sequenced regardless of its organismal source or genetic context. In some studies, mapping of the sequencing reads to reference genomes has enabled identification of pathogens (Wilson et al., 2018). However, the reads can be assembled, and new methods aid in reconstructing eukaryotic genomes from these datasets (West et al., 2018), enabling understanding of these organisms in the context of their entire communities, which also include bacteria, archaea, bacteriophage, viruses and plasmids. Relative to amplicon sequencing, genome assembly has several distinct advantages for understanding communities that contain eukaryotes. First, genomes provide information about *in situ* ploidy (number of distinct chromosomes sets per cell), heterozygosity (for diploid organisms, here used to refer to the fraction of alleles in a genome that have two versus one abundant sequence types), and extent of population microdiversity (here used to refer to additional sequence types that constitute low abundance alleles). Second, strain-tracking can be performed using high-resolution genomic comparisons. Last, newly assembled eukaryotic sequences expand the diversity of genomically defined eukaryotes in public databases, enabling comparative and evolutionary studies.

Here, we used genome-resolved metagenomics to study eukaryote-containing microbiomes of premature infants and their NICU environment. We evaluated the incidence of eukaryotes in room and infant samples and investigated the time period during which infant microbiomes contained eukaryotes. Genomes were assembled for fourteen eukaryotic populations and their ploidy, zygosity, and population microdiversity defined. The same species of eukaryotes were found in infant microbiome and the NICU environment, and a subset of other microbial eukaryotes in NICU rooms were classified as types that can cause nosocomial infections.

4.2 Materials and Methods

4.2.1 *Subject recruitment, sample collection, and metagenomic sequencing*

This study made use of many different previously analyzed infant datasets. These datasets have previously published descriptions of the study design, patient selection, and sample collection, and are referred to as NIH1 (Brown et al., 2018; Raveh-Sadka et al., 2016), NIH2 (Brooks et al., 2017), NIH3 (Raveh-Sadka et al., 2015), NIH4 (Rahman et al., 2018), Sloan2 (Brooks et al., 2017), and SP_CRL (Sharon et al., 2013). Infants were chosen for inclusion in this study irrespective of fungal disease state. Negative extraction controls were performed and sequenced during the sequencing of the Sloan2 cohort. The last well of the extraction block (H12) was left empty, and this well was treated the same as all other samples throughout the extraction protocol. It is therefore a control for the kit reagents, the sterility of the kit tubes/plates and the aseptic technique of the technician who performed the extraction. S2_CON_001E1, S2_CON_002E1, and S2_CON_003E1 were all on different extraction blocks, and S2_CON_002E2 was a second well on the same block as S2_CON_002E1.

This study also involved the collection and processing of an additional 269 samples from 53 infants. Newly collected infant fecal samples followed the same sample collection and DNA extraction protocol as described previously (Olm et al., 2017b; Raveh-Sadka et al., 2015). Metagenomic sequencing of newly collected infant fecal samples was performed in collaboration with the Functional Genomics and Vincent J. Coates Genomics Sequencing Laboratories at the University of California, Berkeley. Library preparation on all samples was performed using the following basic protocol: 1) gDNA shearing to target a 500bp average fragment size was performed with the Diagenode Bioruptor Pico, 2) end repair, A-tailing, and adapter ligation with an Illumina universal stub with Kapa Biosystems Hyper Plus Illumina library preparation reagents, 3) a double AMPure XP bead cleanup, followed by indexing PCR with dual-matched 8bp Illumina compatible primers. Final sequence ready libraries were visualized and quantified on the Advanced Analytical Fragment Analyzer, pooled into 11 subpools based on mass and checked for pooling accuracy by sequencing on Illumina MiSeq Nano sequencing runs. Libraries were then further purified using 1.5% Pippin Prep gel size selection assays collecting library pools from 500-700bp. Pippin pools were visualized on fragment analyzer and quantified with Kapa Illumina library quant qPCR reagents and loaded at 3nM. The eleven pools were then sequenced on individual Illumina HiSeq4000 150 paired-end sequencing lanes with 2% PhiX v3 spike-in controls. Post-sequencing bclfiles were converted to demultiplexed fastq files per the original sample count with Illumina's bcl2fastq v2.19 software. New metagenomic data was processed in the same manner as in the prior studies, and as described previously (Rahman et al., 2018).

Environmental metagenomes were described and published previously as part of the Sloan2 cohort study (Brooks et al., 2017). All samples were collected over a roughly one-year period from the same NICU at the University of Pittsburgh Magee Women's hospital. In order to generate enough DNA for metagenomic sequencing, DNA was collected from multiple sites in the NICU and combined into three separate pools for sequencing. Highly-touched surfaces included samples originating from the isolette handrail, isolette knobs, nurses hands, in-room phone, chair armrest, computer mouse, computer monitor, and computer keyboard. Sink samples included samples from the bottom of the sink basin and drain. Counters and floors consisted of the room floor and surface of the isolette. See previous publications for details (Brooks et al., 2017, 2014).

4.2.2 *Eukaryotic genome binning and gene prediction*

Reads from each sample were assembled independently using IDBA-UD (Peng et al., 2012) under default settings. A co-assembly was also performed for each infant, consisting of reads from all samples taken from that infant concatenated together. Binning assembled sequence scaffolds into eukaryotic genomes was performed using a EukRep based pipeline, described in detail in West et al. 2018. In cases where time-series data were available, samples were pre-binned using time-series information and eukaryotic bins were then subsequently identified with EukRep. In cases where multiple genomes of the same organism were recovered from multiple samples from the same infant, the most complete genome was selected for further analysis. In addition to the gene prediction methodology outlined previously (West et al., 2018), a second homology-based gene prediction step was performed. Ribosomal S3 (rpS3) proteins were identified in genomes using a custom ribosomal protein S3 (rpS3) profile HMM, and identified sequences were searched against the NCBI database (NCBI Resource Coordinators, 2017) and UniProt (UniProt Consortium, 2015) using blast (Altschul et al., 1990). For each de novo assembled genome, gene sets for the top 1-3 most similar organisms were used as homology evidence for a second pass gene prediction step with AUGUSTUS (Stanke et al., 2006), as implemented in MAKER (Cantarel et al., 2008). For *Rhabditida* S2_005_001R2 first pass gene predictions were used, as homology evidence decreased overall estimated genome completeness. Genome completeness was estimated using BUSCO (Simão et al., 2015) and is based on the number of detected single-copy orthologs. N50 was calculated using the program checkM (Parks et al., 2015).

To verify bins, the taxonomy of each scaffold was determined by searching gene sequences against the UniProt database (Raveh-Sadka et al., 2015). All bins were found to have a consistent phylogenetic signal, except the bin created from sample S2_009_000R2. Scaffolds had similar GC content and sequencing coverage, but were either dominated by genes with homology to the Class Sordariomycetes or Eurotiomycetes. Scaffolds from the original "megabin" were split into two separate bins based on this phylogenetic signal, resulting in the genomes *Nectria haematococca* S2_009_000R2 and *Exophiala* sp. S2_009_000R2. Gene prediction was run again for both of these genomes, as described above.

4.2.3 *Phylogenetic Analyses*

In order to construct a phylogenetic tree, rpS3 proteins from each de novo genome were detected as described above and searched against the NCBI database using blast. Protein sets of the 3-5 most similar organisms on NCBI were downloaded for inclusion. Other phylogenetically important genomes, such as *A. thaliana*, were included as well. For each protein set, 16 ribosomal proteins (bacterial ribosomal protein names L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, L24,

S3, S8, S10, S17, and S19) were identified using custom built hidden markov models (HMMs) with HMMER (Finn et al., 2011), using the noise cutoff (NC). The 16 ribosomal protein datasets were then aligned with MUSCLE (Edgar, 2004) and trimmed by removing columns containing 90% or greater gaps. The alignments were then concatenated. A maximum likelihood tree was constructed using RAxML v.8.2.10 (Stamatakis, 2006) on the CIPRES web server (Miller et al., 2010) with the LG plus gamma model of evolution (PROTGAMMALG) and with the number of bootstraps automatically determined with the MRE-based bootstrapping criterion. The constructed tree was visualized with Interactive Tree of Life (ITOL) (Letunic and Bork, 2007).

Average nucleotide identity (ANI) between binned genomes and reference genomes was determined with dRep (Olm et al., 2017a). Resulting whole genome ANI values were used in combination with a 16 ribosomal protein phylogenetic tree to determine the taxonomy of de novo genomes. For genomes without a species-level taxonomy, genomes were searched against the entire NCBI nucleotide database using blast. This resulted in a species-level call for *Malassezia restricta* S2_018_000R1. For genomes without a genus-level taxonomy (*Rhabditida* S2_005_001R2 and *Diptera* S2_005_002R2) an additional step was taken. Mitochondrial COI genes were identified by searching *D. melanogaster* and *C. elegans* COI genes against our PRODIGAL (Hyatt et al., 2010) predicted genes sets with UBLAST (Edgar, 2010). Significant hits from our protein sets were then searched against the Barcode Of Life Database (BOLD) (Ratnasingham and Hebert, 2007) and NCBI in order to identify sequences with high identity to our novel genomes. No significant hits were identified.

4.2.4 Mapping-based genome detection

To detect eukaryotes in an assembly-free manner, reads were mapped to a curated genome collection. This genome collection consists of all fungal genomes in RefSeq (accessed 9/14/17) (Pruitt and Maglott, 2001), as well as genomes assembled in this study with no close representatives in RefSeq (average nucleotide identity of 90% or higher according to Mash (Ondov et al., 2016)). The six genomes with no close representatives in RefSeq were *Malassezia restricta* S2_018_000R1, *Diptera* S2_005_002R2, *Exophiala* sp. S2_009_000R2, *Verruconis* sp. S2_005_001R2, and *Rhabditida* S2_005_001R2. *Candida parapsilosis* CDC317 was also included, as there were no genomes of *C. parapsilosis* in RefSeq.

Reads from all samples were mapped to this reference genome list using Bowtie 2 (Langmead and Salzberg, 2012). To determine which organisms were present in each sample, we primarily relied on breadth of coverage as reported by strainProfiler (<https://github.com/MrOlm/strainProfiler>). In NICU samples, all genomes with 50% breadth of coverage or above were considered present. For infant samples, reads resulting from concatenating all samples belonging to the same infant were first used to determine which fungi be reliably detected. Genomes with 50% breadth of coverage or above were considered present with two exceptions, *Malassezia pachydermatis* and *Malassezia sympodialis*, at ~0.2 and 0.4 breadth, respectively. Considering the extensive and distributed breadth of coverage for these genomes (Supplemental Figure S4.1C), they were considered present in the infant despite having low breadth of coverage overall. Reads from each individual sample from each infant were then mapped to all fungi considered to be present in that infant to determine changes over time. Relative abundance of genomes was determined using the formula: (number of reads mapping to genome / total number of reads in sample).

The lowest coverage genome with this breadth threshold was 1.1x coverage. To determine the limit of detection, we first determined the relative abundance needed to achieve 1.1x coverage using the median infant co-assembly depth (27.5 giga-base pairs) and the median eukaryotic genome length in our database of organisms that were detected at least once (13.7 mega-base pairs). We then calculated the limit of detection using the formula $((\text{min coverage} * \text{median length}) / \text{median co-assembly depth})$. This led to an estimated limit of detection of 0.05% relative abundance for infant fungi detection, through this number has significant variability depending on how deep each individual infant was sequenced.

4.2.5 *Negative extraction control analysis*

Sequences resulting from negative extraction controls were computationally processed in an identical manner to other samples. Reads from all control samples were mapped to the curated genome collection described above, and the relative abundance of all genomes with at least 10% breadth was plotted in Supplemental Figure S4.1. The program strainProfiler (<https://github.com/MrOlm/strainProfiler>) was used to compare reads in sample S2_CON_000E3 to *P. lilacinum* genomes assembled in this study and all publically available *P. lilacinum* genomes. Version 0.2 of the program was run with default settings, resulting in an average nucleotide identity measure between sample S2_CON_000E3 and all *P. lilacinum* genomes. Next, dRep v1.4.3 (Olm et al., 2017a) was used to compare the *P. lilacinum* genomes with each other using the command “dRep cluster --SkipMash”. The resulting distance matrix was merged with the values generated from strainProfiler to generate the dendrogram in Supplemental Figure S4.1B. Full code for implementation available at <https://github.com/MrOlm/InfantEukaryotes>.

All publically available *Malassezia* genomes were acquired by searching for the term “*Malassezia*” in the assembly section of NCBI and downloading them manually. Genomes were compared to each other and representative genomes were chosen using dRep v1.4.3 and the commands “dRep compare --SkipMash” and “dRep choose --noQualityFiltering -sizeW 0.5”. A concatenation of all negative extraction control sequences was then mapped to the resulting genomes using Bowtie2. Custom scripts were used to determine the breadth of coverage of each 10,000bp window of each fungal genome in each sample, and each window with at least 50% breadth was marked with a tick using Circos (Krzywinski et al., 2009) to visualize. Open source code detailing this analysis available at <https://github.com/MrOlm/InfantEukaryotes>.

4.2.6 *Statistical analyses and generation of MDS plot*

To compare the eukaryotic communities present in NICU room samples, multidimensional scaling (MDS) based on Bray-Curtis distance was performed. Bray-Curtis distance was calculated based on the relative abundance of each eukaryote present a sample using the python library SciPy (command `scipy.spatial.distance.braycurtis`) (Jones et al., 2001). Eukaryotes with at least 50% breadth of coverage were considered present in a sample. MDS was performed on the resulting all-vs-all distance matrix using the python library sklearn (command `sklearn.manifold.MDS`) (Pedregosa et al., 2011). MDS was plotted using a custom function in Matplotlib (Hunter, 2007). Stress was calculated using sklearn. Open source code detailing this analysis available at <https://github.com/MrOlm/InfantEukaryotes>.

We tested for significant associations between samples containing eukaryotes and various forms of metadata using the python SciPy package (Jones et al., 2001). Included were six pieces of

continuous metadata (DOL, infant birth weight, ect.), twenty-three pieces of categorical metadata (specific antibiotics given and specific NICU room locations), and the phyla-level abundance of all bacterial genomes (seven total phyla) (Supplemental Table S4.4). Bacterial phyla-level abundance was determined by summing the relative abundance of all bacterial genomes present in a sample. Bacterial genomes for previously sequenced samples are available in a previous publication (Rahman et al., 2018), and bacterial genomes for newly sequenced genomes were binned using the same methods. Metadata was filtered such that between 20-80% of values were non-zero in both samples containing eukaryotes and samples not containing eukaryotes. This resulted in a total of 13 pieces of meta-data for statistical testing (Supplemental Table S4.4).

In order to eliminate statistical bias introduced through sampling the same infant multiple times, one sample from each infant was chosen for statistical tests. If the infant was not colonized by a eukaryote, the sample was chosen at random. If the infant was colonized by a eukaryote, the sample with the highest eukaryotic abundance was chosen. Samples were considered to have a eukaryote present if the sum of the relative abundance of eukaryotes with at least 50% breadth was at least 0.1% relative abundance. Fisher's exact test was used for categorical metadata, and Wilcoxon rank-sum test was used for continuous data. Benjamini-Hochberg p-value correction (Benjamini and Yekutieli, 2001) was performed to account of multiple hypothesis testing. The results of all statistical tests are provided in Supplemental Table S4.5. Open source code detailing this statistical analysis available at <https://github.com/MrOlm/InfantEukaryotes>.

A permutation test was performed to determine the probability of observing 1.3 fungi per infant.. First, 100,000 trials were run where each trial consisted of randomly selecting 13 individuals with replacement from a total population of 161 individuals. The average number of times each infant was chosen was calculated for each trial, and an empirical p-value was determined based on how many trials had that number of average infants or higher.. Open source code detailing this statistical analysis available at <https://github.com/MrOlm/InfantEukaryotes>.

4.2.7 Ploidy, heterozygosity, and population microdiversity

In order to identify variants, reads from the sample a particular genome was binned from were mapped back to the de novo assembled genome using Bowtie2 (Langmead and Salzberg, 2012) with default parameters. The PicardTools (<http://broadinstitute.github.io/picard/>) functions “SortSam” and “MarkDuplicates” were used to sort the resulting sam file and remove duplicate reads. Freebayes (Garrison and Marth, 2012) was used to perform variant calling with the options ‘--pooled-continuous -F 0.01 -C 1’. Variants were filtered downstream to include only those with support of at least 10% of total mapped reads in order to avoid false positives. Furthermore, to avoid including variants as a result of mismapping reads, variants were filtered to include only those with coverage depth within a range of the average genome coverage plus or minus half of the genomes mean coverage. SNP read counts were calculated using the ‘AO’ and ‘RO’ fields in the freebayes vcf output file. Multiallelic sites were defined as sites with two or more non-reference alleles. Variants were called using the same methodology for both simulated read datasets and isolate genomes. Variants were used to determine ploidy, heterozygosity, and population microdiversity as described in the results section. Source code with full implementation details available at <https://github.com/MrOlm/InfantEukaryotes>.

To confirm that multiallelic sites are not the result of non-specifically mapped reads from the bacterial community, we fragmented with pIRS (<https://github.com/galaxy001/pirs>) a diploid C.

parapsilosis genome into simulated reads and added these reads to an infant gut metagenome sample without *C. parapsilosis*. The resulting read data set along with a separate data set comprised of only the simulated reads were then mapped to the original *C. parapsilosis* genome. No additional variants were detected between the sample with metagenomic reads and the sample without, indicating non-specifically mapped reads from bacterial community members have a minimal effect.

In order to determine the effect of stochastic read coverage on variant frequencies, simulated haploid, diploid, and triploid genomes were generated using the pIRS (<https://github.com/galaxy001/pirs>) diploid command with the *C. albicans* P57072 reference genome. The command was used once to generate a diploid genome, and twice to generate a triploid genome. Simulated reads were then generated for each genome using the pIRS simulate command at 10x, 50x, and 100x coverage. Assemblies and raw reads were downloaded for both *C. albicans* A48 and *C. parapsilosis* CDC317 from NCBI to be used as example isolate genomes for comparison. Based on this analysis, only the two genomes with at least 50x coverage were included in peak allele frequency analysis.

Genome aneuploidy was analyzed in two ways. First, reads from each sample were mapped back to genomes assembled from that sample. The coverage of each scaffold was determined in 10 kbp windows, and the coverage of all windows for each scaffold over 10 kbp was plotted. Plots were then analyzed for scaffolds with differing coverage, indicative of the presence of multiple copies of a subset of the chromosomes (Supplemental Figure S4.6). Second, reads from samples with genomes assembled from them were mapped to the closest available reference genome. The same procedure was then performed with these reference genomes in all cases where at least 80% of the genome was covered by reads. This allowed determination of aneuploidy on the whole-chromosome level (Supplemental Figure S4.7). Both methods agreed that in all cases, and no aneuploidy was detected.

4.3 Results

4.3.1 Recovery of novel eukaryotic genomes from metagenomes

In this study we analyzed 1,174 fecal metagenomes and 24 metagenomes from the NICU environment, totaling 5.31 terra-base pairs of DNA sequence (Supplemental Table S4.1). Fecal samples were collected from 161 premature infants primarily during the first 30 days of life (DOL) (full range of DOL 5 - 121; median 18), with an average of 7 samples per infant. NICU samples were taken from six patient rooms within the hospital housing the infants (Magee-Womens Hospital of UPMC, Pittsburgh, PA, USA). Three NICU locations were sampled in each room: swabs from frequently touched surfaces, wipes from other surfaces, and swabs from sinks (Brooks et al., 2017). Eukaryotic genomes were assembled from all samples using a EukRep based pipeline ((West et al., 2017); see methods section for details). The bacterial component of some of the datasets was analyzed previously (see methods).

Fourteen novel eukaryotic genomes were recovered in total, with a median estimated completeness of 91% (Table 4.1). Detailed genome assembly information is available in Supplemental Table S4.2. Genomes were assembled from organisms of a wide phylogenetic breadth, and four are the first genome sequences for their species (Figure 4.1). Twelve of the genomes are classified as

fungal, and are described in more detail below. The two other genomes (both recovered from hospital sink samples) represent the first genomes of their phylogenetic Families. *Diptera* S2_005_002R2 is within the phylogenetic clade of Diptera (true flies), and is equally related to *Drosophila melanogaster* (fruit fly) and *Lucila cuprina* (Australian sheep blowfly). *Rhabditida* S2_005_001R2 is within the Family Rhabditida (nematode), and is related to both pathogenic and non-pathogenic roundworms. In both cases BLAST searches of the rpS3 protein sequence against NCBI revealed no significant hits, and furthermore, comparing the mitochondrial cytochrome c oxidase subunit I gene and protein against the Barcode Of Life Database (BOLD) (Ratnasingham and Hebert, 2007) and NCBI revealed no hits with high identity. Thus, we are unable to tie our genomes to any morphologically described species.

4.3.2 Fungal contaminants in extraction controls

Four negative extraction controls were subjected to metagenomic sequencing to detect sequences resulting from reagent contamination. Reads from one of the four extraction controls mapped to *Purpureocillium lilacinum* (>50% of sample reads with a genome breadth of coverage of 87%) (Supplemental Figure S4.1A). The average nucleotide identity (ANI) was calculated between *P. lilacinum* reads in the extraction control, *P. lilacinum* genomes assembled in the study, and all previously sequenced *P. lilacinum* genomes in NCBI (Supplemental Figure S4.1B). *P. lilacinum* reads from the extraction control were extremely similar to genomes assembled from the NICU and infant gut, and divergent from previously sequenced genomes (Supplemental Figure S4.1B). Thus, *P. lilacinum* genomes assembled from room and gut samples are probably due to reagent contamination and not actually present in the environment.

Reads from three of the four extraction controls mapped to *Malassezia restricta* S2_018_000R1, all at low abundance (<3% of reads with a genome breadth of coverage of 1.3 - 14.2% using reads from the four samples) (Supplemental Figure S4.1C). It was not possible to calculate the ANI between the genomes in samples and controls due to the low sequencing coverage of *Malassezia restricta* S2_018_000R1 in the extraction controls. *Malassezia* is a near-ubiquitous skin-associated fungus (Gaitanis et al., 2012). The low breadth of coverage indicates that the genome sampled from the hospital surface is different to that of the *Malassezia* that contaminated the reagents. For this reason the *Malassezia* in infant and room samples were not excluded from further analysis.

4.3.3 Fungal microbiome of the premature infant gut

Excluding *P. lilacinum*, fungi were detected in 10 of the 161 premature infants profiled in this study (6%) (Figure 4.2A; Supplemental Table S4.3). The limit of detection for eukaryotic organisms was calculated as 0.05% of the total community (Supplemental Figure S4.2) (see methods for details). Eukaryotes were detected significantly more often early in life, and significantly more often when antibiotics were recently administered (Figure 4.2B). Antibiotics were given significantly more often early in life ($p = 5.3E-8$; Wilcoxon rank sum test), making it difficult to determine which of these two variables is driving the association.

Fungal colonization was not significantly associated with gestational age, twin status, birth weight, mode of delivery, or other clinical metadata. (Supplemental Tables S4.4, S4.5). Further, fungal colonization was not associated with bacterial community composition. *P. lilacinum*, presumed to be a metagenomic contaminant (Supplemental Figure S4.1), decreases in abundance as infants age (Figure 4.2), probably because increased bacterial biomass in later collected samples overwhelms

the contaminant DNA, as shown previously (Salter et al., 2014). Given this, we infer that the decrease in relative abundance of fungi present in the microbiomes of later-collected samples is due to bacterial growth.

All seven species detected colonizing the premature infants have been previously implicated as agents of nosocomial infection (Table 4.2), yet no infants colonized by eukaryotes in this study received antifungals or showed any symptoms consistent with acute fungal infection. However, asymptomatic colonization has been shown to be a risk factor for future fungemia (Huang et al., 1998). Eight different eukaryotic species were detected in at least one infant, with only *Candida albicans* and *Candida parapsilosis* colonizing more than one infant (Figure 4.2A). Infant N2_070 was colonized by two fungi, and infant N5_275 was colonized by three. A permutation test was performed to determine the probability of observing 13 fungi in ≤ 10 unique individuals (the probability of observing 1.3 fungi per infant) (Figure 4.2C), with a resulting p-value of 0.008. Thus, in this study multiple fungi colonized the same infant more often than expected random chance.

4.3.4 *Fungal microbiome of the neonatal intensive care unit*

Eukaryotic organisms were detected in 18 of the 24 metagenomes of the NICU room environment (Figure 4.3). Eukaryotic DNA made up an average of 1.23%, 1.22%, and 0.03% of the communities in highly-touched surfaces, sinks, and counters and floors, respectively. In order to compare the influence of room occupants and sampling location on the room mycobiome, we performed a multidimensional scaling (MDS) analysis (Figure 4.3A). Communities were differentiated based on sampling location rather than infant room.

The mycobiome of the NICU surfaces is dominated by species of *Malassezia* (Figure 4.3B). The eukaryotic organisms found in NICU sinks are distinct from, and more diverse than, those found on surfaces. Sink communities contained *Necteria haematococca*, *Candida parapsilosis*, *Exophiala*, and *Verruconis*, all of which were detected in multiple rooms and samples. Additionally, sinks in three separate NICU rooms contain DNA from *Rhabditidia* S2_005_000R1 (a novel nematode; see previous section for details). *Diptera* S2_005_002R2 (fly) also makes up about 2% of the entire community for single time-point in the sink in infant S2_005's room (Figure 4.3B). No macroscopic organisms were noted during the sample collection process. It remains to be seen whether these organisms contribute to dispersal of organisms throughout the NICU or affect the communities themselves.

4.3.5 *Similar fungi are found in the NICU and premature infant gut*

Candida parapsilosis was detected in both the NICU and in a premature infant, as were organisms of the genus *Malassezia*. To contextualize the similarity between *C. parapsilosis* strains in both communities, genomes assembled from both the infant and room environments were compared to all available reference genomes and each other using dRep (Olm et al., 2017a). *C. parapsilosis* genomes from the NICU sink of infant S2_005 and gut of infant N3_182 were more similar to reference genomes than each other (Supplemental Figure S4.3), and thus do not represent direct strain transfer events. However, the finding of similar fungi in the infant and built environment highlights the potential for colonization of premature infants by hospital-associated fungi, especially as many fungi detected in the hospital are known opportunistic pathogens (Table 4.2).

4.3.6 Sequence analysis of new genomes

De novo assembly of eukaryotic genomes from metagenomes not only allows for detailed genomic comparison and detection of novel organisms, but also the determination of ploidy, aneuploidy (abnormal number of chromosomes in a cell), heterozygosity, and population microdiversity of organisms in vivo. Changes in ploidy and aneuploidy have been observed in many eukaryotes, especially yeasts (Hirakawa et al., 2015; Peter et al., 2018), and are thought to be a strategy for relatively quick adaptation to shifts in environmental conditions. To determine the ploidy of genomes reconstructed in this study (Table 4.1), we examined the read count for each allele at a given variant site. For a diploid genome, alleles are expected to have a read count of 50%; for a triploid genome, alleles are expected to have a read count of either 33% or 67%. At low coverage, determining allele frequency with read mapping has more stochasticity relative to high coverage. Simulated reads for haploid, diploid, and triploid genomes at 10x and 100x coverage suggest it is possible to determine ploidy in even our low coverage genomes (Supplemental Figure S4.4). Based upon this analysis, all but one of our reconstructed genomes are diploid (Figure 4.4A, Supplemental Figure S4.5). *C. lusitaniae* is likely haploid. Similarly, aneuploidy can be detected by searching for regions where allele frequencies and/or read coverage differ from the rest of the genome. Given the possibility of a parasexual cycle in *C. albicans* (Bennett and Johnson, 2003), detecting aneuploidy was of particular interest. We searched for evidence of aneuploidy using both our reconstructed genomes and reference genomes, but did not see evidence for aneuploidy in any of our genomes using either method. (Supplemental Figures S4.6, S4.7).

For diploid genomes reconstructed from metagenomes, the sequences for each chromosome are a composite of sequences from the two alleles. Population microdiversity can be detected based on read counts that exceed the expected ratio of 50%. Measuring population microdiversity in this way can be confounded by sequencing error and stochastic read coverage variation (Supplemental Figure S4.4). Genomic datasets for isolates are not expected to have population microdiversity but will display sequencing error and stochastic read coverage variation. Consequently, we could separate sequencing noise from true population microdiversity by comparing the patterns we observed in our population genomic data to microdiversity found in isolate genomic datasets (Jones et al., 2004). For *C. parapsilosis* N3_182_000G1, the peak of allele frequencies is wider than that of the sequenced *Candida parapsilosis* isolate (Figure 4.4A), suggesting considerable population microdiversity. The *P. lilacinum* contaminant also displayed substantial microdiversity (Supplemental Figure S4.10). To avoid the stochasticity introduced by low sequencing coverage (Supplemental Figure S4.4), only genomes with over 50x sequencing coverage were analyzed for population microdiversity in this way.

Another method of measuring population microdiversity involves determining the number of multi-allelic sites (sites with more than 2 sequence variants). Tests with simulated reads were performed to confirm that non-specific mapping of reads from unrelated species do not bias results (see methods). All of our genomes have more multi-allelic sites than isolate sequenced genomes (Figure 4.4B), suggesting that all of our genomes have appreciable population microdiversity. Further, genomes from the room had higher microdiversity than those from the gut, although this comparison is not statistically significant ($p = 0.09$).

Finally, overall heterozygosity for each genome was measured by calculating the number of heterozygous SNPs per kilo-base pair (Figure 4.4C). A wide range of heterozygosity was observed

within genomes. For most organisms there was low heterozygosity, and for *C. albicans* and *C. parapsilosis*, comparable to that of reference isolates. *Malassezia restricta* S2_018_000R1 has both a particularly high rate of SNPs per kilo-base pair and high population microdiversity.

