Title

Metagenomics adds unrecognized lineages to the tree of life and enables a genome-resolved view of microbiome dynamics

Author

Brown, Christopher Thomas

Publication Date

2016

Metagenomics adds unrecognized lineages to the tree of life and enables a genome-resolved view of microbiome dynamics

By

Christopher Thomas Brown

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
Adjunct Professor Eoin L. Brodie
Professor Jennifer A. Doudna

Fall 2016

Abstract

Metagenomics adds unrecognized lineages to the tree of life and enables a genome-resolved view of microbiome dynamics

by

Christopher Thomas Brown

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor Jillian F. Banfield, Chair

Microbes live in complex communities that have shaped the planet for billions of years. However, much is not known about their diversity and metabolic potential due to biases in methods that require cultivation or PCR amplification. Metagenomics circumvents these issues and can be used to obtain genome sequences for microbial community members. Approximately 800 metagenome-derived complete and draft-quality genomes were reconstructed for groundwater-associated bacteria from a radiation of previously unrecognized and little-known phyla with essentially no isolated representatives. Unlike most other bacteria, these organisms consistently have small genomes, lack highly conserved ribosomal proteins, frequently have rRNA gene introns, and have significant metabolic limitations indicative of an obligate symbiotic lifestyle. Combined phylogenetic and genomic analyses enabled recognition of this group as the Candidate Phyla Radiation (CPR), a major feature of domain Bacteria that was subsequently determined to comprise >50% of all bacterial diversity. Using a newly developed method called iRep, it was determined that CPR organisms typically replicate slowly, although they did replicate rapidly under some conditions. These *in situ* measurements were possible because iRep uses draft-quality genomes and metagenome sequencing to determine replication rates based on changes in genome copy number that occur during genome replication.

In contrast to groundwater ecosystems, the human microbiome typically contains microorganisms from only a few phyla. Application of metagenomics enabled strain-level resolution of the human microbiome, measurement of iRep replication rates, and proteomic analyses of activity. Microbiome samples were collected from premature infants during the first months of life, and both metagenomics and metaproteomics were used to detect shifts in the gastrointestinal tract microbiome. Results showed that genetically similar bacteria behave differently depending on community context, leading to substantial changes in overall proteome composition. The metagenomic approach enabled identification of considerable genomic novelty. Analysis of the first genome sequence for a member of the genus *Varibaculum* uncovered a diverse repertoire of sugar utilization pathways and anaerobic respiration capacity. iRep analysis documented highly variable replication rates during initial colonization, and significantly higher rates following antibiotic administration. This work has added large and small branches to the tree of life with corresponding genomic and metabolic information, linked microbial responses and metabolism to changing environmental conditions, and provided previously unobtainable information on *in situ* replication rates.

Table of Contents

Introduction

Microorganisms encompass the vast majority of life's diversity, and have shaped the development of the planet over the last 3.5 billion years. These organisms drive biogeochemical cycles and contribute to human health and disease; however, much remains to be known about the organisms and metabolisms critical to these processes. This is in part due to the fact that only a small fraction of microbes have been isolated in the lab, the traditional method for studying them. PCR amplification and sequencing of marker genes, such as the 16S rRNA gene, have made it possible to inventory and conduct phylogenetic analyses of organisms without the need for cultivation. However, this approach does not provide information on metabolic potential, and both PCR primer biases (Brown et al., 2015) and gene copy number variation (Perisin et al., 2015) obscure organism abundance measurements. Genome-resolved metagenomics can be used to obtain genomes for organisms without the need for cultivation, even for diverse microbial communities (Anantharaman et al., 2016b; Baker et al., 2010; Brown et al., 2015; Castelle et al., 2013; Eloe-Fadrosh et al., 2016; Hug et al., 2013; Iverson et al., 2012; Nielsen et al., 2014; Seitz et al., 2016; Sharon et al., 2012; Tyson et al., 2004; Wrighton et al., 2012). Recovered genomes can be used to infer both the phylogeny and metabolic potential of the organisms, and in combination with metagenomics can enable accurate measurement of community composition.

In the 1990s, 16S rRNA gene sequencing conducted in the Obsidian Pool at Yellowstone National Park identified several organisms that could not be grouped into any previously studied phylum (Hugenholtz et al., 1998). Notable amongst these phylum-level groups with no isolated representatives, so called "candidate phyla" or "candidate divisions," were the OP11. Additional surveys expanded the known diversity of the OP11 to the extent that the group was subdivided into several phyla, including the OD1 (Harris et al., 2004). Decreases in the cost of DNA sequencing enabled a plethora of surveys that demonstrated that the OP11 and OD1 reside in marine and freshwater systems, sediments, groundwater, and a variety of other environments (Harris et al., 2004; Luef et al., 2015). However, it was not until 2012 that anything was known about the metabolism of these enigmatic organisms, when genome-resolved metagenomics was used to reconstruct 49 genome sequences (Wrighton et al., 2012). All of these genomes were small, and metabolic analysis indicated that the surveyed organisms participate in sulfur and hydrogen cycling, and have a fermentation-based metabolism. Additional metabolic and phylogenetic study of the OP11 and OD1 based on single-cell genome sequencing lead to the recognition of these groups as the Microgenomates and the Parcubacteria, respectively (Rinke et al., 2013). However, much of the diversity of these groups had yet to be explored.

Since 2007, an alluvial aquifer adjacent to the Colorado River near the town of Rifle, CO has been a prominent site for subsurface research (**Appendix 1.1, Appendix 1.2, Appendix 1.3, Appendix 1.4, Appendix 1.5, Appendix 1.6, Appendix 1.7, and Appendix 1.8**) (Long et al., 2016). The site began as a vanadium mill in the early 1920s, and both vanadium and uranium milling efforts persisted intermittently through the 1960s. This is the site where the first complete and draft-quality genome sequences for members of the Microgenomates (OP11) and Parcubacteria (OD1) were recovered (Kantor et al., 2013; Wrighton et al., 2012). The finding that these organisms have small genomes suggested that they may also have small cell sizes. In order to investigate this possibility, a cryogenic transmission electron microscopy study was conducted on ultra-small cells capable of passing through 0.2 µm filters, which are often used for sterilization purposes. Metagenomic analysis showed that CPR bacteria dominated small-cell filtrates collected on 0.1 µm filters. Images of members of the Microgenomates, Parcubacteria, and a related group known as Katanobacteria (WWE3) documented striking morphological features of these organisms, including the presence of pili predicted based on genomic analysis, and demonstrated their ultra-small cell volumes (0.009±0.002 mm^3) (Luef et al., 2015). Additional analysis of organism abundances on 0.2 and 0.1 µm filters suggests that particular CPR lineages are more likely than others to have ultra-small cell sizes (**Appendix 2**).

To further investigate these organisms, we conducted metagenome sequencing of cells collected from groundwater on both the 0.2 and 0.1 µm filters. Metagenomics enabled reconstruction of ~800 genomes from members of the Microgenomates, Parcubacteria, and other candidate phyla (**Chapter 2**) (Brown et al., 2015). Our phylogenetic and genomic analyses showed that these organisms are from a group of phyla comprising >15% of bacterial diversity, which we described as the Candidate Phyla Radiation (CPR). This more expansive sampling of CPR genomes showed that they are consistently small, and that the organisms have significant metabolic limitations strongly suggestive of a symbiotic lifestyle. In addition, rRNA gene introns were unusually common throughout this radiation, and specific CPR lineages were found to have unusual ribosome compositions. Phylogenetic analysis of both 16S rRNA gene sequences and concatenated ribosomal proteins showed that the Microgenomates and Parcubacteria are superphyla, and enabled identification of 25 phyla within these two groups (Brown et al., 2015). In subsequent collaborative work that involved phylogenetic analysis of organisms from all domains of life, it was determined that the CPR comprises >50% of all bacterial genetic diversity (**Appendix 1.1**) (Hug et al., 2016). In an additional study of multiple field experiments conducted at the Rifle site, we identified 47 new phyla from analysis of ~2,500 genomes recovered from metagenomes, 30 of which were from the CPR (**Appendix 1.8**) (Anantharaman et al., 2016b). We also identified and conducted a genome-informed metabolic analysis of organisms from novel phyla within the archaeal DPANN superphylum that, like CPR, consistently have small genomes with limited metabolic potential (**Appendix 1.2**) (Castelle et al., 2015).

Unlike other systems, the human gut microbiome is composed of organisms from a small number of phyla. However, considerable species and strain level diversity exists and may have a substantial influence on microbiome function (Sharon et al., 2012). The human microbiome has been implicated in obesity (Ley et al., 2005), inflammatory bowel disease (Xavier and Podolsky, 2007), necrotizing enterocolitis in premature infants (Mai et al., 2011; Morrow et al., 2013; Mshvildadze et al., 2010), and other chronic diseases such as type 1 and type 2 diabetes (Brown et al., 2011; Gilbert et al., 2016; Heintz-Buschart et al., 2016; Qin et al., 2012). Consequently, microbial colonization of the human gut at birth may be important to short and long-term health. Surveys of this period have characterized shifts in community composition related to early-life events (Bokulich et al., 2016; Koenig et al., 2011), but little is known about the implications of these changes due to a lack of a functional understanding of the organisms. This is in part due to the fact that commonly used 16S rRNA gene sequencing methods are limited in their ability to resolve metabolic differences between organisms. In contrast, genome-resolved metagenomics studies of the first weeks of life of premature infants uncovered microbial strain and phage patterns critical to understanding the colonization process (Raveh-Sadka et al., 2016; Sharon et al., 2012).

Using genome-resolved methods, we documented genomic novelty, including the first genome from a member of the genus *Varibaculum*, and community shifts during the third week of life for a premature infant (**Chapter 1**) (Brown et al., 2013). Metabolic analysis indicated that *Varibaculum cambriense* make use of a wide variety of carbon sources and electron acceptors for respiration. Comparative genomics uncovered important differences between related organisms with respect to respiratory metabolism and motility. Species and strain diversity was also present within the microbiome, further emphasizing the need for genome-resolved methods.

In collaborative work, genotypic information was used to distinguish strains in co-hospitalized infants (**Appendix 1.9**). Results indicate that although infants are typically colonized by distinct strains (Raveh-Sadka et al., 2015; Sharon et al., 2012), some specific genotypes persist and colonize different infants at times separated by multiple years (Raveh-Sadka et al., 2016). In other collaborative work, it was found that identical bacterial strains were among the initial colonists of premature infant mouth, skin, and gut, and that strains associated with the mouth and skin were replicating faster than those in the gut (Olm et al., 2016). Subsequently, we have analyzed the microbiome of additional infants using both metagenomics and metaproteomics, in conjunction with our newly developed method for determining *in situ* replication rates. These techniques enabled investigation of organism-specific activity and dynamics during the process of microbial colonization, and showed that organisms in the premature infant gut can behave differently depending on their environment, and that these differences can drive overall microbiome function (**Chapter 4**).

Culture-independent methods have revolutionized our understanding of microbial diversity, but a lack of information about *in situ* activity has limited our ability to identity the contributions of microbes to human and environmental health. We developed a new method, iRep, for determining *in situ* replication rates for bacteria, and used this method to obtain measurements from human and groundwater microbiomes (**Chapter 3**) (Brown et al., 2016). Application to CPR organisms showed that they sometimes replicate quickly, which was not predicted based on genomic features. Combined with our finding that human-associated organisms grow faster following antibiotic administration, this emphasizes that replication rates can be highly variable even for single organisms, and are not constant factors that can be determined in the lab. Overall, this work has added large and small branches to the tree of life, provided additional metabolic information for groups of bacteria for which little was previously known, and linked microbial responses and metabolism to changing environmental conditions.

Acknowledgements

There are so many people that have helped me throughout graduate school. I must first thank Professor Jill Banfield for being an inspiring and supportive mentor. Without her advice and encouragement this dissertation would not have been possible. Jill has guided me through everything from metagenomics analysis to scientific writing and publishing, and even gave me my first pottery lesson. I will always remind myself to think like Jill.

Although the people have changed over the years, the Banfield lab has always been a fun and enlightening environment for doing research and learning to be a scientist. There is not enough space to thank everyone, but I must give special thanks to Brian Thomas, Laura Hug, Cindy Castelle, Kelly Wrighton, Itai Sharon, Nicholas Justice, David “Dudu” Burstein, Sue Spaulding, Karthik Anantharaman, Alex Probst, Brandon “Bubba” Brooks, Rose Kantor, Tyler Arbour and Matt Olm. I would also like to thank our collaborators Dr. Michael Morowitz at the University of Pittsburgh School of Medicine, Dr. Robert Hettich at Oak Ridge National Laboratory, and Dr. Kenneth Williams at Lawrence Berkeley National Laboratory. I have also been fortunate to have great colleagues at UC Berkeley. Time talking about science with Matt Shurtleff and David Hershey is always well spent.

I am also grateful for my undergraduate mentors Professor Eric Triplett and Professor Wayne Nicholson at the University of Florida. They not only showed me that it is possible to be a scientist, but that it is an endlessly rewarding occupation.

Both my parents have always encouraged me to do what I am interested in. My mom has supported my interest in science from a young age. I will never forget the microscope I got for Christmas, or how I was allowed to take over the dining room table for weeks at a time conducting “experiments.” Those experiences left a lasting impression. My brother Danny has always been a best friend to me. I have also been fortunate to have the support of my Aunt Lynn and Uncle Mark Frikker. I cannot imagine having accomplished as much without all of their support.

I am grateful for Emily Thompson, whose love and support make life fun and wonderful. Thank you for going on so many adventures with me. Biking across the country and finishing this degree would have been impossible without you by my side. You have always been there for me, through both the fun and difficult times. I look forward to all of our future adventures.

Moving across the country would have been a daunting task without our California family: Maggie, Jessica, Anne, Margaret, Doug, John, and Faustene. Thank you for all of the times we have shared over the years.

Chapter 1

Genome resolved analysis of a premature infant gut microbial community reveals a *Varibaculum cambriense* genome and a shift towards fermentation-based metabolism during the third week of life

C. T. Brown, I. Sharon, B. C. Thomas, C. J. Castelle, M. J. Morowitz,
and J. F. Banfield

Published in *Microbiome*, December 2013.

Abstract

The premature infant gut has low individual but high inter-individual microbial diversity compared with adults. Based on prior 16S rRNA gene surveys, many species from this environment are expected to be similar to those previously detected in the human microbiota. However, the level of genomic novelty and metabolic variation of strains found in the infant gut remains relatively unexplored. To study the stability and function of early microbial colonizers of the premature infant gut, nine stool samples were taken during the third week of life of a premature male infant delivered via Caesarean section. Metagenomic sequences were assembled and binned into near-complete and partial genomes, enabling strain-level genomic analysis of the microbial community. We reconstructed eleven near-complete and six partial bacterial genomes representative of the key members of the microbial community. Twelve of these genomes share >90% putative ortholog amino acid identity with reference genomes. Manual curation of the assembly of one particularly novel genome resulted in the first essentially complete genome sequence for *Varibaculum cambriense* (strain Dora), a medically relevant species that has been implicated in abscess formation. During the period studied, the microbial community undergoes a compositional shift, in which obligate anaerobes (fermenters) overtake *Escherichia coli* as the most abundant species. Other species remain stable, probably due to their ability to either respire anaerobically or grow by fermentation, and their capacity to tolerate fluctuating levels of oxygen. Metabolic predictions for *V. cambriense* suggest that, like other members of the microbial community, this organism is able to process various sugar substrates and make use of multiple different electron acceptors during anaerobic respiration. Genome comparisons within the family *Actinomycetaceae* reveal important differences related to respiratory metabolism and motility. Genome-based analysis provided direct insight into strain-specific potential for anaerobic respiration and yielded the first genome for the genus *Varibaculum*. Importantly, comparison of these *de novo* assembled genomes with closely related isolate genomes supported the accuracy of the metagenomic methodology. Over a one-week period, the early gut microbial community transitioned to a community with a higher representation of obligate anaerobes, emphasizing both taxonomic and metabolic instability during colonization.

Introduction

The human adult microbiota consists of 10-fold more cells than the human body (the majority reside in the gut) and 100-fold more genes than the human genome (Ley et al., 2006; Qin et al., 2010; Whitman et al., 1998). The gut microbiota are involved in host nutrient acquisition (Turnbaugh et al., 2006), regulation and development of the host immune system (Lathrop et al., 2011; Maslowski et al., 2009), and the modulation of host gene expression [7](Hooper et al., 2001). All of these influences have the potential to seriously affect human health. Aberrations in gut microbiota membership and community structure, termed microbial dysbiosis, have been associated with obesity (Ley et al., 2005) and diseases such as inflammatory bowel disease (Xavier and Podolsky, 2007), both type 1 and type 2 diabetes (Brown et al., 2011; Qin et al., 2012), and necrotizing enterocolitis in premature infants (Mai et al., 2011; Morrow et al., 2013; Mshvildadze et al., 2010). Although previous studies have focused on gut colonization (Koenig et al., 2011; Palmer et al., 2007), few have shown the process in a high-resolution manner (Morowitz et al., 2011; Sharon et al., 2012). Thus, much is still not known about the diversity, metabolic potential, or roles of early gut colonizers.

Although the gut microbiota of infants is characterized by high levels of inter-individual diversity (beta diversity), community composition begins to look like that of adults within the first year of life (Palmer et al., 2007). In comparison with both adults and infants, premature infants have especially low individual diversity (alpha diversity), making them ideal subjects for high-resolution (species or strain-level) community genomics approaches (Morowitz et al., 2011; Sharon et al., 2012). Continued study of microbial colonization in the gut of premature infants may yield further insights into the details of this process and the implications of disease-associated microbial dysbiosis.

Community genomics, the use of genomes sequenced from natural microbial communities to understand the structure and metabolism of the community, has been successful in environments with varying levels of diversity (Castelle et al., 2013; Chivian et al., 2008; Hess et al., 2011; Hug et al., 2013; Morowitz et al., 2011; Sharon et al., 2012; Tyson et al., 2004; Wrighton et al., 2012). Recently, this approach has been applied to the human microbiome, where the genomes of abundant bacterial species were assembled from a premature infant (Morowitz et al., 2011) and, most recently, where increased sequencing depth allowed for genomes to be assembled for both high-abundance and low-abundance members of the microbial community found in the gut of another premature infant (including genomes for members that make up less than 0.05% of the microbial community) (Sharon et al., 2012). Both of these studies involved analysis of strain-level variation within the human gut microbiome. In human adults, a draft genome of Shiga-Toxigenic *Escherichia coli* O104:H4 was assembled from metagenome data taken from individuals involved in an outbreak, providing strain-level resolution of this pathogen (Loman et al., 2013). Strain-level analysis of microbial communities contrasts strongly with 16S rRNA-based fingerprinting methods that characterize communities at a phylum to genus-level of resolution. This is primarily due to the added benefit of being able to directly determine the metabolic potential of strains in a particular community (which need not have been previously studied), and to identify metabolic variation between strains that may have highly similar or even identical 16S rRNA gene sequences (Sharon et al., 2011). In general, the study of infants enables development of an understanding of microbial colonization in humans, and can provide genomes for biologically and medically relevant, and oftentimes novel, species directly from their source environments (without the bias of isolation or cell sorting, or single cell manipulation and genome amplification steps).

Here, we investigate gut colonization in a relatively healthy premature infant during the third week of life with the objectives of comparing genomic novelty between the natural consortia and isolate strains, recovery and analysis of genomes from previously uncharacterized community members, and metabolic analysis of the microbial community. This period of early gut colonization was targeted for intensive sample collection because it is believed that aberrant colonization near this time can contribute to the pathogenesis of necrotizing enterocolitis (which was not observed in this infant). Our approach involves reconstructing complete and near-complete genomes from DNA extracted from fecal samples to enable prediction of the roles of specific species and strains in the community. Time series abundance analysis is a key component of the approach because shifts in community composition can be detected, and also because organism abundance patterns greatly increase the accuracy with which assembled fragments can be assigned to specific organisms (binning; (Sharon et al., 2012)). We show that, even in the human gut, where many species can be represented by reference genomes, there are organisms with genomic potential not represented by reference sequences. Specifically, we report the first genome for the genus *Varibaculum*, a genus that has been implicated in human abscess formation, but that has not been associated with the human gut (Hall et al., 2003).

Results

Metagenome sequencing, assembly, binning, and annotation

The nine samples collected on days of life 14 through 20 from the infant in our study resulted in 35 gigabase pairs (Gbp) of paired-end Illumina DNA sequences with a length of 100 nucleotides. Filtering out human DNA and quality trimming resulted in 27.8 Gbp of Illumina reads with an average length of 93 bp. The iterative metagenome assembly method used resulted in 89.29% of high quality reads being assembled into 12,184 scaffolds longer than 400 bp (40.8 megabase pairs (Mbp), N50: 13,265 bp, longest scaffold: 608,611 bp). From the scaffolds larger than 400 bp, 46,156 ORFs were predicted (average amino acid length: 254), 94.4% of which had a match to the UniRef90 database with an *E*-value less than or equal to 0.001.

Scaffolds were clustered based on their time series abundance patterns using an ESOM, resulting in 25 bins (**Figure 1.1, Supplementary Table 1.1, and Supplementary File 1.1**). These bins represent complete, near-complete, and partial bacterial, plasmid, and viral genomes (**Table 1.1**) and 85.98% of high quality sequencing reads (**Supplementary Table 1.2 and Supplementary Table 1.3**). Six complete (circular) plasmids were assembled along with five putative phage fragments. Plasmid and phage fragments account for only 0.27% and 0.23% of the total sequence data, respectively; however, they account for the majority of the community in terms of relative abundance (41.7% and 16.6%; **Supplementary Table 1.4 and Supplementary Table 1.5**).

Microbial genome identification and curation

Well-defined genomes were binned for *Clostridium butyricum, Enterococcus faecalis, Streptococcus anginosus*, *Streptococcus sp.,* and *Varibaculum cambriense.* However, some bins were not clearly delineated owing to the low abundance of the associated species (at the limits of sequencing detection), similar abundance patterns between species, or coverage miscalculations due to strain variation. This was the case for the bins of *Actinomyces urogenitalis*, *Clostridium* *bartlettii*, two *Escherichia coli* strains, *Leuconostoc sp.*, *Negativicoccus succinicivorans, Propionibacterium sp., Staphylococcus sp., Streptococcus parasanguinis*, *Veillonella dispar*, and two additional *Veillonella* species. Manual curation of these bins resulted in near-complete and partial genome reconstructions for these species, and revealed strain-resolved genomic novelty within the *E. coli* population (**Table 1.1**).

We evaluated the completeness of each genome using a list of 26 single copy marker genes (**Supplementary File 1.2** and (Raes et al., 2007)), revealing that the genomes for *Actinomyces urogenitalis, Clostridium bartlettii, Clostridium butyricum, Enterococcus faecalis, Negativicoccus succinicivorans, Streptococcus anginosus, Streptococcus parasanguinis, Varibaculum cambriense,* and *Veillonella dispar* (9 out of the 17 total genomes) are near-complete (over 75% of marker genes could be identified) (**Table 1.1**).