4.4 Discussion

4.4.1 Eukaryotic genome recovery from metagenomes augments information from isolate studies

In contrast with prior studies that have investigated microbial eukaryote genomes via sequencing of isolates, we employed a whole community sequencing approach and could detect population microdiversity in both NICU and infant-derived samples. *Malassezia* on NICU surfaces has particularly high population microdiversity. Given that *Malassezia* are skin associated fungi (Gaitanis et al., 2012), their high population microdiversity may be the consequence of the accumulation of numerous strains throughout the hospital via shedding of skin from different individuals. This could also reflect naturally large population variation present within the skin of a single individual, as has been reported for skin-associated bacteria (Oh et al., 2014; Tsai et al., 2016).

In the current analysis, most of the samples contained one dominant eukaryotic genotype, presumably one well adapted to the habitat, but other allele variants indicate the presence of lower abundance genotypes (Figure 4.4B). Given this dominance, it was possible to directly estimate genome heterozygosity. Prior studies have reported that *C. albicans* grows clonally in vivo [51], yet *Candida*, when expressing a certain phenotype, undergoes mating (Hull et al., 2000), most likely via a parasexual cycle (Bennett and Johnson, 2003). For *C. albicans*, the measured heterozygosity was comparable to that of previously sequenced isolate genomes [35,48]. Despite high heterozygosity of *C. albicans*, we see low strain heterogeneity. It has been hypothesized *C. albicans* mating may occur primarily on the skin (Lachke et al., 2003). We speculate there may be more strain heterogeneity on the skin or other areas of the human microbiome besides in the gut, as it is probable that heterozygosity in *Candida* populations in the human and room microbiomes arises due to mating with distinct coexisting strains.

The heterozygosity measurements of all other fungi except *Malassezia* were low, possibly indicating diversity reduction due to inbreeding and/or strong selection for specific alleles. We speculate that this reflects a long history of colonization of a habitat type that strongly selects for a specific genotype, so genome structure reflects the relatively low probability of recombination with strains with divergent alleles (in other words, the presence of gut-adapted and sink-adapted strains). However, without the availability of similar genomes to compare to from other habitats, we cannot rule out genetic bottlenecks that took place prior to introduction to the hospital.

An important aspect of the current study is the sequencing of reagent controls, which allowed us to identify *P. lilacinum* as a likely contaminant. It is interesting to note that peak allele frequency analysis indicated high population microdiversity for the contaminant. Genomic microdiversity of the reagent-associated population may indicate its long term persistence in the reagents, analogous to that shown for *Delftia* metagenome contamination that was present in Pippin size selection cassettes for many years (Olm et al., 2017c). Given the increasing use of metagenomic sequencing for pathogen detection and prior reports of *P. lilacinum* as both a contaminant and

disease agent (Luangsa-ard et al., 2011; Shivaprasad et al., 2013), it will be important to rule out a reagent source of *P. lilacinum* in future diagnostic studies.

4.4.2 *Premature infants are colonized by eukaryotes early in life*

Six percent of infants in this study were colonized by fungi, lower than most previous studies of infants (Baley et al., 1986; LaTuga et al., 2011; Stewart et al., 2013, 2012). Compared to shotgun sequencing, DGGE and ITS methods should be more sensitive due to the use of PCR, and thus may be more suitable for broad ecological surveys. However, the ability to amplify very rare sequences from organisms present at exceedingly low abundance levels complicates interpretation of the measured colonization frequencies. Our shotgun sequencing-based methods provide a more balanced view of community composition than methods that rely on PCR, and detection of populations that comprise more than ~0.05 % of the community DNA is possible with read-mapping (Supplemental Table S4.1; Supplemental Figure S4.2). Further, whole-community sequencing measures the relative abundance of eukaryotes in the context of the whole community, something that cannot be done using ITS, DGGE, or culturing based methods. Fungi are generally considered low abundance members of the gut microbiome (Schulze and Sonnenborn, 2009), yet in this study they reached levels as high as 55%, 78%, and 96% of the entire community (Figure 4.2). Differences in fungal communities during early life are known to have effects on infants' health later in life (Fujimura et al., 2016), and it remains to be seen if extreme abundance levels like this have long-lasting effects.

All infants profiled in this study received 2-7 days of prophylactic antibiotics upon birth, meaning antibiotic use is highly correlated with earlier days of life (Supplemental Table S4.4). While both antibiotic administration and DOL were significantly correlated with eukaryote abundance, consistent with previous studies of fungal colonization of low birth weight infants (Baley et al., 1986; Huang et al., 2000), infants who received antibiotics later in life were not colonized by eukaryotes. This suggests that day of life is the more important factor. However, eukaryotes may have not been detected in later collected microbiome samples from those infants due to increased relative abundance of bacteria. In other words, the sensitivity of the shotgun sequencing method may be insufficient to detect fungi that persist at low abundance.

Interestingly, permutation testing revealed that fungi colonized the same infants more often than expected by random chance. There may be several explanations for this phenomenon. For example, some infants may be more genetically predisposed to fungal colonization. Alternatively, fungi may interact synergistically, with the first colonizing species establishing a niche in the gut that makes it more suitable for other fungi. Should this effect prove to be important, it may help to explain how fungal colonization contributes to development of asthma or allergies (Fujimura et al., 2016).

4.4.3 *Differences in colonization patterns of NICU sinks and surfaces*

Because yeasts of the genus *Malassezia* are the most common eukaryotic member of the healthy skin microbiome (Gaitanis et al., 2012; Parfrey et al., 2011). This result is analogous to findings of previous studies, which showed that typically skin-associated bacteria dominate consortia associated with hospital surfaces and parts of other built environments (Brooks et al., 2018, 2017; Chase et al., 2016; Hewitt et al., 2013; Shin et al., 2015).

The same eukaryotes were never detected in sinks and surfaces, and the sinks hosted a comparatively diverse and variable eukaryotic community (Figure 4.3). Sinks are inherently heterogeneous environments with different moisture levels and chemical conditions. Punctuated cleaning events may also give rise to temporal variation. *Diptera* S2_005_002R2 (fly), which was present in present in only one sink sample, may be explained by sequencing of sink-associated eggs, as no macroscopic organisms were detected during the collection process. Recent studies have suggested that insects play significant roles in the dispersal of fungi, and this may occasionally occur in the NICU (Madden et al., 2018).

The other metazoan detected, the worm *Rhabditida* S2_005_001R2, was found in sinks from multiple rooms and samples collected months apart. These organisms may also be a source of fungi, and like the fly, could impact the overall NICU microbiome. Intriguingly, the partial genome appears to derive from an organism that is equally related to a bovine lungworm and *Caenorhabditis elegans*, and is potentially novel at the class level (Figure 4.1). Although we cannot evaluate its medical importance, the organism may have been macroscopically described but lack of a reference genome prevents identification.

4.4.4 *Eukaryotes in the NICU have the genomic potential to colonize hospitalized infants*

Almost all fungal species detected in the infant and NICU environments have been previously implicated as pathogens of immunocompromised individuals (Table 4.2). Some species, like *Candida albicans*, are almost exclusively found in warm-blooded animals and are thus adapted to growth in humans. Other species, like *Nectria haematococca*, are usually associated with soil and the rhizosphere. However, previous studies have shown that strains that colonize humans are similar to environmental strains (Luangsa-ard et al., 2011). Thus, while no direct cases of nosocomial infection were uncovered in this study, the presence of these strains in the NICU environment could pose risks to other immunocompromised premature infants.

C. parapsilosis was identified in two premature infants and four NICU samples, whereas *C. albicans* was identified in six premature infants but not identified in NICU samples. Interestingly, previous studies have linked neonatal *C. albicans* acquisition to vertical maternal transmission, but were unable to identify the source of *C. parapsilosis* (Waggoner-Fountain et al., 1996). In this study different strains of *C. parapsilosis* were found in the NICU and in infants, meaning that either the infant colonizing strain did not come from the NICU or was undetected by our sampling strategy. However the species-resolved detection of *C. parapsilosis* in four of eight sink samples (Figure 4.3), combined with previously reported evidence of non-vertical colonization [57], points to the NICU sink as a possible reservoir for *C. parapsilosis* strains that colonize infants.

We applied genome-resolved metagenomics to study eukaryotes in the gut microbiomes of infants and their NICU rooms and detected eukaryotes associated with pathogenesis of immunocompromised humans, commensals of human skin, and fungi typical of environments such as soil and drain pipes. Genomic analysis of diploid organisms found low rates of heterozygosity that may be explained by persistence of hospital-associated lineages in environments that impose strong selective pressure. The application of this approach in other contexts should greatly expand what is known about eukaryotic genomic diversity, population variation, and strain-level dissemination pathways.

4.5 Figures

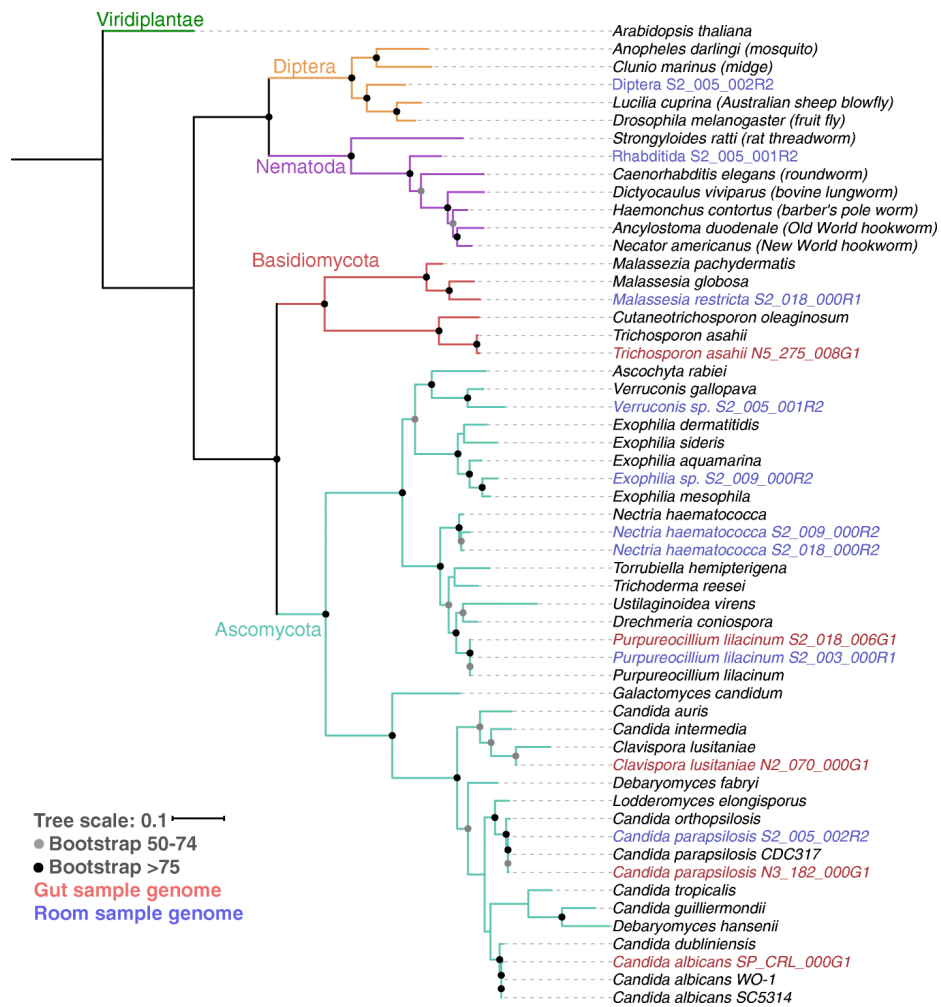


Figure 4.1 Phylogenetic tree of recovered eukaryote genomes. Genomes from infant-derived fecal samples (red) and NICU samples (blue) were classified using a phylogenetic tree based on the concatenation of the sequences of 16 ribosomal proteins (see the “Methods” section). Branches with greater than 50% bootstrap support are labeled with their bootstrap support range. Reference ribosomal protein sequences were obtained from NCBI and the Candida Genome Database

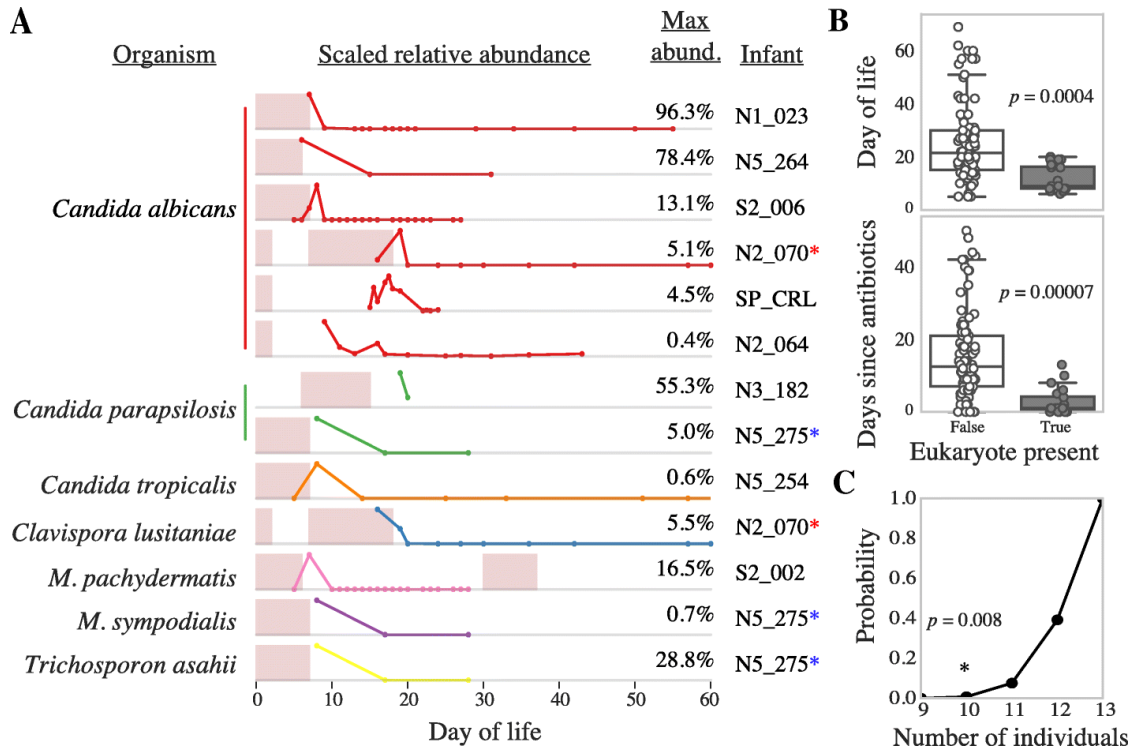


Figure 4.2 Abundance of eukaryotes colonizing infants. **a** The scaled relative abundance of each eukaryote colonizing an infant. Numbers on the right indicate the maximum relative abundance of the organism in that infant, and gray dividing lines indicate 0% relative abundance. Dots on the line-plots indicate days of life on which fecal samples were collected and sequenced. Infants colonized by multiple eukaryotes are marked with a colored asterisk. Pink bars indicate periods of antibiotic administration. **b** Metadata significantly associated with eukaryote abundance. The distribution of values for all samples in which eukaryotes are not present (left; white box plot) compared to values of samples in which eukaryotes are present (right; gray box plot). The p values were calculated using the Wilcoxon rank-sum test with Benjamini-Hochberg multiple testing p value correction. *P. lilacinum* was excluded from statistical tests due to its likely contaminant status. **c** Fungi are distributed among fewer individuals than expected by random chance. A permutation test was performed to determine the probability of observing 10 or less unique individuals colonized by 13 fungi from a population of 161 individuals. The number of unique individuals colonized is shown on the x -axis, and the empirical p value based on 100,000 trials is shown on the y -axis. An asterisk marks the true number of unique infants colonized in this study (10) and the associated p value

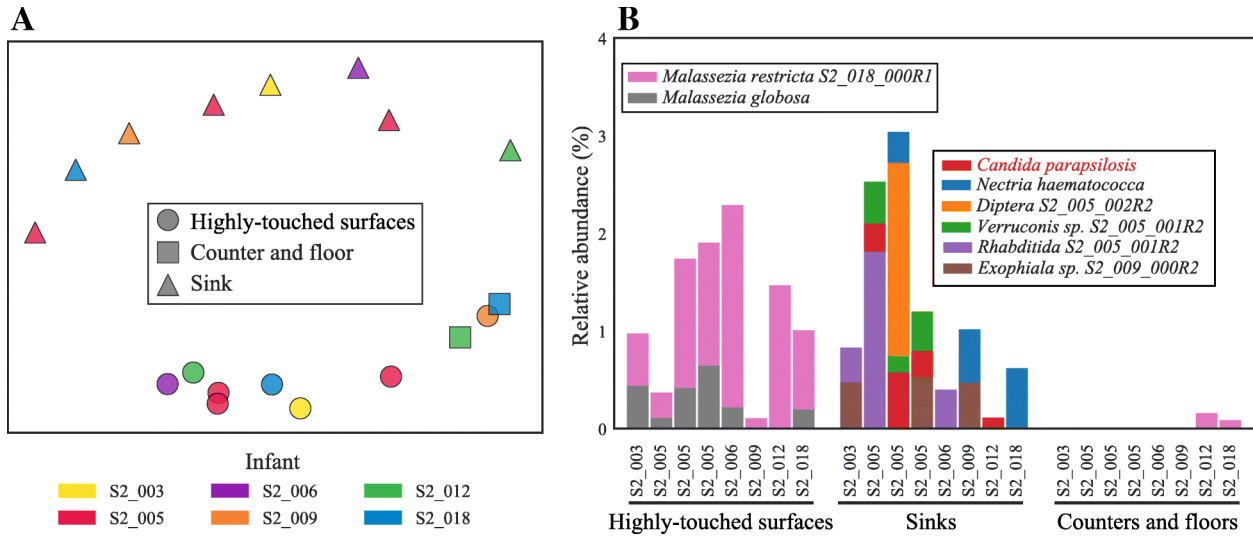


Figure 4.3 Eukaryotic microbiome of the neonatal intensive care unit (NICU). **a** Multidimensional scaling (MDS) of the Bray-Curtis dissimilarity between all NICU samples. Samples cluster by environment type rather than the room or occupant. The stress of the MDS was calculated to be 0.23. **b** Compositional profile of eukaryotic organisms detected in the NICU. Each colored box represents the percentage of reads mapping to an organism's genome, and the stacked boxes for each sample show the fraction of reads in that dataset accounted for by different eukaryotic genomes in each sample

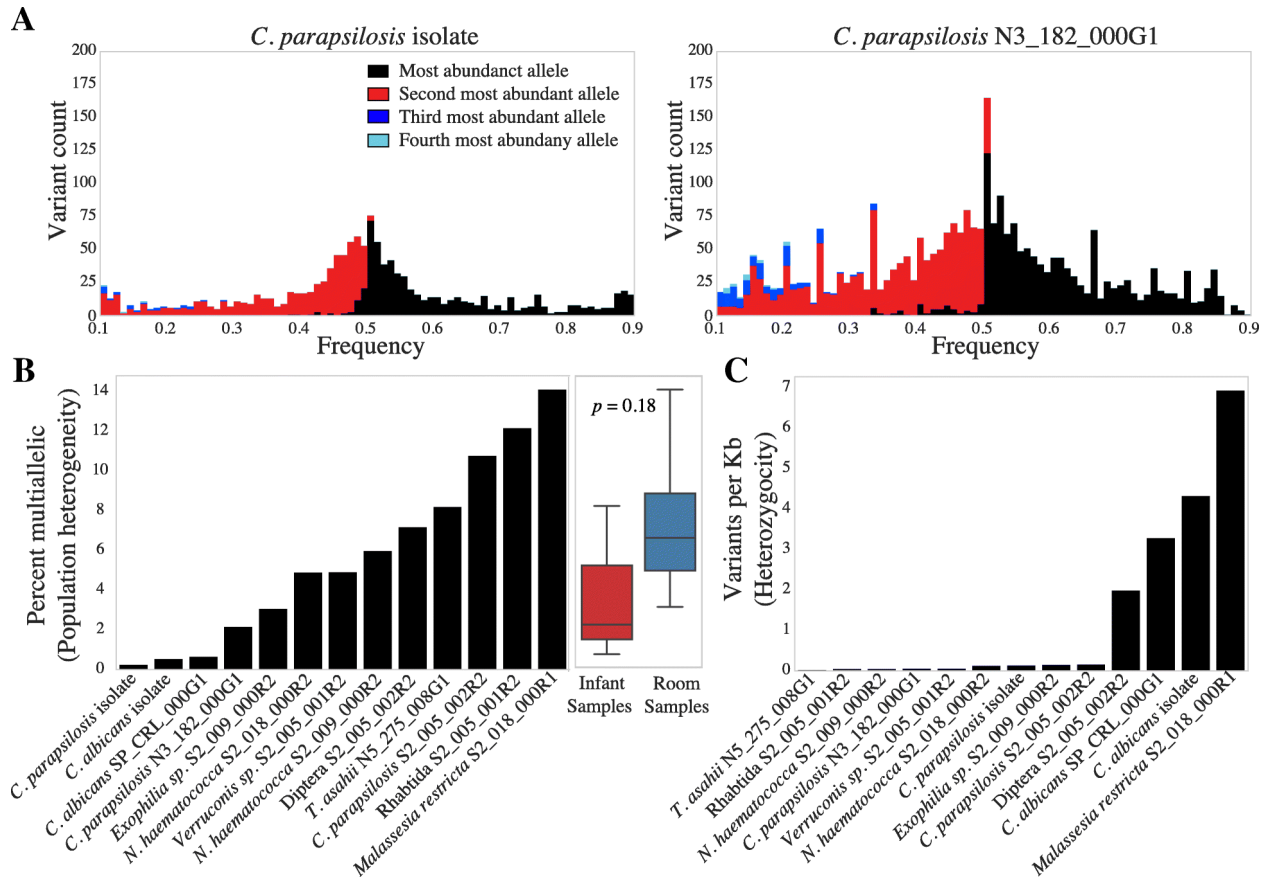


Figure 4.4 Ploidy, zygosity, and microdiversity of recovered eukaryotic genomes. **a** Histogram of the frequencies of the four most abundant variants at each variant site in an isolate genome of *C. parapsilosis* and in a genome of *C. parapsilosis* recovered in this study. Black, red, dark blue, and light blue bars indicate the abundances of the most abundant, second, third, and fourth most abundant variant, respectively. **b** For each genome, black bars indicate the percentage of variant sites that are multiallelic (contain more variants at a site than would be expected based upon ploidy alone). Haplotypes with more than two alleles are also considered to be multiallelic. A box plot compares the values from genomes originating from infant guts vs. the NICU room. **c** For each genome, black bars indicate the number of heterozygous variants per kb across the entire assembled genome

Table 4.1 Description of de novo assembled eukaryotic genomes

Source	Genome	Completeness (%)	Length (bp)	N50 (bp)	Coverage
Infant gut	<i>Purpureocillium lilacinum</i> S2_018_006G1	98.4	35,688,710	422,361	20×

Source	Genome	Completeness (%)	Length (bp)	N50 (bp)	Coverage
Infant gut	<i>Clavispora lusitaniae</i> N2_070_000G1	95.8	11,907,650	89,311	18×
Infant gut	<i>Candida parapsilosis</i> N3_182_000 G1	96.7	12,563,647	65,710	182×
Infant gut	<i>Trichosporon asahii</i> N5_275_008G1	90.1	23,419,590	32,912	13×
Infant gut	<i>Candida albicans</i> SP_CRL_000G1	91.1	12,561,678	22,840	30×
NICU room	<i>Purpureocillium lilacinum</i> S2_003_000R1	98.4	35,724,498	520,486	67×
NICU room	<i>Malassezia restricta</i> S2_018_000R1	72.6	6,457,898	4912	18×
NICU sink	<i>Nectria haematococca</i> S2_018_00 0R2	96.7	44,952,822	24,418	10×
NICU sink	<i>Candida parapsilosis</i> S2_005_002 R2	92.8	11,573,959	14,507	9×
NICU sink	<i>Rhabditida</i> S2_005_001R 2	74.9	50,505,025	8214	8×
NICU sink	<i>Nectria haematococca</i> S2_009_00 0R2	73.6	31,143,909	8000	7×
NICU sink	<i>Exophiala</i> sp. S2_009_000R2	75.9	24,670,482	7386	7×

Source	Genome	Completeness (%)	Length (bp)	N50 (bp)	Coverage
NICU sink	<i>Diptera</i> S2_005_002R2	52.5	43,769,201	6834	10×
NICU sink	<i>Verruconis</i> sp. S2_005_001R2	52.8	15,639,153	5112	6×

Table 4.2 Description of detected fungal taxa

Taxa	Common habitats	Pathogenicity	Number of infants	Locations In NICU	Refs
<i>Candida albicans</i>	Warm blooded animals	Common nosocomial pathogen	6	Undetected	[1]
<i>Candida parapsilosis</i>	Warm blooded animals	Common nosocomial pathogen (especially neonates)	2	Sink	[82]
<i>Candida tropicalis</i>	Warm blooded animals	Common nosocomial pathogen	1	Undetected	[83]
<i>Nectria haematococca</i>	Soil, rhizosphere	Pathogen of immunocompromised patients	0	Sink	[84]
<i>Malassezia sympodialis</i>	Human skin	Opportunistic pathogen	1	Undetected	[85]
<i>Malassezia globosa</i>	Human skin	Common commensal; implicated in dandruff	0	Surfaces	[86]
<i>Malassezia pachydermatis</i>	Skin of mammals	Opportunistic pathogen	1	Undetected	[87]

Taxa	Common habitats	Pathogenicity	Number of infants	Locations In NICU	Refs
<i>Trichosporon asahii</i>	Soil, human skin and GI tract	Rare opportunistic pathogen	1	Undetected	[88]
Verruconis	Soil, decaying vegetation	Verruconis includes black yeasts; human pathogens	0	Sink	[89]
Exophiala	Sinks, drain pipes, swimming pools	Exophiala contains pathogens of vertebrates	0	Sink	[90]

For supplemental figures, tables, and information for Chapter 4, see <https://doi.org/10.1186/s40168-019-0638-1>

5 Necrotizing enterocolitis is preceded by increased gut bacterial replication, *Klebsiella*, and fimbriae-encoding bacteria

Olm, Matthew R; Bhattacharya, Nicholas; Crits-Christoph, Alexander; Firek, Brian A; Baker, Robyn; Song, Yun S; Morowitz, Michael J; Banfield, Jillian F

Published in *bioRxiv*, February 2019

Necrotizing enterocolitis (NEC) is a devastating intestinal disease that occurs primarily in premature infants. We performed genome-resolved metagenomic analysis of 1,163 fecal samples from premature infants to identify microbial features predictive of NEC. Features considered include genes, bacterial strain types, eukaryotes, bacteriophages, plasmids and growth rates. A machine learning classifier found that samples collected prior to NEC diagnosis harbored significantly more *Klebsiella*, bacteria encoding fimbriae, and bacteria encoding secondary metabolite gene clusters related to quorum sensing and bacteriocin production. Notably, replication rates of all bacteria, especially Enterobacteriaceae, were significantly higher two days before NEC diagnosis. The findings uncover biomarkers that could lead to early detection of NEC and targets for microbiome-based therapeutics.

5.1 Introduction

Necrotizing enterocolitis (NEC) is widely studied yet poorly understood. First described in the early 1800s (Obladen, 2009), NEC is a disorder of intestinal inflammation that can progress to bowel necrosis, sepsis, and death (Neu and Walker, 2011). NEC affects 7% of very low birthweight infants born in the United States each year, and mortality rates have remained around 20 - 30% for several decades (Neu and Walker, 2011). The direct cause or causes of NEC remain unknown.

The primary risk factor for NEC is preterm birth (Neu and Walker, 2011). Immature enterocytes exhibit hyperactive immune responses through the TLR4 pathway in response to bacterial lipopolysaccharide (LPS) (Claud et al., 2004), which can lead to bowel damage (Denning and Prince, 2018). Experimental NEC occurs in conventionally raised animals but not those reared in a germ-free environment (Afrazi et al., 2011; Lawrence et al., 1982). These observations suggest that the intestinal microbiome plays a role in the disease and led to the prevailing hypothesis that an excessive immune response to abnormal gut microbes is the most likely basis for the pathogenesis of NEC. Although no single microbe has been consistently identified as a biomarker for NEC (Hosny et al., 2017), increased abundance of bacteria in the phylum Proteobacteria is a frequently reported microbial pattern in NEC infants (Pammi et al., 2017). Most fecal microbiome-based profiling studies of NEC utilize 16S rRNA amplicon sequencing, which provides a general overview of the bacteria present, but does not reveal metabolic features that could contribute to NEC pathogenesis.

Genome-resolved methods may provide new insights into NEC development. The approach has several advantages over 16S rRNA amplicon sequencing. All DNA in a sample is sequenced,

allowing detection of bacteriophages, plasmids, eukaryotes, and viruses. Bioinformatic techniques can also infer *in situ* bacterial replication rates directly from metagenomic data (Brown et al., 2016; Korem et al., 2015), an important metric, as some microbiome-related diseases have a signal related to bacterial replication but not relative abundance (Korem et al., 2015). Importantly, genome assembly and annotation can provide functional information about organisms present and possibly reveal genes associated with NEC. Further, whole-genome comparisons provide strain discrimination, and thus detailed testing of Koch's postulates. Finally, mapping to reference genomes is not required for genome detection, allowing for the discovery of novel bacterial clades (Brown et al., 2015). While identification of a single causative strain, virus, or toxin would be the most actionable result for clinicians, any associations could potentially be used as biomarkers to identify early warning signs of NEC, and microbial communities associated with NEC could be targeted with microbiome-altering techniques such as probiotics, prebiotics, or other approaches (Ronda et al., 2019).

5.2 Materials and Methods

5.2.1 Subject recruitment, sample collection, and metagenomic sequencing

This study was reviewed and approved by the University of Pittsburgh Institutional Review Board (IRB PRO12100487 and PRO10090089). This study made use of many different previously analyzed infant datasets. These datasets have previously published descriptions of the study design, patient selection, and sample collection, and are referred to as NIH1 (Brown et al., 2018; Raveh-Sadka et al., 2016), NIH2 (Brooks et al., 2017), NIH3 (Raveh-Sadka et al., 2015), NIH4 (Rahman et al., 2018), NIH5 (Olm et al., 2018), and Sloan2 (Brooks et al., 2017). Collated sequencing and health information for all infants and samples is provided in the supplemental materials of this manuscript (**Supplemental Tables S5.1, S5.2**).

5.2.2 Metagenomic profiling

Read processing and assembly

Reads from all samples were trimmed using Sickle (Joshi and Fass, 2011), and reads that mapped to the human genome with Bowtie 2 (Langmead and Salzberg, 2012) under default settings were discarded. Reads from all samples were assembled independently using IDBA-UD (Peng et al., 2012) under default settings. Co-assemblies were performed for each infant as well, where reads from all samples from that infant were combined and assembled together. Scaffolds <1 kb in length were discarded, and remaining scaffolds were annotated using Prodigal (Hyatt et al., 2010) to predict open reading frames using default metagenomic settings.

Recovery of de novo bacterial genomes

DasTool (Sieber et al., 2018) was used to select the best bacterial bins from the combination of 3 programs for automatic binning- abawaca (<https://github.com/CK7/abawaca>), concoct (Alneberg et al., 2014), and maxbin2 (Wu et al., 2016). Cross-mapping was performed between samples for each infant to generate differential abundance signals, and each sample was binned independently. For each infant, dRep v1.4.2 (Olm et al., 2017) was then used on all bins created from all samples from that infant to generate an infant-specific genome set, using the command "dRep -comp 50 -con 15 --S_algorithm ANImf -sa .99 -nc .25 --checkM_method taxonomy_wf".

To determine the taxonomy of bins, the amino-acid sequences of all predicted genes were searched against the uniprot database using the command "usearch64 -ublast \$aa_file -db uniprot_cp.fasta.udb -maxhits 1 -evaluate 0.0001 -threads 6 -blast6out \$b6", and tRep (<https://github.com/MrOlm/tRep/tree/master/bin>) was used in combination with ETE 3 (Huerta-Cepas et al., 2016) to convert the list of identified taxIDs into taxonomic levels. Briefly, this assigns a call to each taxonomy level when at least 50% of protein hits reach that taxonomic level.

Bacterial growth rates

iRep values (Brown et al., 2016) were calculated by first mapping reads from all samples in each infant to the de-replicated genome set from that infant using Bowtie 2. iRep was then run using the command "samtools view \$bam | iRep -s - ; iRep_filter.py --long", and values resulting from genomes with less than 0.9 breadth of coverage were discarded.

To visualize growth rates over time (**Figure 5.3a**), all iRep values from all bacteria were averaged together for each DOL relative to NEC and plotted using the seaborn command "sns.pointplot(showfliers=False, ci=68)" (<https://seaborn.pydata.org/>). Days of life in which less than 5 infants were profiled were manually removed. Boxplots in **Figure 5.3b** were also created using seaborn.

Bacteriophages, plasmids, and Eukaryotes

For all assemblies, circular contigs were identified using VICA (Crits-Christoph et al., 2016), and bacteriophages were identified using VirSorter (Roux et al., 2015) and VirFinder (Ren et al., 2017). Bacteriophages were defined as scaffolds that were considered "level 2" or "level 1" by VirSorter, or less than 0.01 p-value by VirFinder. Plasmids were defined as scaffolds which were circular, but not identified as bacteriophage according to the above definition. Bacteriophages and plasmids over 10kb in length were then each de-replicated separately on a per-infant basis using dRep version 2.0.5, with the command "dRep dereplicate -pa .9 --S_algorithm ANImf -nc .5 -l 3000 -N50W 0 -sizeW 1 --noQualityFiltering --clusterAlg singleOverlap". All plasmid and bacteriophage genomes were then compared to each other using the dRep command "dRep dereplicate -pa .9 --S_algorithm ANImf -nc .5 -l 10000 -N50W 0 -sizeW 1 --noQualityFiltering -clusterAlg single -d". Eukaryotes were assembled and binned from the gut samples of premature infants as previously reported (Olm et al., 2018).

Eukaryotic viruses

Eukaryotic viruses were analyzed using the vFam collection (Skewes-Cox et al., 2014), a set of HMMs designed for the identification of eukaryotic viruses within metagenomic sequence data. Hmmssearch (Finn et al., 2011) was used to search the HMM set "vFam-A_2014.hmm" against each assembly. All hits with e-values less than 1e-5 were considered significant and retained. Reads were also mapped to a previously curated list of human viruses (Rampelli et al., 2016). This led to the identification of no viruses when individual samples were used, and a very small number of viruses when combined sets of reads from each infant were used (Torque teno midi virus 2, Torque teno virus 14, and Macaca mulatta polyomavirus 1). This line of work was not followed up on due to lack of signal.

Diversity

Shannon diversity and overall bacteria richness were calculated for each sample. Shannon diversity was calculated using the command `skbio.diversity.alpha.shannon` (<http://scikit-bio.org/>). Richness was calculated as the number of bacteria with relative abundances over 0.1%.

KEGG Modules

KEGG modules were annotated by using HMMER against an in house HMM database built from the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology groups (KOs) (Kanehisa et al., 2014). Briefly, all KEGG database proteins with KOs were compared with all-v-all global similarity search using USEARCH (Edgar, 2010). MCL was then used to sub-cluster KOs (`inflation_value = 1.1`). Each sub-cluster was aligned using MAFFT (Kato and Standley, 2013), and HMMs were constructed from sub-cluster alignments. HMMs were then scored against all KEGG sequences with KOs and a score threshold was set for each HMM at the score of the highest scoring hit outside of that HMMs sub-cluster. KEGG modules were considered present in a genome if all necessary KOs were present in that genome.

Secondary metabolite gene clusters

In order to identify secondary metabolites, `antismash-4.0.2` was run on each infant co-assembly (Weber et al., 2015). The results were parsed using the custom script `parse_antismash.py` (<https://github.com/MrOlm/Public-Scripts>) and resulting key proteins were clustered using `diamond` (Buchfink et al., 2015) (commands: `diamond makedb --in $coassemblies.faa -p 6 -d $coassemblies.faa.db; diamond blastp -q $coassemblies.faa -d $coassemblies.faa.db.dmnd -o $coassemblies.faa.out --id 50`). Alignments were filtered to only retain those with >75% amino acid identity and 50% alignment. Hierarchical clustering was then performed using average amino acid identity (AAI) and resolved using a distance threshold of .5 to assign each secondary metabolite gene cluster to a gene cluster family. Next, for each infant the nucleotide sequences of all genes in a representative for each gene cluster family were concatenated together. The reads from each sample from that infant were mapped to this concatenation of genes in order to determine the dynamics of these genes in all samples from that infant. The breadth of each cluster was calculated as the weighted breadth (considering length) for all genes in that cluster.

Virulence Factors

Virulence Factors Database (VFDB) was used to search for virulence factors (Chen et al., 2005). The database used was from Mar 17, 2017, containing 2597 sequences. `Abriicate` was used to search all predicted protein sequences against the VFDB (<https://github.com/tseemann/abriicate>). A metadata file from VFDB website (<http://www.mgc.ac.cn/VFs/>) named “FVs.xls” was used to get additional information about the virulence factors. About 15% of virulence factors were not included in this metadata file and were excluded from additional analysis.

Botulinum toxin

A blast database of all subtypes of BoNTs toxin was downloaded from <https://bontbase.org/> (as accessed on February 15 2018). `Blastp` was used to search predicted amino acid sequences of all genes against the database. Hits with an e-value less than $1e-5$ were considered valid.

Pathogenic E. coli

It was previously reported that pathogenic *E. coli* may be associated with NEC development; specifically the clades 73, 95, 127, 131, 144, 998, abd 69 (Ward et al., 2016). To identify *E. coli* genomes of these sequencing types in our dataset, all genomes were MLST profiled using PubMLST (Jolley and Maiden, 2010) and the program “mlst” (<https://github.com/tseemann/mlst>). The MLST definition requires having 7 genes; in cases where only 6 genes could be identified, if only one sequence type (ST) existed with those 6 gene types, the ST was inferred. Each sample with an *E. coli* genome of the above STs at over 1% relative abundance were considered to have a “pathogenic” *E. coli*.

Proteins

Three protein clustering methods were evaluated for use in this study- MMseqs2 (Steinegger and Söding, 2017) (run using default settings), cd-hit (command: “cd-hit -c 0.7 -M 200000 -T 10”) (Huang et al., 2010), and a previously described hybrid Markov Cluster approach (Meheust et al., 2018). Algorithms were evaluated based on their ability to reconstruct known protein clusters (specifically a previously described set of 16 universal ribosomal proteins (Hug et al., 2016)), and the hybrid Markov Cluster approach performed best (**Supplemental Table S5.11**). This method was used to cluster the amino acid sequences of all predicted genes from all assembled scaffolds.

5.2.3 The average microbiome of NEC and control infants

To calculate the relative abundance of all microbes in each infant, a full “genome inventory” was generated for each infant by resolving the overlap between the recovered bacteria, eukaryote, bacteriophage, and plasmid genomes. Bacteriophage and plasmid genomes were first aligned using nucmer (Delcher et al., 2003), and in all cases where scaffolds aligned with over 95% ANI on over 50% of the scaffold, the scaffold was removed from the plasmid list. The resulting scaffolds were next aligned to bacterial genomes, and all phage/plasmid scaffolds that aligned to bacterial genomes with the same thresholds were removed. Finally eukaryotic genomes were aligned to the remaining scaffolds, and in cases where similar scaffolds were detected, the scaffold was removed from the eukaryotic genome. Reads from all samples were then mapped to that infant’s genome inventory using Bowtie 2, and the relative abundance of each organism was calculated as the percentage of total samples reads that map to that genome (**Supplemental Table S5.12**).

In order to compare the microbiome between NEC and control infants, the microbiome of each cohort was averaged across all infants in that cohort (**Figure 5.2a**) using the relative abundance values described in the previous paragraph. For each day of life, the average relative abundance of each taxa was first calculated. A 5-day sliding window was next applied, and values from samples in each window were averaged. For example, DOL 10 represents the average abundances from DOL 8 to 12.

5.2.4 Strain-level differences between NEC and control infants

In order to calculate the relative abundance of each bacterium in each sample, each sample was mapped to the infant-specific bacterial genome set for that infant using Bowtie 2. Relative abundances of all bacteria were calculated as the percentage of total sample reads mapping to each genome. Bacteria assembled from all infants were then compared to each other using dRep, and

bacterial genomes with at least 99% ANI were considered to be the same “strain”. A bacterium was considered present in a sample if it had over 0.1% relative abundance, and the fraction of pre-NEC and control samples in which each strain was present was calculated and plotted in **Figure 5.2c**.

A similar procedure was performed for the bacteriophage and plasmid genome sets of each infant. Mapping was done to each infant set separately, and genomes were considered to be the same “strain” if they had 99% ANI over at least 50% of their genomes. Organisms were considered present in a sample if they were present at over 50% genome breadth.

5.2.5 *Principal component analysis*

Principal component analysis (PCA) was performed based on the relative abundance of bacteria in each sample as assessed using weighted UniFrac distance (Lozupone and Knight, 2005). A phylogenetic tree was created by comparing all assembled bacterial genomes to each other using dRep (command: “dRep --SkipSecondary -ms 100000”), the weighted UniFrac distance between all samples was calculated using scikit-bio (<http://scikit-bio.org/>), and PCA was performed using scikit-learn (Pedregosa et al., 2011).

5.2.6 *Machine learning Algorithm development*

2,119 features were calculated for each sample and used as the input to a machine learning classifier to predict pre-NEC and control samples (**Supplemental Table S5.3**). See above methods for how individual features were calculated.

Three machine learning methods were evaluated for their ability to classify pre-NEC vs. control samples- a random forest classifier (`sklearn.ensemble.RandomForestClassifier(n_estimators=460, max_features=10)`) balanced using SMOTE (`imblearn.combine.SMOTEENN()`), a gradient boosting classifier (`sklearn.ensemble.GradientBoostingClassifier(learning_rate = 0.1, max_depth = 10, max_features = 'sqrt', min_samples_split = 0.7, n_estimators = 200)`) balanced using SMOTE, and the same gradient boosting classifier without balancing (Chawla et al., 2002; Pedregosa et al., 2011). Hyperparameters were empirically determined using `sklearn.model_selection.RandomizedSearchCV`, and in general many different combinations of hyperparameters gave similar results. Models were trained and evaluated using cross-validation for 5 iterations each (`sklearn.model_selection.StratifiedKFold(n_splits = 10)`; `sklearn.model_selection.cross_val_predict`; `sklearn.metrics.accuracy_score`), and all achieved similar prediction ability (**Supplemental Table S5.10**).

To determine the accuracy of the gradient boosting classifier, 100 iterations were performed where each iteration consisted of 1) randomly balancing the input to include 21 pre-NEC samples and 21 control samples, 2) classifying each sample in the input using 10 fold cross validation (same methods as above), and 3) calculating the percentage of samples that were correctly classified. The median accuracy value was reported.

Feature importance analysis

Feature importances were determined by 100 iterations of training the gradient boosted classifier on the full dataset of pre-NEC and control samples. Importance values were scaled for each

iteration such that the overall sum equals 1. The median importance value for each feature is reported (**Supplemental Table S5.4**).

KEGG and secondary metabolite enriched genomes

Each bacterial genome was assigned a metabolic importance value by summing the median feature importances of each KEGG pathway encoded by that genome (see above methods for how KEGG pathways were determined). A distribution of KEGG genomes importances was generated (**Supplemental Figure S5.4a**), and based on this distribution, genomes with importance values over 15 were considered “organisms of interest”. Each bacterial genome was also assigned an importance value equivalent to the highest importance value of all secondary metabolite clusters encoded by that genome. A distribution was generated (**Supplemental Figure S5.4b**), and genomes with importances over 0.5 were considered enriched in important secondary metabolite clusters.

5.2.7 Phylogenetic tree

A phylogenetic tree was made to visualize the distributions of organisms of interest and organisms enriched in important secondary metabolite clusters (**Figure 5.4f**). Ribosomal protein S3 was identified in bacterial genomes using pFam PF00189.19 and HMMER with a score cutoff of 50 (El-Gebali et al., 2018; Finn et al., 2011). An anchael outgroup was added, and all sequences were aligned using MAFFT (Kato and Standley, 2013) under default parameters. All positions with gaps in over 50% of sequences were trimmed from the alignment, and FastTree was used with default parameters to generate a phylogenetic tree (Price et al., 2009). The tree was visualized and annotated using iTol (Letunic and Bork, 2007).

5.2.8 Protein clustering

Protein association with NEC

Each protein cluster was considered present in a sample if a protein from that cluster had been assembled from the sample. Fisher's exact test was run on each protein cluster to determine if it was enriched in pre-NEC or control samples, and after Benjamini-Hochberg correction no *p*-values were statistically significant.

Protein association with organisms of interest

Each protein cluster was considered present in an organism of interest if a protein from that cluster was encoded in the organism's genome. The recall and precision of each cluster with organisms of interest was calculated as follows: recall = the number of organisms of interest the cluster is in / the total number of organisms of interest; precision = the number of organisms of interest the cluster is in / the total number of genomes the cluster is in. The recall and precision of each protein cluster was plotted (**Figure 5.5b**), and clusters with recall and precision over 0.7 were considered enriched in organisms of interest.

The 85 protein clusters enriched in organisms of interest were profiled using the pFam database (El-Gebali et al., 2018) with provided noise cutoffs, and the two most common pFams were PF00419.19 (Fimbrial) and PF00005.26 (ABC transporter) with four proteins each. We next determined if organisms encoding these proteins were enriched in pre-NEC samples. For each

pFam with at least three proteins enriched in organisms of interest, we compared the total relative abundance of all bacteria encoding that pFam in pre-NEC vs. control samples, as well as all iRep values of bacteria encoding that pFam in pre-NEC vs. control samples using the Wilcoxon rank-sum test with Benjamini-Hochberg *p*-value correction (**Supplemental Table S5.8**).

5.2.9 *Fimbriae*

Chaperone-usher fimbriae were identified in our dataset using pFam PF00577.19 (usher protein) and clustered using usearch (Edgar, 2010) with the command “usearch -cluster_fast -id 0.9”. The taxonomic profile of each fimbriae cluster was determined based on the taxonomy of organisms encoded by that cluster, and relative abundance and iRep associations with pre-NEC vs. control samples were calculated using the Wilcoxon rank-sum test applied to all bacterial genomes encoding each cluster. A similar procedure was performed using genomes which were not classified as organisms of interest but did encode fimbriae cluster 49, comparing between pre-NEC and control samples and between all samples from NEC infants and all samples from control infants (**Supplemental Figure S5.6**).

A phylogenetic tree was made in order to establish the type of usher proteins identified in our study. Three reference sequences from each previously established type (Nuccio and Bäumlér, 2007) were aligned with three representatives of each of our clusters using MAFFT. All columns with gaps in over 50% of sequences were trimmed from the alignment, IQtree was used with default parameters to generate a phylogenetic tree (Nguyen et al., 2015), and tree annotation was performed using iTol (Letunic and Bork, 2007).

5.2.10 *Effect size calculations*

To determine when signals first become apparent relative to NEC diagnosis, control samples were compared to samples collected over different sliding 3-day windows (**Figure 5.6**). To compare the signal at 5 days prior to NEC diagnosis, for example, a rarefied set of samples was chosen from 4-6 days prior to diagnosis where one sample from each infant that has a sample in that window was randomly chosen. This procedure was repeated 10 times, and the average effect size and 95% confidence intervals were plotted. The effect size was calculated based on the Wilcoxon rank-sum test statistic (as calculated by SciPy (scipy.stats.ranksums) (Jones et al., 2001)) using the formula: effect size = (test statistic / square root ((observations in population 1) + (observations in population 2))). For *iRep* all iRep values were compared between the two sets, for *secondary metabolite gene clusters* the total relative abundance of genomes encoding secondary metabolite gene clusters classified as producing sactipeptides, bacteriocins, or butyrolactones was compared, for *Klebsiella* the total relative abundances of all genomes classified as the genus *Klebsiella* were compared, and for *Fimbriae cluster 49* the total relative abundances of all genomes encoding fimbriae cluster 49 were compared.

5.3 Results

5.3.1 *Metagenomic characterization of premature infant fecal samples*

We analyzed 1,163 fecal metagenomes from 34 preterm infants that developed NEC and 126 preterm infants without NEC (**Figure 5.1**). Premature infant subjects were matched for gestational age and calendar date, and recruited from the UPMC Magee-Womens Hospital (Pittsburgh, PA)

over a 5 year period. Fecal samples were banked and specific samples were later chosen for DNA extraction and sequencing to preferentially study samples immediately prior to NEC. An average of 7.2 samples per infant, mostly from the first month of life, were sequenced and a total of 4.6 terabase pairs of shotgun metagenomic sequencing generated (**Supplemental Table S5.1**). Detailed sequencing information (**Supplemental Table S5.1**) and patient metadata (**Supplemental Table S5.2**) are provided.

Extensive computational analyses were performed on all samples to recover genomes *de novo*, and determine their phylogeny, metabolic potential, and replication rates (iRep (Brown et al., 2016)). We also searched samples for eukaryotic viruses, virulence factors, secondary metabolite gene clusters, and previously implicated pathogens (Ward et al., 2016; Zhang et al., 2018) (**Figure 5.1a**). This analysis resulted in 36 gigabase pairs of assembled sequence, 2,425 de-replicated bacterial genomes (average of 92% completeness and 1.1% contamination), 5,218 bacteriophage genomes, 1,183 plasmid genomes, 7 eukaryotic genomes, and 804,185 *de novo* protein clusters (**Figure 5.1b**; **Supplemental Table S5.6**). As NEC can be a rapidly progressive disorder, for most statistical tests we defined NEC samples as those taken within 2 days prior to NEC diagnosis (“pre-NEC” samples). For infants that did not develop NEC, only one sample from the period associated with NEC onset was used (“control” samples). Pre-NEC and control samples were matched for day of life (DOL), gestational age, and recent antibiotic administration (**Figure 5.1c**; **Supplemental Figure S5.1**). For other analyses, when explicitly stated, all samples were used.

5.3.2 *Klebsiella pneumoniae* is enriched in samples from infants with NEC

The gut microbiomes of all infants were dominated by Proteobacteria, regardless of NEC development (**Figure 5.2a,b**). As compared to previous studies of full-term infants (Bokulich et al., 2016; Penders et al., 2006), the premature infants in this study had increased Enterobacteriaceae (a family of Proteobacteria to which many nosocomial pathogens belong (Khan et al., 2017)) and notably low abundances of Actinobacteria and Bacteroidetes. Factors that could select for these organisms include prophylactic antibiotics given to all premature infants at birth, high rates of birth by cesarean-section, predominance of formula feeding and immaturity of the intestine and immune system. Only the NEC microbiomes contained Fusobacteria and Tenericutes (**Figure 5.2**). Compared to control infants, the NEC infant microbiomes exhibited less stability, lower levels of Firmicutes, and higher levels of Enterobacteriaceae than the microbiomes in control infants ($p = 8.9E-7$; Wilcoxon rank sums test; **Figure 5.2a**). The general association of Enterobacteriaceae and infants that go on to develop NEC has been described previously (Morrow et al., 2013), but this prior analysis was not restricted to the period immediately prior to NEC detection. In our study, the gut microbiomes of infants that developed NEC were not significantly enriched in Enterobacteriaceae in pre-NEC vs. control samples ($p = 0.15$; Wilcoxon rank sums test), so the association of Enterobacteriaceae and NEC infants overall may be due to the proliferation of these bacteria after administration of antibiotics to treat NEC (**Supplemental Figure S5.2**).

A principal component analysis based on weighted UniFrac distance was performed to compare the microbiomes of all samples from all time points (**Figure 5.2b**). The first two principal components explained 73% of the overall variance, but samples collected from NEC infants (red) did not cluster separately from control infants (black dots). Consideration of higher principal

components (up to the 5th principal component) did not separate pre-NEC and control samples, and samples coded by clinical metadata also did not cluster together (**Supplemental Figure S5.7**).

To identify strains enriched in pre-NEC samples, the percentage of pre-NEC vs. control samples carrying each assembled bacterial, bacteriophage, and plasmid genome was calculated (**Figure 5.2c,d**). *Klebsiella pneumoniae* strain 242_2 was the most associated with NEC, and was present above the threshold of detection in 52% of pre-NEC samples vs. 23% of control samples ($p = 0.008$; Fisher's exact test) (**Supplemental Table S5.12**). Interestingly, closely related bacteria (>99% average nucleotide identity (ANI)) colonized up to 35% of all infants (**Figure 5.2c**). This is likely the result of colonization of multiple infants by the same hospital-associated bacteria (Brooks et al., 2017). Importantly, no organisms in this study satisfied Koch's postulate that a disease causing organism should be found in all NEC infants and no healthy patients.

5.3.3 Bacterial replication rates are higher prior to NEC development

Bacterial replication rates are measured from metagenomic data by determining the difference in DNA sequencing coverage at the origin vs. terminus of replication, yielding an index of replication (iRep) that correlates with traditional doubling time measurements (Brown et al., 2016; Korem et al., 2015). Remarkably, iRep values of bacteria overall were significantly higher in pre-NEC vs. control samples ($p = 0.0003$; Wilcoxon rank sums test), in a cohort balanced for DOL, gestational age, and recent antibiotic administration (**Figure 5.3**). Further, iRep values followed a striking pattern in relation to NEC diagnosis: bacterial replication was stable four or more days prior to NEC diagnosis, increased daily in the three days prior to diagnosis, and crashed following diagnosis (probably due to resulting antibiotic administration) (**Figure 5.3a**). Individual species did not have enough data-points to be plotted confidently (minimum of 5 measurements per DOL), but genomes of the family Enterobacteriaceae displayed similar but more dramatic patterns than other bacteria overall (**Figure 5.3ab**). Increased bacterial replication prior to NEC could promote disease onset or merely be a reaction to changing conditions in the gut that led to NEC.

5.3.4 Machine learning identifies additional differences between NEC and control cases

2,119 features (e.g., the abundance of organisms encoding secondary metabolite gene clusters and each of 600 KEGG modules (specific metabolic pathways)) were measured for each of the 1,163 metagenomic samples (**Figure 5.1**; **Supplemental Table S5.3**). In order to evaluate which features are most different between pre-NEC and control samples, a machine learning (ML) classifier was developed. Multiple ML algorithms were evaluated, and although all performed with similar accuracy (**Supplemental Table S5.10**), the boosted gradient classifier was ultimately chosen due to its known ability to handle class imbalance. The classifier was trained on all 2,119 features to predict if samples were pre-NEC or control, and accuracy was measured through cross-validation over 100 iterations. The classifier achieved a median accuracy of 64% on balanced sets; 14% better than random chance. While a classifier with this accuracy may have limited utility in a clinical setting, it allowed us to interrogate which features were most informative for differentiating pre-NEC and control samples.

The most important individual features used by the ML classifier were replication rates (iRep values), KEGG modules, secondary metabolite gene clusters, and overall plasmid abundance (**Figure 5.4**). iRep values of both specific bacterial taxa and median iRep values overall were some of the most important features (**Figure 5.4b**), while KEGG modules accounted for over 50% of

the total feature importance (**Figure 5.4a**) (**Supplemental Table S5.4**). A similar number of KEGG modules were positively and negatively associated with NEC (**Figure 5.4c**), but descriptions of the pathways associated with NEC (e.g., erythritol and galactitol transport systems) and anti-associated with NEC (e.g., sodium and capsular polysaccharide transport systems) bear no obvious relationship to the disease (**Supplemental Table S5.4**). Secondary metabolite gene clusters were the second most important category overall (**Figure 5.4a**), but unlike KEGG modules, very few were negatively associated with NEC (**Figure 5.4c**). The most significant secondary metabolite gene cluster encodes an unusual operon of biosynthetic genes found in *Klebsiella* (cluster 416). In other species, similar operons are implicated in biosynthesis of quorum sensing butyrolactones (Du et al., 2011). The second most significant cluster of genes occurs in *Enterococcus* and is involved in biosynthesis of a sactipeptide resembling subtilisin A1, an antimicrobial agent with known hemolytic activity (Huang et al., 2009) (cluster 438) (**Supplemental Table S5.5**). Interestingly, another cryptic secondary metabolite gene cluster with a high feature importance (cluster 432) is closely related to a previously characterized cluster on a plasmid of Enterotoxin-producing *Clostridium perfringens* adjacent to the enterotoxin gene (*cpe*) and beta2 toxin gene (*cpb2*) (Miyamoto et al., 2006). Overall, high plasmid abundance was correlated with pre-NEC samples (**Figure 5.4b**) and *K. pneumoniae* plasmids in particular were significantly more abundant in pre-NEC samples ($p = 0.03$) (**Supplemental Figure S5.3**).

Feature importances were also analyzed in combination. Each bacterial strain was assigned an importance value based on the sum of the importance scores for the KEGG modules encoded by its genome. 150 genomes have high KEGG importance values (hereinafter referred to as “organisms of interest”) (**Supplemental Table S5.7; Supplemental Figure S5.4**). Interestingly, the organisms of interest were significantly more abundant in pre-NEC samples as compared to control samples ($p = 0.004$) (**Figure 5.4d**), and they cluster phylogenetically (**Figure 5.4f**). 97% were in the family Enterobacteriaceae, and of those, 90% were in the genus *Klebsiella*. The prevalence of *K. pneumoniae* in pre-NEC samples (**Figure 5.2c**) may explain the high abundance of *K. pneumoniae* plasmids in these samples.

Secondary metabolite biosynthetic gene clusters identified to be important by the ML classifier occur in 218 organisms that are significantly associated with pre-NEC samples (**Figure 5.4e**). Several types of secondary metabolite gene clusters were enriched in these genomes ($p < 0.01$; Fisher's exact test), including sactipeptides, bacteriocins, and butyrolactones (encoded by 382, 286, and 11 genomes, respectively) (**Supplemental Table S5.7**). As opposed to organisms of interest, these bacteria were dispersed around the phylogenetic tree (**Figure 5.4f**). This may indicate that the clusters themselves are associated with pre-NEC samples. Overall, the results point to quorum sensing and anti-microbial peptide production as being associated with NEC onset.

5.3.5 Bacteria associated with NEC encode specific types of fimbriae

We leveraged the gene content information provided by genome-resolved metagenomics to search for proteins associated with 1) pre-NEC samples and 2) organisms of interest. Three clustering algorithms were evaluated for their ability to reconstruct known clusters of ribosomal proteins (**Supplemental Table S5.11**), and a hybrid Markov Cluster algorithm approach (Meheust et al., 2018) performed best. Application of the algorithm to the 36,701,491 proteins reconstructed in this study yielded 804,277 protein clusters, none of which was statistically associated with NEC

(Fisher's exact test with false discovery rate correction) (**Figure 5.5a**). However, 85 protein clusters were associated with organisms of interest with high precision and recall (>0.7) (**Figure 5.5b**). The most common pFam annotations for these clusters were fimbriae and ABC transport proteins (**Supplemental Table S5.8**). However, only genomes encoding fimbrial proteins also had a significant association with NEC ($p = 0.02$; Wilcoxon rank-sum with Benjamini-Hochberg FDR correction; **Supplemental Table S5.8**).