The genome for *Varibaculum cambriense* was reassembled and manually curated. Before reassembly, the genome was represented by 35 scaffolds with a total length of 2.25 Mbp and an N50 of 240,417 bp. Following reassembly and manual curation, the genome was assembled into three scaffolds, each terminated by a repeat sequence corresponding to a transposase gene. The three scaffolds include completely assembled 16S rRNA and 23S rRNA genes, a total length of 2.28 Mbp, an N50 of 1,648,569 bp, and 105-fold sequencing coverage. The *V. cambriense* genome has a GC content of 52.5%. Approximately 70% of ORFs could be assigned to a putative function. All three scaffolds are connected to each other but their order cannot be determined; thus, all connections are resolved and we consider this genome to be essentially complete. Furthermore, all of the single copy marker genes used to assess genome completeness could be identified along with all 20 aminoacyl tRNA synthetase genes in the *V. cambriense* genome (**Supplementary File 1.2**).

The assembled *E. coli* plasmid has the highest copy number of any assembled plasmid, phage, or bacterial member of the community (**Table 1.1 and Supplementary Table 1.4**). The plasmid contains a replication protein distinct from other *Enterobacteriaceae* plasmids, suggesting that it is novel (**Figure 1.2**). The largest plasmid (47.63 Kbp) is associated with *Veillonella sp. - species A*, and also contains a novel replication protein, indicating that this plasmid has not been previously studied. Two plasmids are related to known *Staphylococcus* plasmids and are highly correlated with one another (Pearson coefficient = 0.89), although they are not closely related to one another (based on their replication proteins). Based on the annotations and abundance patterns for these plasmids, their host is the *Staphylococcus sp.* for which we reconstructed a near-complete genome (Pearson coefficients of 0.45 and 0.62). Additionally, a complete plasmid genome was assembled and associated with *S. parasanguinis* (based on protein annotations and time series abundance patterns; Pearson coefficient = 0.99).

Phylogenetic placement of EMIRGE 16S rRNA genes

EMIRGE reconstructed 77 candidate 16S rRNA gene sequences, 14 of which could be associated with reconstructed genomes (**Table 1.2 and Supplementary File 1.3**). The discrepancy between the number of EMIRGE sequences and genomes suggests that EMIRGE is overestimating the number of OTUs. Genes constructed by EMIRGE that could not be assigned to genomes were related to *Shigella* (probably represented by binned *E. coli* genomes), *Okadaella, Buttiauxella, Brevibacterium,* and *Citrobacter*. Of these, low-abundance ORFs from the community could be assigned to the genus *Brevibacterium* at greater than 90% amino acid identity, and to *Citrobacter*, but at less than 90% amino acid identity, suggesting the possible presence of these genera in low abundance in the community. EMIRGE sequences that were not connected with bins were not analyzed further.

The 16S rRNA gene sequences reconstructed with EMIRGE were used to build a phylogenetic tree to classify organisms in the community (**Figure 1.3**). The tree shows that many of the reconstructed genomes are from organisms very closely related to those with sequenced genomes, based on their 16S rRNA gene sequences. The tree highlights the lack of reference genomes for *Varibaculum*, although there are numerous 16S rRNA gene sequences from isolates and from clone libraries. Phylogenetic placement of the EMIRGE sequence for bin 20 confirms that this genome is from a member of the species *V. cambriense* (99.2% 16S rRNA gene sequence identity).

Comparison of reconstructed genomes to reference genomes

The 25 ESOM bins represent the genomes of 17 unique organisms from the microbial community. Of the nine reconstructed near-complete genomes, seven share over 90% ortholog amino acid identity with reference strains, while five share over 95% ortholog amino acid identity (with at least 60% of ORFs being defined as orthologs) (**Table 1.2**). Despite this high-level of similarity, 14% of the ORFs predicted for near-complete reconstructed genomes do not have orthologs within their most closely related reference genomes. At the extreme, 57% of the predicted ORFs for *Negativicoccus succinicivorans* are not orthologous with genes found in the most closely related reference genome. Although the level of genomic divergence observed here does not fall outside of the range previously observed (Konstantinidis and Tiedje, 2005), this finding underscores the importance of genome reconstructions, as opposed to 16S rRNA gene sequence analysis, for inferring microbial metabolic potential. Of particular note, the genome for *V. cambriense* has an average ortholog amino acid sequence identity of 54% with the genome of its closest sequenced relative, *Mobiluncus mulieris*.

Evidence for the importance of anaerobic metabolism and oxygen tolerance

Several genomes encode cytochrome *bd* oxidase (**Figure 1.4 and Table 1.3**), a high oxygen affinity enzyme indicative of an ability to grow in the presence of low levels of oxygen, either by providing protection from reactive oxygen species or by using oxygen as a terminal electron acceptor during respiration (Das et al., 2005; Morris and Schmidt, 2013). The presence of fumarate, TMAO, DMSO, nitrate, nitrite, and nitric oxide reductase genes supports the notion that members of the community are capable of using several terminal electron acceptors to respire anaerobically (**Figure 1.5**). *E. coli* was the only organism found to encode heme-copper cytochrome oxidase genes, indicating its ability to use oxygen as an electron acceptor when present. To further assess the oxygen utilization capacity of the community, all binned and unbinned ORFs were searched for cytochrome *c* oxidase genes, which would indicate aerobic, or possibly aero-tolerant metabolism (Morris and Schmidt, 2013), but none were found. Taken together, we conclude that all organisms in the community are either obligate or facultative anaerobes.

Microbial abundance and community shifts

Escherichia coli - strain A accounts for the greatest percentage of high quality sequence data and *Propionibacterium sp.* accounts for the least amount (33.45% and 0.03%, respectively; **Supplementary Table 1.2**). Species abundance patterns over the third week of life defined two phases of microbial community composition (**Figure 1.6 Figure 1.7**). The first phase is observed during DOL 14 to DOL 15, whereas the second covers DOL 18 to DOL 20. The first phase is defined by a dominant *E. coli* strain (a facultative anaerobe), and the second phase is dominated by obligate anaerobes (*Streptococcus anginosus, Clostridium butyricum,* and *Veillonella dispar*). Early in the second phase (first time point on DOL 18) there is an increase in the relative abundance of *Streptococcus anginosus,* followed by a stable abundance afterwards. This is followed by a spike in the abundance of *Clostridium butyricum* observed on the second time point taken on DOL 18, after which the relative abundance immediately decreases. There is no apparent clinical variable (for example, change in diet or medication) that accounts for the shift between phase one and phase two. The distinct difference in community composition between theses phases corresponds with a shift towards fermentation-based metabolism in the successors of initially dominant *E. coli*.

V. cambriense is nearly undetectable during the first time point (0.15%) and remains at a low abundance throughout the time series (always ≤3%). It is interesting to note that *Streptococcus, Escherichia, Veillonella, Actinomyces,* and *Enterococcus* dominate the microbial community and that, similar to observations in other premature infants, no *Bacteroides*, *Bifidobacterium*, or *Lactobacillus* were observed throughout the time series (Caplan, 2009; Morowitz et al., 2011; Sharon et al., 2012).

Metabolism of V. cambriense based on genomic analysis

V. cambriense (strain Dora; **Table 1.3 and Supplementary File 1.4**) became the focus of further analysis for several reasons, including (i) the lack of reference genomes available for *Varibaculum* (**Table 1.2 and Figure 1.3**), (ii) the availability of a new, essentially complete genome (**Table 1.1**), and (iii) the fact that members of this genus are medically relevant and drastically understudied in the human gut (Chu et al., 2009; Hall, 2008; Hall et al., 2003). Further, *Varibaculum* have never been studied in the human gut at a species (or genome) level of resolution. Finally, there is some metabolic information in *Bergey's Manual of Systematic Bacteriology* for cultured strains of *V. cambriense* (Whitman et al., 2012).

V. cambriense cell wall and motility

Genome analysis shows that genes involved in the lipopolysaccharide biosynthesis pathway are missing, confirming that *V. cambriense* does not have a Gram-negative cell envelope. The peptidoglycan biosynthesis pathway containing meso-diaminopimelate is complete. However, the β-lactam resistance pathway for peptidoglycan synthesis is incomplete, suggesting sensitivity to β-lactam antibiotics. No genes for flagella, pili, or chemotaxis were identified, indicating that this strain, like cultured members of this species, is not motile.

V. cambriense transporters and resistance

Twenty different sugar transport ORFs were identified, indicating that *V. cambriense* has the ability to use many different types of sugars. A putative sialic acid transporter is present, along with genes required for metabolizing *N*-acetylneuraminate (discussed later), suggesting that this abundant component of both human and non-human cell-surface glycoproteins, human breast milk glycans, and intestinal mucins can be used as a nutrient. Additionally, a glycerol-3-phosphate transporter, a putative phosphotransferase IIA system, a sodium-galactoside symporter, and multiple sugar transport system permease genes were identified. No acetate transporter was found, although several pathways were identified for converting pyruvate into acetate for the bidirectional conversion between acetyl-CoA and acetate. Complete KEGG modules suggest that *V. cambriense* can transport ribose, phosphate, nickel, lipopolysaccharides, and fructose.

Although no complete antibiotic resistance pathways were identified, a drug resistance transporter (EmrB/QacA subfamily) and a methicillin resistance protein are coded for in the genome. *V. cambriense* has various resistance mechanisms, including an arsenate resistance pathway, the pathway for glycine betaine biosynthesis (a compound capable of protecting against osmotic stress; (Boch et al., 1997)), and a P-type ATPase for translocating copper and silver (suggesting Cu^{2+} tolerance).

The genome contains the enzyme trehalose synthase, which is necessary for trehalose synthesis from β-maltose. The enzymes for synthesizing trehalose from glycogen were also identified, but not the enzymes for degrading trehalose. Trehalose has several biological roles, including that of structural component (Leslie et al., 1994) and stress protector (Strøm and Kaasen, 1993). The pathway for ppGpp is also encoded by the genome; this pathway is known for its role in regulating responses to nutrient or energy starvation and environmental stresses (Traxler et al., 2008).

V. cambriense nutrient sources

Based on the genome, *V. cambriense* is able to use acetoacetate, ammonia, arabinose, ethanol, fructose, glucose, glycerol, glycogen, lactose, mannose, melibiose, ribose, sialic acid, starch, sucrose, and xylose as nutrient sources. It is interesting to note that cultured representatives of *V. cambriense* have not been shown to use either starch or xylose (Whitman et al., 2012). Fructose, glucose, glycerol, glycogen, lactose, mannose, melibiose, starch, sucrose, and xylose can all be directed to glycolysis, for which the complete pathway was identified. Unlike other members of the community, including *A. urogenitalis* (the other Actinomycetaceae), *V. cambriense* does not have the enzymes for the Entner-Doudoroff pathway (**Figure 1.4**).

The genome encodes several neuraminidases (also known as sialidases), suggesting that *V. cambriense* is able to cleave various sialic acid species from host-derived substrates. Sialic acids are found terminally bound to cell-surface glycoproteins, human breast milk glycans, and mucins, but are only accessible by microbes once they have been cleaved from their substrate (David, 2012; Lewis and Lewis, 2012; Vimr, 2013). The genome also contains a sialic acid transporter and the enzymes necessary for converting the predominant form of sialic acid found in humans, *N*-acetylneuraminate, to d-fructose-6-phosphate, which can in turn be fed into glycolysis. Thus, unlike many bacterial species, *V. cambriense* is probably able both to liberate sialic acids and to make use of them as a nutrient source. However, the genes for the pathway that converts *N*-acetylneuraminate to CMP-*N*-acetylneuraminate are not present, making it unlikely that *V. cambriense* can coat its outer membrane with sialic acids as other species do to evade the host immune system (Severi et al., 2007).

The pathway for degrading *N*-acetylglucosamine, a derivative of glucose that is found in chitin, fungal, and prokaryote cell walls, was identified. A near-complete pathway was identified for converting myo-inositol to dihydroxyacetone phosphate, acetyl-CoA, and CO_2 (while reducing two NAD^+ to NADH). However, no pathways for butyrate or cellulose metabolism could be found, nor genes involved in mucin protein degradation, nor evidence for CO_2 fixation (no evidence could be found for the presence of pyruvate formate-lyase).

V. cambriense pentose phosphate pathway

The *V. cambriense* genome is missing both glucose-6-phosphate dehydrogenase and 6-phosphogluconolactonase, primary components of the oxidative branch of the pentose phosphate pathway. However, as is expected for a facultative anaerobe, the non-oxidative pathway is complete. The presence of a gluconate transporter and ribose transport system suggests that the pentose phosphate pathway could use these precursors to produce d-glyceraldehyde-3-phosphate, which in turn could enter the methylerythritol phosphate pathway and create isopentenyl diphosphate and dimethylallyl diphosphate (fundamental units of isoprenoid biosynthesis), and geranyl diphosphate (a crucial precursor of menaquinone biosynthesis). Likewise, d-fructose-6-phosphate produced in the pentose phosphate pathway can be fed into glycolysis or can participate in the synthesis of UDP-N-acetyl-d-glucosamine, a necessary precursor of cell wall peptidoglycan. Another branch from d-erythrose-4-phosphate in the pentose phosphate pathway could lead to the biosynthesis of chorismate, an important biochemical intermediate. Overall, these pathways indicate the sources of several key metabolic precursors.

V. cambriense fermentation and degradation reactions

The presence of lactate dehydrogenase suggests that pyruvate produced from glycolysis can be fermented to lactate; however, consistent with isolate metabolic data, no pathways were found to consume lactate (Whitman et al., 2012). We identified that the α, β, and γ subunits of pyruvate-ferredoxin oxidoreductase (EC:1.2.7.1) are successively encoded on the *V. cambriense* genome (the δ subunit could not be identified). This enzyme uses an oxidized ferredoxin to ferment pyruvate and produce H^+, CO_2, and acetyl-CoA, which can subsequently be converted into either acetate or ethanol. Although the directionality of ethanol interconversion is difficult to infer from protein sequences alone, it is possible that the reverse of the ethanol fermentation reaction can occur without additional enzymes, resulting in the formation of acetyl-CoA from ethanol with a gain of two NADH. Additionally, the pathway for acetoacetate degradation through the intermediate acetoacetyl-CoA, which has a net yield of one molecule of acetyl-CoA, also exists. In *E. coli*, acetoacetate can function as a total source of carbon and energy through this pathway (Pauli and Overath, 1972), and this may be the case for *V. cambriense*.

V. cambriense tricarboxylic acid cycle

There is strong evidence for a tricarboxylic acid (TCA) cycle and respiratory capacity in the *V. cambriense* genome. Two of the three components of the pyruvate dehydrogenase complex are encoded by the genome (the E1 component could not be identified). If functional, *V. cambriense* could convert pyruvate into acetyl-CoA (which could be used in the TCA cycle) using either this enzyme or pyruvate-ferredoxin oxidoreductase. The genome encodes the enzymes for converting acetyl-CoA into succinate. The TCA cycle can then continue by converting succinate into fumarate using succinate dehydrogenase/fumarate reductase (EC:1.3.99.1). Four subunits are required for this enzyme, but only three were identified (the iron-sulfur subunit, flavoprotein subunit, and cytochrome b556 subunit are co-localized on the genome). No evidence could be found for the membrane anchor subunit, which may be due to the presence of a small scaffolding gap in this region of the genome, or may indicate the existence of a divergent form of this enzyme. The presence of fumarate lyase provides a way for fumarate to be converted into malate, thus continuing the cycle. The form of malate dehydrogenase that converts malate into oxaloacetate by reducing NAD^+ to NADH and H^+ is present, but not the form of the enzyme that uses a quinone. Taken together, *V. cambriense* encodes a complete TCA cycle.

V. cambriense anaerobic respiration

Several components of an anaerobic respiratory chain were identified. The large and small subunits of the hydrogenase enzyme (containing iron-sulfur clusters, EC:1.12.99.6) and all 14 subunits of NADH dehydrogenase (EC:1.6.5.3) were identified, indicating that both hydrogen acquired from the environment and NADH produced by glycolysis, the TCA cycle, and substrate degradation reactions (ethanol degradation, for example) can be used as electron donors during anaerobic respiration.

Identification of fumarate reductase (EC:1.3.99.1), nitrate reductase (EC:1.7.99.4), and dimethyl sulfoxide (DMSO) reductase (EC:1.8.5.3), suggests that fumarate, nitrate, and DMSO can all be used as terminal electron acceptors. Phylogenetic analysis of the nitrate reductase and DMSO reductase catalytic subunits supports the functional roles of these genes (**Figure 1.5**). Furthermore, owing to the presence of a TAT signal sequence in the *V. cambriense* nitrate reductase, the active site is located on the outside of the cytoplasmic membrane, as is common in Archaea (Pauli and Overath, 1972). Also, the five signature residues suggested as being involved in nitrite and nitrate binding are conserved (Martinez-Espinosa et al., 2007). We also identified genes encoding a nitrate-nitrite transporter, but no other reactions that produce or consume nitrate or nitrite (the organism does not fix nitrogen, for example). No reactions for forming DMSO, or transporters for DMSO could be identified. The nitrite resulting from nitrate reduction could be further reduced by several other community members, several of which have the capacity to further reduce nitric oxide to nitrous oxide (*Escherichia coli - strain A, Staphylococcus sp., Veillonella dispar,* and *Veillonella sp. - species A*), although no species is predicted to be able to reduce nitrous oxide (**Table 1.3**).

As is common in gram-positive bacteria, *V. cambriense* does not have the genes required for the formation of ubiquinones. However, a near-complete pathway is present for the biosynthesis of menaquinones, electron mediators essential during fumarate, DMSO, and nitrate reduction (Wissenbach et al., 1990). All subunits of the F-type H^+-transporting ATPase were identified, indicating that *V. cambriense* is able to produce ATP from the generated proton gradient.

As noted, *V. cambriense* is not capable of aerobic respiration. However, both subunits of the cytochrome *bd* complex were identified. This cytochrome along with cysteine synthase and superoxide dismutase (also identified) can protect against oxidative stress and contribute to limited oxygen tolerance (Das et al., 2005; Rolfe et al., 1978). Consistent with cultured strains, no evidence was found for catalase production (Hall et al., 2003; Whitman et al., 2012).

Other metabolic pathways found in V. cambriense

The genome contains complete amino acid biosynthesis and degradation pathways and all 20 aminoacyl tRNA synthetase genes. Complete pathways were found for riboflavin (vitamin B2) and vitamin K2 biosynthesis. The pathway for the synthesis of folate (vitamin B6) is incomplete; however, some crucial and unique enzymes to the pathway were identified, suggesting that this organism may also be able to synthesize this vitamin (approximately 30% of ORFs do not have a predicted function and could be responsible for this and other pathways). The combination of these functions indicates that *V. cambriense* may exist symbiotically with its human host under certain conditions.

Abundance of Varibaculum in the healthy adult human microbiota

We searched data from the HMP, surveying the V3-5 region of the 16S rRNA gene in order to assess the abundance and distribution of *Varibaculum* in the human microbiota (**Figure 1.8**) (The NIH HMP Working Group et al., 2009). Out of the 5,000 samples taken from 235 healthy human subjects, only 90 had hits for *Varibaculum* (0% oral, 0.31% stool, 1.42% nasal, 3.92% skin, and 10.56% vaginal samples encompassing 24.68% of subjects). In only 29 of these (from 25 different individuals) was the relative abundance of *Varibaculum* greater than 0.05%, while this genus never represented more than 2.5% of any sample. On average, the most abundant organism in communities studied by the HMP represent 14.42% of the total community (standard deviation of 14.11%), and in communities with *Varibaculum*, the most abundant organism represented 20.48% (standard deviation 10.96%). *Varibaculum* was most abundant in samples from the antecubital fossa (skin) and the vagina. Only one stool sample had hits for *Varibaculum*, where it represented only 0.02% relative abundance. Although *Varibaculum* is not uncommon, it is never a dominant community member in the large, healthy, adult population surveyed by the HMP.

Comparative genomics of V. cambriense

Comparing the genome for *V. cambriense* with available genomes for members of the family *Actinomycetaceae* revealed few unique genes, most of which are phage-associated or are not annotated. Several of these unique genes corresponded with folate, butanoate, and benzoate metabolism (among other pathways), but no complete pathways could be established from this set. However, there is considerable metabolic variation within the *Actinomycetaceae* (**Figure 1.4 and Supplementary File 1.5**). Only members of the genus *Mobiluncus* are motile, encoding genes for both flagella and chemotaxis. Nitrate reductase is common in the family, but not encoded by *Actinomyces coleocanis, Actinomyces graevenitzii, Arcanobacterium haemolyticum,* some strains of *Mobiluncus curtisii,* nor *Mobiluncus mulieris,* while nitrite reductase and nitric oxide reductase are found only in the *Actinomyces*. DMSO reductase is found only in *Mobiluncus curtisii*, *Actinomyces urogenitalis, Arcanobacterium haemolyticum,* and *V. cambriense.* Taken together, the *Actinomycetaceae* rely on several different terminal electron acceptors for anaerobic respiration.

Although genes were identified for riboflavin biosynthesis in *V. cambriense*, these genes are not common in the *Actinomycetaceae*. Several species of *Actinomyces* and members of the microbial community of the premature infant in our study (but not the *A. urogenitalis* genome reconstructed from the community) encode the genes for trehalose biosynthesis, suggesting a possible interrelationship between these species and *V. cambriense*, which only encodes the genes for trehalose degradation (**Figure 1.4**). In the *Actinomycetaceae*, pyruvate-ferredoxin oxidoreductase, mentioned previously for its importance in converting pyruvate into acetyl-CoA, is found only in *V. cambriense* and members of *Mobiluncus*. Neuraminidases, which are required for cleaving sialic acids from glycoproteins, human breast milk glycans, and intestinal mucins, are distributed throughout the *Actinomycetaceae*. However, most members of the family are missing transporters for sialic acids, although the presence of the enzymes for their degradation suggests that a currently uncharacterized transporter exists for these species (**Supplementary File 1.6**). Members of the *Actinomycetaceae* and the microbial community in the gut of this infant engage in diverse metabolisms. Clustering based on select metabolic characters shows that *V. cambriense* is more metabolically similar to other *Actinomycetaceae* than to other members identified in the gut of this infant, despite significant metabolic overlap among community members (**Figure 1.4**).

Discussion

Genome reconstructions facilitated prediction of the metabolic roles of individual bacterial members in the context of their community. Applied to the gut microbiome of a premature male infant, the time series abundance information also provided by this method revealed strain-specific dynamics during the third week of life. Comparison of reconstructed genomes to the genomes of isolate strains revealed genomic novelty, even among members of this relatively simple microbial community. However, overall similarities between most reconstructed and reference genomes validated the *de novo* genome binning strategy.

Metabolic analysis revealed a community consisting of facultative anaerobes and obligate (fermentative) anaerobes. The facultative anaerobe *Escherichia coli* was initially dominant in the time series, but was replaced by obligate anaerobes (*Streptococcus anginosus* and *Clostridium butyricum*) during a switch to a community dominated by fermentation-based metabolism. This shift emphasizes the instability of the infant gut in terms of both membership and metabolism, and could be the result of several factors previously observed in the human gut. For example, dominance of species from the family *Enterobacteriaceae* (including *E. coli*) has been associated with either high oxygen levels or the availability of nitrate (a natural byproduct of the host immune response) (Winter et al., 2013a). Thus, this shift could be the result of decreased availability of either nitrate or oxygen, either of which could be depleted by *E. coli* during respiration. Decreased inflammation could also decrease available nitrate and decrease the competitive advantage of *E. coli* over obligate fermenters (Winter et al., 2013b). In terms of the gut environment, succession during early life is driven by the presence of oxygen (Eckburg et al., 2005), and replacement of *E. coli* with obligate fermenters is suggestive of a decrease in oxygen; however, we cannot rule out the hypothesis that this shift in relative abundance is stochastic. Regardless of the mechanism, this represents a dramatic shift in community composition with the potential to affect host metabolism.

Although the abundance of *S. anginosus* and *E. coli* appeared to equilibrate by the end of the time series, the drop in abundance of *C. butyricum* after its initial spike suggests a potential competition between the two obligate anaerobes. Interestingly, *S. anginosus* and *C. butyricum* have different clinical presentations. *S. anginosus* is commonly observed in association with abscess formation (Takahashi et al., 2011), while *C. butyricum* is usually considered a beneficial, butyrate-producing commensal (Woo et al., 2011). In this case, the microbe more likely to be beneficial in the gut environment is outcompeted.

To explore the potential role of a species not commonly observed or previously characterized from the human gut, we manually curated and metabolically analyzed the genome of *Varibaculum cambriense* (strain Dora), resulting in the first genomic sampling of a member of the genus *Varibaculum*. Strains of *V. cambriense* isolated from human cerebral and skin abscesses, intrauterine contraceptive devices, and the human vagina have been used to show that it is an anaerobic, catalase-negative, gram-positive, diphtheroid-shaped bacterium (Hall et al., 2003; Whitman et al., 2012). However, genome analysis revealed additional insight into the metabolic potential of this organism and informed which substrates *V. cambriense* may use for anaerobic respiration. *V. cambriense* is metabolically similar to common gut inhabitants, and is predicted to use various carbon sources, respire anaerobically (using fumarate, nitrate, and DMSO), and produce lactate during fermentation.

Several community members (including *V. cambriense*) are predicted to use myo-inositol as a nutrient source. This is interesting because myo-inositol plays a role in eukaryotic cell messaging and is found in breast milk and infant formula, although it is generally not found in solutions used for intravenous feeding (Pereira et al., 1990). Inositol has been shown to benefit premature infants with respiratory ailments (Hallman et al., 1992), suggesting that microbial degradation of this compound would decrease its health benefits. In contrast, the potential for *V. cambriense* to produce essential vitamins suggests a beneficial contribution by this organism to its human host.

Although this infant was fed fortified breast milk (an abundant source of glycan-bound sialic acids) during the time period studied, only *V. cambriense*, *Streptococcus sp.*, and *Streptococcus parasanguinis*, all low-abundance members of the gut community, have neuraminidases (enzymes that cleave sialic acids) and enzymes for sialic acid degradation. Although the *E. coli* genome does not encode a neuraminidase, it has a sialic acid transporter and degradation machinery. Thus, low-abundance community members may be making sialic acids available to *E. coli*. It has been shown that some pathogenic species incapable of accessing bound sialic acids are able to make use of sialic acids cleaved by commensal organisms to promote their own growth (Ng et al., 2013).

The ability of *V. cambriense* to degrade, but not produce, trehalose suggests a possible dependency on other members of the microbial community able to produce this disaccharide. Furthering community interrelationships, nitrite produced by *V. cambriense* during anaerobic respiration can by further utilized by community members capable of nitrite and nitric oxide reduction.

Conclusion

This study underlines the higher resolution insight that can be obtained using genome-centric metagenomic approaches. Strain-resolved community dynamics revealed two phases in colonization during the third week of life of a premature infant. The phases were distinguishable based on the dominance of either respiratory or fermentation-based metabolism. Comparison of *V. cambriense* with other members of the microbial community revealed similarities with traditional gut inhabitants, while comparisons with other members of the family

Actinomycetaceae illustrate how the metabolic diversity of this family could mislead species-level functional analysis based on 16S rRNA gene sequencing. Analysis of reconstructed genomes enabled strain-specific metabolic potential to be determined for the microbial community, and suggested potential community interdependencies.

Methods

Patient, samples, and sequencing

We studied the colonization of the gut of a male (birth weight 1,205 g) born via Caesarean section during the 31st week of pregnancy to a mother with chronic hypertension and superimposed pre-eclampsia. The patient was born at Comer Children's Hospital at the University of Chicago. He was administered total parenteral nutrition (TPN) soon after admission, but started bolus nasogastric feeding on his second day of life (DOL). The patient was weaned from TPN as he began increasing feeds with fortified breast milk. TPN was discontinued on DOL 6. The patient reached full feeds on DOL 8, continuing to be fed on fortified breast milk. The patient was never intubated but did briefly receive supplemental oxygen. He received antibiotics (ampicillin and gentamicin) only during the first 48 hours of life. Stool samples were collected on the following days of life: 14, 15, 18, 19, and 20. Samples were collected twice daily, except for DOL 14, and stored at −80°C. The patient was discharged in good health on DOL 53.

Microbial DNA was extracted from frozen fecal samples using the QIAamp DNA Stool mini-Kit (Qiagen) with modifications (Zoetendal et al., 2006). DNA was sequenced on an Illumina HiSeq2000 sequencer for 101 cycles from each DNA fragment end using the TruSeq SBS sequencing kit (version 2). Sequencing data were handled with pipeline 1.7 according to the manufacturer's instructions (Illumina, San Diego, CA). The protocol for sample collection and processing was approved by the Institutional Review Board of The University of Chicago (IRB #15895A). All samples were collected with the consent of the infant's mother.

Metagenome assembly, binning, and annotation

Environmental shotgun DNA sequences for all samples were processed and assembled as previously described (Sharon et al., 2012). Sequences were quality trimmed and human DNA was filtered out prior to assembly. Sequences from all samples were co-assembled in a multistep, iterative approach, in which optimal parameters (coverage and k-mer length) for assembly of genomes from specific populations or groups of populations were selected. Velvet (Zerbino and Birney, 2008) was used to assemble the data and the resulting assembly was subjected to quality controls that detect miss-assemblies based on regions of zero insert coverage.

All scaffolds longer than 400 base pairs (bp) were annotated by first predicting open reading frames (ORFs) using the metagenome implementation of Prodigal (Hyatt et al., 2010) and then searching translated ORFs against the UniProt UniRef90 database (Suzek et al., 2007) using USEARCH (Edgar, 2010) with an *E*-value threshold of 0.001. The coverage of each scaffold was determined by mapping reads using BowTie2 (Langmead and Salzberg, 2012) with the parameters -best and -e 200. Coverage was calculated as the total number of sequence bases mapped to a scaffold divided by the length of the scaffold.

Clustering of scaffolds into genome bins was conducted as previously described (Sharon et al., 2012). Specifically, the Databionic implementation of an emergent self-organizing map (ESOM; (Ultsch, 2005)) was used to cluster scaffolds longer than 400 bp based on their time series abundance patterns (after first breaking scaffolds into 1.5 Kbp fragments) (**Supplementary File 1.1**). This allowed us to bin scaffolds into near-complete and partial genomes. Genome bins were assessed in part by coloring fragments clustered on the ESOM based on the best BLAST (Altschul et al., 1990) hit of each scaffold against the NCBI NT database (Pruitt et al., 2012) (**Figure 1.9**). Bins were manually extracted by contouring fields on the ESOM. Genome completeness was determined by comparing the length of each putative genome with the most closely related reference genome and by searching for 26 universal single copy marker genes (**Supplementary File 1.2** and (Raes et al., 2007)).

Manual curation of microbial genome bins

Each genome bin was evaluated based on its size, coverage, and the presence of single copy genes. Single copy genes were used to estimate genome completeness and to determine whether multiple genomes were being clustered into a single bin. Incomplete bins may be the result of several factors: (i) low coverage can prevent an entire genome from being assembled; (ii) sequence variation can cause genomic regions unique to strains to assemble separately, sometimes generating very small contigs that cannot be binned; (iii) strain-specific genomic regions will be binned separately from shared regions if the time series abundance patterns of the strains differ; and (iv) inherent noise in coverage calculations can result in scaffolds representing a genome being placed in separate bins. Bins with similar taxonomic affiliations were evaluated in order to determine whether their scaffolds were split into different bins owing to strain variation or noise in coverage calculations. Such bins were subsequently combined into a single genome bin if they did not contain redundant single copy genes. Other bins with similar taxonomic affiliations but with overall different time series abundance patterns were considered to be the result of strain variation.

The genome assembled and binned for *V. cambriense* was reassembled and manually curated. This involved analyzing the read mapping for the entire metagenome assembly and capturing the reads (along with their pairs) that mapped to scaffolds in the *V. cambriense* bin. These reads were then used in several Velvet assemblies, in which the parameters for k-mer length and expected coverage were altered. Genome size was estimated based on the size of the bin and used to evaluate the Velvet assemblies. The best assembly based on expected length and N50 was checked for miss-assemblies (scaffolds were split at regions with zero insert coverage) and then manually curated.

The genome for *V. cambriense* was manually curated using a suite of in-house scripts designed to extend scaffolds by recruiting paired-reads that extend from existing scaffolds. These paired-reads were assembled independently using Velvet. Their resulting contigs were compared with existing scaffolds, based on their sequence similarity determined by BLAST. Regions of high similarity between the newly assembled contigs and existing scaffolds oftentimes reveal scaffolds that could be combined with one another (Sharon et al., 2012).

Open reading frames for the manually curated *V. cambriense* genome were annotated using an in-house pipeline that includes BLAST-based homology searches against the NCBI NR (Pruitt et al., 2012), KEGG (Kanehisa et al., 2012), UniRef90, and COG (Tatusov et al., 2001) databases, in addition to HMM-based functional domain recognition searches using InterProScan (Zdobnov and Apweiler, 2001). Metabolic analysis of the functional predictions for the *V. cambriense* genome was completed using Pathway Tools (Karp et al., 2002) and KEGG (**Supplementary File 1.4 and Supplementary File 1.5**).

Plasmid and phage genomes

Plasmid genomes were identified by searching for potentially circular scaffolds by computing the Needleman-Wunsch (Needleman and Wunsch, 1970) global alignment for the first and last 100 bp of the scaffold. Scaffolds with high overlap identity were further analyzed by searching those scaffolds for genes indicative of plasmids (such as plasmid replication and maintenance proteins). Putative phage fragments were identified by searching for phage-related genes (such as the capsid or tail fibers) on unbinned scaffolds and on scaffolds binned along with a genome. Phage scaffolds binned with bacterial genomes have significantly higher coverage than the bacterial-associated scaffolds.

Plasmids were associated with individual species in the community by comparing the abundance patterns of each plasmid with each species and by leveraging plasmid phylogenetic annotations. To narrow down the list of possible host organisms for each plasmid, Pearson correlation coefficients were calculated on the abundance pattern of each plasmid compared with each bacterial species. Putative phage fragments were associated with species based on initial ESOM binning, searching for integration sites in reconstructed genomes, and by comparing the relative abundance patterns of phage with bacterial species (assisted by the Pearson correlation coefficient).

Plasmid novelty and diversity were determined by building a phylogenetic tree of plasmid replication proteins. Sequences representative of closely related plasmid replication proteins were acquired by searching the NCBI NR database using BLAST. The amino acid sequences were aligned using MUSCLE (Edgar, 2004) and a phylogenetic tree was reconstructed using FastTree2 (Price et al., 2010) with the Jones-Taylor-Thornton model of amino acid evolution and by assuming a single rate of evolution for each site (known as the CAT approximation) (Stamatakis, 2006). Local support values were calculated with the Shimodaira-Hasegawa test (Shimodaira, 2001) and the tree was formatted using FigTree (tree.bio.ed.ac.uk/software/figtree/).

Coverage and abundance calculations

Coverage was calculated for each scaffold based on mapped reads. Absolute abundance, average coverage, and relative abundance were calculated for reconstructed genomes at each time point, in order to represent changes in microbial community structure (**Supplementary Table 1.2, Supplementary Table 1.3, Supplementary Table 1.4, and Supplementary Table 1.5**). Genome coverage was calculated as the number of bases mapped to the genome divided by the total length of the genome. Relative abundance was calculated for a genome by taking the average coverage of the genome and normalizing it by the sum of the average coverage values for all genomes. Thus, relative abundance is the abundance of a genome taken as a percent of the total abundance of all genomes. Absolute abundance was calculated for each genome by dividing the total number of sequence bases that mapped to the genome by the total number of bases associated with reads used in the assembly. Rank abundance was calculated from the relative abundance of each genome from the combined read mapping of the time points in each phase of community colonization (phases were defined after observing community abundance patterns across the time series). Plots of relative abundance were created using the R plot function r-project.org/).

Microbial community composition based on EMIRGE 16S rRNA genes

EMIRGE (Miller et al., 2011) was used to reconstruct 16S rRNA gene sequences from the metagenomic data (**Supplementary File 1.3**). The closest relatives of each sequence were found by searching a TaxCollector (Giongo et al., 2010) version of the Ribosomal Database Project (RDP) database (Maidak et al., 1997) and the GreenGenes (DeSantis et al., 2006) database using BLAST. Reconstructed 16S rRNA gene sequences were connected with genomes based on several criteria. First, paired-end sequences were used to link scaffolds carrying fragments of 16S rRNA gene sequences to genome scaffolds. Then, coverage and taxonomic information, from both marker genes and for the genome overall, were used to refine associations when paired-read connections were inconclusive. More 16S rRNA gene sequences were reconstructed by EMIRGE than could be represented by genome bins; thus, a subset of low-abundance 16S rRNA gene sequences were assumed to be either incorrectly reconstructed or from very rare community members, and thus were disregarded during further analyses.

Reconstructed 16S rRNA gene sequences were aligned along with sequences from their closest relatives and additional species previously reported to be in the infant gut. Sequences were aligned with PyNAST (Caporaso et al., 2010a) using the GreenGenes alignment of operational taxonomic units (OTUs) classified at 97% sequence similarity as a template. FastTree2 was used to construct the phylogenetic tree using the generalized time-reversible model for nucleotide evolution and the CAT approximation. Local support values were calculated with the Shimodaira-Hasegawa test. The tree was rooted with the 16S rRNA gene sequence for *Halobacterium salinarum* and formatted using FigTree.

Comparison of genomes with reference genomes

Each complete, near-complete, and partial bacterial and plasmid genome was compared to the genome of its closest sequenced relative. Both complete and draft genomes from NCBI were used in the comparison. The most closely related bacterial genomes were determined using reconstructed 16S rRNA gene sequences, ribosomal protein L5, ribosomal protein S15, and hits to other protein sequences in UniRef90. Aligning reconstructed plasmid genomes to all available sequenced plasmid genomes identified their most closely related relatives. Once selected, each genome was compared to its reference genome. The shared amino acid identity between each reconstructed and reference genome was calculated as the average amino acid identity of reciprocal best USEARCH hits between the two genomes (putative orthologs).

Evaluating community oxygen tolerance and respiration capability

Community oxygen tolerance and respiration capacity were evaluated based on the presence of specific genes in the metagenome. To assess the oxygen utilization capacity of the community, all predicted ORFs were searched for cytochrome *c* oxidase, cytochrome *bd* oxidase, and heme-copper cytochrome oxidase genes based on assignments from UniRef90. To evaluate the potential for the community to use various terminal electron acceptors in anaerobic respiration, UniRef90 annotations were searched for the presence of fumarate, trimethylamine *N*-oxide

(TMAO), dimethyl sulfoxide (DMSO), nitrate, nitrite, and nitric oxide reductase genes. To confirm the annotations for these genes in the *V. cambriense* genome, a phylogenetic tree was reconstructed with the amino acid sequences of the putative catalytic subunits for the nitrate reductase and DMSO reductase genes. The tree was reconstructed using MEGA5 (Tamura et al., 2011) to produce a maximum-likelihood phylogeny calculated with 100 bootstrap replicates based on the Jones-Taylor-Thornton model of amino acid evolution. All positions containing alignment gaps and missing data were eliminated based on pairwise sequence comparisons (pairwise deletion option).

Analysis of Human Microbiome Project data

The Human Microbiome Project (HMP) (The NIH HMP Working Group et al., 2009) hosts QIIME (Caporaso et al., 2010b) output files for the HMP 16S rRNA Clinical Production Phase I and the HMP 454 Clinical Production Pilot studies (NCBI SRA projects SRP002395 and SRP002012, respectively), which together consist of over 5,700 samples. The 16S rRNA gene variable region 3 to 5 (V3-5) was sequenced for all samples and the 16S variable region 1 to 3 (V1-3) was sequenced for a subset of 2,911 samples. The OTU abundance matrices were downloaded for each dataset (V3-5 and V1-3) and the abundance of each OTU was calculated as a percent of total reads for each sample. These tables were used to evaluate the relative abundance of *Varibaculum* across samples and body sites in the HMP data collected for healthy human adults.

Comparative genomics

Although there are no previously sequenced genomes for any member of the genus *Varibaculum*, several complete and draft genomes are available for members of the family *Actinomycetaceae*. These genomes, along with the genomes reconstructed from the microbial community of this premature infant, were used in a comparative analysis. Each genome was annotated by finding reciprocal best USEARCH hits between each genome and a subset of the KEGG database containing only prokaryotic protein sequences with KOs (with a minimum bit score of 40 and maximum *E*-value of 0.01). Metabolic functional potential was compared across genomes by identifying gene sequences associated with specific metabolic functions in each genome (see **Supplementary File 1.6** for a complete list of these proteins and their associations). These findings were visualized by normalizing the number of genes identified for each function and then using the R pheatmap library to produce a heatmap clustered using the complete linkage method on a Euclidean distance matrix.

Data Availability

All data have been made publically available and can be accessed through NCBI GenBank Short Read Archive (SRS470507), DDBJ/EMBL/GenBank (AZMA00000000, AZMB00000000, AZMC00000000, AZMD00000000, AZME00000000, AZMF00000000, AZMG00000000, AZMH00000000, AZMI00000000, AZMJ00000000, AZMK00000000, AZML00000000, AZMM00000000), and ggKbase (ggkbase.berkeley.edu/DORA/organisms).

Author Contributions

CTB carried out the binning and genome curation, functional, time series abundance and phylogenetic analysis, and drafted the manuscript. IS assembled the sequence data and contributed to the bioinformatics analysis. BCT performed the functional annotation. CJC assisted with biochemical analysis. MJM oversaw sample collection and handled medical aspects of the research. MJM and JFB designed and oversaw the study and data analysis. All authors contributed to manuscript preparation. All authors read and approved the final manuscript. The authors declare that they have no competing interests.

Acknowledgements

This work was supported by NIH Grant 1R01AI092531-01.

Supplementary Tables

Supplementary Table 1.1 | ESOM bins. Assembly and classification of scaffolds clustered into bins using an ESOM. Each bin was compared to a sequenced microbial genome by identifying orthologs. Orthologs were identified by finding reciprocal best hits from pair-wise protein sequence searches between each bin and the genome of a sequenced relative.

Supplementary Table 1.2 | Absolute abundance of reconstructed genomes. Percent of total reads used in the metagenome assembly that could be mapped to reconstructed genomes at each time point and as a total.

Supplementary Table 1.3 | Coverage of reconstructed genomes. Coverage of reconstructed genomes at each time point and as a total.

Supplementary Table 1.4 | Relative abundance of reconstructed genomes. Relative abundance of reconstructed genomes at each time point and as a total.

Supplementary Table 1.5 | Relative abundance of bacterial genomes. Relative abundance of reconstructed bacterial genomes at each time point and as a total.

Supplementary Files

Supplementary File 1.1 | ESOM data.

Supplementary File 1.2 | Single copy genes in reconstructed genomes.

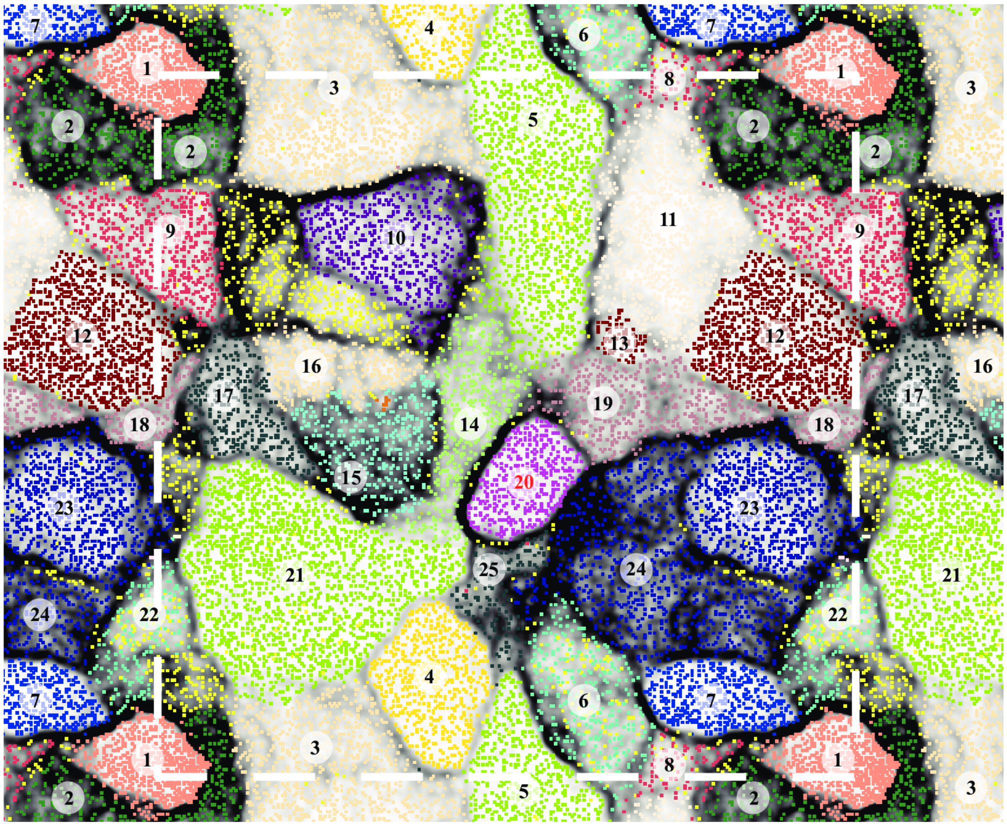
Supplementary File 1.3 | EMIRGE reconstructed 16S rRNA gene sequences.