Comparison of fimbrial operons against public databases revealed that the majority encode chaperone-usher (CU) type fimbriae. A classification scheme exists for CU fimbriae based on usher protein pFam PF00577.19 (Nuccio and Bäumlner, 2007) (El-Gebali et al., 2018). The 32,646 usher proteins identified in our sequencing data (**Supplemental Table S5.9**) were clustered into groups based on amino acid sequence identity, and the ten most prevalent groups were placed in a phylogenetic tree with reference sequences from each subtype of CU fimbriae (**Figure 5.5d**). All ten fimbriae clusters fit into the established CU fimbriae taxonomy, with 9/10 falling in the γ super-clade and one into the π clade (**Figure 5.5d**). Four fimbriae clusters identified in this study were significantly more abundant in pre-NEC samples, and genomes encoding cluster 49 (γ 4 clade) also had significantly higher iRep values in pre-NEC samples (**Figure 5.5c**). 27 genomes that encode fimbrial cluster 49 were not identified as genomes of interest, yet they were at significantly higher abundance, and have significantly higher iRep values, when considering all samples from NEC vs. control infants (**Supplemental Figure S5.6**) ($p < 0.01$; Wilcoxon rank-sums test). This suggests fimbrial cluster 49 itself may be associated with NEC and not incidentally associated with metabolically important genomes.

5.3.6 Biomarkers of NEC are most informative closer to NEC diagnosis

Statistical tests uncovered four factors significantly associated with pre-NEC samples (samples taken within day days prior to NEC diagnosis): iRep values overall (**Figure 5.3b**), genomes encoding specific types of secondary metabolite gene clusters (sactipeptides, bacteriocins, and butyrolactones) (**Supplemental Table S5.7**), *Klebsiella* (**Figure 5.2c**), and fimbriae cluster 49 (**Figure 5.5c**). We performed a similar analysis each day up to eight days prior to NEC diagnosis (**Figure 5.6**). Genomes encoding specific types of secondary metabolite gene clusters and *Klebsiella* genomes were always significantly more abundant in NEC samples, though the effect size of the difference became slightly higher closer to NEC diagnosis. iRep values and the abundance of genomes encoding fimbriae cluster 49, on the other hand, were only significantly higher 3 days and 1 day prior to diagnosis, respectively.

5.4 Discussion

Given that we found no single predictor of NEC and identified several factors as important by machine learning, our results support prior indications that NEC is a complex and likely multifactorial disease (Ballance et al., 1990; Neu and Walker, 2011). Of the four aspects of the gut microbiome that differ in pre-NEC compared to control samples (**Figure 5.6**), the iRep values of all organisms in each sample had the highest effect size. Given that iRep is a measure of bacterial replication rather than relative abundance, the result highlights that reliance on relative abundance alone could be misleading. This is largely due to the fact that relative abundance metrics are themselves misleading because an organism can increase in relative abundance simply due to the decline in relative abundances of other organisms. The higher bacterial replication rate prior to

NEC diagnosis could be sustained by nutrient release from the breakdown of gut tissue. Alternatively, increased bacterial replication may trigger onset of NEC, possibly because high activity of a specific organism leads to imbalance in concentrations of compounds in the gut environment.

Secondary metabolite gene clusters of specific types (bacteriocins, sactipeptides, and butyrolactones) were significantly enriched in pre-NEC compared to control samples (**Supplemental Table S5.5**). Bacteriocins are small peptides that kill closely related bacteria, and when produced, cell lysis could contribute to onset NEC via release of immunostimulatory compounds such as LPS. Sactipeptides are a class of posttranslationally modified peptides with diverse bioactivities (Arnison et al., 2013). The sactipeptide with the highest overall importance is related to a subtilisin (antimicrobial agent) with known hemolytic activity. Interestingly, all sactipeptides identified in this study were encoded by Firmicutes, including *Clostridium perfringens* and *Clostridium difficile* (**Supplemental Figure S5.5; Supplemental Table S5.5**). Production of sactipeptides by these species could trigger NEC through direct toxicity to human cells or via release of immunostimulatory bacterial compounds following bacterial cell lysis. This phenomenon could explain previous reports that implicate *Clostridium* in development of NEC (de la Cochetiere et al., 2004; Dittmar et al., 2008; Morowitz et al., 2010).

Butyrolactones are generally involved in quorum sensing in Actinobacteria (Du et al., 2011), but in this study were mostly found encoded in genomes of Proteobacteria, and over half were identified in *Klebsiella* genomes. Whereas known quorum sensing systems in Proteobacteria are responsible for the production of virulence factors, including fimbriae (Rutherford and Bassler, 2012; Sturbelle et al., 2015), the functions of butyrolactones in Proteobacteria remain unstudied. Higher proportions of *Klebsiella* were found in infants that went on to develop NEC, and their capacity to produce secondary metabolites and fimbriae could explain this association.

Organisms with genomes encoding fimbriae cluster 49 were at significantly higher abundances on both the day of and the day before NEC diagnosis. Fimbriae are known stimulants of TLR4 receptors (Fischer et al., 2006), immune receptors that are overexpressed in premature infants and previously linked to NEC in animal studies (Jilling et al., 2006; Leaphart et al., 2007). Fimbriae are the hallmark pathogenicity factors of uropathogenic *E. coli* (Wiles et al., 2008), a group of organisms that have been previously implicated as a causative agent of NEC (Ward et al., 2016). Uropathogenic *E. coli* were specifically evaluated in this study and not found to be significantly enriched in pre-NEC compared to control samples (**Supplemental Table S5.3; Figure 5.2c**). The associations in prior work and the current study may instead reflect a general link between fimbriae and TLR4 receptor stimulation.

An advantage of genome-resolved metagenomics is that it provides whole community information, going far beyond what can be deduced from 16S rRNA gene surveys that are the hallmark of most prior and much current human microbiome research. Here we applied this approach to a sufficiently large dataset to achieve statistical power unprecedented in a genome-resolved metagenomic study, and find that there is likely no single bacteriophage, plasmid, eukaryote, virus, or even protein that is responsible for NEC. However, we identify several promising associations through machine learning, many of which have previously been proposed to explain NEC onset but none of which alone can explain all cases. Bacteria of the genus *Klebsiella* emerged from our analyses as organisms of potential importance, with secondary metabolite, LPS and fimbriae

production all being possible contributors. The association of these bacteria, as well as bacteria of the *Clostridium* genera, with NEC and their presence in the NICU (Brooks et al., 2017) supports prior reports proposing that colonization of premature infants by nosocomial microbes may be clinically significant. Overall, we provide insight into how previously proposed but distinct explanations for development of NEC are interconnected, and identify bacterial growth rates as the strongest predictor of disease onset.

5.5 Figures

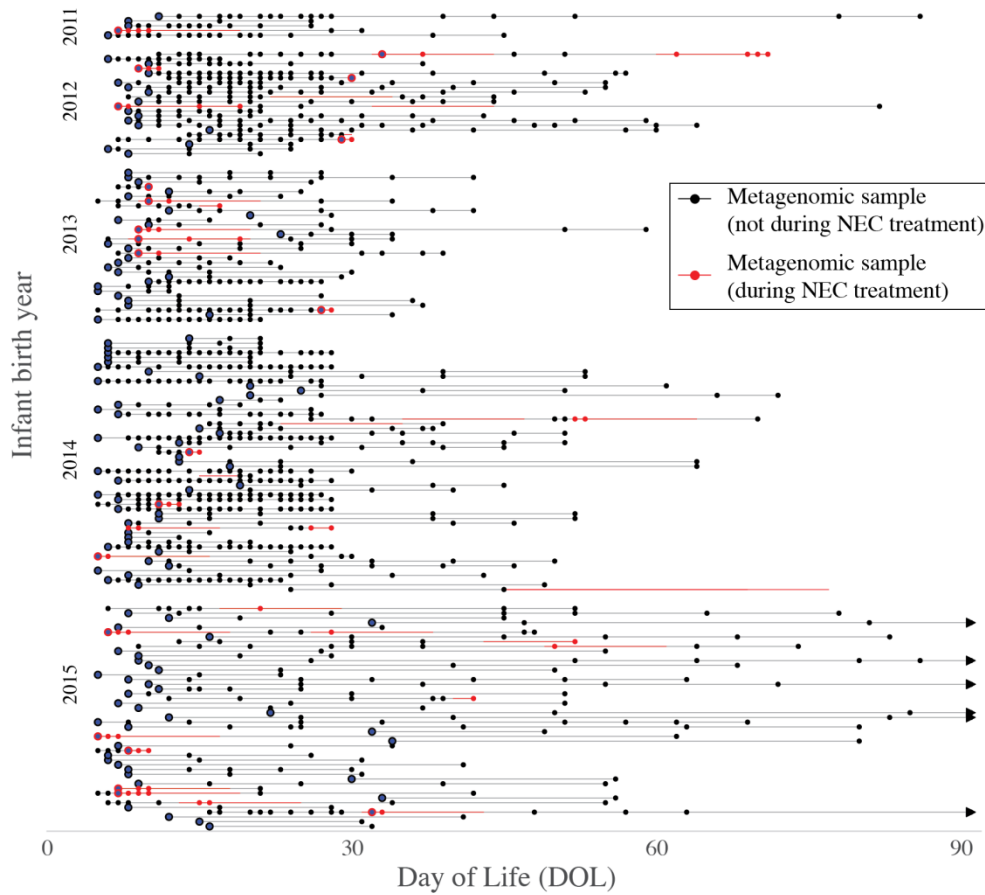


Figure 5.1 Metagenomic characterization of 1,163 samples from 160 premature infants. Each infant is represented by a horizontal line, and dots on the line represent sequenced metagenomic samples. Red sections indicate periods in which the infant was undergoing treatment for necrotizing enterocolitis. For some statistical tests one sample was chosen for each infant (pre-NEC and control samples); these samples are marked with larger circles.

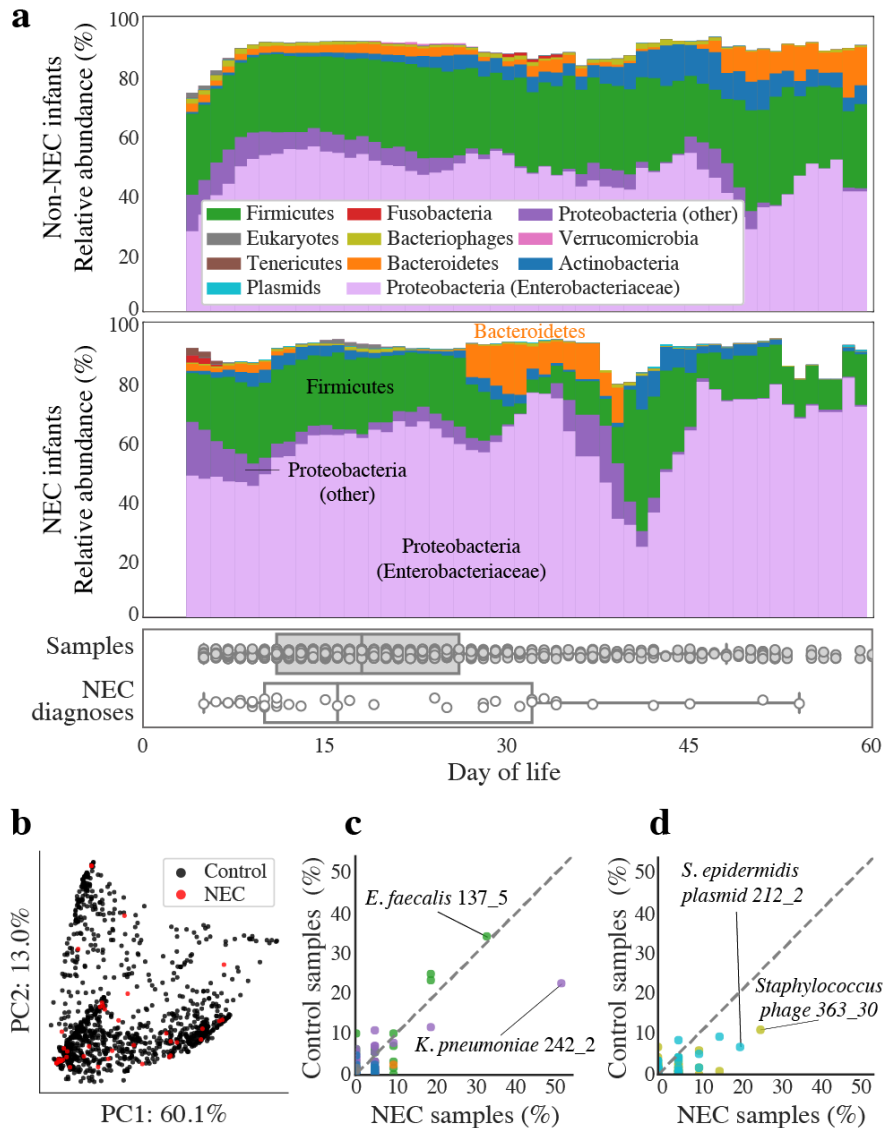


Figure 5.2 Comparison of microbes in premature infants that do and do not develop NEC. **(a)** The compositional profile of microbes colonizing infants that were and were not diagnosed with NEC. Bacteria were classified based on their phyla and other microbes were classified based on their domain. Each color represents the percentage of reads mapping to all organisms belonging to a taxon, and the stacked boxes for each sample show the fraction of reads in that dataset accounted for by the genomes assembled from the sample. Proteobacteria were subdivided into the family Enterobacteriaceae and other. All relative abundance values were averaged over a 5-day sliding window. Boxplots show the days of life in which samples were collected (top) and in which infants were diagnosed with NEC (bottom). **(b)** Principal component analysis (PCA) based on weighted UniFrac distance for all samples from NEC infants (red) and control infants (black). **(c, d)** The percentage of NEC infants vs. the percentage of non-NEC infants colonized by strains of **(c)** bacteria or **(d)** bacteriophage (gold) and plasmids (blue). Colonization by bacteria is defined as the presence of a strain at $\geq 0.1\%$ relative abundance. Plasmid and bacteriophage detection required read-based genome breadth of coverage of $\geq 50\%$. Each dot represents a strain, and dashed lines show a 1:1 colonization rate.

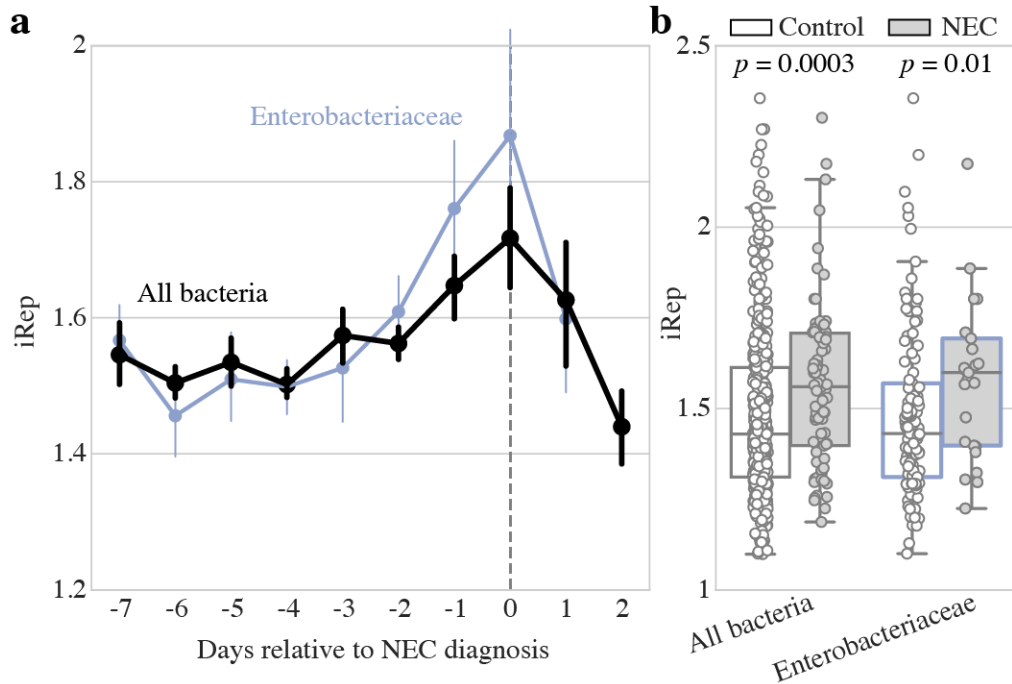


Figure 5.3 Bacterial replication rates are significantly higher prior to NEC development. **(a)** Replication rates for bacterial groups relative to day of NEC diagnosis. Dots represent the mean value for each group on each day, and error bars represent standard error of the mean. Days of life in which growth rates were calculated from at least 5 infants are shown. **(b)** Growth rates in control (white) vs pre-NEC (grey) samples. p -values shown from Mann–Whitney U test.

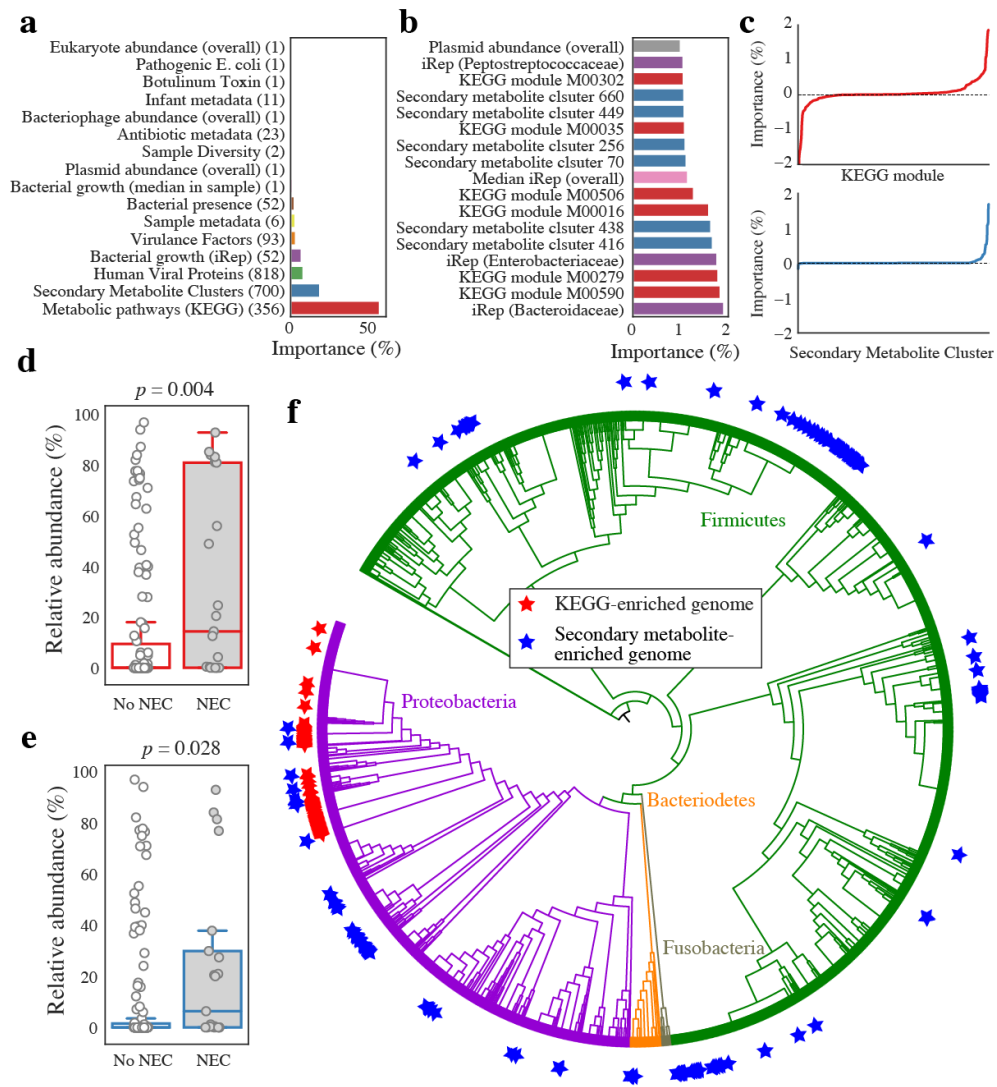


Figure 5.4 Machine learning identifies differences between pre-NEC and control samples. **(a)** The sum of all individual importances for each feature category. The number of features in each category is listed in parentheses. **(b)** Importance of all individual features associated with NEC with classifier importances over 1%. **(c)** Signed importances of all individual KEGG modules (top, red) and secondary metabolite clusters (bottom, blue). Negative values are negatively associated with pre-NEC samples, positive values are positively associated with pre-NEC samples. **(d, e)** The relative abundance of genomes enriched in important KEGG modules **(d)** and important secondary metabolite enriched genomes **(e)** in pre-NEC vs. control samples. p -values shown from Mann–Whitney U test. **(f)** The distribution of genomes enriched in important KEGG modules (red star) and important secondary metabolite clusters (blue star) around a phylogenetic tree of all recovered bacterial genomes. Genomes enriched in important KEGG modules are more clustered on the tree than those enriched in important secondary metabolite clusters.

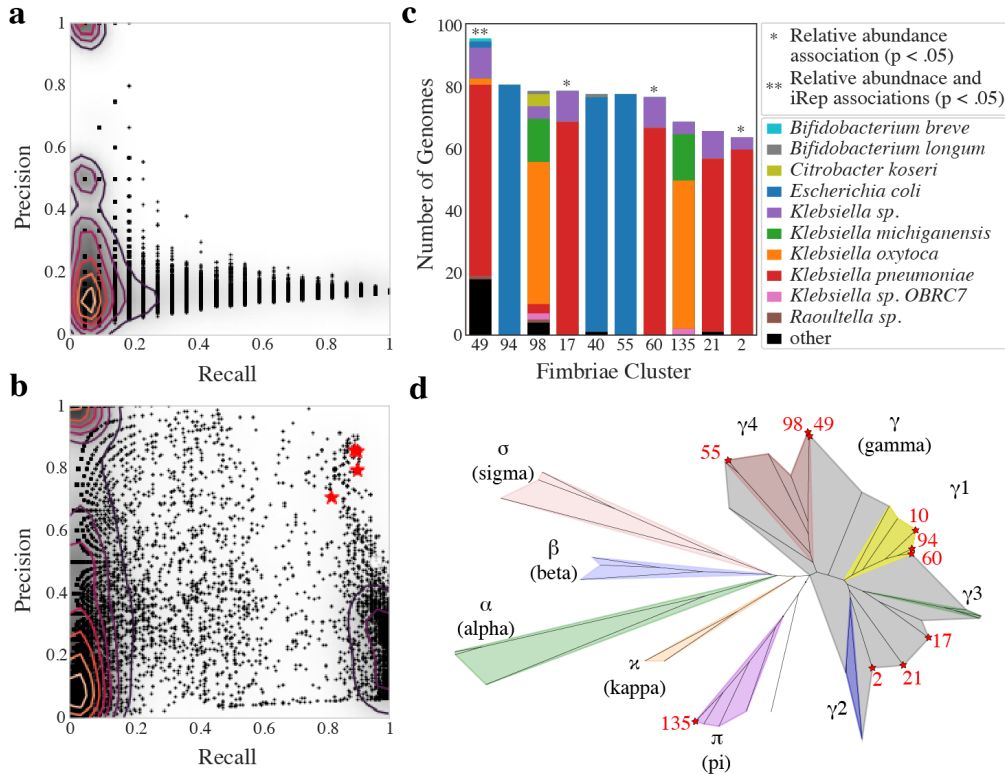


Figure 5.5 Genomes encoding fimbriae are associated with necrotizing enterocolitis development. **(a, b)** Association of protein clusters with pre-NEC2 samples **(a)** and organisms of interest **(b)**. Each dot represents a protein cluster. Recall is **(a)** the number of pre-NEC samples the cluster is in / the total number of pre-NEC samples and **(b)** the number of organisms of interest the cluster is in / the total number of organisms of interest. Precision is **(a)** the number of pre-NEC samples the cluster is in / the number of total pre-NEC samples, and **(b)** the number of organisms of interest the cluster is in / the total number of genomes the cluster is in. Clusters annotated as fimbriae are marked with a red star. Contour lines are drawn to indicate density. **(c)** The number of bacterial genomes encoding each fimbriae cluster, the special-level phylogenetic profile of genomes encoded by each fimbriae cluster, and each cluster's association with NEC. **(d)** Phylogenetic tree of CU usher proteins built using IQtree. Three amino acid sequences from each *de novo* CU cluster and three reference amino acid sequences from each defined CU clade was included in the tree. Colors mark the phylogenetic breadth spanned by reference sequences, and stars represent *de novo* CU clades. For all *de novo* clusters the three randomly chosen sequences fell extremely close to each other on the tree.

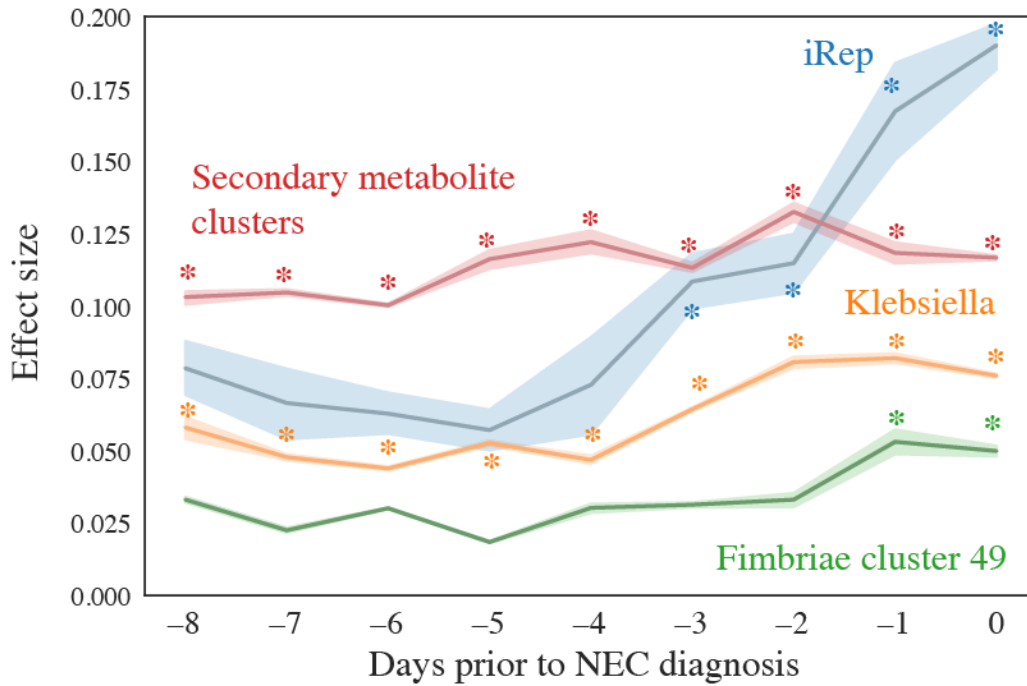


Figure 5.6 Biomarkers of NEC are most informative closer to NEC diagnosis. The effect size for difference of each feature in pre-NEC vs. control samples is shown based on a Mann Whitney rank sums test over a 2 day sliding window (e.g., -5 compares samples collected from -6 to -4 days relative to NEC diagnosis to control samples). Comparisons with $p < 0.05$ are marked with an asterisk.

For supplemental figures, tables, and information for Chapter 5, see <https://doi.org/10.1101/558676>

6 Consistent metagenome-derived metrics verify and define bacterial species boundaries

Longstanding questions in microbiology relate to the existence of naturally distinct bacterial species and genetic approaches to distinguish them. Bacterial genomes in public databases form distinct groups, but these databases have isolation and deposition biases. Here we compared 5,203 bacterial genomes obtained from 1,457 environmental metagenomic samples to test the existence of naturally distinct clouds of diversity and evaluated metrics that could be used to define the species boundary. Bacterial genomes from the human gut microbiome, soil, and the ocean all exhibited gaps in average nucleotide identities (ANI) near the previously suggested species threshold of 95% ANI. While pairwise genome-wide dN/dS ratio levels off at around 98% pairwise ANI, estimates for homologous recombination approached zero at ~95% ANI, supporting the idea that sequence divergence-based breakdown in homologous recombination is a species-forming force. We next evaluated 45 genome-based metrics for their ability to distinguish species in cases where full genomes are not recovered. Full length 16S rRNA genes were least able to distinguish species and were under-recovered from metagenomes, but many ribosomal proteins displayed both high recoverability and species-discrimination power. Taken together our results verify the existence microbial species in metagenome-derived genomes and highlight the usefulness of ribosomal genes for gene-level species discrimination.

6.1 Introduction

A fundamental question of microbiology is whether bacterial diversity is a genetic continuum or divided into distinct clusters (Cohan, 2019, 2002; Shapiro and Polz, 2015). The existence of sequence discrete populations have been identified in public databases (Goris et al., 2007; Konstantinidis and Tiedje, 2005), most recently in a study using the ~90,000 bacterial genomes available in the public NCBI Genome database as of 15 March 2017 (Jain et al., 2018). These studies all describe a gap in average nucleotide identity (ANI) values at 95%, and based on this have proposed a 95% ANI species threshold. However it is still unclear whether the observed population structures are confounded by biases of microbial cultivation or whether they reflect true relationships between microbial cells across different natural environments, as comparison of biased genome sets could form spurious patterns of sequence clusters.

Over 75% of the genomes with taxonomy in the NCBI Genome database are from the Proteobacteria and Firmicutes phyla, and over 10% are from the genus *Streptococcus* alone (Jain et al., 2018). Attempts have been made to un-bias the genome set when looking for sequence identity gaps: for example by sampling five genomes from each named species with at least 5 genomes present in the database (Jain et al., 2018), but all biases of selective cultivation and sequencing are difficult to account for. These include a historical bias to sequence and deposit isolates that meet the expected phenotypic criteria of target species, or cultivation biases against alternative genotypes using defined media. Sets of genomes without selection and cultivation biases can be acquired through the direct sequencing of environmental DNA (genome-resolved metagenomics). While metagenomic methods do suffer from their own set of biases like better DNA extraction from gram positive than gram negative bacteria (Albertsen et al., 2015; Guo and

Zhang, 2013), it is unlikely that this kind of broad bias would contribute to patterns of species-level sequence groups.

If distinct microbial species do exist, a relatively comprehensive analysis of public data may uncover the roles of recombination and selection in their origin. Several hypotheses have been proposed to explain genetic discontinuities, including a drop in homologous recombination at the species threshold (Majewski and Cohan, 1999; Vulić et al., 1997), periodic selection events that purge genetic diversity (Gevers et al., 2005), and neutral processes (Wilmes et al., 2009). Computer simulations suggest that both homologous recombination and selection are needed to form genotypic clusters (Fraser et al., 2007), and leading theory points to the declining rates of homologous recombination concurrent with sequence divergence as the force behind the clustering. Quantitative population genomic analyses of metagenomics data support this concept (Eppley et al., 2007). While compelling descriptions of speciation have been shown for a limited number of organisms (Cadillo-Quiroz et al., 2012; Shapiro et al., 2012), these evolutionary forces have not been measured and analyzed at scale across thousands of genomes nor in direct relation to the recently proposed 95% ANI species threshold.

Here, we analyzed thousands of bacterial genomes recovered directly from the sequencing of environmental DNA to test for the existence of a sequence identity gap, and developed software to estimate the strength of recombination and selection forces between these genomes. Discrete sequence gaps were identified in all environments tested, and both estimated recombination rates and genome-wide dN/dS ratios showed clear patterns in relation to the 95% ANI species threshold. Whole genome ANI methods were compared to various marker gene alignments (including 16S rRNA) for the ability to create species-level groups, and recommended species thresholds are provided for each method. Overall our results support the existence of discrete species-level groups for bacteria in the three divergent environments tested, provide sequence-based evidence for the likely evolutionary forces at play, and provide metrics for species delineation in metagenomics studies.

6.2 Materials and Methods

6.2.1 Preparation of genome sets

All publicly available genomes available in RefSeq as of February 21, 2018 were downloaded using `ncbi-genome-download` (<https://github.com/kblin/ncbi-genome-download>) with the command “`ncbi-genome-download --format genbank -p 4 bacteria`”. Taxonomy of all genomes was determined using ETE3 (Huerta-Cepas et al., 2016). A genome set consisting of a subset of the entire RefSeq set was generated to balance taxonomic representation- ten genomes were randomly chosen from of the 480 species in RefSeq that contained at least 10 species, leading to a total of 4,800 genomes. Genomes from metagenomes assembled from the ocean, soil, and premature infant fecal samples were accessed as deposited from the following publications (Diamond et al., 2018; Olm et al., 2019; Tully et al., 2018). CheckM (Parks et al., 2015) was run on all genome sets and only those with greater than or equal to 70% completeness and less than 5% contamination were retained.

6.2.2 Visualization of average nucleotide identity gap

All genomes in each genome set were compared to each other in a pairwise manner using FastANI (Jain et al., 2018)), and the genome alignment fraction was calculated by dividing the count of bidirectional fragment mappings by the number of total query fragments. ANI values and genome alignment fraction values were averaged for reciprocal comparisons and comparisons of genomes to themselves were removed. The density of each combination of ANI and alignment fraction was calculated using `scipy.stats.kde` (Jones et al., 2001). The density was plotted in a 3-dimensional histogram using `matplotlib` (Hunter, 2007).

6.2.3 Calculation of dN/dS and estimated homologous recombination

dRep (Olm et al., 2017) was used to compare each genome set in a pairwise manner on a gene-by-gene basis using the command “dRep dereplicate \$wd --S_algorithm goANI -pa 0.8 -con 5 -comp 70”. Briefly, this identifies open reading frames using Prodigal (Hyatt et al., 2010) and compares their nucleic acid sequences using NSimScan (Novichkov et al., 2016). The script “dnds_from_drep.py” was then used to calculate the dN/dS ratio among aligned sequences (<https://github.com/MrOlm/bacterialEvolutionMetrics>). This involved first aligning the amino acid sequences encoded by pairs of genes which at least 70% of the genes aligned with at least 70% sequence identity, and which were reciprocal best hits. Sequences were aligned globally using the BioPython Align.PairwiseAligner (Cock et al., 2009), using a `blosum62` substitution matrix, -12 open gap score, and -3 extend gap score. The alignment was then converted into a codon alignment using `biopython`, and the number of synonymous sites, synonymous substitutions, non-synonymous sites, and nonsynonymous substitutions recorded. Finally, the overall dN/dS was calculated for each genome alignment using the following formula: $((\text{non-synonymous substitutions} / \text{non-synonymous sites}) / (\text{synonymous substitutions} / \text{synonymous sites}))$.

Homologous recombination between genome pairs was calculated as the bias towards identical genes based on the overall ANI between genome pairs, similar to previously described methods (Brito et al., 2016). First, the set of aligned genes was filtered to only those with at least 500 bp aligned. The probability of each gene alignment being identical by chance was determined using the formula $(\text{overall genome ANI}^{(\text{length of genome alignment})})$. The genome-wide number of expected identical genes was calculated as the sum of the probabilities of each individual gene alignment being the same. The actual number of identical genes for each genome pair was calculated as the number of alignments with a percent identity greater than 99.99%. Finally, the genome-wide bias towards identical genes was calculated using the formula $((\text{number of identical genes} - \text{expected number of identical genes}) / \text{number of aligned genes})$.

6.2.4 Marker gene analyses

Bacterial single copy genes were identified based on a previously curated set of Hidden-Markov Models (HMMs) for 16 ribosomal proteins (Albertsen et al., 2013), as accessed on GitHub at the following link on April 10, 2019 <https://github.com/MadsAlbertsen/multi-metagenome/blob/master/R.data.generation/essential.hmm>. The amino acid sequences of all genomes were annotated using `prodigal` (Hyatt et al., 2010) and searched against the single copy gene HMMs using the command “`hmmsearch -E .001 --domE .001`” (`hmmsearch` (hmmsearch.org)). All hits with scores above the trusted cutoff for each HMM were retained. Both nucleic acid and amino acid sequences for each hit were compared using `usearch` (Edgar, 2010) with the command “`usearch -calc_distmx`”.

16S rRNA genes were identified using SEARCH_16S (Edgar, 2017), with the specific command “usearch -search_16s \$loc -bitvec gg97.bitvec”. gg97.bitvec was created using the commands “usearch -makeudb_usearch 97_otus.fasta -wordlength 13” and “usearch -udb2bitvec” based on the Greengenes reference database (as accessed at https://github.com/biocore/qiime-default-reference/blob/master/qiime_default_reference/gg_13_8_otus/rep_set/97_otus.fasta.gz) (DeSantis et al., 2006). Identified 16S rRNA genes were aligned to each other using Mothur ((Schloss et al., 2009) with RDP release 11, update 5 (Cole et al., 2014) used as the template. Distance matrices were calculated using the Mothur command dist.seqs.

Species delineation scores were calculated based the ability to recreate species-level clusters defined by RefSeq. A pairwise matrix was established listing each pair of genomes in our RefSeq genome subset and whether or not the pair belonged to the same species. The recall of a given genome clustering was defined as the number of genome pairs correctly identified as belonging to the same species divided by the total number of pairs of genomes belonging to the same species. The precision of a genome clustering was defined as the total number of pairs of genomes correctly identified as belonging to the same species divided by the total number of pairs of genomes correctly or incorrectly identified as belonging to the same species. The species delineation score was calculated as the sum of the recall and precision.

Optimal thresholds for species delineation were empirically determined based on a distance matrix between all genomes in our RefSeq genome subset (distance matrix generation described above). For each tested genome comparison method, all distance thresholds between 50% and 100% were tested, incrementing by 0.1% (50%, 50.1%, 50.2%, etc.). Each pair of genomes at least as similar as the threshold were considered belonging to the same species, and remaining pairs are considered to belong to different species. A species delineation score was calculated for each threshold, and the threshold with the highest score was considered optimal.

Recoverability was based on the total number of comparable units that could be recovered from a given metagenomic assembly. For each compared set of MAGs, the total number of recovered genomes was set as 100% recoverability. The recoverability of each compared protein was calculated as the number proteins that could be identified from each set of assemblies that genomes were binned from, divided by the number of quality-filtered genomes recovered from that set of assemblies. For example, if 100 genomes were recovered from a set of samples, and 300 proteins were recovered from the same set of samples, the recoverability of the protein would be 300%. The recoverability of all 16 ribosomal proteins concatenated together was calculated as the percentage of recovered genomes that contained all 16 ribosomal proteins.

6.3 Results

6.3.1 Generation of unbiased genome sets

Four criteria were used to identify sets of genomes unbiased by isolation and selection biases. 1) Genomes must be assembled from DNA extracted directly from the environment without enrichment or culturing. 2) There must be no preference for particular taxa during metagenomic genome binning and/or curation. 3) Genomes must be available from at least 50 samples from the same or similar environments, and there must be at least 1000 genomes total. 4) All genomes must be publically available for download, not just the de-replicated genome set. Many potential

metagenomic studies were disqualified based on criteria (3) and (4), leading to the ultimate selection of three genome sets for follow up analysis (Diamond et al., 2018; Olm et al., 2019; Tully et al., 2018). Recent studies involving large-scale genome binning (Parks et al., 2017; Pasolli et al., 2019) were disqualified, both because their de-replicated genome sets used a cutoff of 95% ANI (thus excluded natural strain clusters) and because their pre-de-replication sets included identical genomes from the same time series, leading to artificial genome clusters.

In this study, the first analysis set contains 2,178 bacterial genomes from 1,163 premature infant fecal samples, all of which were collected from infants born into the same neonatal intensive care unit (Olm et al., 2019) (**Supplemental Table S1**). These samples are low diversity and Proteobacteria and Firmicutes account for >80% of the bacteria (and for most samples, >90% of the reads could be assigned to genomes). The second set contains 1,166 genomes from the ocean, including Bacteria and Archaea (Tully et al., 2018). The third set contains 1,859 genomes from a meadow soil ecosystem (Diamond et al., 2018) with extremely diverse and complex microbial communities. We also included 4,800 genomes from NCBI GenBank, accessed February 2018, where we randomly selected 10 genomes from each of the 480 bacterial species with at least 10 genomes.

6.3.2 *Discrete sequence groups exist in all analyzed genome sets*

All genomes within each set were compared to each other in a pairwise manner using the FastANI algorithm (Jain et al., 2018). Discrete sequence groups based on both ANI and genome alignment percentage were found in all genome sets (**Figure 6.1**). Notably, species identity gaps were even more prominent in genome sets based on MAGs (metagenome assembled genomes) than GenBank (which mainly consists of cultured isolate genomes). Comparison of GenBank genomes marked as belonging to the same vs. different bacterial species showed the identity gap is largely consistent with annotated NCBI species taxonomy, and most genome clusters segregate from each other with a cluster boundary at around 95%. Thus, the analysis is consistent with prior suggestions that this cutoff delineates the species boundary. MAGs from the human microbiome were often very similar to each other (>98% ANI), whereas MAG clusters from the ocean included more divergent strain types. In contrast, most of the comparisons involving genomes from soil involved distinct species.

6.3.3 *Gaps in ANI spectra are consistent with measurements of recombination and selection*

We next looked for clues of the evolutionary forces that could lead to discrete sequences clusters. An open-source program was written to estimate the rate of horizontal gene transfer and the dN/dS ratio from each pairwise genome alignment in a rapid and high-throughput manner (see methods for details). Rates of homologous recombination were estimated based on the presence of more identical genes than would be expected by random chance based on the genome-wide ANI, similar to previously described methods (Brito et al., 2016). Genome-wide average dN/dS ratios, a measurement of bias towards nonsynonymous vs synonymous substitutions, were calculated using a python implementation of the Nei equation (Nei and Gojobori, 1986). Genes with a bias towards non-synonymous substitutions (potentially reflecting diversifying selection) have high dN/dS ratios (>1), while genes with a bias against non-synonymous mutations (potentially reflecting purifying selection) have low dN/dS ratios (<1).

Measurements of estimated homologous recombination and *dN/dS* both followed consistent patterns in relation to the 95% ANI species threshold in all three measured genome sets (**Figure**

6.2). Estimated homologous recombination rates followed showed a sharp decline from 100% ANI to around 95% ANI. This could be due to decreasing efficiency of homologous recombination with decreasing sequence similarity.

All dN/dS ratios were below 1, as expected for whole-genome dN/dS comparisons (Rocha et al., 2006). Values were highest (~ 0.4) between organisms with high sequence similarity and decreased with decreasing ANI, reaching a bottom plateau of about 0.1 (**Figure 6.2**). Interestingly, the dN/dS plateau did not tend to occur at 95% ANI, like homologous recombination, but earlier at around 98% ANI. It is well documented that whole-genome dN/dS values tend to be higher in recently diverged genomes (i.e., those with high ANI values) (Castillo-Ramírez and Feil, 2013; Rocha et al., 2006), and it is hypothesized that this is because it takes time for selection to purge non-synonymous mutations that are only slightly deleterious (nearly neutral). MAG clusters from soil had a slower decline in dN/dS with increasing divergence than was observed in other environments.

6.3.4 *Evaluating alternative methods for bacterial species delineation*

Species distinctions needed to generate an overview of the species composition of an environment could be generated using whole genome ANI comparisons if genomes were reconstructed comprehensively from metagenomes. However, this is not possible when genomes are not reconstructed for most organisms, a common outcome for high-complexity environments like soil. For example, (Howe et al., 2014) only 11% of the reads from soil were reconstructed from soil, essentially precluding genome recovery. Even in better cases, only 36.4% of reads were assembled into binned contigs and genomes were reconstructed for only $\sim 23\%$ of the detected bacteria (Diamond et al. 2019). Thus, we investigated thresholds for taxonomic species delineation using 17 marker genes that occur in single copy in all genomes, benchmarked using RefSeq. Marker gene thresholds were generally above 99% ANI (**Figure 6.3a**; **Supplemental Table S6.2**). Whole-genome alignments performed best (**Figure 6.3b**) and full-length 16S rRNA alignments performed worst for species discrimination. All ribosomal protein genes performed reasonably well, with the best result for rpL6.

It is important that genes used to generate species inventories are well reconstructed from metagenomes. We compared the number of marker genes that could be assembled from each dataset as to the number of genomes that were assembled from the same dataset, and found that on average five times more ribosomal genes than genomes were recovered (**Figure 6.3b**). This finding is generally consistent with results from soil, where Diamond et al. reconstructed 795 genomes and 3325 rpS3 genes (although the dereplication of the rpS3 genes was not as stringent as is now recommended). 16S rRNA genes were recovered much less often than ribosomal protein genes. 16S rRNA sequences were assembled in $<15\%$ of genomes from metagenomes, whereas genes encoding ribosomal proteins were reconstructed for $>80\%$ of genomes (**Supplemental Figure S6.3**). Thus, while whole genome comparison methods are most accurate, ribosomal proteins are a good option for species-level marker gene analysis in studies when genomes were not comprehensively recovered.

6.4 Discussion

In line with previous studies using reference databases, here we show that bacterial diversity in natural communities is clustered in all three environments studied. Estimated rates of horizontal

gene transfer fell to near zero at the 95% ANI boundary in all tested environments, and genome-wide dN/dS ratios consistently leveled near values of 0.15 at around 98% ANI in most environments. The three independent metrics support the existence of natural “microbial species”.

The observed drop in estimated homologous recombination with decreasing DNA similarity suggests that sequence-dependent homologous recombination is likely a homogenizing force preventing dissolution of bacterial species, in line with previous experimental laboratory studies, computer simulations, (Fraser et al., 2007; Majewski and Cohan, 1999; Vulić et al., 1997), and direct measurements of recombination vs. mutation rates in natural populations (Eppley et al., 2007). The drop in dN/dS values at ANIs significantly above 95% suggests that purifying selection is not likely a species-preserving evolutionary force, and is evidence against the model of speciation positing that species clusters result from series of genetic sweeps (Gevers et al., 2005). Taken together, these observations support the notion that bacteria that share >95% ANI can recombine due to shared sequence similarity, but the rates decline as this threshold is approached. In combination, the observations support the applicability of the eukaryotic biological species concept to bacteria.

Given that an increasing number of genomes derive from metagenomic DNA without culturing or isolation, a sequence-based method for species delineation is a practical necessity. While thresholds are always prone to exceptions, a genome-wide 95% ANI threshold for species delineation appears to be optimal given the data presented here and previously (Goris et al., 2007; Jain et al., 2018; Konstantinidis and Tiedje, 2005), as well as current species-level taxonomic assignments in NCBI. Importantly, we identified many single copy genes that are well reconstructed for metagenomes and provide metrics that can proxy for ANI values, and thus are useful for descriptions of community composition.

6.5 Figures

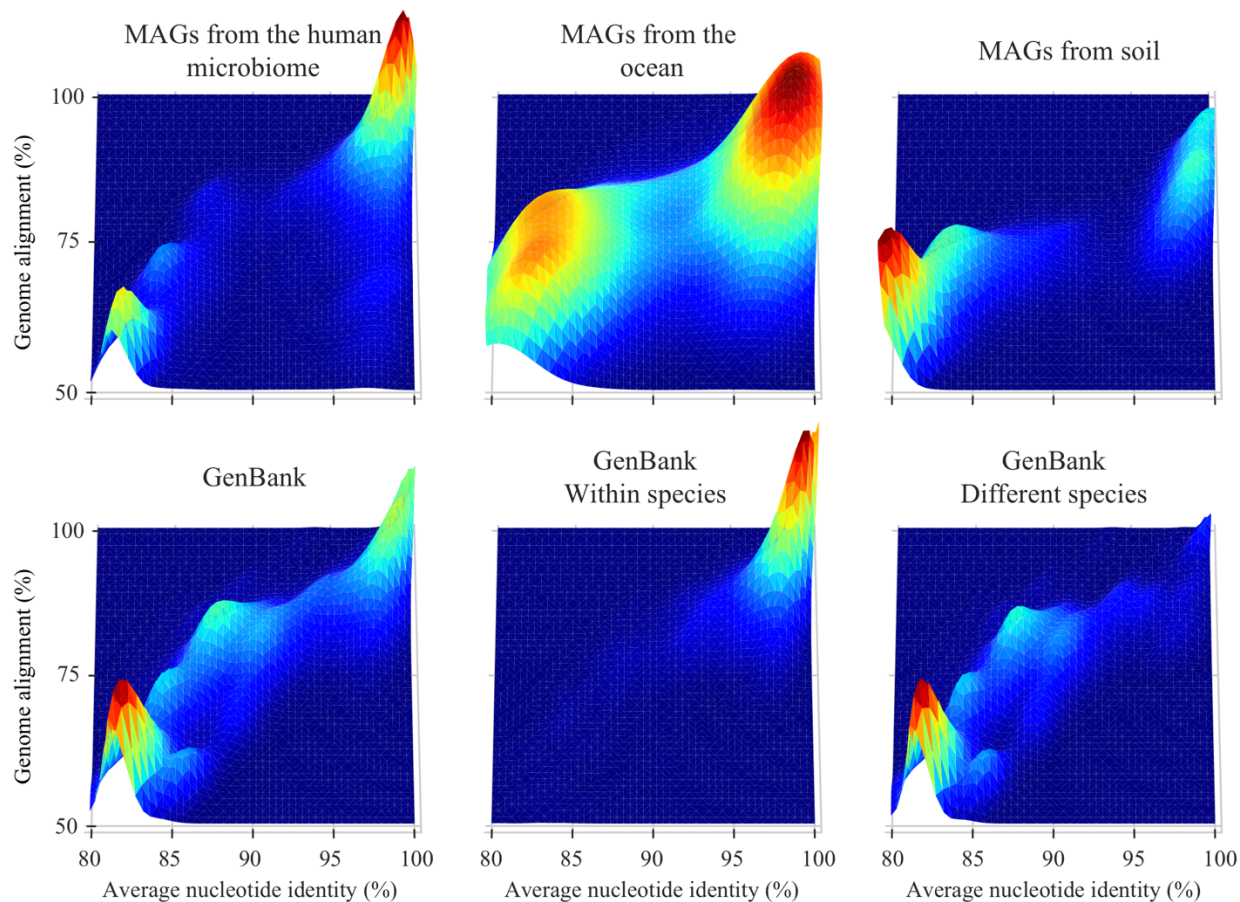


Figure 6.1 Average nucleotide identity gaps exist near ~95% ANI in all tested genome sets. Each plot is a histogram of average nucleotide identity and genome alignment percentage values resulting from pairwise comparison within a genome set. Height and color represent density, with higher peaks and hotter colors representing higher numbers of comparisons with that particular ANI and genome alignment percentage. The top row contains three sets of metagenome assembled genomes (MAGs) from different environments. On the bottom row is NCBI GenBank (rarefied to reduce taxonomic bias; see methods), GenBank only including comparisons between genomes annotated as the same species, and GenBank only including comparisons between genomes annotated as different species.

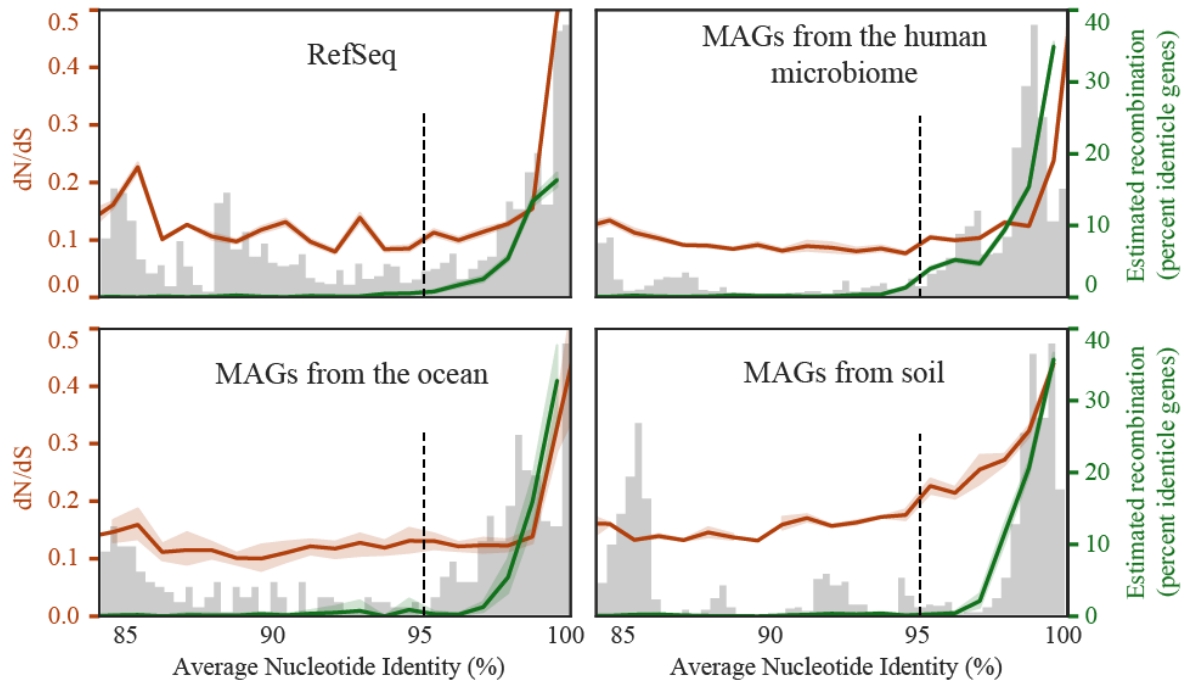


Figure 6.2 Markers of recombination and selection follow patterns related to the proposed 95% ANI species threshold. Each plot displays a histogram of ANI values resulting from pairwise comparison with a genome set (light grey bars), the median estimated recombination rate at each ANI level (green line), and the median dN/dS ratio at each ANI level (orange line). A dotted line is drawn at at 95% ANI to mark the commonly proposed threshold for species delineation, and 95% confidence intervals are shown shaded around green and red lines.

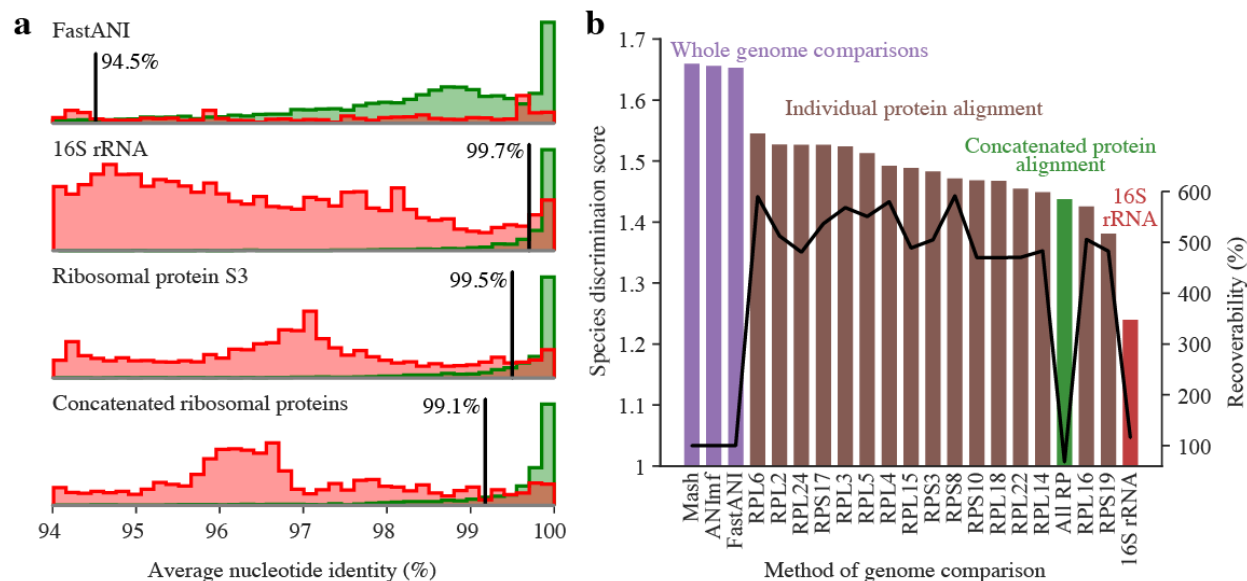


Figure 6.3 Whole genome alignment outperforms marker genes for species discrimination. **(a)** Histograms of ANI values between bacteria from GenBank annotated as belonging to the same species (green) or different species (red). Each row is a different method of nucleotide sequence alignment, and vertical black lines indicate the ANI value with the highest species discrimination score for that method. **(b)** The highest species discrimination score, a metric for how well microbes of the same vs. different species can be separated, for various nucleotide alignment methods. Shown are three algorithms for whole genome alignment (purple), nucleotide alignments of 16 ribosomal proteins (brown), an alignment resulting from concatenating all ribosomal proteins (green), and a full length 16S rRNA gene alignment (red). A black line indicates the average recoverability of each comparison method across all three tested datasets. Recoverability describes the ability of the feature being compared by the metric to be assembled. The number of genomes recovered from a dataset is defined as 100%, and the recoverability of a feature that was recovered twice as often as genomes would have a recoverability of 200%, for example.