Supplementary File 1.4 | *V. cambriense* Pathway Tools data.

Supplementary File 1.5 | KEGG annotations.

Supplementary File 1.6 | Metabolic features for comparative genomics.

Figure 1.1 | Emergent self-organizing map (ESOM) binning of the metagenome assembly. ESOM showing the clustering and binning of *de novo* assembled metagenomic data. Each point represents a fragment of an assembled scaffold. Clustering of data points is based on the time series abundance pattern of each assembled scaffold. Dark lines between clusters show definitive separation of genome bins. Colors designate the genome bin for each scaffold fragment.



Unbinned		🟨	🟫	12, 13	***Actinomyces urogenitalis***
Clostridium butyricum	1	🟧	🟥	12	***Actinomyces urogenitalis - Phage***
Leuconostoc sp.	2	🟩	🟩	14	***Escherichia coli - Strain B***
Veillonella sp. - Species A	3, 16	⬜	🟨	14	***Enterococcus faecalis - Plasmid***
Enterococcus faecalis	4	🟨	🟩	14	***Escherichia coli - Strain A/B - Plasmid***
Escherichia coli - Strain A	5, 21	🟩	🟩	15	***Propionibacterium sp.***
Escherichia coli - Strain A - Phage	5	🟩	🟧	16	***Veillonella sp. - Species A - Plasmid***
Staphylococcus sp.	6, 22	🟩	⬜	22	***Veillonella sp. - Species A - Phage***
Staphylococcus sp. - Plasmid A	unbinned	🟦	⬛	17, 25	***Negativicoccus succinicivorans***
Staphylococcus sp. - Plasmid B	unbinned	🟦	🟪	18, 19	***Veillonella sp. - Species B***
Staphylococcus sp. - Phage	unbinned	🟩	⬜	23	***Veillonella sp. - Species B - Phage***
Streptococcus anginosus	7	🟦	🟪	20	***Varibaculum cambriense***
Clostridium bartlettii	8, 9	🟥	🟦	23, 24	***Streptococcus parasanguinis***
Streptococcus sp.	10	🟪	🟪	23	***Streptococcus parasanguinis - Plasmid***
Veillonella dispar	11	⬜			

Figure 1.2 | Phylogeny of plasmid replication protein genes. Plasmid novelty and diversity is shown using plasmid replication proteins. Protein sequences were aligned using MUSCLE and a phylogenetic tree was constructed using FastTree2 with the Jones-Taylor-Thornton model of amino acid evolution and the CAT approximation. Local support values were calculated with the Shimodaira-Hasegawa test. Sequences from plasmids reconstructed from the microbial community are shown in red.

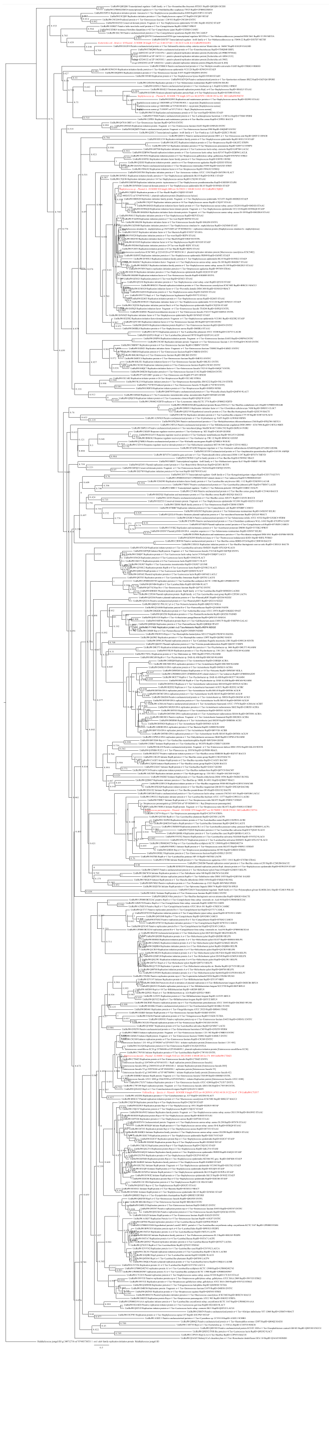


Figure 1.3 | Phylogeny of EMIRGE 16S rRNA genes. Reconstructed 16S rRNA gene sequences were aligned along with sequences from their closest relatives in addition to species previously identified in the infant gut. Sequences were aligned with PyNAST using the GreenGenes alignment of OTUs classified at 97% sequence similarity as a template. The phylogenetic tree was constructed with FastTree2 using the generalized time-reversible model for nucleotide evolution and the CAT approximation. Local support values were calculated with the Shimodaira-Hasegawa test. The tree was rooted with the 16S sequence for *Halobacterium salinarum* and formatted using FigTree. Sequences reconstructed with EMIRGE from the microbial community are shown in blue, and reference sequences with an associated sequenced genome are shown in red.

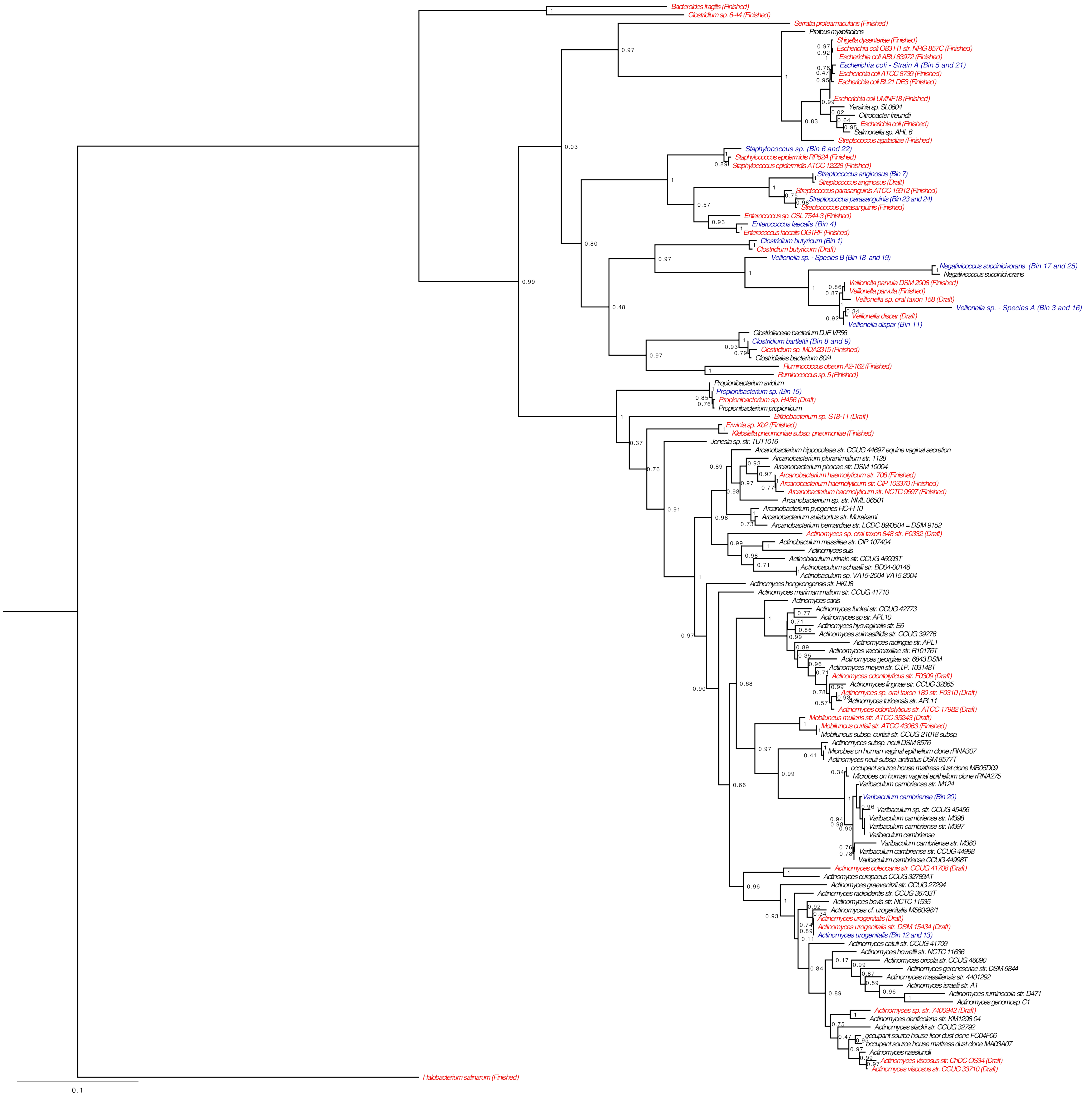
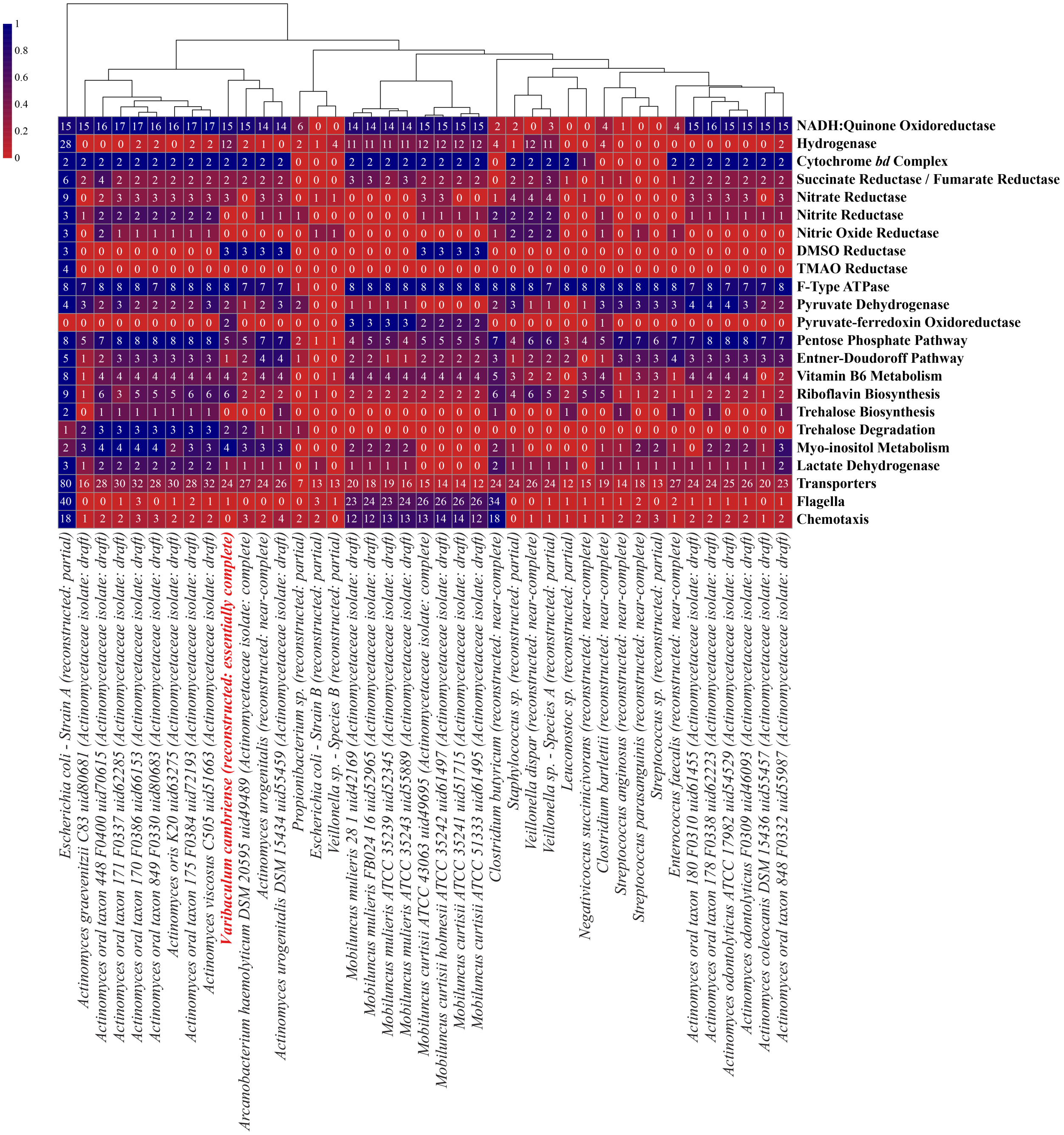

Bacteroides fragilis (Finished)
Clostridium sp. 6-44 (Finished)
Serratia proteamaculans (Finished)
Proteus myxofaciens
Shigella dysenteriae (Finished)
Escherichia coli O83 H1 str. NRG 857C (Finished)
Escherichia coli ABU 83972 (Finished)
Escherichia coli - Strain A (Bin 5 and 21)
Escherichia coli ATCC 8739 (Finished)
Escherichia coli BL21 DE3 (Finished)
Escherichia coli UMNF18 (Finished)
Yersinia sp. SL0604
Citrobacter freundii
Escherichia coli (Finished)
Salmonella sp. AHL 6
Streptococcus agalactiae (Finished)
Staphylococcus sp. (Bin 6 and 22)
Staphylococcus epidermidis RP62A (Finished)
Staphylococcus epidermidis ATCC 12228 (Finished)
Streptococcus anginosus (Bin 7)
Streptococcus anginosus (Draft)
Streptococcus parasanguinis ATCC 15912 (Finished)
Streptococcus parasanguinis (Bin 23 and 24)
Streptococcus parasanguinis (Finished)
Enterococcus sp. CSL 7544-3 (Finished)
Enterococcus faecalis (Bin 4)
Enterococcus faecalis OG1RF (Finished)
Clostridium butyricum (Bin 1)
Clostridium butyricum (Draft)
Veillonella sp. - Species B (Bin 18 and 19)
Negativicoccus succinicivorans (Bin 17 and 25)
Negativicoccus succinicivorans
Veillonella parvula DSM 2008 (Finished)
Veillonella parvula (Finished)
Veillonella sp. oral taxon 158 (Draft)
Veillonella sp. - Species A (Bin 3 and 16)
Veillonella dispar (Draft)
Veillonella dispar (Bin 11)
Clostridiaceae bacterium DJF VP56
Clostridium bartlettii (Bin 8 and 9)
Clostridium sp. MDA2315 (Finished)
Clostridiales bacterium 80/4
Ruminococcus obeum A2-162 (Finished)
Ruminococcus sp. 5 (Finished)
Propionibacterium avidum
Propionibacterium sp. (Bin 15)
Propionibacterium sp. H456 (Draft)
Propionibacterium propionicum
Bifidobacterium sp. S18-11 (Draft)
Erwinia sp. Xb2 (Finished)
Klebsiella pneumoniae subsp. pneumoniae (Finished)
Jonesia sp. str. TUT1016
Arcanobacterium hippocoleae str. CCUG 44697 equine vaginal secretion
Arcanobacterium pluranimalium str. 1128
Arcanobacterium phocae str. DSM 10004
Arcanobacterium haemolyticum str. 708 (Finished)
Arcanobacterium haemolyticum str. CIP 103370 (Finished)
Arcanobacterium haemolyticum str. NCTC 9697 (Finished)
Arcanobacterium sp. str. NML 06501
Arcanobacterium pyogenes HC-H 10
Arcanobacterium suiabortus str. Murakami
Arcanobacterium bernardiae str. LCDC 89/0504 = DSM 9152
Actinomyces sp. oral taxon 848 str. F0332 (Draft)
Actinobaculum massiliae str. CIP 107404
Actinomyces suis
Actinobaculum urinale str. CCUG 46093T
Actinobaculum schaalii str. BD04-00146
Actinobaculum sp. VA15-2004 VA15 2004
Actinomyces hongkongensis str. HKU8
Actinomyces marimammalium str. CCUG 41710
Actinomyces canis
Actinomyces funkei str. CCUG 42773
Actinomyces sp str. APL10
Actinomyces hyovaginalis str. E6
Actinomyces suimastitidis str. CCUG 39276
Actinomyces radingae str. APL1
Actinomyces vaccimaxillae str. R10176T
Actinomyces georgiae str. 6843 DSM
Actinomyces meyeri str. C.I.P. 103148T
Actinomyces odontolyticus str. F0309 (Draft)
Actinomyces lingnae str. CCUG 32865
Actinomyces sp. oral taxon 180 str. F0310 (Draft)
Actinomyces turicensis str. APL11
Actinomyces odontolyticus str. ATCC 17982 (Draft)
Mobiluncus mulieris str. ATCC 35243 (Draft)
Mobiluncus curtisii str. ATCC 43063 (Finished)
Mobiluncus subsp. curtisii str. CCUG 21018 subsp.
Actinomyces subsp. neuii DSM 8576
Microbes on human vaginal epithelium clone rRNA307
Actinomyces neuii subsp. anitratus DSM 8577T
occupant source house mattress dust clone MB05D09
Microbes on human vaginal epithelium clone rRNA275
Varibaculum cambriense str. M124
Varibaculum cambriense (Bin 20)
Varibaculum sp. str. CCUG 45456
Varibaculum cambriense str. M398
Varibaculum cambriense str. M397
Varibaculum cambriense
Varibaculum cambriense str. M380
Varibaculum cambriense str. CCUG 44998
Varibaculum cambriense CCUG 44998T
Actinomyces coleocanis str. CCUG 41708 (Draft)
Actinomyces europaeus CCUG 32789AT
Actinomyces graevenitzii str. CCUG 27294
Actinomyces radicidentis str. CCUG 36733T
Actinomyces bovis str. NCTC 11535
Actinomyces cf. urogenitalis M560/98/1
Actinomyces urogenitalis (Draft)
Actinomyces urogenitalis str. DSM 15434 (Draft)
Actinomyces urogenitalis (Bin 12 and 13)
Actinomyces catuli str. CCUG 41709
Actinomyces howellii str. NCTC 11636
Actinomyces oricola str. CCUG 46090
Actinomyces gerencseriae str. DSM 6844
Actinomyces massiliensis str. 4401292
Actinomyces israelii str. A1
Actinomyces ruminocola str. D471
Actinomyces genomosp. C1
Actinomyces sp. str. 7400942 (Draft)
Actinomyces denticolens str. KM1298 04
Actinomyces slackii str. CCUG 32792
occupant source house floor dust clone FC04F06
occupant source house mattress dust clone MA03A07
Actinomyces naeslundii
Actinomyces viscosus str. ChDC OS34 (Draft)
Actinomyces viscosus str. CCUG 33710 (Draft)
Halobacterium salinarum (Finished)
0.1

Figure 1.4 | Metabolic analysis of reconstructed community and isolate genomes. Genomes reconstructed from the microbial community were compared with each other and with the genomes of cultured isolates previously sequenced for members of the family *Actinomycetaceae*. Each genome was annotated with KEGG and the genes that matched specific metabolic features were counted (**Supplementary File 1.5**). The number of genes identified for each group was normalized across genomes to facilitate coloring and clustering. The number of genes identified for each feature in each genome is shown.