Supplemental Table S6.1. Information on MAG sets

Set	Genomes	Samples	DOI
infantMAGs	2178	1163	https://doi.org/10.1101/558676
soilMAGs	1859	60	https://doi.org/10.1101/445817

oceanMAGs	1166	234	https://doi.org/10.1038/sdata.2017.203
-----------	------	-----	---

Supplemental Table S6.2. Optimal species delineation thresholds

Method	Sequence type	Threshold	Precision	Recall	Score
Mash	DNA	94.9	0.80936106	0.85215909	1.66152015
ANImf	DNA	94.4	0.80715056	0.85087121	1.65802177
FastANI	DNA	94.5	0.80790634	0.84689394	1.65480028
RPL6	DNA	99.3	0.79069307	0.75625	1.54694307
RPL2	DNA	99.2	0.75317177	0.77579545	1.52896723
RPL24	DNA	99.4	0.77576988	0.75287879	1.52864866
RPS17	DNA	99.3	0.73539121	0.7932197	1.5286109
RPL3	DNA	99.1	0.72483635	0.80113636	1.52597272
RPL5	DNA	99.1	0.69889019	0.81579545	1.51468564

RPL4	DNA	98.7	0.65225812	0.84193182	1.49418994
RPL15	DNA	99.1	0.718292	0.77227273	1.49056472
RPS3	DNA	99.5	0.74522001	0.73965909	1.4848791
RPS8	DNA	99	0.62661546	0.84666667	1.47328212
RPS10	DNA	99.4	0.70115657	0.7692803	1.47043687
RPL18	DNA	99.2	0.67850759	0.79079545	1.46930304
RPL22	DNA	99.4	0.65202713	0.80473485	1.45676198
RPL14	DNA	99.2	0.61561444	0.83511364	1.45072808
All RP	DNA	99.1	0.73423015	0.705	1.43923015
RPL16	DNA	99.6	0.69596251	0.73128788	1.42725039
RPS19	DNA	100	0.68101452	0.7017803	1.38279482
16S rRNA	DNA	99.7	0.65569738	0.58568182	1.2413792
RPL6	AminoAcid	99.7	0.73725194	0.73878788	1.47603982

All RP	AminoAcid	99.9	0.86037332	0.57791667	1.43828998
RPL3	AminoAcid	99.6	0.65177335	0.76640152	1.41817486
RPS8	AminoAcid	99.3	0.54718347	0.85878788	1.40597135
RPL5	AminoAcid	99.5	0.54341432	0.83280303	1.37621735
RPL2	AminoAcid	99.9	0.57972268	0.78708333	1.36680601
RPL24	AminoAcid	99.1	0.52917235	0.83723485	1.3664072
RPL15	AminoAcid	98.7	0.46445273	0.89159091	1.35604364
RPS17	AminoAcid	98.9	0.48545339	0.86590909	1.35136248
RPL18	AminoAcid	99.2	0.470282	0.82121212	1.29149412
RPS3	AminoAcid	99.3	0.31017773	0.91162879	1.22180651
RPL16	AminoAcid	99.6	0.32937547	0.88159091	1.21096638
RPL22	AminoAcid	99.1	0.29092691	0.91056818	1.2014951
RPL14	AminoAcid	99.7	0.25883927	0.92340909	1.18224836

RPL4	AminoAcid	98.1	0.24711558	0.93136364	1.17847921
RPS19	AminoAcid	99.4	0.21774692	0.90655303	1.12429996
RPS10	AminoAcid	98.1	0.07505528	0.97333333	1.04838861

Supplemental Table S6.3. Marker gene recoverability from metagenomes

Set	Gene	Percentage Recovered (%)	Average copy number	Median length (bp)
GenBank	16s	85.1666667	2.37059687	1502
GenBank	RPL14	99.5833333	1.00439331	369
GenBank	RPL15	99.125	1.0023119	441
GenBank	RPL16	99.0416667	1.00694152	417
GenBank	RPL18	97.5208333	1.00640889	357
GenBank	RPL2	99.6458333	1.00188166	831

GenBank	RPL22	99.625	1.00418235	342
GenBank	RPL24	99.1875	1.00231044	315
GenBank	RPL3	99.2291667	1.00209952	636
GenBank	RPL4	99.7291667	1.01044496	621
GenBank	RPL5	99.8958333	1.00208551	540
GenBank	RPL6	99.4791667	1.00586387	537
GenBank	RPS10	99.7291667	1.00376018	309
GenBank	RPS17	99.5833333	1.00251046	261
GenBank	RPS19	97.8333333	1.00212947	279
GenBank	RPS3	99.7708333	1.00313218	699
GenBank	RPS8	99.8958333	1.00479666	399
GenBank	concat_16	93.9791667	1	7356
infantMAGs	16s	13.4891349	1.37575758	1490

infantMAGs	RPL14	84.2148421	1.01605839	369
infantMAGs	RPL15	88.601886	1.01248844	441
infantMAGs	RPL16	84.3378434	1.00777454	432
infantMAGs	RPL18	83.1898319	1.00935961	360
infantMAGs	RPL2	83.6818368	1.00391773	831
infantMAGs	RPL22	83.1488315	1.01084278	339
infantMAGs	RPL24	83.2718327	1.00442913	315
infantMAGs	RPL3	85.895859	1.00572519	630
infantMAGs	RPL4	85.5268553	1.01724964	624
infantMAGs	RPL5	84.0918409	1.00633528	540
infantMAGs	RPL6	84.4608446	1.01358564	537
infantMAGs	RPS10	86.797868	1.003305	309
infantMAGs	RPS17	83.3948339	1.00737101	261

infantMAGs	RPS19	82.1648216	1.00498753	279
infantMAGs	RPS3	83.8458385	1.00537634	669
infantMAGs	RPS8	84.4608446	1.01213003	399
infantMAGs	concat_16	67.6916769	1	7347

Concluding remarks and future work

An overarching conclusion from this work is that techniques for cultivation independent research of the human microbiome are rapidly advancing, with large-scale changes in the research landscape occurring over the five years during which this thesis was written. It has been known for over a century that there are microbes that inhabit humans and likely impact health (Gordon, 2008; Shulman et al., 2007), but the vast complexity and diversity of the bacteria that make up human oral and fecal microbiomes were not established until the mid-2000s through the use of 16S rRNA gene sequencing (Aas et al., 2005; Gill et al., 2006). These discoveries in part led to the establishment of the NIH human microbiome project in 2007, a large-scale multi-center effort to characterize the healthy human microbiome using 16S rRNA sequencing and limited shotgun metagenomic sequencing.

The first major results from the Human Microbiome Project, published in 2012, used an OTU-based analysis pipeline to analyze 16S rRNA sequencing data, QIIME, and methods that involve mapping metagenomic reads to reference genomes to analyze metagenomic data, MetaPhlAn and HUMAnN (Abubucker et al., 2012; Caporaso et al., 2010; Human Microbiome Project Consortium, 2012; Segata et al., 2012). 16S rRNA sequencing is still heavily in use today, but there have been many significant changes to algorithm specifics in recent years. These include a move away from OTU-based pipelines, which have low specificity, to the use of oligotypes and exact sequence variants (ESVs), which involve more computational complexity during analysis but result in more detailed taxonomic calls (Callahan et al., 2017; Eren et al., 2013; Nayfach et al., 2015)). These methods are still not able to achieve species-level resolution, however, as presented in Chapter 6 of this thesis. Mapping-based shotgun metagenomic pipelines are also still in use today, and thanks to computational improvements and the increasing size of reference databases, these methods can achieve reasonable specificity and accuracy when reference genomes are available (Nayfach et al., 2016). However, large fractions of bacterial strains in the human microbiome do not have representatives in public databases to this day (Nayfach et al., 2019; Pasolli et al., 2019), making genome-resolved metagenomics the best option for strain-resolved analyses and functional potential prediction. This resolution is especially important when profiling human microbiome samples, where single genes or mutations can alter the pathogenicity of an organism (Schloissnig et al. 2012; Franzosa et al. 2015).

The biggest hurdle to genome-resolved metagenomic analysis is the assembly step. While the first genome-resolved metagenomic study was published in 2004 using DNA extracted from acid mine drainage biofilms (Tyson et al., 2004), the higher complexity of the human gut microbiome precluded genome-resolved metagenomic study until sequencing technologies improved. The first culture-independent genomes recovered from the human microbiome were described in 2011, and were based on relatively simple premature infant fecal samples (Morowitz et al., 2011). Thanks to increases in sequencing throughput, genome-resolved metagenomic studies have since been performed on many aspects of the adult human microbiome (Di Rienzi et al., 2013; Bäckhed et al., 2015; Vineis et al., 2016), including the thousands of genomes recovered in this thesis, and recent studies have assembled and binned hundreds of thousands of microbial genomes from the human microbiome (Nayfach et al., 2019; Pasolli et al., 2019).

The increasing popularity of genome-resolved metagenomics has led to a number of new analysis methods and pipelines for genome-resolved metagenomic data. Arguably the most important development is the ability to measure *in situ* genome replication rates from metagenomic data. First proposed for use with complete reference genomes (Korem et al., 2015), improvements have been made to allow its use in fragmented draft genomes (Brown et al., 2016 and Appendix 1.1) and to improve measurement accuracy (Gao and Li, 2018). New algorithms have also been developed to compare large number of genomes at speeds that are orders of magnitude faster than before (Jain et al., 2018; Olm et al., 2017, see Chapter 3; Ondov et al., 2016), and a plethora of automatic genome binning algorithms have been recently developed as well (Lu et al., 2017; Sedlar et al., 2017; Sieber et al., 2018). An exciting prospect afforded by the increasing number of samples generated per study is the ability to perform statistical and machine-learning based analyses. The use of classic machine learning algorithms have led to significant improvements in our ability to predict microbiome responses (see Chapter 5, Rahman et al., 2018; see Appendix 1.4; Smillie et al., 2018), and the full exploitation of these and future algorithms will likely lead to our next level understanding of the complex dynamics that govern human microbiome assembly.

Another overarching conclusion from this work is the power of custom metagenomic methods. Many studies restrict their sequencing-based analysis to what can be performed by a pre-developed pipeline, like QIIME for 16S rRNA data or MIDAS for mapping based metagenomic data (Caporaso et al., 2010; Nayfach et al., 2015). While this is sufficient to address questions like “are these microbiomes different?” and “which microbes are significantly more abundant in specific samples?”, many other detailed questions can only be answered by customizing analysis methods to address specific questions. For example, the progressive fixation of SNPs of identical genomes in different body-sites, described in Chapter 2 of this thesis, was only possible by linking together several previously described algorithms and developing a significant amount of new methodologies as well. As shown throughout this thesis, analysis of the sequencing reads and generation of inventories of polymorphic sites can address many interesting questions, such as patterns of microbial strain dispersal, *in situ* evolutionary rates, strain-specific growth rates, strain-specific gene content and zygosity and ploidy of eukaryotic genomes.

A conclusion of this work is that there is vast amount left to learn about the human microbiome. In this thesis, significant effort was made to fully characterize the microbiomes studied, including consideration of bacteria, bacteriophages, plasmids, and eukaryotes. While this led to assignment of ~90% of the reads to genomes, the remaining 10% of reads probably derive from important, but low abundance community members, and deeper sequencing would reveal additional such rare members. Further, while *in situ* evolution was characterized for relatively high-abundance microbial community members, rare evolutionary events that could be insightful about the evolutionary pressures faced by other microbiome members may be missed. For example, acquisition of mobile elements known to heavily modulate the virulence of common bacterial pathogens like *Vibrio cholera* (Reidl and Mekalanos, 1995) may only occur in a few cells.

Gaps in knowledge about microbial communities can also be due to fragmentation of assemblies. Recently, it was shown that large bacteriophages are common in many different microbiome types (Al-Shayeb et al., 2019; Devoto et al., 2019). However, these are likely far more widespread than currently realized because their genomes are almost always fragmented. New methods for

automatic genome curation would be both to solve this problem and to greatly improve the reliability of microbial genomes.

Human microbiome research could benefit enormously from the use of models. For example, a model describing the sequences from microbial populations as clouds, rather than genomes with specific sequence variants, would more closely reflect reality and could be useful for both between-sample with within-sample comparisons. This model should include consideration of linkage and clonality, and would immediately be useful for identifying positions within the genome that are experiencing unique selective pressures. Machine learning models may also find application. Such models have been successfully applied to microbial genomic data, but significant improvements could be made by altering the models to better handle the complex inputs from human microbiome, including genes, growth rates, and species relationships.

There are several readily identifiable next research steps to follow up on this work. First, efforts should be made to follow cohorts of infants over longer time periods. This would allow determination of how long initial strains persist, what life events most effect strain acquisition, and patterns of microbiome assembly. Work presented in this thesis shows the short term-effects of things like antibiotic administration and acute disease, but it is unclear how long these changes persist. Additional consideration of the initial colonization sources of strains is needed as well. The hospital room is a likely prominent source of initial infant-colonizing strains, as discussed in Appendix 1, but the maternal microbiome likely has a large influence as well. Characterization of the relative impact of these sources, as well as the how long strains from different environments persist over time, would provide greater clarity into which exposures are most important to develop a healthy microbiome. These questions can all be addressed as research transitions to rely on genome-resolved metagenomics methods that leverage the important information contained in the sequencing reads.

References

- Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I., & Dewhirst, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.*, *43*(11), 5721–5732.
<https://doi.org/10.1128/JCM.43.11.5721-5732.2005>
- Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., ... Huttenhower, C. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.*, *8*(6), e1002358.
<https://doi.org/10.1371/journal.pcbi.1002358>
- Achtman, M., & Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*, *6*(6), 431–440.
<https://doi.org/10.1038/nrmicro1872>
- Acland, A., Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., ... Zbicz, K. (2013). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *41*(D1), D8–D20. <https://doi.org/10.1093/nar/gks1189>
- Afrazi, A., Sodhi, C. P., Richardson, W., Neal, M., Good, M., Siggers, R., & Hackam, D. J. (2011). New insights into the pathogenesis and treatment of necrotizing enterocolitis: Toll-like receptors and beyond. *Pediatr. Res.*, *69*(3), 183–188.
<https://doi.org/10.1203/PDR.0b013e3182093280>
- Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, *31*(6), 533–538.
<https://doi.org/10.1038/nbt.2579>
- Albertsen, M., Karst, S. M., Ziegler, A. S., Kirkegaard, R. H., & Nielsen, P. H. (2015). Back to Basics--The Influence of DNA Extraction and Primer Choice on Phylogenetic Analysis of Activated Sludge Communities. *PLoS One*, *10*(7), e0132783.
<https://doi.org/10.1371/journal.pone.0132783>
- Aliaga, S., Clark, R. H., Laughon, M., Walsh, T. J., Hope, W. W., Benjamin, D. K., ... Smith, P. B. (2014). Changes in the Incidence of Candidiasis in Neonatal Intensive Care Units. *Pediatrics*, *133*(2), 236–242. <https://doi.org/10.1542/peds.2013-0671>
- Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., ... Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods*, *11*(11), 1144–1146. <https://doi.org/10.1038/nmeth.3103>
- Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., ... Banfield, J. F. (2019). *Clades of huge phage from across Earth's ecosystems*.
<https://doi.org/10.1101/572362>
- ALTSCHUL, S., GISH, W., MILLER, W., MYERS, E., & LIPMAN, D. (1990). BASIC LOCAL ALIGNMENT SEARCH TOOL. *Journal of Molecular Biology*, *215*(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](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*, 13219.
<https://doi.org/10.1038/ncomms13219>

- Arnison, P. G., Bibb, M. J., Bierbaum, G., Bowers, A. A., Bugni, T. S., Bulaj, G., ... van der Donk, W. A. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.*, 30(1), 108–160. <https://doi.org/10.1039/c2np20085f>
- Arrieta, M.-C., Stiemsma, L. T., Dimitriu, P. A., Thorson, L., Russell, S., Yurist-Doutsch, S., ... Finlay, B. B. (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Science Translational Medicine*, 7(307), 307ra152. <https://doi.org/10.1126/scitranslmed.aab2271>
- Assefa, S., Keane, T. M., Otto, T. D., Newbold, C., & Berriman, M. (2009). ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, 25(15), 1968–1969. <https://doi.org/10.1093/bioinformatics/btp347>
- Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., ... Jun, W. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe*, 17(5), 690–703. <https://doi.org/10.1016/j.chom.2015.04.004>
- BALEY, J., KLIEGMAN, R., BOXERBAUM, B., & FANAROFF, A. (1986). FUNGAL COLONIZATION IN THE VERY-LOW-BIRTH-WEIGHT INFANT. *Pediatrics*, 78(2), 225–232.
- Ballance, W. A., Dahms, B. B., Shenker, N., & Kliegman, R. M. (1990). Pathology of neonatal necrotizing enterocolitis: a ten-year experience. *J. Pediatr.*, 117(1 Pt 2), S6-13. [https://doi.org/10.1016/S0022-3476\(05\)81124-2](https://doi.org/10.1016/S0022-3476(05)81124-2)
- Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., ... Zhang, J. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1), D204–D212. <https://doi.org/10.1093/nar/gku989>
- Bendall, M. L., Stevens, S. L. R., Chan, L.-K., Malfatti, S., Schwientek, P., Tremblay, J., ... Malmstrom, R. R. (2016). Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME Journal*, 10(7), 1589–1601. <https://doi.org/10.1038/ismej.2015.241>
- Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4), 1165–1188.
- Bennett, R. J., & Johnson, A. D. (2003). Completion of a parasexual cycle in *Candida albicans* by induced chromosome loss in tetraploid strains. *Embo Journal*, 22(10), 2505–2515. <https://doi.org/10.1093/emboj/cdg235>
- Bhaya, D., Grossman, A. R., Steunou, A.-S., Khuri, N., Cohan, F. M., Hamamura, N., ... Heidelberg, J. F. (2007). Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J.*, 1(8), 703–713. <https://doi.org/10.1038/ismej.2007.46>
- Bjedov, I., Tenaillon, O., Gerard, B., Souza, V., Denamur, E., Radman, M., ... Matic, I. (2003). Stress-induced mutagenesis in bacteria. *Science*, 300(5624), 1404–1409. <https://doi.org/10.1126/science.1082240>
- Bokulich, N. A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., ... Blaser, M. J. (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.*, 8(343), 343ra82-343ra82. <https://doi.org/10.1126/scitranslmed.aad7121>
- Bokulich, N. A., Mills, D. A., & Underwood, M. A. (2013). Surface Microbes in the Neonatal Intensive Care Unit: Changes with Routine Cleaning and over Time. *Journal of Clinical Microbiology*, 51(8), 2617–2624. <https://doi.org/10.1128/JCM.00898-13>

- Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., ... Alm, E. J. (2016). Mobile genes in the human microbiome are structured from global to individual scales. *Nature*. <https://doi.org/10.1038/nature18927>
- 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. <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, 1814. <https://doi.org/10.1038/s41467-017-02018-w>
- Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., ... Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, 523(7559), 208-U173. <https://doi.org/10.1038/nature14486>
- 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/10.1038/nbt.3704>
- Brown, C. T., Xiong, W., Olm, M. R., Thomas, B. C., Baker, R., Firek, B., ... Banfield, J. F. (2018). Hospitalized Premature Infants Are Colonized by Related Bacterial Strains with Distinct Proteomic Profiles. *Mbio*, 9(2), e00441-18. <https://doi.org/10.1128/mBio.00441-18>
- Browne, H. P., Forster, S. C., Anonye, B. O., Kumar, N., Neville, B. A., Stares, M. D., ... Lawley, T. D. (2016). Culturing of “unculturable” human microbiota reveals novel taxa and extensive sporulation. *Nature*, 533(7604), 543-+. <https://doi.org/10.1038/nature17645>
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 12(1), 59–60. <https://doi.org/10.1038/nmeth.3176>
- Budowle, B., Connell, N. D., Bielecka-Oder, A., Colwell, R. R., Corbett, C. R., Fletcher, J., ... Minot, S. (2014). Validation of high throughput sequencing and microbial forensics applications. *Investigative Genetics*, 5, 9–9. <https://doi.org/10.1186/2041-2223-5-9>
- Bush, R. K., & Portnoy, J. M. (2001). The role and abatement of fungal allergens in allergic diseases. *Journal of Allergy and Clinical Immunology*, 107(3), S430–S440. <https://doi.org/10.1067/mai.2001.113669>
- Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. A. S., Sakthikumar, S., Munro, C. A., ... Cuomo, C. A. (2009). Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, 459(7247), 657–662. <https://doi.org/10.1038/nature08064>
- Butterfield, C. N., Li, Z., Andeer, P. F., Spaulding, S., Thomas, B. C., Singh, A., ... Banfield, J. F. (2016). Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ*, 4, e2687. <https://doi.org/10.7717/peerj.2687>
- Cadillo-Quiroz, H., Didelot, X., Held, N. L., Herrera, A., Darling, A., Reno, M. L., ... Whitaker, R. J. (2012). Patterns of gene flow define species of thermophilic Archaea. *PLoS Biol.*, 10(2), e1001265. <https://doi.org/10.1371/journal.pbio.1001265>
- Cahenzli, J., Koeller, Y., Wyss, M., Geuking, M. B., & McCoy, K. D. (2013). Intestinal Microbial Diversity during Early-Life Colonization Shapes Long-Term IgE Levels. *Cell Host & Microbe*, 14(5), 559–570. <https://doi.org/10.1016/j.chom.2013.10.004>

- Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.*, *11*(12), 2639–2643. <https://doi.org/10.1038/ismej.2017.119>
- Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., ... Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196. <https://doi.org/10.1101/gr.6743907>
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, *7*(5), 335–336. <https://doi.org/10.1038/nmeth.f.303>
- Castillo-Ramírez, S., & Feil, E. J. (2013). Covering All the Bases: The Promise of Genome-Wide Sequence Data for Large Population Samples of Bacteria. In G. Trueba & C. Montúfar (Eds.), *Evolution from the Galapagos* (Vol. 2, pp. 41–62). Retrieved from http://link.springer.com/10.1007/978-1-4614-6732-8_5
- Chan, M. S., Maiden, M. C. J., & Spratt, B. G. (2001). Database-driven multi locus sequence typing (MLST) of bacterial pathogens. *Bioinformatics*, *17*(11), 1077–1083. <https://doi.org/10.1093/bioinformatics/17.11.1077>
- Chang, H. J., Miller, H. L., Watkins, N., Arduino, M. J., Ashford, D. A., Midgley, G., ... Jarvis, W. R. (1998). An epidemic of *Malassezia pachydermatis* in an intensive care nursery associated with colonization of health care workers pet dogs. *New England Journal of Medicine*, *338*(11), 706–711. <https://doi.org/10.1056/NEJM199803123381102>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *1*, *16*, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.*, *33*(Database issue), D325-8. <https://doi.org/10.1093/nar/gki008>
- Chen, T. A., & Hill, P. B. (2005). The biology of *Malassezia* organisms and their ability to induce immune responses and skin disease. *Veterinary Dermatology*, *16*(1), 4–26. <https://doi.org/10.1111/j.1365-3164.2005.00424.x>
- Cilieborg, M. S., Boye, M., & Sangild, P. T. (2012). Bacterial colonization and gut development in preterm neonates. *Early Human Development*, *88*, S41–S49. <https://doi.org/10.1016/j.earlhumdev.2011.12.027>
- Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., ... DiCuccio, M. (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.*, *68*(7), 2386–2392. <https://doi.org/10.1099/ijsem.0.002809>
- Claud, E. C., Lu, L., Anton, P. M., Savidge, T., Allan Walker, W., & Cherayil, B. J. (2004). Developmentally regulated IκB expression in intestinal epithelium and susceptibility to flagellin-induced inflammation. *Proc. Natl. Acad. Sci. U. S. A.*, *101*(19), 7404–7408. <https://doi.org/10.1073/pnas.0401710101>
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... Wilczynski, B. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423.
- Cohan, F. M. (2002). What are Bacterial Species? *Annu. Rev. Microbiol.*, *56*(1), 457–487. <https://doi.org/10.1146/annurev.micro.56.012302.160634>
- Cohan, F. M. (2019). Systematics: The Cohesive Nature of Bacterial Species Taxa. *Curr. Biol.*, *29*(5), R169–R172. <https://doi.org/10.1016/j.cub.2019.01.033>

- Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., ... Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.*, 42(Database issue), D633-42. <https://doi.org/10.1093/nar/gkt1244>
- 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), e00782-13. <https://doi.org/10.1128/mBio.00782-13>
- 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(6086), 1255–1262. <https://doi.org/10.1126/science.1224203>
- Crits-Christoph, A., Gelsinger, D. R., Ma, B., & others. (2016). Functional interactions of archaea, bacteria and viruses in a hypersaline endolithic community. *Environmentalist*. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1111/1462-2920.13259>
- Dawson, T. L. (2007). Malassezia globosa and restricta: Breakthrough understanding of the etiology and treatment of dandruff and seborrheic dermatitis through whole-genome analysis. *Journal of Investigative Dermatology Symposium Proceedings*, 12(2), 15–19. <https://doi.org/10.1038/sj.jidsymp.5650049>
- de la Cochetiere, M.-F., Piloquet, H., des Robert, C., Darmaun, D., Galmiche, J.-P., & Roze, J.-C. (2004). Early intestinal bacterial colonization and necrotizing enterocolitis in premature infants: the putative role of Clostridium. *Pediatr. Res.*, 56(3), 366–370. <https://doi.org/10.1203/01.PDR.0000134251.45878.D5>
- Deatherage, D. E., & Barrick, J. E. (2014). Identification of Mutations in Laboratory-Evolved Microbes from Next-Generation Sequencing Data Using breseq. In L. Sun & W. Shou (Eds.), *Engineering and Analyzing Multicellular Systems: Methods and Protocols* (Vol. 1151, pp. 165–188).
- Delcher, A. L., Salzberg, S. L., & Phillippy, A. M. (2003). Using MUMmer to identify similar regions in large sequence sets. *Curr. Protoc. Bioinformatics*, Chapter 10, Unit 10.3. <https://doi.org/10.1002/0471250953.bi1003s00>
- Denning, N.-L., & Prince, J. M. (2018). Neonatal intestinal dysbiosis in necrotizing enterocolitis. *Mol. Med.*, 24(1), 4. <https://doi.org/10.1186/s10020-018-0002-0>
- DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, 72(7), 5069–5072. <https://doi.org/10.1128/AEM.03006-05>
- Devoto, A. E., Santini, J. M., Olm, M. R., Anantharaman, K., Munk, P., Tung, J., ... Banfield, J. F. (2019). Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nature Microbiology*. <https://doi.org/10.1038/s41564-018-0338-9>
- Di Rienzi, S. C., Sharon, I., Wrighton, K. C., Koren, O., Hug, L. A., Thomas, B. C., ... Ley, R. E. (2013). The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *Elife*, 2, e01102.
- Diamond, S., Andeer, P., Li, Z., Crits-Christoph, A., Burstein, D., Anantharaman, K., ... Banfield, J. F. (2018). *Processing of grassland soil C-N compounds into soluble and volatile molecules is depth stratified and mediated by genomically novel bacteria and archaea*. <https://doi.org/10.1101/445817>
- Ding, T., & Schloss, P. D. (2014). Dynamics and associations of microbial community types across the human body. *Nature*, 509(7500), 357-+. <https://doi.org/10.1038/nature13178>

- Dittmar, E., Beyer, P., Fischer, D., Schäfer, V., Schoepe, H., Bauer, K., & Schlösser, R. (2008). Necrotizing enterocolitis of the neonate with *Clostridium perfringens*: diagnosis, clinical course, and role of alpha toxin. *Eur. J. Pediatr.*, *167*(8), 891–895. <https://doi.org/10.1007/s00431-007-0614-9>
- Dombrowski, N., Donaho, J. A., Gutierrez, T., Seitz, K. W., Teske, A. P., & Baker, B. J. (2016). Reconstructing metabolic pathways of hydrocarbon-degrading bacteria from the Deepwater Horizon oil spill. *Nature Microbiology*, *1*(7), 16057. <https://doi.org/10.1038/NMICROBIOL.2016.57>
- Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(26), 11971–11975. <https://doi.org/10.1073/pnas.1002601107>
- Dominguez-Bello, M. G., De Jesus-Laboy, K. M., Shen, N., Cox, L. M., Amir, A., Gonzalez, A., ... Clemente, J. C. (2016). Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nature Medicine*, *22*(3), 250–253. <https://doi.org/10.1038/nm.4039>
- Donaldson, G. P., Lee, S. M., & Mazmanian, S. K. (2016). Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology*, *14*(1), 20–32. <https://doi.org/10.1038/nrmicro3552>
- Du, Y.-L., Shen, X.-L., Yu, P., Bai, L.-Q., & Li, Y.-Q. (2011). Gamma-butyrolactone regulatory system of *Streptomyces chattanoogensis* links nutrient utilization, metabolism, and development. *Appl. Environ. Microbiol.*, *77*(23), 8415–8426. <https://doi.org/10.1128/AEM.05898-11>
- 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>
- Edgar, Robert C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. <https://doi.org/10.1093/bioinformatics/btq461>
- Edgar, Robert C. (2017). *SEARCH_16S: A new algorithm for identifying 16S ribosomal RNA genes in contigs and chromosomes*. <https://doi.org/10.1101/124131>
- Edwards, R. A., Matlock, B. C., Heffernan, B. J., & Maloy, S. R. (2001). Genomic analysis and growth-phase-dependent regulation of the SEF14 fimbriae of *Salmonella enterica* serovar Enteritidis. *Microbiology*, *147*(Pt 10), 2705–2715. <https://doi.org/10.1099/00221287-147-10-2705>
- El Tahir, Y., & Skurnik, M. (2001). YadA, the multifaceted *Yersinia* adhesin. *International Journal of Medical Microbiology*, *291*(3), 209–218. <https://doi.org/10.1078/1438-4221-00119>
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., ... Finn, R. D. (2018). The Pfam protein families database in 2019. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gky995>
- Eppley, J. M., Tyson, G. W., Getz, W. M., & Banfield, J. F. (2007). Genetic exchange across a species boundary in the archaeal genus *Ferroplasma*. *Genetics*, *177*(1), 407–416. <https://doi.org/10.1534/genetics.107.072892>

- Eren, A. M., Esen, O. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for omics data. *PeerJ*, 3, e1319. <https://doi.org/10.7717/peerj.1319>
- Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., & Sogin, M. L. (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.*, 4(12). <https://doi.org/10.1111/2041-210X.12114>
- Erkus, O., de Jager, V. C. L., Spus, M., van Alen-Boerrigter, I. J., van Rijswijk, I. M. H., Hazelwood, L., ... Smid, E. J. (2013). Multifactorial diversity sustains microbial community stability. *ISME Journal*, 7(11), 2126–2136. <https://doi.org/10.1038/ismej.2013.108>
- Faith, J. J., Colombel, J.-F., & Gordon, J. I. (2015). Identifying strains that contribute to complex diseases through the study of microbial inheritance. *Proceedings of the National Academy of Sciences of the United States of America*, 112(3), 633–640. <https://doi.org/10.1073/pnas.1418781112>
- Fierer, N., Ladau, J., Clemente, J. C., Leff, J. W., Owens, S. M., Pollard, K. S., ... McCulley, R. L. (2013). Reconstructing the Microbial Diversity and Function of Pre-Agricultural Tallgrass Prairie Soils in the United States. *Science*, 342(6158), 621–624. <https://doi.org/10.1126/science.1243768>
- Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39, W29–W37. <https://doi.org/10.1093/nar/gkr367>
- Fischer, H., Yamamoto, M., Akira, S., Beutler, B., & Svanborg, C. (2006). Mechanism of pathogen-specific TLR4 activation in the mucosa: fimbriae, recognition receptors and adaptor protein selection. *Eur. J. Immunol.*, 36(2), 267–277. <https://doi.org/10.1002/eji.200535149>
- Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., & Huttenhower, C. (2015). Sequencing and beyond: integrating molecular “omics” for microbial community profiling. *Nat. Rev. Microbiol.*, 13(6), 360–372. <https://doi.org/10.1038/nrmicro3451>
- Fraser, C., Hanage, W. P., & Spratt, B. G. (2007). Recombination and the nature of bacterial speciation. *Science*, 315(5811), 476–480. <https://doi.org/10.1126/science.1127573>
- Fridkin, S. K., & Jarvis, W. R. (1996). Epidemiology of nosocomial fungal infections. *Clinical Microbiology Reviews*, 9(4), 499-. <https://doi.org/10.1128/CMR.9.4.499>
- Friedman, R., Drake, J. W., & Hughes, A. L. (2004). Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. *Genetics*, 167(3), 1507–1512. <https://doi.org/10.1534/genetics.104.026344>
- Fujimura, K. E., Johnson, C. C., Ownby, D. R., Cox, M. J., Brodie, E. L., Havstad, S. L., ... Lynch, S. V. (2010). Man's best friend? The effect of pet ownership on house dust microbial communities. *Journal of Allergy and Clinical Immunology*, 126(2). <https://doi.org/10.1016/j.jaci.2010.05.042>
- Fujimura, K. E., Sitarik, A. R., Haystad, S., Lin, D. L., Levan, S., Fadrosch, D., ... Lynch, S. V. (2016). Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nature Medicine*, 22(10), 1187–1191. <https://doi.org/10.1038/nm.4176>

- Gaitanis, G., Magiatis, P., Hantschke, M., Bassukas, I. D., & Velegraki, A. (2012). The *Malassezia* Genus in Skin and Systemic Diseases. *Clinical Microbiology Reviews*, 25(1), 106-+. <https://doi.org/10.1128/CMR.00021-11>
- Gao, Y., & Li, H. (2018). Quantifying and comparing bacterial growth dynamics in multiple metagenomic samples. *Nat. Methods*, 15(12), 1041–1044. <https://doi.org/10.1038/s41592-018-0182-0>
- Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., ... Swings, J. (2005). Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.*, 3, 733. <https://doi.org/10.1038/nrmicro1236>
- 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(4), 16024. <https://doi.org/10.1038/NMICROBIOL.2016.24>
- Gilchrist, C. A., Turner, S. D., Riley, M. F., Petri, W. A., & Hewlett, E. L. (2015). Whole-Genome Sequencing in Outbreak Analysis. *Clinical Microbiology Reviews*, 28(3), 541–563. <https://doi.org/10.1128/CMR.00075-13>
- Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., ... Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778), 1355–1359. <https://doi.org/10.1126/science.1124234>
- Giraldo, A., Sutton, D. A., Samerpitak, K., de Hoog, G. S., Wiederhold, N. P., Guarro, J., & Gene, J. (2014). Occurrence of *Ochroconis* and *Verruconis* Species in Clinical Specimens from the United States. *Journal of Clinical Microbiology*, 52(12), 4189–4201. <https://doi.org/10.1128/JCM.02027-14>
- Gordon, S. (2008). Elie Metchnikoff: father of natural immunity. *Eur. J. Immunol.*, 38(12), 3257–3264. <https://doi.org/10.1002/eji.200838855>
- Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007a). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.*, 57(1), 81–91. <https://doi.org/10.1099/ijs.0.64483-0>
- Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007b). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.*, 57(1), 81–91. <https://doi.org/10.1099/ijs.0.64483-0>
- Greenblum, S., Carr, R., & Borenstein, E. (2015). Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species. *Cell*, 160(4), 583–594. <https://doi.org/10.1016/j.cell.2014.12.038>
- Grigoriev, A. (1998). Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Research*, 26(10), 2286–2290. <https://doi.org/10.1093/nar/26.10.2286>
- Grissa, I., Vergnaud, G., & Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, 35, W52–W57. <https://doi.org/10.1093/nar/gkm360>
- Guo, F., & Zhang, T. (2013). Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. *Appl. Microbiol. Biotechnol.*, 97(10), 4607–4616. <https://doi.org/10.1007/s00253-012-4244-4>
- Hagedorn, C., Blanch, A. R., & Harwood, V. J. (Eds.). (2011). *Microbial Source Tracking: Methods, Applications, and Case Studies*.