	Escherichia coli - Strain A (reconstructed: partial)	*Actinomyces graevenitzii C83 uid80681 (Actinomycetaceae isolate: draft)*	*Actinomyces oral taxon 448 F0400 uid70615 (Actinomycetaceae isolate: draft)*	*Actinomyces oral taxon 171 F0337 uid62285 (Actinomycetaceae isolate: draft)*	*Actinomyces oral taxon 170 F0386 uid66153 (Actinomycetaceae isolate: draft)*	*Actinomyces oral taxon 849 F0330 uid80683 (Actinomycetaceae isolate: draft)*	*Actinomyces oris K20 uid63275 (Actinomycetaceae isolate: draft)*	*Actinomyces oral taxon 175 F0384 uid72193 (Actinomycetaceae isolate: draft)*	*Actinomyces viscosus C505 uid51663 (Actinomycetaceae isolate: draft)*	***Varibaculum cambriense (reconstructed: essentially complete)***	*Arcanobacterium haemolyticum DSM 20595 uid49489 (Actinomycetaceae isolate: complete)*	*Actinomyces urogenitalis (reconstructed: near-complete)*	*Actinomyces urogenitalis DSM 15434 uid55459 (Actinomycetaceae isolate: draft)*	*Propionibacterium sp. (reconstructed: partial)*	*Escherichia coli - Strain B (reconstructed: partial)*	*Veillonella sp. - Species B (reconstructed: partial)*	*Mobiluncus mulieris 28 1 uid42169 (Actinomycetaceae isolate: draft)*	*Mobiluncus mulieris FB024 16 uid52965 (Actinomycetaceae isolate: draft)*	*Mobiluncus mulieris ATCC 35239 uid52345 (Actinomycetaceae isolate: draft)*	*Mobiluncus mulieris ATCC 35243 uid55889 (Actinomycetaceae isolate: draft)*	*Mobiluncus curtisii ATCC 43063 uid49695 (Actinomycetaceae isolate: complete)*	*Mobiluncus curtisii holmesii ATCC 35242 uid61497 (Actinomycetaceae isolate: draft)*	*Mobiluncus curtisii ATCC 35241 uid51715 (Actinomycetaceae isolate: draft)*	*Mobiluncus curtisii ATCC 51333 uid61495 (Actinomycetaceae isolate: draft)*	*Clostridium butyricum (reconstructed: near-complete)*	*Staphylococcus sp. (reconstructed: partial)*	*Veillonella dispar (reconstructed: near-complete)*	*Veillonella sp. - Species A (reconstructed: partial)*	*Leuconostoc sp. (reconstructed: partial)*	*Negativicoccus succinicivorans (reconstructed: near-complete)*	*Clostridium bartlettii (reconstructed: near-complete)*	*Streptococcus anginosus (reconstructed: near-complete)*	*Streptococcus parasanguinis (reconstructed: near-complete)*	*Streptococcus sp. (reconstructed: partial)*	*Enterococcus faecalis (reconstructed: near-complete)*	*Actinomyces oral taxon 180 F0310 uid61455 (Actinomycetaceae isolate: draft)*	*Actinomyces oral taxon 178 F0338 uid62223 (Actinomycetaceae isolate: draft)*	*Actinomyces odontolyticus ATCC 17982 uid54529 (Actinomycetaceae isolate: draft)*	*Actinomyces odontolyticus F0309 uid46093 (Actinomycetaceae isolate: draft)*	*Actinomyces coleocanis DSM 15436 uid55457 (Actinomycetaceae isolate: draft)*	*Actinomyces oral taxon 848 F0332 uid55987 (Actinomycetaceae isolate: draft)*
NADH:Quinone Oxidoreductase	15	15	16	17	17	16	16	17	17	15	15	14	14	6	0	0	14	14	14	14	15	15	15	15	2	2	0	3	0	0	4	1	0	0	4	15	16	15	15	15	15
Hydrogenase	28	0	0	0	2	2	0	2	2	12	2	1	0	2	1	4	11	11	11	11	12	12	12	12	4	1	12	11	0	0	4	0	0	0	0	0	0	0	0	0	2
Cytochrome *bd* Complex	2	2	2	2	2	2	2	2	2	2	2	2	2	0	0	0	2	2	2	2	2	2	2	2	0	2	2	2	2	1	0	0	0	0	2	2	2	2	2	2	2
Succinate Reductase / Fumarate Reductase	6	2	4	2	2	2	2	2	2	2	2	2	2	0	0	0	3	3	2	3	2	2	2	2	0	2	2	3	1	0	1	1	0	0	1	2	2	2	2	2	2
Nitrate Reductase	9	0	2	3	3	3	3	3	3	3	0	3	3	0	1	1	0	0	0	0	3	3	0	0	1	4	4	4	0	1	0	0	0	0	0	3	3	3	3	0	3
Nitrite Reductase	3	1	2	2	2	2	2	2	2	0	0	1	1	1	0	0	0	0	0	0	1	1	1	1	2	2	2	2	0	0	1	0	0	0	0	1	1	1	1	1	1
Nitric Oxide Reductase	3	0	2	1	1	1	1	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	2	2	2	0	0	1	0	1	0	1	0	0	0	0	0	0
DMSO Reductase	3	0	0	0	0	0	0	0	0	3	3	3	3	0	0	0	0	0	0	0	3	3	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TMAO Reductase	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F-Type ATPase	8	7	8	8	8	7	8	8	8	8	7	7	7	1	0	0	8	8	8	8	8	8	8	8	8	8	8	7	8	8	8	8	8	8	8	7	8	7	7	7	8
Pyruvate Dehydrogenase	4	3	2	3	2	2	2	2	3	2	1	2	3	2	0	0	1	1	1	1	0	0	0	0	2	3	1	1	0	1	3	3	3	3	3	4	4	4	3	2	2
Pyruvate-ferredoxin Oxidoreductase	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	3	3	3	3	2	2	2	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Pentose Phosphate Pathway	8	5	7	8	8	8	8	8	8	5	5	7	7	2	1	1	4	5	5	4	5	5	5	5	7	4	6	6	3	4	5	7	7	6	7	7	8	8	8	7	7
Entner-Doudoroff Pathway	5	1	2	3	3	3	3	3	3	1	2	4	4	1	0	0	2	2	1	1	2	2	2	2	3	1	2	2	2	0	1	3	3	3	4	3	3	3	3	3	3
Vitamin B6 Metabolism	8	1	4	4	4	4	4	4	4	4	2	4	4	0	0	1	4	4	4	4	4	4	4	4	5	3	2	2	0	3	4	1	3	3	1	4	4	4	4	0	2
Riboflavin Biosynthesis	9	1	6	3	5	5	5	6	6	6	2	2	2	0	1	0	2	2	2	2	2	2	2	2	6	4	6	5	2	5	5	1	1	2	1	1	2	2	2	1	2
Trehalose Biosynthesis	2	0	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	1
Trehalose Degradation	1	2	3	3	3	3	3	3	3	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Myo-inositol Metabolism	2	3	4	4	4	4	2	3	3	4	3	3	3	0	0	0	2	2	2	2	0	0	0	0	2	1	0	0	0	0	1	1	2	2	1	0	2	2	2	1	3
Lactate Dehydrogenase	3	1	2	2	2	2	2	2	2	1	1	1	1	0	1	0	1	1	1	1	0	0	0	0	2	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	2
Transporters	80	16	28	30	32	28	30	28	32	24	27	24	26	7	13	13	20	18	19	16	15	14	14	12	24	24	26	24	12	15	19	14	18	13	27	24	24	25	26	20	23
Flagella	40	0	0	1	3	0	1	2	1	1	0	0	0	0	3	1	23	24	23	24	26	26	26	26	34	0	0	1	1	1	1	1	2	0	2	1	1	1	1	0	1
Chemotaxis	18	1	2	2	3	3	2	2	2	0	3	2	4	2	2	0	12	12	13	13	13	14	14	12	18	0	1	1	1	1	1	2	2	3	1	2	2	2	2	1	2

Figure 1.5 | Phylogenetic analysis of the catalytic subunits of the dimethyl sulfoxide (DMSO) reductase superfamily. Proteins assigned to the DMSO reductase superfamily in this study are indicated by red stars.

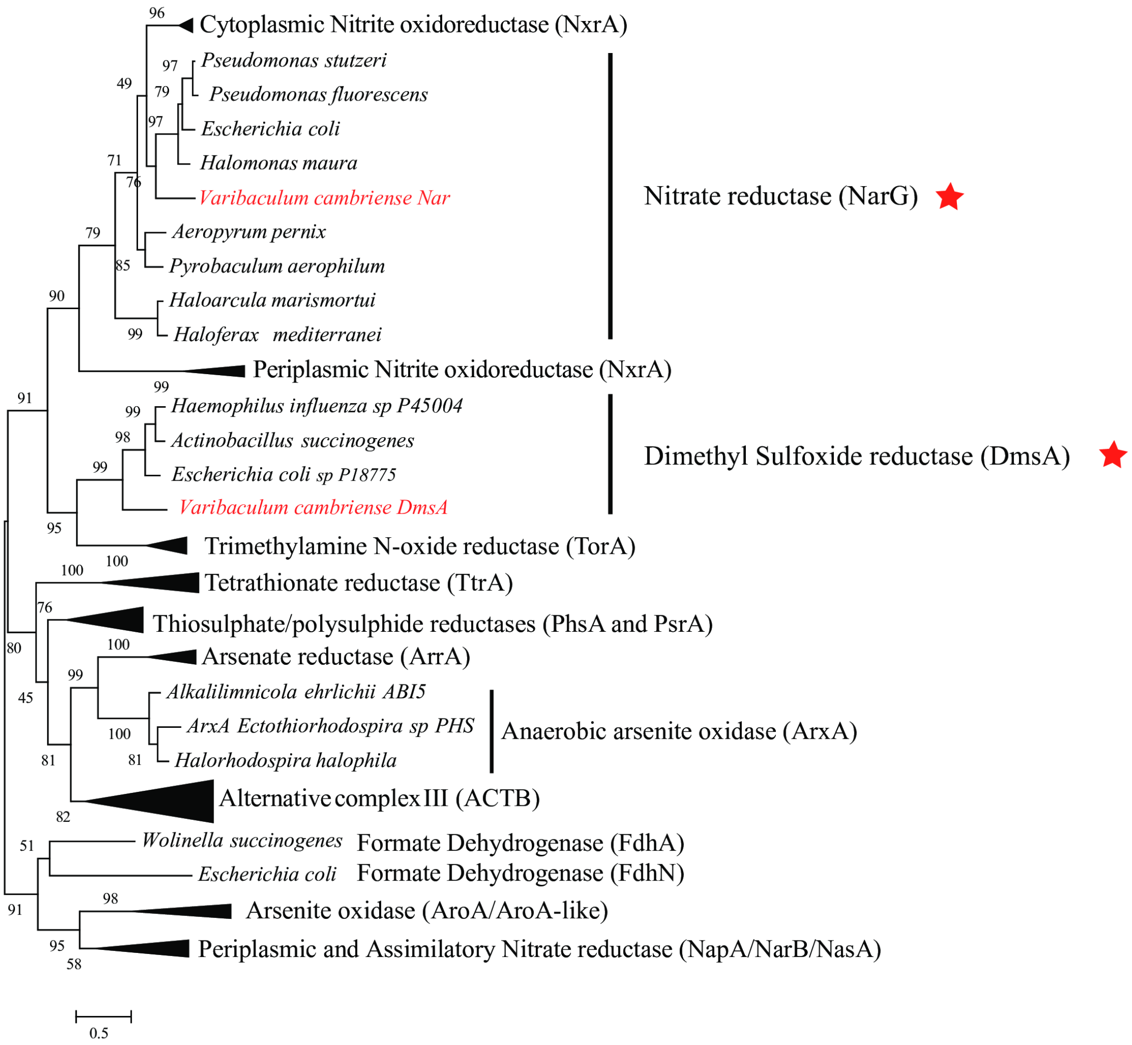

96
Cytoplasmic Nitrite oxidoreductase (NxrA)
97
Pseudomonas stutzeri
49
79
Pseudomonas fluorescens
97
Escherichia coli
71
Halomonas maura
76
Varibaculum cambriense Nar
Nitrate reductase (NarG) ★
79
Aeropyrum pernix
85
Pyrobaculum aerophilum
90
Haloarcula marismortui
99
Haloferax mediterranei
Periplasmic Nitrite oxidoreductase (NxrA)
91
99
99
Haemophilus influenza sp P45004
98
Actinobacillus succinogenes
99
Escherichia coli sp P18775
Dimethyl Sulfoxide reductase (DmsA) ★
Varibaculum cambriense DmsA
95
Trimethylamine N-oxide reductase (TorA)
100
100
Tetrathionate reductase (TtrA)
76
Thiosulphate/polysulphide reductases (PhsA and PsrA)
80
100
Arsenate reductase (ArrA)
99
45
Alkalilimnicola ehrlichii ABI5
100
ArxA Ectothiorhodospira sp PHS
Anaerobic arsenite oxidase (ArxA)
81
81
Halorhodospira halophila
Alternative complex III (ACTB)
82
51
Wolinella succinogenes
Formate Dehydrogenase (FdhA)
Escherichia coli
Formate Dehydrogenase (FdhN)
98
91
Arsenite oxidase (AroA/AroA-like)
95
Periplasmic and Assimilatory Nitrate reductase (NapA/NarB/NasA)
58
0.5

Figure 1.6 | Relative abundance of bacterial species over time. Relative abundances were calculated for bacterial species at nine different time points during the third week of life of a premature male infant. **(a)** Shows dominant taxa and **(b)** shows low-abundance species across the time series. During this period, the colonization process is defined by two distinct phases based on the dominance of either facultative (phase 1) or obligate (phase 2) anaerobes.

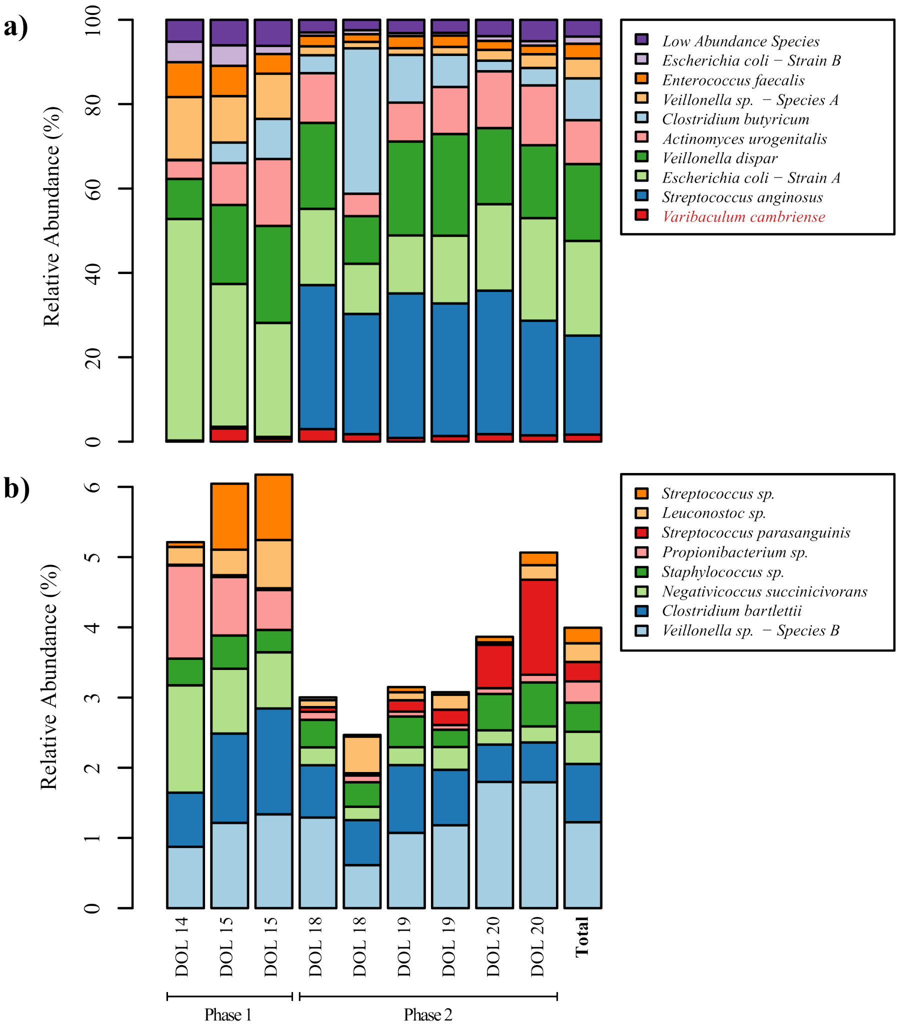


Figure 1.7 | Rank abundance of bacterial species during phases of colonization. Rank abundance was determined from the relative abundance of each species during each phase of microbial colonization. Taxonomic identification and metabolic analysis was completed based on genome reconstructions from the shotgun-sequenced microbial community. The colonization process is broken into two distinct phases defined by the dominance of either **(a)** facultative anaerobes during phase one or **(b)** obligate anaerobes during phase two.

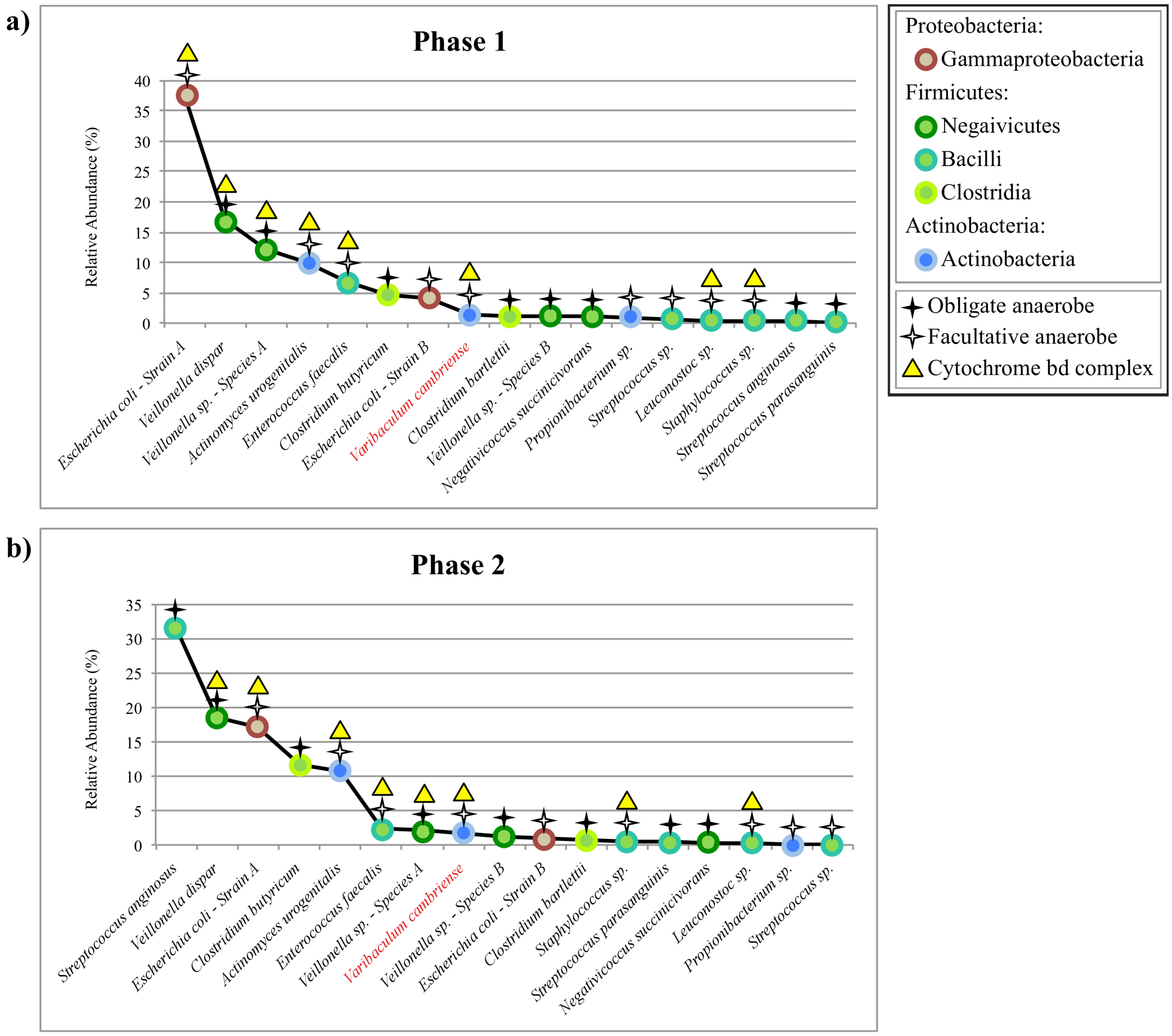


Figure 1.8 | Relative abundance of organisms from the genus *Varibaculum* in Human Microbiome Project samples. The relative abundance of *Varibaculum* organisms was determined for each sample analyzed as part of the Human Microbiome Project. These samples were characterized by sequencing the V1-3 (a) and V3-5 (b) regions of the 16S rRNA gene.

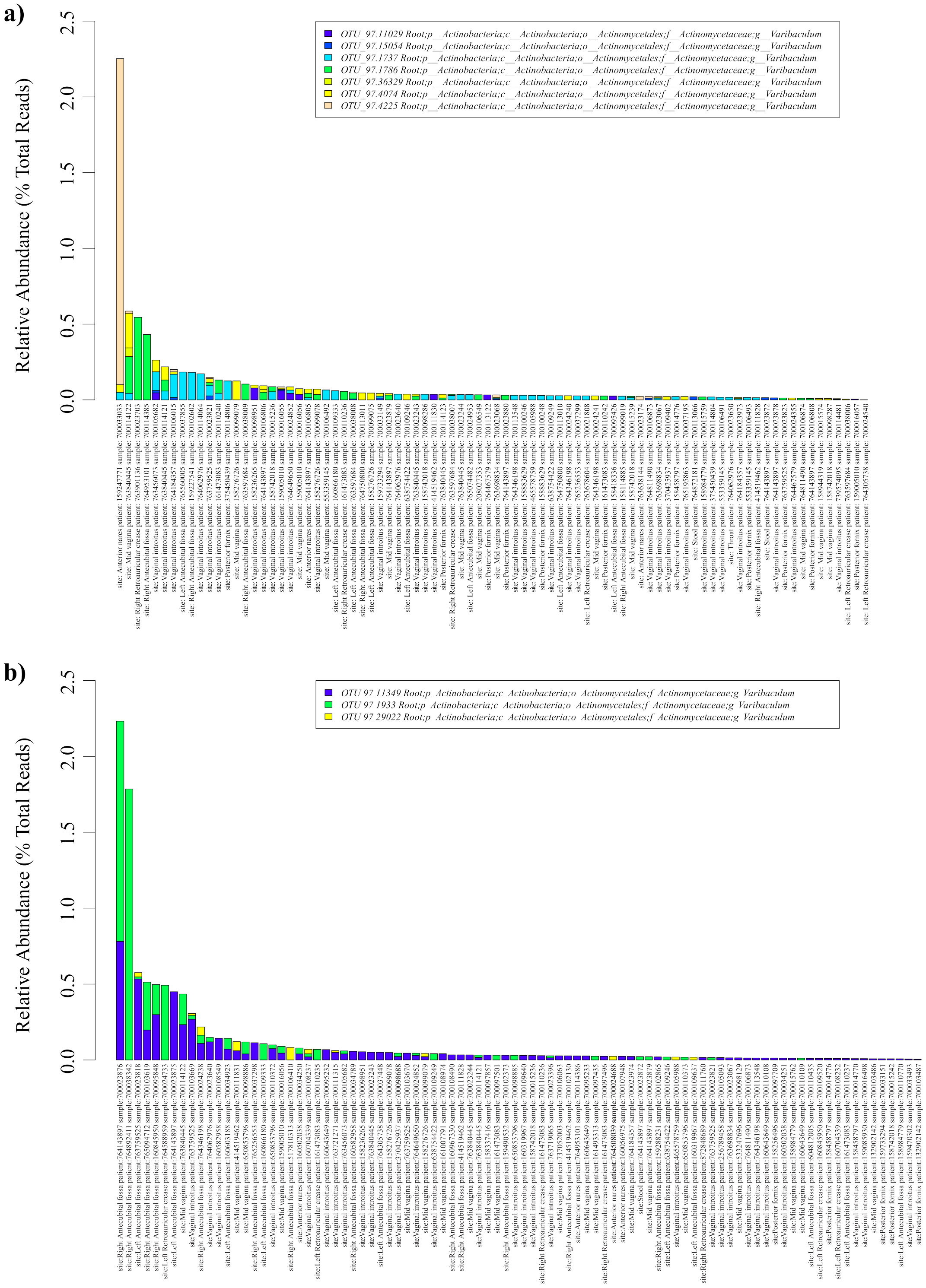
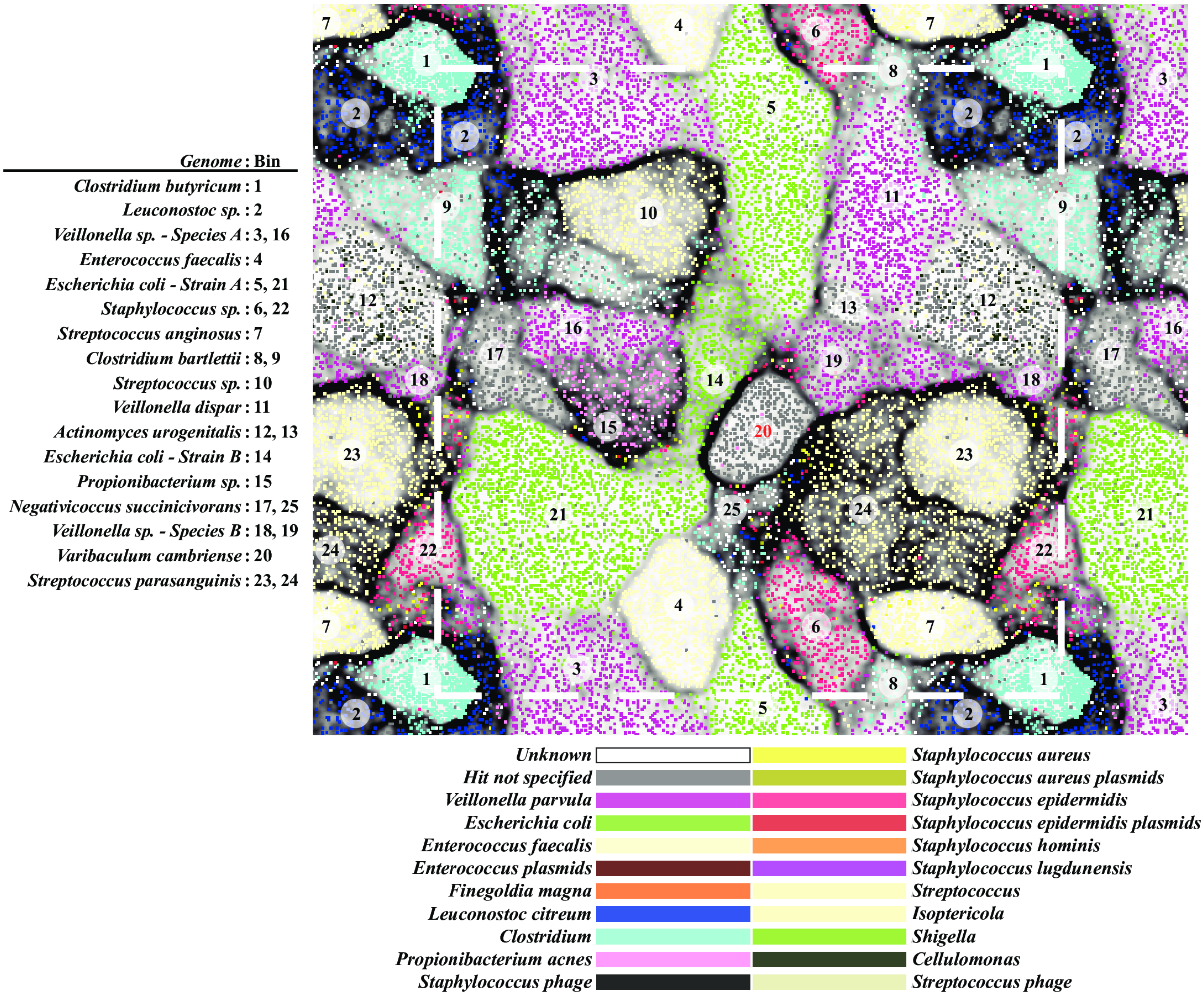


Figure 1.9 | ESOM clustering of *de novo* assembled metagenomic data. Each point represents a fragment of an assembled scaffold. Clustering of data points is based on the time series abundance pattern of each assembled scaffold. Dark lines between clusters show definitive separation of genome bins. Data points are colored based on the best BLAST hit of each scaffold compared against the NCBI NT database (coloring is independent of the assembly and binning).