- Harris, S. R., Feil, E. J., Holden, M. T. G., Quail, M. A., Nickerson, E. K., Chantratita, N., ... Bentley, S. D. (2010). Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science*, 327(5964), 469–474. <https://doi.org/10.1126/science.1182395>
- Harwood, V. J., Staley, C., Badgley, B. D., Borges, K., & Korajkic, A. (2014). Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. *Fems Microbiology Reviews*, 38(1), 1–40. <https://doi.org/10.1111/1574-6976.12031>
- Hewitt, K. M., Mannino, F. L., Gonzalez, A., Chase, J. H., Caporaso, J. G., Knight, R., & Kelley, S. T. (2013). Bacterial Diversity in Two Neonatal Intensive Care Units (NICUs). *Plos One*, 8(1), e54703. <https://doi.org/10.1371/journal.pone.0054703>
- Hirakawa, M. P., Martinez, D. A., Sakthikumar, S., Anderson, M. Z., Berlin, A., Gujja, S., ... Cuomo, C. A. (2015). Genetic and phenotypic intra-species variation in *Candida albicans*. *Genome Research*, 25(3), 413–425. <https://doi.org/10.1101/gr.174623.114>
- Hoffmaster, A. R., Fitzgerald, C. C., Ribot, E., Mayer, L. W., & Popovic, T. (2002). Molecular subtyping of *Bacillus anthracis* and the 2001 bioterrorism-associated anthrax outbreak, United States. *Emerging Infectious Diseases*, 8(10), 1111–1116. <https://doi.org/10.3201/eid0810.020394>
- Horvath, P., & Barrangou, R. (2010). CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science*, 327(5962), 167–170. <https://doi.org/10.1126/science.1179555>
- Horz, H. P., Vianna, M. E., Gomes, B., & Conrads, G. (2005). Evaluation of universal probes and primer sets for assessing total bacterial load in clinical samples: General implications and practical use in endodontic antimicrobial therapy. *Journal of Clinical Microbiology*, 43(10), 5332–5337. <https://doi.org/10.1128/JCM.43.10.5332-5337.2005>
- Hosny, M., Cassir, N., & La Scola, B. (2017). Updating on gut microbiota and its relationship with the occurrence of necrotizing enterocolitis. *Human Microbiome Journal*, 4, 14–19. <https://doi.org/10.1016/j.humic.2016.09.002>
- Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., & Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences of the United States of America*, 111(13), 4904–4909. <https://doi.org/10.1073/pnas.1402564111>
- Hu, Y., Yang, X., Li, J., Lv, N., Liu, F., Wu, J., ... Zhu, B. (2016). The Bacterial Mobile Resistome Transfer Network Connecting the Animal and Human Microbiomes. *Appl. Environ. Microbiol.*, 82(22), 6672–6681. <https://doi.org/10.1128/AEM.01802-16>
- Huang, T., Geng, H., Miyyapuram, V. R., Sit, C. S., Vederas, J. C., & Nakano, M. M. (2009). Isolation of a variant of subtilisin A with hemolytic activity. *J. Bacteriol.*, 191(18), 5690–5696. <https://doi.org/10.1128/JB.00541-09>
- Huang, Y. C., Li, C. C., Lin, T. Y., Lien, R. I., Chou, Y. H., Wu, J. L., & Hsueh, C. (1998). Association of fungal colonization and invasive disease in very low birth weight infants. *Pediatric Infectious Disease Journal*, 17(9), 819–822. <https://doi.org/10.1097/00006454-199809000-00014>
- Huang, Y. C., Lin, T. Y., Lien, R. I., Chou, Y. H., Kuo, C. Y., Yang, P. H., & Hsieh, W. S. (2000). *Candidaemia* in special care nurseries: Comparison of *Albicans* and *Parapsilosis* infection. *Journal of Infection*, 40(2), 171–175. <https://doi.org/10.1053/jinf.2000.0638>

- Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, *26*(5), 680–682. <https://doi.org/10.1093/bioinformatics/btq003>
- Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.*, *33*(6), 1635–1638. <https://doi.org/10.1093/molbev/msw046>
- Huffnagle, G. B., & Noverr, M. C. (2013). The emerging world of the fungal microbiome. *Trends in Microbiology*, *21*(7), 334–341. <https://doi.org/10.1016/j.tim.2013.04.002>
- Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., ... Banfield, J. F. (2016). A new view of the tree of life. *Nat Microbiol*, *1*, 16048. <https://doi.org/10.1038/nmicrobiol.2016.48>
- Hull, C. M., Raisner, R. M., & Johnson, A. D. (2000). Evidence for mating of the “asexual” yeast *Candida albicans* in a mammalian host. *Science*, *289*(5477), 307–310. <https://doi.org/10.1126/science.289.5477.307>
- Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Hyatt, D., Chen, G.-L., 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. <https://doi.org/10.1186/1471-2105-11-119>
- Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, *9*(1), 5114. <https://doi.org/10.1038/s41467-018-07641-9>
- Jaspers, E., & Overmann, J. (2004). Ecological significance of microdiversity: Identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysologies. *Applied and Environmental Microbiology*, *70*(8), 4831–4839. <https://doi.org/10.1128/AEM.70.8.4831-4839.2004>
- Jilling, T., Simon, D., Lu, J., Meng, F. J., Li, D., Schy, R., ... Caplan, M. S. (2006). The roles of bacteria and TLR4 in rat and murine models of necrotizing enterocolitis. *J. Immunol.*, *177*(5), 3273–3282.
- Jolley, K. A., & Maiden, M. C. J. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, *11*, 595. <https://doi.org/10.1186/1471-2105-11-595>
- Jones, E., Oliphant, T., & Peterson, P. (2001). SciPy: Open source scientific tools for Python. URL [Http://Scipy.Org](http://Scipy.Org).
- Jones, T., Federspiel, N. A., Chibana, H., Dungan, J., Kalman, S., Magee, B. B., ... Scherer, S. (2004). The diploid genome sequence of *Candida albicans*. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(19), 7329–7334. <https://doi.org/10.1073/pnas.0401648101>
- Joshi, N. A., & Fass, J. N. (2011). Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files. Available from: *Github. Com/Najoshi/Sickle*.
- Jovel, J., Patterson, J., Wang, W., Hotte, N., O’Keefe, S., Mitchel, T., ... Wong, G. K.-S. (2016). Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in Microbiology*, *7*, 459. <https://doi.org/10.3389/fmicb.2016.00459>

- Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, 42(D1), D199–D205.
- KARL, D. (1979). MEASUREMENT OF MICROBIAL ACTIVITY AND GROWTH IN THE OCEAN BY RATES OF STABLE RIBONUCLEIC-ACID SYNTHESIS. *Applied and Environmental Microbiology*, 38(5), 850–860.
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4), 772–780.
<https://doi.org/10.1093/molbev/mst010>
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647–1649.
<https://doi.org/10.1093/bioinformatics/bts199>
- Khan, H. A., Baig, F. K., & Mehboob, R. (2017). Nosocomial infections: Epidemiology, prevention, control and surveillance. *Asian Pac. J. Trop. Biomed.*, 7(5), 478–482.
<https://doi.org/10.1016/j.apjtb.2017.01.019>
- King, K. C., Brockhurst, M. A., Vasieva, O., Paterson, S., Betts, A., Ford, S. A., ... Hurst, G. D. D. (2016). Rapid evolution of microbe-mediated protection against pathogens in a worm host. *ISME Journal*, 10(8), 1915–1924. <https://doi.org/10.1038/ismej.2015.259>
- Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., ... Kelley, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*, 8(9), 761–U107. <https://doi.org/10.1038/nmeth.1650>
- Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., ... Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17), 2283–2285.
<https://doi.org/10.1093/bioinformatics/btp373>
- Konstantinidis, K. T., & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.*, 102(7), 2567–2572.
<https://doi.org/10.1073/pnas.0409727102>
- Korem, T., Zeevi, D., Suez, J., Weinberger, A., Avnit-Sagi, T., Pompan-Lotan, M., ... Segal, E. (2015a). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science*. <https://doi.org/10.1126/science.aac4812>
- Korem, T., Zeevi, D., Suez, J., Weinberger, A., Avnit-Sagi, T., Pompan-Lotan, M., ... Segal, E. (2015b). MICROBIOME Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science*, 349(6252), 1101–1106.
<https://doi.org/10.1126/science.aac4812>
- Kothavade, R. J., Kura, M. M., Valand, A. G., & Panthaki, M. H. (2010). *Candida tropicalis*: its prevalence, pathogenicity and increasing resistance to fluconazole. *Journal of Medical Microbiology*, 59(8), 873–880. <https://doi.org/10.1099/jmm.0.013227-0>
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., ... Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, 19(9), 1639–1645. <https://doi.org/10.1101/gr.092759.109>
- Lachke, S. A., Lockhart, S. R., Daniels, K. J., & Soll, D. R. (2003). Skin facilitates *Candida albicans* mating. *Infection and Immunity*, 71(9), 4970–4976.
<https://doi.org/10.1128/IAI.71.9.4970-4976.2003>

- Laforest-Lapointe, I., & Arrieta, M.-C. (2018). Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies. *Msystems*, 3(2), e00201-17.
<https://doi.org/10.1128/mSystems.00201-17>
- Lang, G. I., Rice, D. P., Hickman, M. J., Sodergren, E., Weinstock, G. M., Botstein, D., & Desai, M. M. (2013). Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464), 571-+. <https://doi.org/10.1038/nature12344>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357-U54. <https://doi.org/10.1038/NMETH.1923>
- LaTuga, M. S., Ellis, J. C., Cotton, C. M., Goldberg, R. N., Wynn, J. L., Jackson, R. B., & Seed, P. C. (2011). Beyond Bacteria: A Study of the Enteric Microbial Consortium in Extremely Low Birth Weight Infants. *Plos One*, 6(12), e27858.
<https://doi.org/10.1371/journal.pone.0027858>
- Lawrence, G., Bates, J., & Gaul, A. (1982). PATHOGENESIS OF NEONATAL NECROTISING ENTEROCOLITIS. *Lancet*, 319(8264), 137–139.
[https://doi.org/10.1016/S0140-6736\(82\)90383-X](https://doi.org/10.1016/S0140-6736(82)90383-X)
- Lax, S., Sangwan, N., Smith, D., Larsen, P., Handley, K. M., Richardson, M., ... Gilbert, J. A. (2017). Bacterial colonization and succession in a newly opened hospital. *Science Translational Medicine*, 9(391), eaah6500. <https://doi.org/10.1126/scitranslmed.aah6500>
- Leaphart, C. L., Cavallo, J., Gribar, S. C., Cetin, S., Li, J., Branca, M. F., ... Hackam, D. J. (2007). A critical role for TLR4 in the pathogenesis of necrotizing enterocolitis by modulating intestinal injury and repair. *J. Immunol.*, 179(7), 4808–4820.
<https://doi.org/10.4049/jimmunol.179.7.4808>
- Lee, S. T. M., Kahn, S. A., Delmont, T. O., Shaiber, A., Esen, O. C., Hubert, N. A., ... Eren, A. M. (2017). Tracking microbial colonization in fecal microbiota transplantation experiments via genome-resolved metagenomics. *Microbiome*, 5, 50.
<https://doi.org/10.1186/s40168-017-0270-x>
- Letunic, I., & Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1), 127–128.
<https://doi.org/10.1093/bioinformatics/btl529>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079.
<https://doi.org/10.1093/bioinformatics/btp352>
- Liu, B., Faller, L. L., Klitgord, N., Mazumdar, V., Ghodsi, M., Sommer, D. D., ... Amar, S. (2012). Deep Sequencing of the Oral Microbiome Reveals Signatures of Periodontal Disease. *Plos One*, 7(6), e37919. <https://doi.org/10.1371/journal.pone.0037919>
- Liu, L., Chen, X., Skogerbø, G., Zhang, P., Chen, R., He, S., & Huang, D.-W. (2012). The human microbiome: a hot spot of microbial horizontal gene transfer. *Genomics*, 100(5), 265–270. <https://doi.org/10.1016/j.ygeno.2012.07.012>
- Loman, N. J., Constantinidou, C., Christner, M., Rohde, H., Chan, J. Z.-M., Quick, J., ... Pallen, M. J. (2013). A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic *Escherichia coli* O104:H4. *Jama - Journal of the American Medical Association*, 309(14), 1502–1510.
<https://doi.org/10.1001/jama.2013.3231>
- 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. *Appl. Environ. Microbiol.*, *71*(12), 8228–8235. <https://doi.org/10.1128/AEM.71.12.8228-8235.2005>
- Lu, Y. Y., Chen, T., Fuhrman, J. A., & Sun, F. (2017). COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics*, *33*(6), 791–798. <https://doi.org/10.1093/bioinformatics/btw290>
- Luangsa-ard, J., Houbraken, J., van Doorn, T., Hong, S.-B., Borman, A. M., Hywel-Jones, N. L., & Samson, R. A. (2011). *Purpureocillium*, a new genus for the medically important *Paecilomyces lilacinus*. *Fems Microbiology Letters*, *321*(2), 141–149. <https://doi.org/10.1111/j.1574-6968.2011.02322.x>
- Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R. J., & Gevers, D. (2015). ConStrains identifies microbial strains in metagenomic datasets. *Nature Biotechnology*, *33*(10), 1045–+. <https://doi.org/10.1038/nbt.3319>
- Luo, C., Walk, S. T., Gordon, D. M., Feldgarden, M., Tiedje, J. M., & Konstantinidis, K. T. (2011). Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(17), 7200–7205. <https://doi.org/10.1073/pnas.1015622108>
- Majewski, J., & Cohan, F. M. (1999). DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics*, *153*(4), 1525–1533.
- Manzoni, P., Mostert, M., & Castagnola, E. (2015). Update on the management of *Candida* infections in preterm neonates. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, *100*(5), F454–F459. <https://doi.org/10.1136/archdischild-2012-303350>
- Markel, T. A., Crisostomo, P. R., Wairiuko, G. M., Pitcher, J., Tsai, B. M., & Meldrum, D. R. (2006). Cytokines in necrotizing enterocolitis. *Shock*, *25*(4), 329–337. <https://doi.org/10.1097/01.shk.0000192126.33823.87>
- McGhee, G. C., & Sundin, G. W. (2012). *Erwinia amylovora* CRISPR Elements Provide New Tools for Evaluating Strain Diversity and for Microbial Source Tracking. *Plos One*, *7*(7), e41706. <https://doi.org/10.1371/journal.pone.0041706>
- Meheust, R., Burstein, D., Castelle, C. J., & Banfield, J. F. (2018). *Biological capacities clearly define a major subdivision in Domain Bacteria*. <https://doi.org/10.1101/335083>
- Mesquita-Rocha, S., Godoy-Martinez, P. C., Goncalves, S. S., Daniel Urrutia, M., Carlesse, F., Seber, A., ... Colombo, A. L. (2013). The water supply system as a potential source of fungal infection in paediatric haematopoietic stem cell units. *Bmc Infectious Diseases*, *13*, 289. <https://doi.org/10.1186/1471-2334-13-289>
- Miller, C. S., Baker, B. J., Thomas, B. C., Singer, S. W., & Banfield, J. F. (2011). EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol.*, *12*(5), R44. <https://doi.org/10.1186/gb-2011-12-5-r44>
- Miyamoto, K., Fisher, D. J., Li, J., Sayeed, S., Akimoto, S., & McClane, B. A. (2006). Complete sequencing and diversity analysis of the enterotoxin-encoding plasmids in *Clostridium perfringens* type A non-food-borne human gastrointestinal disease isolates. *J. Bacteriol.*, *188*(4), 1585–1598. <https://doi.org/10.1128/JB.188.4.1585-1598.2006>
- Morowitz, M. J., Deneff, V. J., Costello, E. K., Thomas, B. C., Poroyko, V., Relman, D. A., & Banfield, J. F. (2011). Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proc. Natl. Acad. Sci. U. S. A.*, *108*(3), 1128–1133. <https://doi.org/10.1073/pnas.1010992108>

- Morowitz, M. J., Poroyko, V., Caplan, M., Alverdy, J., & Liu, D. C. (2010). Redefining the role of intestinal microbes in the pathogenesis of necrotizing enterocolitis. *Pediatrics*, *125*(4), 777–785. <https://doi.org/10.1542/peds.2009-3149>
- Morrow, A. L., Lagomarcino, A. J., Schibler, K. R., Taft, D. H., Yu, Z., Wang, B., ... others. (2013). Early microbial and metabolomic signatures predict later onset of necrotizing enterocolitis in preterm infants. *Microbiome*, *1*(1), 1.
- Mueller, N. T., Bakacs, E., Combellick, J., Grigoryan, Z., & Dominguez-Bello, M. G. (2015). The infant microbiome development: mom matters. *Trends in Molecular Medicine*, *21*(2), 109–117. <https://doi.org/10.1016/j.molmed.2014.12.002>
- Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935. <https://doi.org/10.1093/bioinformatics/btt509>
- Nayfach, S., Rodriguez-Mueller, B., Garud, N., & Pollard, K. S. (2016). An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.*, *26*. <https://doi.org/10.1101/gr.201863.115>
- Nayfach, Stephen, Shi, Z. J., Seshadri, R., Pollard, K. S., & Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature*. <https://doi.org/10.1038/s41586-019-1058-x>
- Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, *3*(5), 418–426. <https://doi.org/10.1093/oxfordjournals.molbev.a040410>
- Neu, J., & Walker, W. A. (2011). Necrotizing enterocolitis. *N. Engl. J. Med.*, *364*(3), 255–264. <https://doi.org/10.1056/NEJMra1005408>
- Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, *32*(1), 268–274. <https://doi.org/10.1093/molbev/msu300>
- Nocker, A., Cheung, C.-Y., & Camper, A. K. (2006). Comparison of propidium monoazide with ethidium monoazide for differentiation of live vs. dead bacteria by selective removal of DNA from dead cells. *Journal of Microbiological Methods*, *67*(2), 310–320. <https://doi.org/10.1016/j.mimet.2006.04.015>
- Novichkov, V., Kaznadzey, A., Alexandrova, N., & Kaznadzey, D. (2016). NSimScan: DNA comparison tool with increased speed, sensitivity and accuracy. *Bioinformatics*, *32*(15), 2380–2381. <https://doi.org/10.1093/bioinformatics/btw126>
- Nuccio, S.-P., & Bäumlner, A. J. (2007). Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek. *Microbiol. Mol. Biol. Rev.*, *71*(4), 551–575. <https://doi.org/10.1128/MMBR.00014-07>
- Oberauner, L., Zachow, C., Lackner, S., Hoegenauer, C., Smolle, K.-H., & Berg, G. (2013). The ignored diversity: complex bacterial communities in intensive care units revealed by 16S pyrosequencing. *Scientific Reports*, *3*, 1413. <https://doi.org/10.1038/srep01413>
- Obladen, M. (2009). Necrotizing enterocolitis--150 years of fruitless search for the cause. *Neonatology*, *96*(4), 203–210.
- Ochman, H., Elwyn, S., & Moran, N. A. (1999). Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(22), 12638–12643. <https://doi.org/10.1073/pnas.96.22.12638>
- Oh, J., Byrd, A. L., Deming, C., Conlan, S., Kong, H. H., & Segre, J. A. (2014). Biogeography and individuality shape function in the human skin metagenome. *Nature*, *514*(7520), 59–+. <https://doi.org/10.1038/nature13786>

- Olive, D. M., & Bean, P. (1999). Principles and applications of methods for DNA-based typing of microbial organisms. *Journal of Clinical Microbiology*, 37(6), 1661–1669.
- Olm, M. R., Bhattacharya, N., Crits-Christoph, A., Firek, B. A., Baker, R., Song, Y. S., ... Banfield, J. F. (2019). *Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria that may stimulate TLR4 receptors*. <https://doi.org/10.1101/558676>
- Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*, 11(12), 2864–2868. <https://doi.org/10.1038/ismej.2017.126>
- Olm, M. R., Brown, C. T., Brooks, B., Firek, B., Baker, R., Burstein, D., ... Banfield, J. F. (2017). Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. *Genome Research*, 27(4), 601–612. <https://doi.org/10.1101/gr.213256.116>
- Olm, M. R., Butterfield, C. N., Copeland, A., Boles, T. C., Thomas, B. C., & Banfield, J. F. (2017). The Source and Evolutionary History of a Microbial Contaminant Identified Through Soil Metagenomic Analysis. *Mbio*, 8(1), e01969-16. <https://doi.org/10.1128/mBio.01969-16>
- Olm, M. R., West, P. T., Brooks, B., Firek, B. A., Baker, R., Morowitz, M. J., & Banfield, J. F. (2018). *Strain-level overlap between infant and hospital fungal microbiomes revealed through de novo assembly of eukaryotic genomes from metagenomes*. <https://doi.org/10.1101/324566>
- 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, 132. <https://doi.org/10.1186/s13059-016-0997-x>
- Ott, S. J., Kuehbach, T., Musfeldt, M., Rosenstiel, P., Hellmig, S., Rehman, A., ... Schreiber, S. (2008). Fungi and inflammatory bowel diseases: Alterations of composition and diversity. *Scandinavian Journal of Gastroenterology*, 43(7), 831–841. <https://doi.org/10.1080/00365520801935434>
- Palmer, K. L., & Gilmore, M. S. (2010). Multidrug-Resistant Enterococci Lack CRISPR-cas. *Mbio*, 1(4), e00227-10. <https://doi.org/10.1128/mBio.00227-10>
- Pammi, M., Cope, J., Tarr, P. I., Warner, B. B., Morrow, A. L., Mai, V., ... Neu, J. (2017). Intestinal dysbiosis in preterm infants preceding necrotizing enterocolitis: a systematic review and meta-analysis. *Microbiome*, 5(1). <https://doi.org/10.1186/s40168-017-0248-8>
- Parfrey, L. W., Walters, W. A., & Knight, R. (2011). Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Frontiers in Microbiology*, 2, 153. <https://doi.org/10.3389/fmicb.2011.00153>
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, 25(7), 1043–1055. <https://doi.org/10.1101/gr.186072.114>
- Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., ... Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2(11), 1533–1542. <https://doi.org/10.1038/s41564-017-0012-7>
- Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., ... Segata, N. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000

- Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, 176(3), 649-662.e20. <https://doi.org/10.1016/j.cell.2019.01.001>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12(Oct), 2825–2830.
- Penders, J., Thijs, C., Vink, C., Stelma, F. F., Snijders, B., Kummeling, I., ... Stobberingh, E. E. (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>
- 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>
- Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergstrom, A., ... Schacherer, J. (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*, 556(7701), 339-+. <https://doi.org/10.1038/s41586-018-0030-5>
- Pfaller, M. A. (1996). Nosocomial candidiasis: Emerging species, reservoirs, and modes of transmission. *Clinical Infectious Diseases*, 22, S89–S94. https://doi.org/10.1093/clinids/22.Supplement_2.S89
- Porteous, N. B., Grooters, A. M., Redding, S. W., Thompson, E. H., Rinaldi, M. G., De Hoog, G. S., & Sutton, D. A. (2003). Identification of *Exophiala mesophila* isolated from treated dental unit waterlines. *Journal of Clinical Microbiology*, 41(8), 3885–3889. <https://doi.org/10.1128/JCM.41.8.3885-3889.2003>
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.*, 26. <https://doi.org/10.1093/molbev/msp077>
- Pride, D. T., Sun, C. L., Salzman, J., Rao, N., Loomer, P., Armitage, G. C., ... Relman, D. A. (2011). Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. *Genome Research*, 21(1), 126–136. <https://doi.org/10.1101/gr.111732.110>
- Probst, A. J., Castelle, C. J., Singh, A., Brown, C. T., Anantharaman, K., Sharon, I., ... Banfield, J. F. (2017). Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO₂ concentrations. *Environmental Microbiology*, 19(2), 459–474. <https://doi.org/10.1111/1462-2920.13362>
- Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., ... Lennon, J. J. (2007). The role of ecological theory in microbial ecology. *Nat. Rev. Microbiol.*, 5(5), 384–392.
- Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., ... Young, J. P. W. (2007). Essay - The role of ecological theory in microbial ecology. *Nature Reviews Microbiology*, 5(5), 384–392. <https://doi.org/10.1038/nrmicro1643>
- Pruitt, K. D., & Maglott, D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 29(1), 137–140. <https://doi.org/10.1093/nar/29.1.137>
- Raes, J., Korb, J. O., Lercher, M. J., von Mering, C., & Bork, P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biology*, 8(1), R10. <https://doi.org/10.1186/gb-2007-8-1-r10>
- 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>
- Rampelli, S., Soverini, M., Turrone, S., Quercia, S., Biagi, E., Brigidi, P., & Candela, M. (2016). ViromeScan: a new tool for metagenomic viral community profiling. *BMC Genomics*, 17(1), 165. <https://doi.org/10.1186/s12864-016-2446-3>
- Rasko, D. A., Worsham, P. L., Abshire, T. G., Stanley, S. T., Bannan, J. D., Wilson, M. R., ... Ravel, J. (2011). Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation. *Proceedings of the National Academy of Sciences of the United States of America*, 108(12), 5027–5032. <https://doi.org/10.1073/pnas.1016657108>
- Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, 7(3), 355–364.
<https://doi.org/10.1111/j.1471-8286.2006.01678.x>
- 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. *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*, 4, e05477.
<https://doi.org/10.7554/eLife.05477>
- Reidl, J., & Mekalanos, J. J. (1995). Characterization of Vibrio cholerae bacteriophage K139 and use of a novel mini-transposon to identify a phage-encoded virulence factor. *Mol. Microbiol.*, 18(4), 685–701.
- Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, 5(1). <https://doi.org/10.1186/s40168-017-0283-5>
- Richter, M., & Rossello-Mora, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45), 19126–19131. <https://doi.org/10.1073/pnas.0906412106>
- Rissman, A. I., Mau, B., Biehl, B. S., Darling, A. E., Glasner, J. D., & Perna, N. T. (2009). Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics*, 25(16), 2071–2073. <https://doi.org/10.1093/bioinformatics/btp356>
- Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H., & Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.*, 239(2), 226–235. <https://doi.org/10.1016/j.jtbi.2005.08.037>
- RODRIGUEZ, G., PHIPPS, D., ISHIGURO, K., & RIDGWAY, H. (1992). USE OF A FLUORESCENT REDOX PROBE FOR DIRECT VISUALIZATION OF ACTIVELY RESPIRING BACTERIA. *Applied and Environmental Microbiology*, 58(6), 1801–1808.
- Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., ... Rohwer, F. (2010). Viral and microbial community dynamics in four aquatic environments. *ISME Journal*, 4(6), 739–751. <https://doi.org/10.1038/ismej.2010.1>
- Ronda, C., Chen, S. P., Cabral, V., Yaung, S. J., & Wang, H. H. (2019). Metagenomic engineering of the mammalian gut microbiome in situ. *Nat. Methods*.
<https://doi.org/10.1038/s41592-018-0301-y>