Genome	Bin
Clostridium butyricum	1
Leuconostoc sp.	2
Veillonella sp. - Species A	3, 16
Enterococcus faecalis	4
Escherichia coli - Strain A	5, 21
Staphylococcus sp.	6, 22
Streptococcus anginosus	7
Clostridium bartlettii	8, 9
Streptococcus sp.	10
Veillonella dispar	11
Actinomyces urogenitalis	12, 13
Escherichia coli - Strain B	14
Propionibacterium sp.	15
Negativicoccus succinicivorans	17, 25
Veillonella sp. - Species B	18, 19
Varibaculum cambriense	20
Streptococcus parasanguinis	23, 24

Table 1.1 | Assessment of genomes reconstructed from the shotgun-sequenced microbial community: assembly, binning, phylogeny, and genome completeness.

Genome	Near-complete genome	Phylum	Bin	Genome size (bp)	Relative abundance (%)	N50 (bp)	ORFs	Percent single copy genes
Escherichia coli - strain A/B - plasmid	Yes		14	3,225	41.67	3,225	5	
Escherichia coli - strain A - phage			5	3,920	16.57	3,920	4	
Enterococcus faecalis - plasmid	Yes		14	5,231	5.98	5,231	5	
Streptococcus anginosus	Yes	Firmicutes	7	2,108,491	5.14	537,826	2,252	100
Escherichia coli - strain A		Proteobacteria	5, 21	5,662,200	4.91	5,304	7,405	48.1
Actinomyces urogenitalis - phage			12	2,996	4.85	2,996	4	
Veillonella dispar	Yes	Firmicutes	11	2,445,194	3.99	53,688	2,693	100
Veillonella sp. - species A - phage			22	20,453	3.49	20,453	32	
Veillonella sp. - species B - phage			23	16,533	2.43	16,533	30	
Actinomyces urogenitalis	Yes	Actinobacteria	12, 13	2,604,957	2.28	3,337	3,035	92.6
Clostridium butyricum	Yes	Firmicutes	1	4,350,784	2.17	103,127	4,094	100
Veillonella sp. - species A - plasmid	Yes		16	47,631	1.2	47,631	56	
Veillonella sp. - species A		Firmicutes	3, 16	2,664,763	1.03	4,180	3,427	37
Enterococcus faecalis	Yes	Firmicutes	4	2,960,721	0.76	235,714	2,906	100
Staphylococcus sp. - plasmid B	Yes		Unbinned	2,539	0.65	2,539	2	
Staphylococcus sp. - plasmid A	Yes		Unbinned	2,556	0.54	2,556	3	
Staphylococcus sp. - phage			Unbinned	2,423	0.42	2,423	5	
Escherichia coli - strain B		Proteobacteria	14	633,084	0.37	4,041	873	3.7
Varibaculum cambriense	Yes	Actinobacteria	20	2,247,641	0.37	240,417	1,954	100
Streptococcus parasanguinis - plasmid	Yes		23	8,975	0.31	8,975	7	
Veillonella sp. - species B		Firmicutes	18, 19	639,180	0.27	2,044	879	29.6
Clostridium bartlettii	Yes	Firmicutes	8, 9	2,685,446	0.18	12,095	2,633	92.6
Negativicoccus succinicivorans	Yes	Negativicoccus	17, 25	1,508,898	0.1	15,236	1,686	100
Staphylococcus sp.		Firmicutes	6, 22	1,509,765	0.09	9,200	1,634	33.3
Propionibacterium sp.		Propionibacterium	15	336,576	0.07	1,117	565	18.5
Streptococcus parasanguinis	Yes	Firmicutes	23, 24	2,822,032	0.06	3,494	3,647	77.8
Leuconostoc sp.		Firmicutes	2	566,369	0.06	1,603	874	25.9
Streptococcus sp.		Firmicutes	10	1,915,777	0.05	11,008	2,114	55.6

Table 1.2 | Comparison of reconstructed genomes and 16S rRNA gene sequences with those from reference databases. Reconstructed genomes were compared with the genomes of isolate strains. Reconstructed 16S rRNA genes were searched against sequences in the RDP and GreenGenes databases to aid classification.

Genome	Genome size (bp)	Closest relative with sequenced genome	Closest relative genome size (bp)	% ORFs orthologous	Average % amino acid ID of orthologs	EMIRGE	EMIRGE 16S % ID
Escherichia coli - strain A/B - plasmid	3,225	*Salmonella enterica subsp. enterica serovar Newport str.*	684	20	93.9		
Escherichia coli - strain A - phage	3,920						
Enterococcus faecalis - plasmid	5,231	*Enterococcus faecalis 62 plasmid*	1,295	100	100		
Streptococcus anginosus	2,108,491	*Streptococcus anginosus 1 2 62CV uid62163*	1,821,055	67.7	95.2	*S. anginosus*	99.8
Escherichia coli - strain A	5,662,200	*Escherichia coli S88 uid62979*	5,032,268	47.7	97.5	*E. coli O83:H1 str. NRG 85*	99.9
Actinomyces urogenitalis - phage	2,996						
Veillonella dispar	2,445,194	*Veillonella dispar ATCC 17748*	2,118,767	60	91.6	*V. dispar*	99.3
Veillonella sp. - species A - phage	20,453						
Veillonella sp. - species B - phage	16,533						
Actinomyces urogenitalis	2,604,957	*Actinomyces urogenitalis DSM 15434*	2,702,812	67.1	98.7	*A. urogenitalis*	99.9
Clostridium butyricum	4,350,784	*Clostridium butyricum 5521*	4,540,699	73.5	96.5	*C. butyricum*	99.3
Veillonella sp. - species A - plasmid	47,631	*Caldicellulosiruptor kristjanssonii 177R1B plasmid*	3,674	3.6	38.8		
Veillonella sp. - species A	2,664,763	*Veillonella dispar ATCC 17748*	2,118,767	47.1	95.6	*Veillonella sp. oral taxon 158*	91.2
Enterococcus faecalis	2,960,721	*Enterococcus faecalis OG1RF*	2,739,625	79.1	98.8	*E. faecalis OG1RF*	98.7
Staphylococcus sp. - plasmid B	2,539	*Staphylococcus haemolyticus JCSC1435 plasmid*	402	100	99.8		
Staphylococcus sp. - plasmid A	2,556	*Macrococcus caseolyticus JCSC5402 plasmid*	997	33.3	80.3		
Staphylococcus sp. - phage	2,423						
Escherichia coli - strain B	633,084	*Escherichia coli S88 uid62979*	5,032,268	52.7	82.3		
Varibaculum cambriense	2,247,641	*Mobiluncus mulieris ATCC 35239*	2,533,633	56	53.9	*V. cambriense*	99.2
Streptococcus parasanguinis - plasmid	8,975	*Enterococcus faecium Aus0004 plasmid*	1,192	14.3	32.4		
Veillonella sp. - species B	639,180	*Veillonella dispar ATCC 17748*	2,118,767	62.5	86.7	*V. parvula DSM 2008*	90.3
Clostridium bartlettii	2,685,446	*Clostridium bartlettii DSM 16795*	2,972,256	81.7	98.1	*Clostridium sp. MDA2315*	99.4
Negativicoccus succinicivorans	1,508,898	*Bacillus coagulans 36D1*	3,552,226	45	43.4	*N. succinicivorans*	98
Staphylococcus sp.	1,509,765	*Staphylococcus epidermidis ATCC 12228*	2,499,279	78.3	98.4	*S. epidermidis ATCC 12228*	96.2
Propionibacterium sp.	336,576	*Propionibacterium 5 U 42AFAA*	2,532,807	46.4	82.4	*Propionibacterium sp. H456*	99.4
Streptococcus parasanguinis	2,822,032	*Streptococcus parasanguinis ATCC 15912*	2,153,652	45	94.3	*S. parasanguinis*	98.9
Leuconostoc sp.	566,369	*Leuconostoc citreum KM20*	1,796,284	66.7	96		
Streptococcus sp.	1,915,777	*Streptococcus M334*	2,207,013	67.6	93.9		

Table 1.3 | Metabolism of bacterial members of the microbial community. Presence (*) and absence of components required for anaerobic respiration and the predicted oxygen requirement of each member of the bacterial community.

Genome	Predicted oxygen requirement	NADH: quinone oxidoreductase	Cytochrome *bd* complex	Fumarate reductase	TMAO reductase	DMSO reductase	Nitrate reductase	Nitrite reductase	Nitric oxide reductase
Streptococcus anginosus	obligate anaerobe								
Escherichia coli - strain A	facultative anaerobe	*	*	*	*	*	*	*	*
Veillonella dispar	obligate anaerobe		*	*			*	*	*
Actinomyces urogenitalis	facultative anaerobe	*	*	*		*	*	*	
Clostridium butyricum	obligate anaerobe					*	*	*	
Veillonella sp. - species A	obligate anaerobe		*	*			*	*	*
Enterococcus faecalis	facultative anaerobe		*						
Escherichia coli - strain B	facultative anaerobe						*		*
Varibaculum cambriense	facultative anaerobe	*	*	*		*	*		
Veillonella sp. - species B	obligate anaerobe						*		*
Clostridium bartlettii	obligate anaerobe								
Negativicoccus succinicivorans	obligate anaerobe						*		
Staphylococcus sp.	facultative anaerobe		*				*	*	*
Propionibacterium sp.	facultative anaerobe	*		*				*	
Streptococcus parasanguinis	obligate anaerobe								*
Leuconostoc sp.	facultative anaerobe		*						
Streptococcus sp.	facultative anaerobe								

Chapter 2

Unusual biology across a group comprising more than 15% of domain Bacteria

C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, and J. F. Banfield

Published in *Nature*, July 2015.

Abstract

A prominent feature of the bacterial domain is a radiation of major lineages that are defined as candidate phyla (CP) because they lack isolated representatives. Bacteria from these phyla occur in diverse environments (Harris et al., 2004) and are suggested to mediate carbon and hydrogen cycles (Wrighton et al., 2012). Genomic analyses of a few representatives suggested that metabolic limitations have prevented their cultivation (Albertsen et al., 2013; Kantor et al., 2013; Rinke et al., 2013; Wrighton et al., 2012; 2014). We reconstructed 8 complete and 789 draft genomes from bacteria representing >35 phyla and documented features that consistently distinguish these organisms from other bacteria. We infer that this group, which may comprise >15% of the bacterial domain, has shared evolutionary history and describe it as the Candidate Phyla Radiation (CPR; **Figure 2.1**). All CPR genomes are small and most lack numerous biosynthetic pathways. Due to divergent 16S rRNA gene sequences, 50-100% of organisms sampled from specific phyla would evade detection in typical cultivation-independent surveys. CPR organisms often have self-splicing introns and proteins encoded within their rRNA genes, a feature rarely reported in bacteria. Further, they have unusual ribosome compositions. All are missing a ribosomal protein often absent in symbionts, and specific lineages are missing ribosomal proteins and biogenesis factors considered universal in bacteria. This implies different ribosome structures and biogenesis mechanisms, and underlines unusual biology across a large part of the bacterial domain.

Introduction

We sampled microbial communities from an aquifer adjacent to the Colorado River near the town of Rifle, CO, USA in 2011. Groundwater was filtered through a 1.2 µm pre-filter and cells collected on serial 0.2 and 0.1 µm filters (**Figure 2.2**). Post-0.2 µm filtrates were targeted because CPR bacteria were predicted to have ultra-small cells based on their small genomes (Wrighton et al., 2012). Groundwater was sampled prior to and during an acetate amendment experiment that reproduced conditions that generated the first genomes from CPR bacteria (Castelle et al., 2015; Luef et al., 2015; Wrighton et al., 2012; 2014) (**Supplementary Table 2.1**). Total DNA and RNA were extracted from filters and sequenced. We obtained 224 Gbp of paired-end metagenomic sequence from 12 samples (150 bp reads, 6 time points, 0.2 and 0.1 µm filters; **Supplementary Table 2.2)**. Sequence assembly generated 3.9 Gbp of contiguous sequences ≥5 Kbp. We also obtained 181 Gbp of metatranscriptomic sequence from six samples (50 bp reads, 0.2 μm filters).

Results and Discussion

Assembled scaffolds were binned into genomes based on their GC content, DNA sequence coverage, abundance pattern across samples, and taxonomic affiliation (binning was validated with a tetranucleotide sequence signature method; **Figure 2.3**). Overall, we reconstructed >1,750 genome bins from microbial community sequence data. Here, we focus on genomes from CPR bacteria and TM6, which represented >60% of bins. Included in our analyses of the CPR are members of the Parcubacteria (OD1), Microgenomates (OP11), WWE3, Berkelbacteria (ACD58), Saccharibacteria (TM7), WS6, Peregrinibacteria (PER), Kazan, and previously unrecognized lineages (CPR1 through 3). In total, 789 draft-quality (≥50% complete) genomes were reconstructed (**Table 2.1**). We manually curated eight genomes to completion: the first three from Microgenomates, two from Parcubacteria, one each from Kazan and Berkelbacteria, and an additional genome from Saccharibacteria. All complete and draft genomes are small and most are <1 Mbp in length (**Supplementary Table 2.3 and Supplementary Table 2.4**).

In total, 1,543 bacterial 16S rRNA genes ≥800 bp were assembled and curated to eliminate assembly errors (713 sequences clustered at 97% identity; **Supplementary File 2.1**). Relative abundance measurements show enrichment of CPR organisms in small-cell filtrates, suggesting they have ultra-small cells (**Figure 2.4**). This finding is supported by a recent microscopy study (Luef et al., 2015). Surprisingly, 31% of 16S rRNA genes encoded a large (≥10 bp) insertion sequence (max: 2,004 bp, mean: 519 bp, standard deviation: 372 bp; **Supplementary Table 2.5**). Insertions are found in phylogenetically diverse members of CPR phyla (**Figure 2.1, Supplementary File 2.2, and Supplementary File 2.3**). Insertion sites are clustered in several distinct locations on the 16S rRNA gene, both in variable and conserved regions (**Figure 2.5**). Most insertions ≥500 bp encode a catalytic RNA intron (group I or II) and/or an open reading frame (ORF), suggesting they are self-splicing. Encoded proteins frequently belong to families of homing endonucleases (LAGLIDAG 1-3 and GIY-YIG). However, 25% are not similar to known protein families or to each other. These may represent novel endonucleases or may no longer be functional, since loss of function is common in homing endonucleases (Burt and Koufopanou, 2004).

Four members of the Thiotrichaceae are the only bacteria known to have self-splicing introns within their 16S rRNA genes (Salman et al., 2012). An extensive search for insertions in genes from our study and the Silva database (Quast et al., 2013) suggests their rarity in bacteria outside the CPR (**Figure 2.6 and Supplementary Table 2.6**). Especially rare are insertions encoding predicted self-splicing introns and/or ORFs. However, these genes need not be functional if the genome encodes additional, insertion-free copies. Importantly, all complete CPR genomes have only one copy of the 16S rRNA gene (this study and others (Albertsen et al., 2013; Kantor et al., 2013)). Sequencing coverage analysis of draft genomes further indicates that a single copy is typical for these lineages (**Figure 2.7 and Supplementary Table 2.7**).

Mapping metatranscriptomic sequences to assembled 16S rRNA genes showed that insertions are not retained in transcribed RNAs, and are likely rapidly degraded (**Supplementary Table**

2.8). However, it is possible that spliced sequences are rendered inaccessible to sequencing after hybridizing, circularizing, or, in some cases, due to their small size. Regardless of their fate, splicing establishes these insertions as introns. Self-splicing is expected if insertions encode a catalytic RNA intron; however, splicing could also occur via an RNase III-mediated mechanism (Evguenieva-Hackenberg, 2005). Several genes contain multiple introns. For example, one of the complete genomes we obtained encodes a 16S rRNA gene with four introns (**Figure 2.8**).

CPR bacteria frequently encode introns in 23S rRNA genes with features similar to those in 16S rRNA genes (**Figure 2.9, Supplementary Table 2.5, Supplementary Table 2.8, and Supplementary File 2.4)**. However, these introns and encoded proteins share little sequence similarity with one another (**Supplementary Table 2.9**). It remains a puzzle as to why introns in critical, highly transcribed rRNA genes do not make these organisms uncompetitive, as their transcription is costly, even though formation of nonfunctional ribosomes is avoided by splicing.

Insertions in rRNA genes are found in *Coxiella* and Rickettsiales-lineage endosymbionts (Baker et al., 2003; Raghavan et al., 2008). Interestingly, one member of the Parcubacteria, *Candidatus Sonnebornia yantaiensis*, is intracellular (Gong et al., 2014), but does not contain an insertion in its 16S rRNA gene (**Figure 2.1**). However, there is no evidence that an intercellular lifestyle is typical across CPR lineages, although a strong dependence on other community members is likely (Kantor et al., 2013; Wrighton et al., 2014).

Metagenomic analyses are PCR-independent and, therefore, not biased by primers designed based on expectations of sequence conservation. As a consequence, our sampling indicated that many CPR organisms would evade detection by 16S rRNA gene amplicon surveys. Primer binding analysis showed that primers extensively used in microbial surveys (515F and 806R (Caporaso et al., 2012)) would likely not bind to 16S rRNA genes of ~50% of Microgenomates, ~50% of Saccharibacteria, 60% of WWE3, and 100% of WS6 sequences sampled here (**Figure 2.10**). In fact, these primers would likely miss ~20% of all bacteria detected in this study, including organisms outside the CPR. Further, introns in these genes would interfere with amplification, both because they occur in regions targeted by primers, and as they increase the length of the target sequence. In addition to being excluded during size-selection of amplicons, intron-containing genes are less likely to amplify compared with shorter, intron-free genes (Salman et al., 2012). Thus, several barriers have prevented identification of many CPR bacteria.

Removal of introns from 16S rRNA gene sequences, followed by structural alignment (Nawrocki, 2009), was critical to establishing a reliable phylogeny. The new phylogenetic analysis shows that the CPR is monophyletic (**Figure 2.1**), a result also evident in concatenated ribosomal protein trees (**Supplementary File 2.2**), and seen in prior analyses (Baker and Dick; Kantor et al., 2013; Quast et al., 2013; Rinke et al., 2013; Wrighton et al., 2012; 2014). Phylogenetic analysis defined 35 phyla within the CPR (see below), which encompasses a proposed superphylum “Patescibacteria,” previously suggested to include just three phyla (Rinke et al., 2013).

Recently, Yarza *et al.* suggested the existence of ~1,500 bacterial phyla using a 75% 16S rRNA gene sequence identity threshold (Yarza et al., 2014). This contrasts with the current view, which includes 29 established phyla and ~60 CP. Using the Yarza *et al.* definition, we estimate that the

CPR consists of >250 phyla (**Figure 2.1 and Supplementary File 2.2**). With the addition of >550 Mbp of CPR genome sequence, there is sufficient sampling to clearly resolve 14 phyla within the Parcubacteria and 11 phyla within the Microgenomates, which have sufficient sequence divergence to account for >120 and >60 phyla, respectively. We propose that these 25 phyla be recognized because i) complete and/or draft genomes are available, ii) they are monophyletic lineages in both 16S rRNA gene and concatenated ribosomal protein trees, and iii) they pass an approximate 75% 16S rRNA gene sequence identity threshold. Importantly, regardless whether previous phyla designations or new criteria (Yarza et al., 2014) are used, the CPR comprises >15% of domain Bacteria.

A striking finding from analysis of complete and draft genomes (see statistical assessment in Methods) is unusual ribosome composition in CPR bacteria. All CPR and TM6 bacteria lack ribosomal protein L30 (rpL30; **Table 2.1, Figure 2.11, Supplementary Table 2.10, and Supplementary File 2.5**). Apparently non-essential in bacteria (Akanuma et al., 2012), this protein is commonly present except in some symbionts, parasites, Cyanobacteria, and throughout the Planctomycetes-Verrucomicrobia-Chlamydiae (PVC) superphylum (Lagkouvardos et al., 2014; Lecompte, 2002; Yutin et al., 2012). Although loss of ribosomal protein L25 is often seen in conjunction with absence of rpL30 (Lecompte, 2002), TM6 (not within the CPR) is the only CP studied here where this is the case. This suggests different trajectories of ribosome evolution between the CPR and other lineages without rpL30.

WS6, WWE3, Saccharibacteria, and almost all Microgenomates are missing ribosomal protein L9 (rpL9; **Table 2.1**). RpL9 is thought to be universal in bacteria (Yutin et al., 2012), and is involved in both initiation of ribosome assembly (Nowotny and Nierhaus, 1982) and maintaining translation fidelity (Atkins and Björk, 2009), yet culture-based studies suggest it does not contribute to fitness (Akanuma et al., 2012). Of the three complete Microgenomates genomes, one encodes rpL9. This rpL9 sequence is phylogenetically related to Parcubacteria sequences (**Supplementary File 2.2**), suggesting acquisition by lateral gene transfer.

Ribosomal protein L1 (rpL1) is absent from a group within the Parcubacteria that potentially includes >90 phyla. We refer to this group as OD1-L1 (**Figure 2.1**). No other organisms are known to lack rpL1, a large protein that forms a prominent feature of the large subunit (Schuwirth, 2005). This ribosome initiator protein (Nowotny and Nierhaus, 1982) controls its own expression (Nevskaya, 2005), and loss of rpL1 results in severe growth defects (Akanuma et al., 2012). Absence of rpL1 in this diverse clade suggests alternative mechanisms of ribosome regulation, possibly involving an analogous protein and/or an alternative ribosome structure.

The ribosomal protein biogenesis factor GTPase Der is missing from almost all organisms lacking either rpL9 or rpL1 (**Figure 2.11**). Der is essential for ribosome production and is conserved throughout bacteria (Shajani et al., 2011). Thus, in addition to having unusual ribosome composition, many CPR bacteria likely employ alternative ribosome assembly methods. Although some CPR bacteria have both atypical ribosomes and rRNA introns, these features are not directly linked and thus are not compensatory.

Conclusion

Typically, bacteria within a phylum have widely varying genome sizes and metabolic capacities. In contrast, organisms throughout the CPR have consistently small genomes and similar metabolic limitations. Specifically, all have incomplete tricarboxylic acid cycles and lack electron transport chain complexes, including terminal oxidases and reductases; some lack ATP synthase (**Figure 2.11**). With the notable exception of the Peregrinibacteria, most have incomplete nucleotide and amino acid biosynthesis pathways. CPR bacteria are likely obligate fermenters dependent on other organisms for survival, although they could support respiring organisms by excreting fermentation end products. Overall, these characteristics, in addition to unusual ribosomes, a high frequency of rRNA introns, and a distinct phylogeny, establish the CPR as a subdivision within domain Bacteria.

Methods

Groundwater sampling and geochemical measurements

We studied groundwater microbial communities from an aquifer adjacent to the Colorado River near Rifle, CO, USA at the Rifle Integrated Field Research Challenge (IFRC) site. Aquifer well CD-01 (39°31'44.69" N, 107°46'19.71" W; 1,617.5 meters above mean sea level) was observed from August 23 to December 22, 2011, during which a 79-day acetate amendment experiment was conducted (**Figure 2.2** and see (Castelle et al., 2015; Luef et al., 2015)). This well had been subjected to an acetate stimulation experiment during the previous year (Luef et al., 2012; Williams et al., 2011). Acetate (15 μM target concentration within the aquifer) was administered to the alluvial aquifer through a series of injection wells, and microbial biomass was sampled from groundwater pumped from a down gradient monitoring well. Approximately 100 liters of groundwater was sampled from a depth of 5 m below ground surface (bgs) through a 1.2 μm pre-filter, and cells were collected on serial 0.2 and 0.1 μm filters (Supor disc filters; Pall Corporation, NY), with the specific objective of enriching for organisms with small cell sizes. Filters were immediately frozen after collection in a dry ice and ethanol bath. See **Supplementary Table 2.1** for sampling dates and the amount of groundwater filtered over the course of the experiment. Geochemical measurements were made on samples collected 5 m bgs. The HACH phenanthroline assay and sulfide reagent kits were used to measure ferrous iron and sulfide concentrations, respectively (Hach Co., Loveland, CO). Acetate and sulfate concentrations were measured by ion chromatography, as previously described (Williams et al., 2011). Briefly, acetate and sulfate concentrations were measured with a Dionex ICS-2100 fitted with an AS-18 guard and analytical column (Dionex Co., Sunnyvale, CA).

Metagenome and metatranscriptome sequencing

Six time points spanning a range of geochemical conditions were chosen for metagenomic and metatranscriptomic analysis (**Figure 2.2 and Supplementary Table 2.1**). DNA was extracted from ~1.5 g of each frozen filter using the PowerSoil DNA Isolation Kit (MO-BIO Labs Inc., Carlsbad, CA) with the following modifications: DNA was concentrated by sodium acetate/ethanol precipitation with glycogen, and DNA was eluted in 50 μl TRIS buffer. DNA library preparation and sequencing was conducted at the Joint Genome Institute (Walnut Creek, CA). Total DNA was sequenced on an Illumina HiSeq (Illumina Inc., San Diego, CA), producing 150 bp paired reads with a targeted insert size of 500 bp. Sequence data were processed using version 1.8 of the Illumina CASAVA pipeline, and all reads were trimmed based on quality scores using Sickle (Joshi) (default parameters; **Supplementary Table 2.2**).

RNA was extracted from the 0.2 µm filters using the Invitrogen TRIzol® reagent, followed by genomic DNA removal and cleaning using the Qiagen RNase-Free DNase Set kit and the Qiagen Mini RNeasy™ kit. An Agilent 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA) was used to assess the integrity of the RNA samples. The Applied Biosystems SOLiD™ Total RNA-Seq kit was used to generate the cDNA template library. The SOLiD™ EZ Bead system (Life Technologies, Grand Island, NY) was used to perform emulsion clonal bead amplification to generate bead templates for SOLiD™ platform sequencing. Samples were sequenced at Pacific Northwest National Laboratory on the 5500XL SOLiD™ platform. The 50 bp single reads were trimmed using Sickle (default parameters; **Supplementary Table 2.2**).

Metagenome assembly, annotation, and genome binning

Total community DNA was assembled individually for each sample using IDBA_UD (Peng et al., 2012) with default parameters (**Supplementary Table 2.2**). 16S and 23S rRNA gene sequences were identified from all assembled sequences and curated using an automated method (see below). Scaffold coverage was calculated by mapping reads back to the assembly using Bowtie2 (Langmead and Salzberg, 2012) with default parameters for paired reads. All scaffolds ≥5 Kbp were included when binning genomes from the metagenome assembly. These scaffolds were annotated by first predicting open reading frames (ORFs) using the metagenome implementation of Prodigal (Hyatt et al., 2010), and then using USEARCH (–ublast) (Edgar, 2010) to search protein sequences against UniRef90 (Suzek et al., 2007), KEGG (Kanehisa et al., 2012; Minoru Kanehisa, 2000), and an in-house database composed of ORFs predicted from genomes of candidate phyla (CP) organisms. The in-house database includes previously published genomes (Castelle et al., 2013; Hug et al., 2013; Kantor et al., 2013; Wrighton et al., 2012; 2014) and genomes from ongoing work. Scaffolds were binned based on their GC content, DNA sequence coverage, abundance pattern across samples, and taxonomic affiliation, both automatically with the ABAWACA algorithm (see below) and manually using ggKbase tools (ggkbase.berkeley.edu). Bins generated by ABAWACA were manually inspected within ggKbase. Reported here are genomes binned for organisms associated with the Candidate Phyla Radiation (CPR; **Figure 2.1 and Supplementary Table 2.3**) and TM6 (a phylum of organisms with similar characteristics).

To test the accuracy of this binning method, 20 draft-quality genomes were randomly selected from a sample with a high proportion of CPR genomes (GWA2). These genomes were fragmented and then re-clustered based on tetranucleotide signatures using an Emergent Self-Organizing Map (ESOM), as previously described (Dick et al., 2009). Tetranucleotide frequencies were calculated for 5-10 Kbp scaffold fragments. The number of occurrences of each tetranucleotide in each fragment was normalized based on the total number of times the tetranucleotide was observed across all fragments, and then these values were log-transformed, standardized so they would follow a normal distribution, and then scaled from 0-1. Normalized tetranucleotide values for each fragment were standardized so they would also follow a normal distribution. The resulting matrix was used to train an ESOM for 100 epochs using esom_train.pl (Norman) (downloaded October 2014). The ESOM was visualized using the Databionic ESOM

Tools software (Ultsch, 2005). Coloring fragments (data points) in the ESOM based on the genome each fragment originated from enabled validation of these genome bins (**Figure 2.3)**.

ABAWACA genome binning

ABAWACA was used to generate preliminary genome bins for each sample. This algorithm assesses different characteristics of assembled scaffolds to bin them into genomes. Here, we used a combination of mono-, di-, and tri- nucleotide frequencies and coverage values calculated by mapping DNA sequences from all samples to the scaffolds from the sample being binned. This algorithm uses the given information in a hierarchical clustering fashion as follows. First, all scaffolds are broken into 5 Kbp segments called data points, and the properties of each data point are computed. The binning process begins with a single bin that contains all scaffolds and proceeds by iteratively splitting this and subsequent bins. All non-final bins are evaluated during each iteration. The algorithm searches for a single value for one of the characteristics that will result in the best separation of the scaffolds into two bins. Separation quality is calculated based on the number of data points that were assigned correctly given the separation of the scaffolds. Once a split has been made, scaffolds are separated into the bin with the majority of the data points representing the scaffold. Bins are approved if the quality score exceeds a predefined threshold, and both bins consist of at least 50 data points. A bin is considered final if no separation can be made; otherwise, it undergoes further rounds of binning.

Genome assessment and finishing

Genome bins were associated with CPR lineages based on phylogenetic analysis of 16S rRNA genes and/or ribosomal proteins (see below). When these phylogenetic markers were not present for a particular genome bin, taxonomic placement was achieved based on a consensus of the taxonomic assignments given to ORFs based on their similarity to ORFs from CPR representatives in the CP database described above. Genome completeness was assessed using a modified version of a previously reported list of universal single copy genes (SCGs) for bacteria (**Supplementary Table 2.3** and see (Raes et al., 2007)). Several SCGs were not included as they were found to be unsuitable for the CPR, either because these genes were too divergent in CPR genomes to be reliably detected, or because members of the CPR do not encode these genes. For example, the genes for ribosomal proteins L1 and L9 are not encoded in the genomes of many CPR organisms (see main text). SCGs were identified based on a reciprocal best BLAST (Altschul et al., 1990) hit procedure using a database of SCG protein sequences from a representative set of genomes. First, SCG proteins from the database were searched against all protein sequences in a given genome to identity SCG candidates (blastall –p blastp –F F –e 1e-2). Then, these candidate proteins were searched against the SCG protein sequence database to confirm the assignment (blastall –p blastp –F F –e 1e-5 –b 1 –v 1). SCGs were considered to be present if they were identified by the reciprocal hit method, and the best alignment with a database sequence covered ≥50% of the protein sequence.

In order to be included in this study as a draft genome, a bin must have contained at least 50% of these SCGs with less than 1.125 copies of the genes (indicating that the bin does not contain significant contamination from other genomes). In order to make consistent comparisons with previously sequenced genomes from the CPR, all available genomes were re-assessed using these methods (**Supplementary Table 2.4**).

Several high-quality genome bins were selected for manual curation and genome finishing. Binned scaffolds were connected with one another by extending scaffolds and searching for overlaps. Scaffold extension was achieved by assembling reads mapped to the ends of scaffolds. Assembly errors were detected by manually inspecting the read mapping for these genomes. Genomes were only considered to be complete if they were circular, did not contain gaps, and were, based on complete visual inspection of mapped reads, free of assembly errors. Assembly errors can be identified as regions that do not have read support (i.e. reads may map but with mismatches, or regions may not be supported by paired reads). These regions can be manually corrected. Genomes were also checked for the presence of "orphaned pairs," which could indicate alternative assembly paths. The complete genome for GWB1_sub10_OD1-complete was obtained by first assembling 1/10 of the sequence data for sample GWB1, binning scaffolds based on GC content, coverage, and taxonomic affiliation, and then genome finishing as described above.

Identification of rRNA genes and insertions

16S and 23S rRNA gene sequences were identified based on Hidden Markov Model (HMM) searches using the cmsearch program from the Infernal package (Nawrocki et al., 2009) (cmsearch –hmmonly –acc –noali –T –1). Importantly, all identified gene sequences were curated to remove assembly errors before any analysis was conducted (see below). To identify 16S rRNA gene sequences, all assembled contigs were searched against the manually curated structural alignment of the 16S rRNA provided with SSU-Align (Nawrocki, 2009). Since the SSU-Align 16S rRNA gene covariance model did not include sequences with insertions, large gaps in the alignment between each sequence and the model revealed the boundaries of insertions. Because no equivalent model existed for the 23S rRNA gene, we built a sequence-only model from the manually curated seed alignment maintained by the Comparative RNA Web (Cannone et al., 2002) (**Supplementary File 2.4**). While this model did not contain secondary structure information, it was appropriate for identifying 23S rRNA genes, and the boundaries of insertion sequences, from sequence-based HMM alignments, as was done for 16S rRNA genes. In order to identify the location of rRNA gene insertions with respect to well-studied *Escherichia coli* sequences, all bacterial rRNA gene sequences found to encode insertions were aligned against models consisting of only the respective rRNA from *E. coli* strain K12 substrain DH10B (**Figure 2.5, Figure 2.9, and Supplementary Table 2.5**).

Similarity of rRNA insertions to previously studied structural RNA families (e.g. group I and group II catalytic RNAs) was determined by searching full rRNA sequences against Rfam (Burge et al., 2012) using cmscan (also from Infernal; **Supplementary Table 2.5**). Regions of the rRNA with significant alignments to a structural RNA family (passed model inclusion threshold) were considered as positive hits if at least 25% of the alignment overlapped with an insertion. These rRNA structural families were of particular interest for determining whether or not insertions encode catalytic RNAs potentially capable of self-splicing from containing RNA sequences (**Figure 2.5 and Figure 2.9**). RNA secondary structure was predicted for selected intervening sequences using the Andronescu 2007 model (Andronescu et al., 2007) implemented in Geneious v. 7.1.5 (Kearse et al., 2012) (**Figure 2.8**).

ORFs encoded within rRNA insertion sequences were identified by first predicting ORFs across full rRNA genes, and then selecting ORFs encoded within insertion regions. In order to exclude false ORF predictions, at least 90% of the ORF had to overlap with an insertion. Insertion-encoded ORFs were searched against Pfam (Finn et al., 2013) in order to associate encoded proteins with known families (**Figure 2.5, Figure 2.9, and Supplementary Table 2.5**). In some cases, Phyre2 (Kelley and Sternberg, 2009) was used to model protein sequences and provide further support for identified homing endonucleases (**Figure 2.8**). Insertions and ORFs identified within 16S and 23S rRNA genes were compared with one another using BLAST (**Supplementary Table 2.9**). In order to assess the prevalence and types of intervening sequences previously sampled in 16S rRNA genes from bacteria, version 115 of non-redundant SILVA (Quast et al., 2013) was analyzed using the same methods (**Figure 2.6 and Supplementary Table 2.6**). Importantly, all insertions ≥10 bp were removed prior to multiple sequence alignment and phylogenetic analysis of 16S rRNA gene sequences.

Bacterial community composition based on assembled 16S rRNA genes

The composition of the bacterial community was determined based on assembled and curated 16S rRNA gene sequences. Each sequence was given a taxonomic assignment based on the phylogenetic analysis described below. Coverage of all assembled 16S rRNA gene sequences was determined for each sample by stringently mapping reads using Bowtie2 (no mismatches allowed). For each sample, the coverage of all sequences belonging to each lineage of interest was summed, and then converted to a percent relative abundance in order to observe the composition of each filtrate, and shifts in the community across the time series (**Figure 2.4**).

16S rRNA gene copy number

16S rRNA gene copy number was estimated for all complete and draft genomes based on two assessments. First, the number of assembled 16S rRNA gene sequences was determined. Second, coverage of 16S rRNA gene regions was compared with the coverage of the rest of the genome in order to determine relative copy number. Relative copy number was calculated because of the likeliness of assembling only one 16S rRNA gene for organisms with multiple, identical copies of the gene. Due to the conserved nature of the 16S rRNA gene, it is common for these regions to have inflated coverage values based on default mapping parameters due to inaccurate assignment of reads to sequences from other organisms. To avoid this, both genome and 16S rRNA gene coverage values were calculated based on reads that mapped with zero mismatches. Relative copy number was calculated as: (16S rRNA gene coverage)/(genome coverage). Copy number for each genome was determined by whichever value was greatest, the number of assembled genes or relative copy number (**Figure 2.7 and Supplementary Table 2.7**). Only ten CPR genomes were found to encode more than one copy of the 16S rRNA gene; however, since these genes were not similar to one another, it is more likely that these rare cases were binning errors.

rRNA gene transcript analysis

In order to determine the fate of rRNA insertion sequences, RNA transcript sequences recovered from 0.2 μm filters were stringently mapped to assembled, curated rRNA genes. In order to prevent short reads from erroneously matching to either rRNA genes or insertions, zero mismatches were allowed between reads and assemblies. Coverage was calculated separately for 16S rRNA gene and predicted insertion regions, and then the values compared with one another (**Supplementary Table 2.8**). Most insertions were found to have zero coverage. However, in some cases very low coverage of insertion regions was found. In almost all cases these low coverage values were the result of a small portion of the insertion region being covered by RNA sequence, likely the result of a small difference between predicted and actual insertion regions, but possibly the result of partial recovery of spliced insertion sequences.

16S rRNA gene primer binding analysis

The level of sequence divergence of the 16S rRNA genes assembled here from metagenome data, compared with sequences from existing databases, suggests that they would elude PCR-based analysis. We assessed the binding affinity of commonly used 16S rRNA gene survey primers 515F and 806R (Caporaso et al., 2012; Gilbert et al., 2010). Assembled 16S rRNA gene sequences were clustered at 97% sequence identity using USEARCH (–cluster_smallmem –query_cov 0.50 –target_cov 0.50 –id 0.97) in order to remove redundant sequences from the analysis. Because some of the sequences are not complete, only those spanning the 515-806 region of the *E. coli* 16S rRNA gene were included. Primer binding was assessed with PrimerProspector (Walters et al., 2011) using default parameters (**Figure 2.10**).

Phylogenetic analysis

Phylogenetic analysis was carried out using several different marker sequences in order to best survey the diversity within the groundwater microbial community, and to robustly assign taxonomy to complete and draft genomes. Markers included the 16S rRNA gene, ribosomal proteins encoded by a syntenic block of genes, and ribosomal protein S3 (rpS3). The syntenic block encodes the genes for ribosomal proteins L - 2, 3, 4, 5, 6, 14, 15, 16, 18, 22, 24 and S - 3, 8, 10, 17, 19, hereafter referred to as rp16. In the rp16 analysis, individual protein sequence alignments were concatenated for phylogenetic inference. Unlike in previous metagenomic studies, near-complete 16S rRNA gene sequences were assembled commonly enough to be able to infer phylogeny for many community members. However, rp16 was also used for phylogenetic analysis because i) it is encoded in genomes as a syntenic block and is found in only one copy, and thus can be used as a proxy for a particular genotype independent of binning, ii) it encodes ribosomal proteins that provide a robust phylogenetic signal, and iii) it is assembled more frequently from metagenome sequence data compared with the 16S rRNA gene (Hug et al., 2013). RpS3 was also independently used as a phylogenetic marker because of its strong phylogenetic signal, despite having a relatively short protein sequence. In cases where a genome did not contain any of these markers (**Supplementary Table 2.3**), taxonomic assignment was made based on whole genome comparisons to the database of reference genomes described above. In all cases, metagenome assembly was necessary for providing a robust phylogenetic analysis.

After removing insertions ≥10 bp from 16S rRNA gene sequences from this and previous studies, sequences were aligned with SSU-Align. SSU-Align classifies sequences as bacteria, archaea, or eukarya, and then generates separate alignments for sequences from each domain. The resulting Stockholm-formatted bacterial multiple sequence alignment was converted to FASTA, and all alignment insert columns were removed. This resulted in a 1,582 bp alignment. All sequences with ≥800 bp of aligned sequence were used for phylogenetic analysis. Several archaeal reference sequences were chosen for the phylogenetic root, aligned to the bacterial 16S rRNA gene model provided with SSU-Align, and concatenated with the bacterial multiple sequence alignment. A maximum-likelihood phylogeny was inferred using RAxML (Stamatakis, 2014) with the GTRCAT model of evolution and 100 bootstrap re-samplings (**Supplementary File 2.2**

and Supplementary File 2.3). A subset of the tree was annotated using GraPhlAn (Segata and Huttenhower) (**Figure 2.1**).

Rp16 ORFs were identified by searching all ORFs encoded on scaffolds ≥5 Kbp against databases of each of these ribosomal proteins. Searches were carried out with USEARCH (–ublast). Syntenic groups of ORFs were selected if at least three of the ribosomal proteins in rp16 could be identified with an E-value ≤1 x 10^{-6}. This allowed for identification of all instances of each ribosomal protein in rp16 encoded within assembled scaffolds. For each ribosomal protein, all identified protein sequences along with reference sequences were aligned to their respective Pfam HMM profile using hmmalign from the HMMER 3.0 package (Eddy, 2011). Protein sequence alignments were converted from Stockholm format to FASTA, alignment insert columns were removed, and the 16 protein alignments concatenated. This resulted in a 1,935 amino acid (aa) alignment. All sequences with ≥1,000 aligned residues were kept for phylogenetic analysis. Because of the size of the multiple sequence alignment, phylogenetic analysis was carried out in two steps. First, FastTree2 (Price et al., 2010) was used to infer the phylogeny of the entire sequence set using the Jones-Taylor-Thornton model of amino acid evolution (JTT) and by assuming a single rate of evolution for each site, the "CAT" approximation (additional options: –spr 4 –mlacc 2 –slownni). Then, sequences associated with the CPR and TM6 were selected, along with representatives of the Archaea and Chloroflexi, in order to infer a maximum-likelihood phylogeny using RAxML with the LG + alpha + gamma model of evolution and 100 bootstrap re-samplings (see (Hug et al., 2013) for choice of evolutionary model). Archaea were included as a root for the tree, and Chloroflexi as a root for the CPR. Notably, the CPR is evident as a monophyletic group in both of these analyses, and in the 16S rRNA gene phylogeny (**Figure 2.1 and Supplementary File 2.2**).

Phylogenies were inferred from individual protein sequences for rpS3 and ribosomal protein L9 (rpL9). All rpS3 protein sequences were identified from metagenome ORFs by searching protein annotation descriptions. The same was done for rpL9, except only sequences associated with CPR genome bins were included. Erroneously annotated sequences were excluded based on the alignment score inclusion threshold for their respective Pfam HMM profiles (aligned using hmmalign), followed by manual removal of non-rpS3 or rpL9 sequences. Sequences were combined with reference sequences and aligned. RpS3 sequences were aligned to Pfam HMM profile PF00189 using the same procedure as was described for the rp16 protein sequences (see above). RpL9 was aligned using MUSCLE (Edgar, 2004). All sequences with ≥50 aligned aa residues were used for phylogenetic analysis using RAxML with 100 bootstrap re-samplings and an evolutionary model chosen using ProtTest (Abascal et al., 2005) (**Supplementary File 2.2**). The ProtTest 2.4 server(Abascal et al., 2005) was run on the Pfam seed alignment for rpS3 and on a random subset of the rpL9 alignment, indicating that the LG + gamma, and the LG + gamma (with fixed base frequencies) evolutionary models should be used for rpS3 and rpL9, respectively.