- Rosen, M. J., Davison, M., Bhaya, D., & Fisher, D. S. (2015). Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science*, 348(6238), 1019–1023. <https://doi.org/10.1126/science.aaa4456>
- Roux, S., Enault, F., Hurwitz, B. L., & Sullivan, M. B. (2015). VirSorter: mining viral signal from microbial genomic data. *PeerJ*, 3, e985. <https://doi.org/10.7717/peerj.985>
- Ruan, S.-Y., Chien, J.-Y., & Hsueh, P.-R. (2009). Invasive Trichosporonosis Caused by *Trichosporon asahii* and Other Unusual *Trichosporon* Species at a Medical Center in Taiwan. *Clinical Infectious Diseases*, 49(1), E11–E17. <https://doi.org/10.1086/599614>
- Rutherford, S. T., & Bassler, B. L. (2012). Bacterial quorum sensing: its role in virulence and possibilities for its control. *Cold Spring Harb. Perspect. Med.*, 2(11). <https://doi.org/10.1101/cshperspect.a012427>
- Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., ... Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *Bmc Biology*, 12, 87. <https://doi.org/10.1186/s12915-014-0087-z>
- Salzberg, S. L., & Yorke, J. A. (2005). Beware of mis-assembled genomes. *Bioinformatics*, 21(24), 4320–4321. <https://doi.org/10.1093/bioinformatics/bti769>
- SANCHEZ, V., VAZQUEZ, J., BARTHJONES, D., DEMBRY, L., SOBEL, J., & ZERVOS, M. (1992). EPIDEMIOLOGY OF NOSOCOMIAL ACQUISITION OF CANDIDA-LUSITANIAE. *Journal of Clinical Microbiology*, 30(11), 3005–3008.
- Schleheck, D., Knepper, T. P., Fischer, K., & Cook, A. M. (2004). Mineralization of individual congeners of linear alkylbenzenesulfonate by defined pairs of heterotrophic bacteria. *Applied and Environmental Microbiology*, 70(7), 4053–4063. <https://doi.org/10.1128/AEM.70.7.4053-4063.2004>
- Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., ... Bork, P. (2013). Genomic variation landscape of the human gut microbiome. *Nature*, 493(7430), 45–50. <https://doi.org/10.1038/nature11711>
- Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, 75(23), 7537–7541. <https://doi.org/10.1128/AEM.01541-09>
- Schulze, J., & Sonnenborn, U. (2009). Yeasts in the Gut: From Commensals to Infectious Agents. *Deutsches Arzteblatt International*, 106(51–52), 837–841. <https://doi.org/10.3238/arztebl.2009.0837>
- Sedlar, K., Kupkova, K., & Provaznik, I. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotechnol. J.*, 15, 48–55. <https://doi.org/10.1016/j.csbj.2016.11.005>
- Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods*, 9(8), 811–814. <https://doi.org/10.1038/nmeth.2066>
- Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., ... Alm, E. J. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336(6077), 48–51. <https://doi.org/10.1126/science.1218198>
- Shapiro, B. J., & Polz, M. F. (2015). Microbial Speciation. *Cold Spring Harb. Perspect. Biol.*, 7(10), a018143. <https://doi.org/10.1101/cshperspect.a018143>

- Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., & Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research*, 23(1), 111–120. <https://doi.org/10.1101/gr.142315.112>
- Shin, H., Pei, Z., Martinez, K. A., Rivera-Vinas, J. I., Mendez, K., Cavallin, H., & Dominguez-Bello, M. G. (2015). The first microbial environment of infants born by C-section: the operating room microbes. *Microbiome*, 3, UNSP 59. <https://doi.org/10.1186/s40168-015-0126-1>
- Shivaprasad, A., Ravi, G. C., Shivapriya, & Rama. (2013). A Rare Case of Nasal Septal Perforation Due to *Purpureocillium lilacinum*: Case Report and Review. *Indian Journal of Otolaryngology and Head & Neck Surgery*, 65(2), 184–188. <https://doi.org/10.1007/s12070-012-0570-1>
- Shulman, S. T., Friedmann, H. C., & Sims, R. H. (2007). Theodor Escherich: the first pediatric infectious diseases physician? *Clin. Infect. Dis.*, 45(8), 1025–1029. <https://doi.org/10.1086/521946>
- Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, 3(7), 836–843. <https://doi.org/10.1038/s41564-018-0171-1>
- 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–F364. <https://doi.org/10.1136/archdischild-2012-302872>
- Simao, 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>
- Skewes-Cox, P., Sharpton, T. J., Pollard, K. S., & DeRisi, J. L. (2014). Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS One*, 9(8), e105067. <https://doi.org/10.1371/journal.pone.0105067>
- Smillie, Chris S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A., & Alm, E. J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, 480(7376), 241–244. <https://doi.org/10.1038/nature10571>
- Smillie, Christopher S, Sauk, J., Gevers, D., Friedman, J., Sung, J., Youngster, I., ... Alm, E. J. (2018). Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. *Cell Host Microbe*, 23(2), 229-240.e5. <https://doi.org/10.1016/j.chom.2018.01.003>
- Snitkin, E. S., Zelazny, A. M., Thomas, P. J., Stock, F., Henderson, D. K., Palmore, T. N., & Segre, J. A. (2012). Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae* with Whole-Genome Sequencing. *Science Translational Medicine*, 4(148), 148ra116. <https://doi.org/10.1126/scitranslmed.3004129>
- Spellerberg, I. F., & Fedor, P. J. (2003). A tribute to Claude Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the “Shannon-Wiener” Index. *Global Ecology and Biogeography*, 12(3), 177–179. <https://doi.org/10.1046/j.1466-822X.2003.00015.x>

- Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21), 2688–2690. <https://doi.org/10.1093/bioinformatics/btl446>
- Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, 34, W435–W439. <https://doi.org/10.1093/nar/gkl200>
- Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, 35(11), 1026–1028. <https://doi.org/10.1038/nbt.3988>
- Stewart, C. J., Marrs, E. C. L., Magorrian, S., Nelson, A., Lanyon, C., Perry, J. D., ... Berrington, J. E. (2012). The preterm gut microbiota: changes associated with necrotizing enterocolitis and infection. *Acta Paediatrica*, 101(11), 1121–1127. <https://doi.org/10.1111/j.1651-2227.2012.02801.x>
- Stewart, Christopher James, Nelson, A., Scribbins, D., Marrs, E. C. L., Lanyon, C., Perry, J. D., ... Berrington, J. E. (2013). Bacterial and fungal viability in the preterm gut: NEC and sepsis. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 98(4), F298–F303. <https://doi.org/10.1136/archdischild-2012-302119>
- Sturbelle, R. T., de Avila, L. F. da C., Roos, T. B., Borchardt, J. L., da Conceição, R. de C. dos S., Dellagostin, O. A., & Leite, F. P. L. (2015). The role of quorum sensing in Escherichia coli (ETEC) virulence factors. *Vet. Microbiol.*, 180(3–4), 245–252. <https://doi.org/10.1016/j.vetmic.2015.08.015>
- Sun, C. L., Thomas, B. C., Barrangou, R., & Banfield, J. F. (2016). Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. *ISME Journal*, 10(4), 858–870. <https://doi.org/10.1038/ismej.2015.162>
- SURAWICZ, C., ELMER, G., SPEELMAN, P., MCFARLAND, L., CHINN, J., & VANBELLE, G. (1989). PREVENTION OF ANTIBIOTIC-ASSOCIATED DIARRHEA BY SACCHAROMYCES-BOULARDII - A PROSPECTIVE-STUDY. *Gastroenterology*, 96(4), 981–988. [https://doi.org/10.1016/0016-5085\(89\)91613-2](https://doi.org/10.1016/0016-5085(89)91613-2)
- Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10), 1282–1288. <https://doi.org/10.1093/bioinformatics/btm098>
- Tamburini, S., Shen, N., Wu, H. C., & Clemente, J. C. (2016). The microbiome in early life: implications for health outcomes. *Nature Medicine*, 22(7), 713–722. <https://doi.org/10.1038/nm.4142>
- Tenaillon, O., Skurnik, D., Picard, B., & Denamur, E. (2010). The population genetics of commensal Escherichia coli. *Nature Reviews Microbiology*, 8(3), 207–217. <https://doi.org/10.1038/nrmicro2298>
- Touchon, M., Charpentier, S., Clermont, O., Rocha, E. P. C., Denamur, E., & Branger, C. (2011). CRISPR Distribution within the Escherichia coli Species Is Not Suggestive of Immunity-Associated Diversifying Selection. *Journal of Bacteriology*, 193(10), 2460–2467. <https://doi.org/10.1128/JB.01307-10>
- Touchon, M., & Rocha, E. P. C. (2010). The Small, Slow and Specialized CRISPR and Anti-CRISPR of Escherichia and Salmonella. *Plos One*, 5(6), e11126. <https://doi.org/10.1371/journal.pone.0011126>

- Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., & Pasolli, E. (2015). MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, *12*. <https://doi.org/10.1038/nmeth.3589>
- Tsai, Y.-C., Conlan, S., Deming, C., Segre, J. A., Kong, H. H., Korfach, J., & Oh, J. (2016). Resolving the Complexity of Human Skin Metagenomes Using Single-Molecule Sequencing. *Mbio*, *7*(1), e01948-15. <https://doi.org/10.1128/mBio.01948-15>
- Tu, Q., He, Z., & Zhou, J. (2014). Strain/species identification in metagenomes using genome-specific markers. *Nucleic Acids Research*, *42*(8), e67. <https://doi.org/10.1093/nar/gku138>
- Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data*, *5*, 170203. <https://doi.org/10.1038/sdata.2017.203>
- Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., ... Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43. <https://doi.org/10.1038/nature02340>
- Tyson, Gene W., & Banfield, J. F. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environmental Microbiology*, *10*(1), 200–207. <https://doi.org/10.1111/j.1462-2920.2007.01444.x>
- Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., & Pati, A. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Research*, *43*(14), 6761–6771. <https://doi.org/10.1093/nar/gkv657>
- VAZQUEZ, J., SANCHEZ, V., DMUCHOWSKI, C., DEMBRY, L., SOBEL, J., & ZERVOS, M. (1993). NOSOCOMIAL ACQUISITION OF CANDIDA-ALBICANS - AN EPIDEMIOLOGIC-STUDY. *Journal of Infectious Diseases*, *168*(1), 195–201. <https://doi.org/10.1093/infdis/168.1.195>
- Vineis, J. H., Ringus, D. L., Morrison, H. G., Delmont, T. O., Dalal, S., Raffals, L. H., ... Sogin, M. L. (2016). Patient-Specific Bacteroides Genome Variants in Pouchitis. *MBio*, *7*(6), e01713-16. <https://doi.org/10.1128/mBio.01713-16>
- Vulić, M., Dionisio, F., Taddei, F., & Radman, M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc. Natl. Acad. Sci. U. S. A.*, *94*(18), 9763–9767.
- Ward, D. V., Scholz, M., Zolfo, M., Taft, D. H., Schibler, K. R., Tett, A., ... Morrow, A. L. (2016). Metagenomic Sequencing with Strain-Level Resolution Implicates Uropathogenic E. coli in Necrotizing Enterocolitis and Mortality in Preterm Infants. *Cell Rep.* <https://doi.org/10.1016/j.celrep.2016.03.015>
- Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Brucoleri, R., ... Medema, M. H. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.*, *43*(W1), W237–W243. <https://doi.org/10.1093/nar/gkv437>
- West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C., & Banfield, J. F. (2018). Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Research*, *28*(4), 569–580. <https://doi.org/10.1101/gr.228429.117>
- Wildschutte, H., Wolfe, D. M., Tamewitz, A., & Lawrence, J. G. (2004). Protozoan predation, diversifying selection, and the evolution of antigenic diversity in Salmonella. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(29), 10644–10649. <https://doi.org/10.1073/pnas.04040284101>

- Wiles, T. J., Kulesus, R. R., & Mulvey, M. A. (2008). Origins and virulence mechanisms of uropathogenic *Escherichia coli*. *Exp. Mol. Pathol.*, *85*(1), 11–19.
<https://doi.org/10.1016/j.yexmp.2008.03.007>
- Williamson, K. E., Radosevich, M., & Wommack, K. E. (2005). Abundance and diversity of viruses in six Delaware soils. *Applied and Environmental Microbiology*, *71*(6), 3119–3125. <https://doi.org/10.1128/AEM.71.6.3119-3125.2005>
- Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., ... Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, *40*(22), 11189–11201. <https://doi.org/10.1093/nar/gks918>
- Wilmes, P., Simmons, S. L., Deneff, V. J., & Banfield, J. F. (2009). The dynamic genetic repertoire of microbial communities. *FEMS Microbiol. Rev.*, *33*(1), 109–132.
<https://doi.org/10.1111/j.1574-6976.2008.00144.x>
- Wilson, M. R., O'Donovan, B. D., Gelfand, J. M., Sample, H. A., Chow, F. C., Betjemann, J. P., ... DeRisi, J. L. (2018). Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing. *Jama Neurology*, *75*(8), 947–955.
<https://doi.org/10.1001/jamaneurol.2018.0463>
- Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607. <https://doi.org/10.1093/bioinformatics/btv638>
- Zaharia, M., Bolosky, W. J., Curtis, K., Fox, A., Patterson, D., Shenker, S., ... Sittler, T. (2011). Faster and more accurate sequence alignment with SNAP. *ArXiv Preprint ArXiv:1111.5572*.
- Zhang, N., O'Donnell, K., Sutton, D. A., Nalim, F. A., Summerbell, R. C., Padhye, A. A., & Geiser, D. M. (2006). Members of the *Fusarium solani* species complex that cause infections in both humans and plants are common in the environment. *Journal of Clinical Microbiology*, *44*(6), 2186–2190. <https://doi.org/10.1128/JCM.00120-06>
- Zhang, S., Lebreton, F., Mansfield, M. J., Miyashita, S.-I., Zhang, J., Schwartzman, J. A., ... Dong, M. (2018). Identification of a Botulinum Neurotoxin-like Toxin in a Commensal Strain of *Enterococcus faecium*. *Cell Host Microbe*, *23*(2), 169-176.e6.
<https://doi.org/10.1016/j.chom.2017.12.018>

Appendix 1: Select co-authored publications

Appendix 1.1 | Measurement of bacterial replication rates in microbial communities

ANALYSIS

Measurement of bacterial replication rates in microbial communities

Christopher T Brown¹, Matthew R Olm¹, Brian C Thomas² & Jillian F Banfield²⁻⁴

Culture-independent microbiome studies have increased our understanding of the complexity and metabolic potential of microbial communities. However, to understand the contribution of individual microbiome members to community functions, it is important to determine which bacteria are actively replicating. We developed an algorithm, iRep, that uses draft-quality genome sequences and single time-point metagenome sequencing to infer microbial population replication rates. The algorithm calculates an index of replication (iRep) based on the sequencing coverage trend that results from bi-directional genome replication from a single origin of replication. We apply this method to show that microbial replication rates increase after antibiotic administration in human infants. We also show that uncultivated, groundwater-associated, Candidate Phyla Radiation bacteria only rarely replicate quickly in subsurface communities undergoing substantial changes in geochemistry. Our method can be applied to any genome-resolved microbiome study to track organism responses to varying conditions, identify actively growing populations and measure replication rates for use in modeling studies.

Dividing cells in a natural population contain, on average, more than one copy of their genome (Fig. 1). In an unsynchronized population of growing bacteria, cells contain genomes that are replicated to different extents, resulting in a gradual reduction in the average genome copy number from the origin to the terminus of replication¹. This decrease can be detected by measuring changes in DNA sequencing coverage across complete genomes². Bacterial genome replication proceeds bi-directionally from a single origin of replication^{3,4}, therefore the origin and terminus of replication can be deduced based on this coverage pattern². GC skew⁵⁻⁷ and genome coverage⁸ analyses of a wide variety of bacteria have shown that this replication mechanism is broadly applicable. Further, early studies of bacterial cultures revealed that cells can achieve faster division by simultaneously initiating multiple

rounds of genome replication⁹, which results in an average of more than two genome copies in rapidly growing cells.

Korem *et al.*⁸ used the ratio of sequencing coverage at the origin compared to the terminus of replication to measure replication rates for bacteria. Because the origin and terminus correspond to coverage peaks and troughs, respectively, the authors named their method PTR (peak-to-trough ratio). They applied PTR to calculate replication rates for specific bacteria in the human microbiome, but the requirement for mapping sequencing reads to a complete, closed, circular reference genome for a bacterium of interest is a major limitation. The vast majority of bacteria remain uncultivated and lack reference genomes.

Metagenomics methods routinely generate draft genomes for bacteria and archaea that lack reference genomes¹⁰⁻²³ (Fig. 1 and Supplementary Fig. 1). Often these organisms are from little known microbial phyla, and are vastly different from organisms for which there are complete genomes in databases^{15-17,24-27}. It is sometimes possible to recover hundreds or thousands of draft or near-complete genomes from a single ecosystem. We introduce a method that can extend coverage-based replication rate analyses to enable measurements based on sequencing coverage trends for these draft genomes. The method works, even though the order of the fragments is unknown. Unlike PTR, our approach can be applied in virtually any natural or engineered ecosystem, including complex systems such as soil, for which complete genomes for the vast majority of bacteria are unavailable.

RESULTS

The iRep metric

The method that we developed determines replication rates based on measuring the rate of the decrease in average sequence coverage from the origin to the terminus of replication. This rate of coverage change can be used to accurately estimate the ratio between the coverage at the origin and terminus of replication, which is proportional to replication rate. The values are comparable to PTR, but are derived differently so we named this method and metric iRep (Index of Replication). With PTR, the origin and terminus of replication must be identified and the calculation requires position-specific coverage values. In contrast, the iRep algorithm is distinct in that it makes use of the total change in coverage across all genome fragments.

iRep values are calculated by mapping metagenome sequencing reads to the collection of assembled sequences that represent a draft genome (Fig. 1, Supplementary Fig. 1, Online Methods and Supplementary Code). The read coverage is evaluated at every nucleotide position across every scaffold. The series of coverage values for the scaffolds are then concatenated, and the average coverage values within 5-Kbp sliding

© 2016 Nature America, Inc., part of Springer Nature. All rights reserved.

¹Department of Plant and Microbial Biology, University of California, Berkeley, California, USA. ²Department of Earth and Planetary Science, University of California, Berkeley, California, USA. ³Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, USA. ⁴Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. Correspondence should be addressed to J.F.B. (jbanfield@berkeley.edu).

Received 11 March; accepted 20 September; published online 7 November 2016; doi:10.1038/nbt.3704

Appendix 1.2 | The developing premature infant gut microbiome is a major factor shaping the microbiome of neonatal intensive care unit rooms

Brooks et al. *Microbiome* (2018) 6:112
<https://doi.org/10.1186/s40168-018-0493-5>


Microbiome

RESEARCH

Open Access



The developing premature infant gut microbiome is a major factor shaping the microbiome of neonatal intensive care unit rooms

Brandon Brooks¹ , Matthew R. Olm¹, Brian A. Firek², Robyn Baker³, David Geller-McGrath⁴, Sophia R. Reimer⁴, Karina R. Soenjoyo⁴, Jennifer S. Yip⁴, Dylan Dahan^{5,6}, Brian C. Thomas⁴, Michael J. Morowitz² and Jillian F. Banfield^{4*}

Abstract

Background: The neonatal intensive care unit (NICU) contains a unique cohort of patients with underdeveloped immune systems and nascent microbiome communities. Patients often spend several months in the same room, and it has been previously shown that the gut microbiomes of these infants often resemble the microbes found in the NICU. Little is known, however, about the identity, persistence, and absolute abundance of NICU room-associated bacteria over long stretches of time. Here, we couple droplet digital PCR (ddPCR), 16S rRNA gene surveys, and recently published metagenomics data from infant gut samples to infer the extent to which the NICU microbiome is shaped by its room occupants.

Results: Over 2832 swabs, wipes, and air samples were collected from 16 private-style NICU rooms housing very low birth weight (< 1500 g), premature (< 31 weeks' gestation) infants. For each infant, room samples were collected daily, Monday through Friday, for 1 month. The first samples from the first infant and the last samples from the last infant were collected 383 days apart. Twenty-two NICU locations spanning room surfaces, hands, electronics, sink basins, and air were collected. Results point to an incredibly simple room community where 5–10 taxa, mostly skin-associated, account for over 50% of the amplicon reads. Biomass estimates reveal four to five orders of magnitude difference between the least to the most dense microbial communities, air, and sink basins, respectively. Biomass trends from bioaerosol samples and petri dish dust collectors suggest occupancy to be a main driver of suspended biological particles within the NICU. Using a machine learning algorithm to classify the origin of room samples, we show that each room has a unique microbial fingerprint. Several important taxa driving this model were dominant gut colonizers of infants housed within each room.

Conclusions: Despite regular cleaning of hospital surfaces, bacterial biomass was detectable at varying densities. A room-specific microbiome signature was detected, suggesting microbes seeding NICU surfaces are sourced from reservoirs within the room and that these reservoirs contain actively dividing cells. Collectively, the data suggests that hospitalized infants, in combination with their caregivers, shape the microbiome of NICU rooms.

Keywords: Infant gut, Microbiome, Built environment, Neonatal intensive care unit

* Correspondence: jbanfield@berkeley.edu

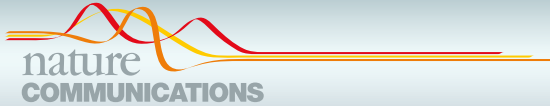
⁴Department of Earth and Planetary Sciences, University of California, Berkeley, CA, USA

Full list of author information is available at the end of the article



© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Appendix 1.3 | Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome



ARTICLE

DOI: 10.1038/s41467-017-02018-w

OPEN

Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome

Brandon Brooks¹, Matthew R. Olm¹, Brian A. Firek², Robyn Baker³, Brian C. Thomas⁴, Michael J. Morowitz² & Jillian F. Banfield⁴

Preterm infants exhibit different microbiome colonization patterns relative to full-term infants, and it is speculated that the hospital room environment may contribute to infant microbiome development. Here, we present a genome-resolved metagenomic study of microbial genotypes from the gastrointestinal tracts of infants and from the neonatal intensive care unit (NICU) room environment. Some strains detected in hospitalized infants also occur in sinks and on surfaces, and belong to species such as *Staphylococcus epidermidis*, *Enterococcus faecalis*, *Pseudomonas aeruginosa*, and *Klebsiella pneumoniae*, which are frequently implicated in nosocomial infection and preterm infant gut colonization. Of the 15 *K. pneumoniae* strains detected in the study, four were detected in both infant gut and room samples. Time series experiments showed that nearly all strains associated with infant gut colonization can be detected in the room after, and often before, detection in the gut. Thus, we conclude that a component of premature infant gut colonization is the cycle of microbial exchange between the room and the occupant.

¹Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA. ²Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. ³Division of Newborn Medicine, Magee-Womens Hospital of Pittsburgh of UPMC, Pittsburgh, PA 15224, USA. ⁴Department of Earth and Planetary Sciences, and Environmental Science, Policy and Management, University of California, Berkeley, CA 94720, USA. Correspondence and requests for materials should be addressed to J.F.B. (email: jbanfield@berkeley.edu)

Appendix 1.4 | Machine Learning Leveraging Genomes from Metagenomes Identifies Influential Antibiotic Resistance Genes in the Infant Gut Microbiome



RESEARCH ARTICLE
Host-Microbe Biology



Machine Learning Leveraging Genomes from Metagenomes Identifies Influential Antibiotic Resistance Genes in the Infant Gut Microbiome

Sumayah F. Rahman,^a Matthew R. Olm,^a Michael J. Morowitz,^b Jillian F. Banfield^c

^aDepartment of Plant and Microbial Biology, University of California, Berkeley, Berkeley, California, USA

^bDepartment of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA

^cDepartment of Earth and Planetary Sciences and Environmental Science, Policy and Management, University of California, Berkeley, Berkeley, California, USA

ABSTRACT Antibiotic resistance in pathogens is extensively studied, and yet little is known about how antibiotic resistance genes of typical gut bacteria influence microbiome dynamics. Here, we leveraged 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 found 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 than *C. difficile* strains lacking this gene. Organisms with genes for major facilitator superfamily drug efflux pumps have higher 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 to five principal components classified by boosted decision trees. Among the genes involved in predicting whether an organism increased in relative abundance after treatment are those that encode 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 treatment and predict how organisms in the gut microbiome will respond to antibiotic administration.

IMPORTANCE The process of reconstructing genomes from environmental sequence data (genome-resolved metagenomics) allows unique insight into microbial systems. We apply this technique to investigate how the antibiotic resistance genes of bacteria affect their ability to flourish in the gut under various conditions. Our analysis reveals that strain-level selection in formula-fed infants drives enrichment of beta-lactamase genes in the gut resistome. Using genomes from metagenomes, we built a machine learning model to predict how organisms in the gut microbial community respond to perturbation by antibiotics. This may eventually have clinical applications.

KEYWORDS *Clostridium difficile*, antibiotic resistance, genome-resolved metagenomics, infant, machine learning, microbiome, resistome

Antibiotic use steadily increased over the past several decades and is correlated with the prevalence of antibiotic resistance in bacteria (1). Widespread antibiotic resistance, in combination with the decline in development of new antibiotics, presents a major threat to human health (2). The gut microbiome is a reservoir for antibiotic

Downloaded from <http://msystems.asm.org/> on October 16, 2018 by guest

Received 7 September 2017 Accepted 14 December 2017 Published 9 January 2018

Citation Rahman SF, Olm MR, Morowitz MJ, Banfield JF. 2018. Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *mSystems* 3:e00123-17. <https://doi.org/10.1128/mSystems.00123-17>.

Editor Nicola Segata, University of Trento

Copyright © 2018 Rahman et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Jillian F. Banfield, jbanfield@berkeley.edu.

Infant #formula feeding selects for certain #Cdiff strains and #MachineLearning on #GutMicrobiota reveals patterns of #AntibioticResistance

Appendix 1.5 | Impacts of microbial assemblage and environmental conditions on the distribution of anatoxin-a producing cyanobacteria within a river network

The ISME Journal
https://doi.org/10.1038/s41396-019-0374-3



ARTICLE



Impacts of microbial assemblage and environmental conditions on the distribution of anatoxin-a producing cyanobacteria within a river network

Keith Bouma-Gregson^{1,2} · Matthew R. Olm³ · Alexander J. Probst^{2,7} · Karthik Anantharaman^{2,8} · Mary E. Power¹ · Jillian F. Banfield^{1,2,4,5,6}

Received: 24 September 2018 / Revised: 28 January 2019 / Accepted: 31 January 2019
© The Author(s) 2019. This article is published with open access

Abstract

Blooms of planktonic cyanobacteria have long been of concern in lakes, but more recently, harmful impacts of riverine benthic cyanobacterial mats been recognized. As yet, we know little about how various benthic cyanobacteria are distributed in river networks, or how environmental conditions or other associated microbes in their consortia affect their biosynthetic capacities. We performed metagenomic sequencing for 22 Oscillatoriales-dominated (Cyanobacteria) microbial mats collected across the Eel River network in Northern California and investigated factors associated with anatoxin-a producing cyanobacteria. All microbial communities were dominated by one or two cyanobacterial species, so the key mat metabolisms involve oxygenic photosynthesis and carbon oxidation. Only a few metabolisms fueled the growth of the mat communities, with little evidence for anaerobic metabolic pathways. We genomically defined four cyanobacterial species, all which shared <96% average nucleotide identity with reference Oscillatoriales genomes and are potentially novel species in the genus *Microcoleus*. One of the *Microcoleus* species contained the anatoxin-a biosynthesis genes, and we describe the first anatoxin-a gene cluster from the *Microcoleus* clade within Oscillatoriales. Occurrence of these four *Microcoleus* species in the watershed was correlated with total dissolved nitrogen and phosphorus concentrations, and the species that contains the anatoxin-a gene cluster was found in sites with higher nitrogen concentrations. Microbial assemblages in mat samples with the anatoxin-a gene cluster consistently had a lower abundance of Burkholderiales (Betaproteobacteria) species than did mats without the anatoxin-producing genes. The associations of water nutrient concentrations and certain co-occurring microbes with anatoxin-a producing *Microcoleus* motivate further exploration for their roles as potential controls on the distributions of toxigenic benthic cyanobacteria in river networks.

Introduction

When cyanobacteria proliferate in freshwaters, their toxins can threaten water quality and public health [1]. Harmful cyanobacterial blooms in lakes have been described for decades (e.g. [2]). In rivers, however, first reports of animal

Supplementary information The online version of this article (<https://doi.org/10.1038/s41396-019-0374-3>) contains supplementary material, which is available to authorized users.

✉ Jillian F. Banfield
jbanfield@berkeley.edu

¹ Department of Integrative Biology, University of California, Berkeley, CA, USA

² Department of Earth and Planetary Science, University of California, Berkeley, CA, USA

³ Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA

⁴ Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA

⁵ Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

⁶ Chan Zuckerberg Biohub, San Francisco, CA, USA

⁷ Present address: Group for Aquatic Microbial Ecology, Biofilm Center, Department for Chemistry, University of Duisburg-Essen, Essen, Germany

⁸ Present address: Department of Bacteriology, University of Wisconsin, Madison, WI, USA

Published online: 26 February 2019

SPRINGER NATURE