All phylogenetic trees were visualized using Dendroscope (Huson and Scornavacca, 2012).

Identification of novel phyla

The number of phyla within the CPR, Parcubacteria (OD1), and Microgenomates (OP11) was estimated by counting 16S rRNA gene sequence clusters created based on a 75% sequence identity threshold. After removing insertions ≥10 bp, sequences were clustered using USEARCH (–cluster_smallmem –query_cov 0.50 –target_cov 0.50 –id 0.75). This threshold and method for estimating the number of phyla were proposed by Yarza *et al*. These authors proposed that phyla could be identified as monophyletic lineages composed of members distinguished by approximately this level of sequence divergence. We classified new phyla based on this and additional, strict criteria. Clusters of 16S rRNA genes that share ≥75% sequence identity were used to assess the divergence and coherence of deep branches of the phylogenetic tree (**Supplementary File 2.2**). Bootstrap support values were often higher for lineages primarily composed of one or few clusters, validating the use of this threshold. Lineages were proposed as phyla if i) they formed a monophyletic group in the 16S rRNA gene phylogeny, ii) 16S rRNA genes were approximately 25% divergent from other lineages, iii) they were also supported by the rp16 concatenated ribosomal protein phylogeny, and iv) representative complete and/or draft genomes were available. Names for these phyla were proposed based on the names of lifetime achievement award recipients in microbiology (**Figure 2.1**, **Table 2.2, and Supplementary File 2.2**). Genomes were associated with these phyla using the 16S rRNA gene and/or rp16 phylogenies (**Supplementary Table 2.3**).

Sequence curation

Assembled 16S rRNA genes, 23S rRNA genes, and scaffold regions encoding rp16 genes were curated in order to identify and fix assembly errors prior to assessment of insertions in rRNA genes and/or phylogenetic analysis. For curation, these genes were extracted along with 2 Kbp of sequence from each side. Assembly errors, typically short regions of misassembled sequence associated with scaffolding contigs with one another, were identified as regions with zero coverage by stringently mapped paired-end reads. Only one mismatch per read was permitted and only paired reads were included in the analysis. Regions with 1x coverage were only allowed if at least 3 bp on either side of the read overlapped with other reads, with zero mismatches in the overlap region. When an assembly error was detected, read pairs mapped (Bowtie2) to a 1 Kbp region surrounding the error were collected and re-assembled using Velvet (Zerbino and Birney, 2008). Reads were collected for re-assembly as long as at least one read in the pair mapped with two or fewer mismatches. Velvet was run by iterating from kmer 21 to 71, increasing by 10 in each iteration. Re-assembled fragments were then merged with the original assembly based on overlap of ≥10 bp. All assembly modifications were verified with a subsequent round of error detection. If an error could not be corrected, the original scaffold was split at the position of the error. In addition to error correction, reads mapped to the ends of scaffolds were re-assembled and used to extend scaffolds, or the ends of broken scaffolds, when possible. Following curation, genes of interest were re-identified on curated scaffolds using the methods described above (**Supplementary File 2.1**). On average, 1.5 assembly errors were corrected for each scaffold region containing a 16S rRNA gene.

Ribosomal protein inventory and metabolic potential of CPR genomes

Metabolic potential of CPR genomes was assessed using ggKbase. In ggKbase, lists related to different proteins or metabolic pathways were generated by searching for specific keywords in gene annotations. Here, lists were created to assess ribosomal protein composition and metabolic potential across the CPR (**Figure 2.11**). Genomes were compared with one another by creating ggKbase genome summaries based on a selection of these lists. This allowed for the simultaneous assessment and comparison of the 8 complete and 789 draft-quality genomes assembled here.

In order to compare genomes based on both their phylogenetic associations and metabolic capacity, and to get the clearest picture of the metabolic potential of the CPR, an additional analysis was conducted with only complete and near-complete genomes (≥75% of single copy genes and ≤1.125 copies, including an assembled 16S rRNA gene). Since similar genotypes were assembled independently from different samples, this set of complete and near-complete genomes was de-replicated by choosing a representative genome for all flat branches on the 16S rRNA gene tree (**Supplementary File 2.2**). The genome summary was then ordered based on the 16S rRNA gene phylogeny, a step that was critical for identifying lineages missing specific ribosomal proteins (**Figure 2.11**). In order to find ribosomal proteins that may have evaded detection due to sequence divergence, six-frame translations (bacterial translation table 11) of all complete and draft CPR genomes were searched against Pfam ribosomal protein HMM profiles using hmmscan; however, this confirmed the initial finding of missing ribosomal proteins in organisms from CPR lineages (**Supplementary Table 2.10**).

Although complete genomes are invaluable for metabolic analyses, this extensive inventory of draft-quality genomes from organisms representing diverse lineages, and assembled from different samples, enabled confident assessment of gene absence. For example, there are no reported complete WS6 genomes, but the 16 reconstructed draft-quality genomes from this study (median estimated completeness of 91%) showed that this lineage is missing rpL9. The probability of the gene being present, but missing in all 16 genome reconstructions, is $(1 - 0.91)^{16}$, i.e., ~ 2 x 10^{-17}. Even if we lower the completion requirement to a very conservative value of 35% complete, 16 such genomes would yield a confidence value of 0.001 for the gene being absent. For lineages where we have hundreds of genomes the probability of missing the gene due to chance is effectively zero.

Code Availability

ABAWACA is maintained under github.com/CK7/abawaca (version 1.00 used in this analysis: github.com/CK7/abawaca/releases/tag/v1.00) and the script used for curating scaffolds, re_assemble_errors.py, is maintained under github.com/christophertbrown/fix_assembly_errors (version 1.00 used in this analysis: github.com/christophertbrown/fix_assembly_errors/releases/tag/1.00).

Data Availability

DNA and RNA sequences are available through NCBI SRA accession SRP050083, and genomes through NCBI BioProject PRJNA273161 (first versions described here). Genomes are also available through ggKbase: ggkbase.berkeley.edu/CPR-complete-draft/organisms. ggKbase is a "live data" site, thus annotations and genomes may be improved after publication.

Author Contributions

Samples and geochemical measurements were taken by MJW, KCW, and KHW. BCT assembled the metagenome data. IS implemented the ABAWACA algorithm. CTB and JFB binned the data and carried out the ESOM binning validation. JFB closed and curated the complete genomes. CTB, LAH, and BCT conducted the rRNA gene insertion analysis. CTB and LAH performed phylogenetic analyses. MJW and KCW conducted the RNA sequencing. CTB carried out the 16S rRNA gene copy number, primer binding, and transcript analyses. CTB and JFB carried out the ribosomal protein analyses. CTB, LAH, CJC, and JFB conducted the metabolic analysis. AS and BCT provided bioinformatics support. CTB and JFB drafted the manuscript. All authors reviewed the results and approved the manuscript. The authors declare that they have no competing interests.

Acknowledgements

We thank J. Cate and S. Moore for input into the ribosomal protein analysis, J. Doudna and E. Nawrocki for suggestions on the rRNA insertion analysis, and M. Markillie and R. Taylor for assistance with RNA sequencing. Research was supported by the US Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research under award number DE-AC02-05CH11231 (Sustainable Systems Scientific Focus Area and DOE-JGI) and award number DE-SC0004918 (Systems Biology Knowledge Base Focus Area). LAH was partially supported by an NSERC postdoctoral fellowship. DNA sequencing was conducted at the DOE Joint Genome Institute, a DOE Office of Science User Facility, via the Community Science Program. RNA sequencing was performed at the DOE-supported Environmental Molecular Sciences Laboratory at PNNL.

Supplementary Table

Supplementary Table 2.1 | Geochemical measurements from acetate amendment field experiment conducted in aquifer well CD-01 at the Rifle IFRC site.

Supplementary Table 2.2 | Metagenomics and metatranscriptomics sequencing and assembly statistics.

Supplementary Table 2.3 | Candidate phyla and Candidate Phyla Radiation (CPR) genomes reconstructed from groundwater metagenomics. The columns for 16S rRNA gene, rp16, and rpS3 phylogeny designate whether or not the specified phylogenetic marker was used to confirm the taxonomy for the organism. Complete genomes are circular and have been manually curated. Draft and partial genome status was determined based on the percent of single copy genes (SCGs) that could be identified. The inventory of these SCGs makes up the right-most columns of the table (rp is an abbreviation for ribosomal protein). Draft-quality genomes have ≥50% of these single copy genes with ≤1.125 copies overall, and must encode at least one of the specified phylogenetic marker genes.

Supplementary Table 2.4 | CPR genomes from previous studies. Previously described CPR genomes (Albertsen et al., 2013; Kantor et al., 2013; Marcy et al., 2007; McLean et al., 2013; Podar et al., 2007; Rinke et al., 2013; Wrighton et al., 2014) were assessed in order to make comparisons with genomes reconstructed in the current study. Genome status was determined in the same way as was done for the genomes presented in **Supplementary Table 2.3**. The 16S rRNA gene column specifies whether or not a 16S rRNA gene was sequenced for the genome, and if so how it was determined. The rp16 column designates whether or not the genes that make up rp16 were assembled. The inventory of these SCGs makes up the right-most columns of the table (rp is an abbreviation for ribosomal protein).

Supplementary Table 2.5 | 16S and 23S rRNA gene insertions in sequences reconstructed from groundwater-associated bacteria. The 97% ID centroid column refers to whether or not a given rRNA gene sequence was the representative sequence for a group of rRNA gene sequences clustered based on a 97% sequence identity threshold (USEARCH –cluster_smallmem –query_cov 0.50 –target_cov 0.50 –id 0.75). When identified, sequences of insertion-encoded open reading frames (ORFs) and catalytic RNA introns (group I or group II introns) are provided.

Supplementary Table 2.6 | 16S rRNA gene insertions in the Silva database. When identified, sequences of insertion-encoded open reading frames (ORFs) and catalytic RNA introns (group I or group II introns) are provided.

Supplementary Table 2.7 | 16S rRNA gene copy number. 16S rRNA gene copy number was estimated for all draft CPR genomes and genome bins for organisms found outside of the CPR. Relative 16S rRNA gene copy number was calculated as: (16S rRNA gene coverage)/(genome coverage). Estimated 16S rRNA gene copy number was determined for each genome based on whichever value was greatest, the number of assembled genes or relative copy number. Shown in red are the few CPR genomes with discrepant copy number estimates, as discussed elsewhere (see section on 16S rRNA gene copy number in Methods and **Figure 2.9**).

Supplementary Table 2.8 | 16S and 23S rRNA gene and insertion transcript analysis. Coverage of rRNA gene and insertion regions by metatranscriptomic sequences were separately determined and compared. Coverage was calculated as (number of mapped bases)/(length of the region). Length coverage refers to the percent of bases in a given region with coverage >0.

Supplementary Table 2.9 | Comparison of 16S and 23S rRNA gene-encoded insertions and ORFs. All insertions identified in 16S rRNA genes were compared with insertions in 23S rRNA genes (reciprocal BLASTn). Likewise, ORFs encoded within these rRNA genes were compared with one another (reciprocal BLASTp). Reported here are the top 10 hits for each search.

Supplementary Table 2.10 | HMM identification of ribosomal proteins in CPR genomes. Six-frame translations of all CPR genomes were searched against Pfam ribosomal protein HMM profiles. Identifiers for the profiles searched for each ribosomal protein are shown above the name of the protein. Shown here is the number of genes assigned to each ribosomal protein for all complete and draft-quality genomes. Ribosomal proteins discussed in the text are highlighted in red.

Supplementary Files

Supplementary File 2.1 | Curated 16S rRNA gene (a), 23S rRNA gene (b), and rp16 encoding contig (c) sequences in FASTA format.

Supplementary File 2.2 | Phylogenetic analyses. Sequences from this study are labeled GW[A-F][1-2] depending on which sample they originated from (e.g GWA1 is the sample taken at time point A from the 0.1 μm filter). Unless otherwise specified, maximum-likelihood phylogenies were inferred using RAxML with 100 bootstrap re-samplings. Also see **Supplementary File 2.3**. (a) Phylogeny inferred from 16S rRNA gene sequences aligned using SSU-Align after having removed insertions ≥10 bp long (sequences with ≥800 aligned bp). Reference sequences from previously assembled genomes from the CPR were included, along with sequences associated with the CPR in the Silva database (clustered at 90% sequence identity), sequences from Silva that were the best-hits of sequences assembled here, and a set of reference sequences representative of major clades within domain bacteria. The 75% sequence identity threshold cluster that each sequence was assigned to is indicated with a unique identifier. Based on the 16S rRNA gene phylogeny, the SR1 are not part of the CPR, although they have previously been associated with this group. A subset of this phylogeny was used in **Figure 2.1** and **Figure 2.11**. (b) Approximate maximum-likelihood phylogeny inferred using FastTree2 for a concatenation of aligned ribosomal protein sequences (rp16) assembled here and from reference genomes (sequences with ≥1,000 aligned aa residues are included). (c) Maximum-likelihood phylogeny inferred for subset of sequences represented in (b). (d) Phylogeny for rpS3 sequences identified here and from reference genomes. (e) Phylogeny for rpL9 sequences associated with CPR genomes.

Supplementary File 2.3 | (a) 16S rRNA gene RAxML phylogeny, (b) rp16 concatenated protein FastTree phylogeny, (c) rp16 concatenated protein RAxML phylogeny, (d) rpS3 protein RAxML phylogeny, and (e) rpL9 protein RAxML phylogeny in Nexus format.

Supplementary File 2.4 | 23S rRNA HMM in Infernal format.

Supplementary File 2.5 | ggKbase summary of draft genomes in SVG format.

Figure 2.1 | Phylogeny and genomic sampling of the Candidate Phyla Radiation (CPR). Subsets of a maximum-likelihood 16S rRNA gene phylogeny (**Supplementary File 2.2**) (**a**) showing the CPR, a monophyletic radiation of CP, and (**b**) genomic sampling of CP. Proposed names for phyla within the superphyla Parcubacteria and Microgenomates are explained in **Table 2.2**. Many CPR 16S rRNA genes encode insertions (length shown by blue bars, combined length for multiple insertions).

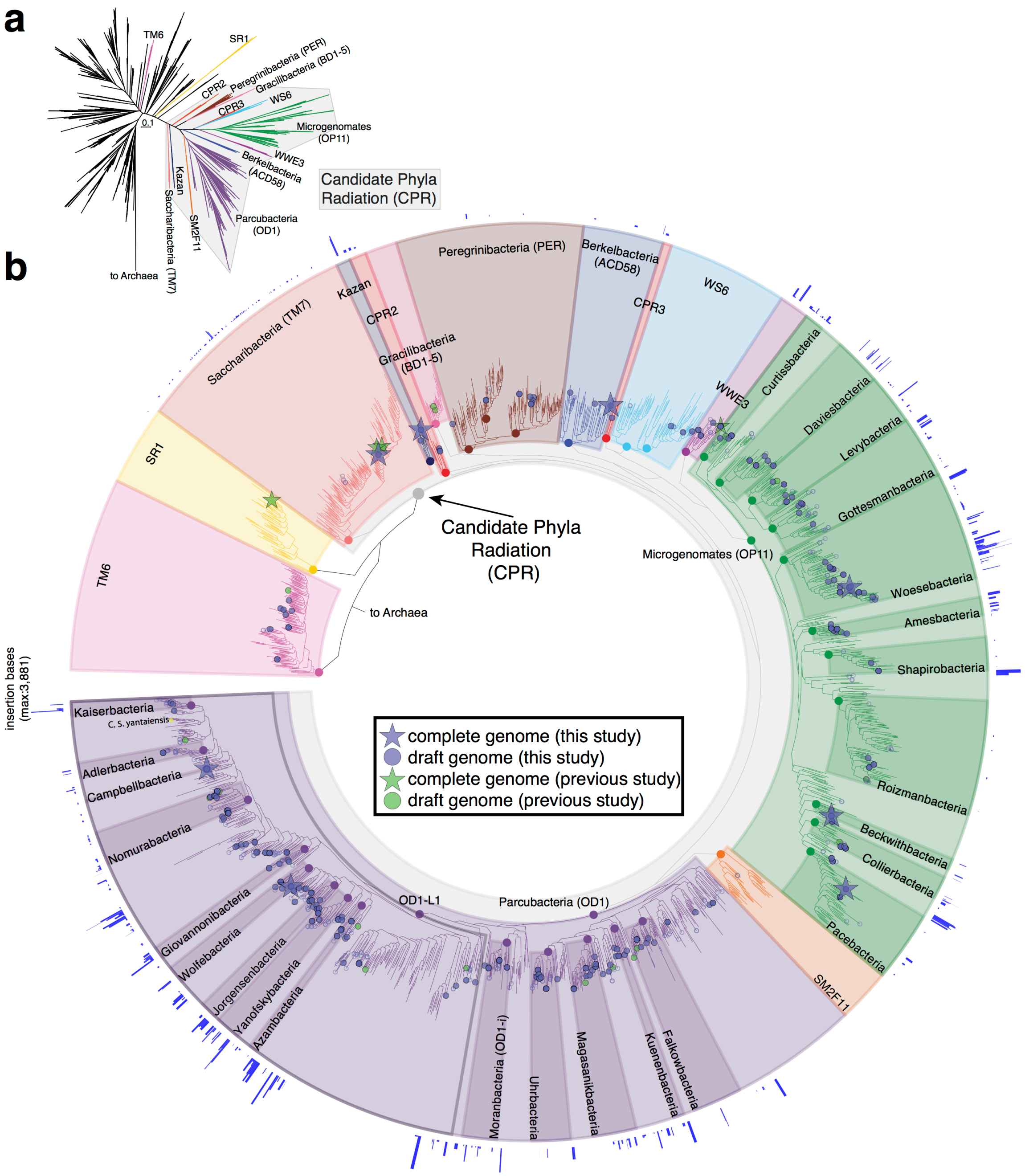


Figure 2.2 | Sampling and geochemical measurements from acetate amendment field experiment conducted in aquifer well CD-01 at the Rifle IFRC site. Samples were collected for metagenomics and metatranscriptomics at six time points (A-F) spanning several redox transitions during acetate stimulation of groundwater microbial communities. **(a)** Groundwater was pumped from the alluvial aquifer and filtered through serial 1.2, 0.2, and 0.1 μm filters (aerial image provided by S.M Stoller for the US DOE under contract DE-AM01-07LM00060). DNA was extracted and sequenced from both the 0.2 and 0.1 μm filters, and RNA extracted and sequenced from the 0.2 μm filters. **(b)** Geochemical measurements were taken throughout the time series, showing a transition from dominant iron reduction to sulfate reduction through to methane production in the sampling environment.

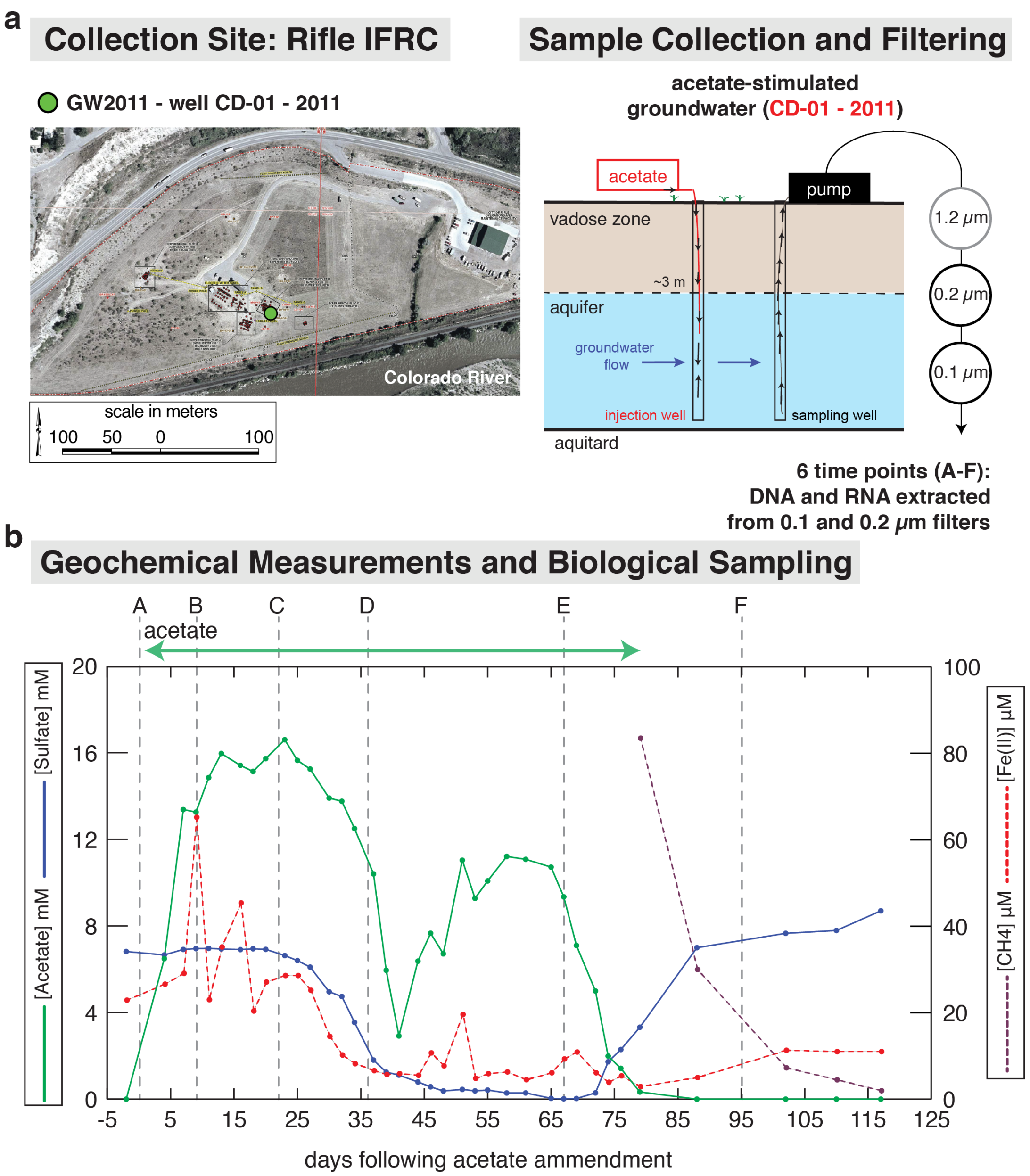


Figure 2.3 | Validation of 20 draft-quality genomes by ESOM clustering of genome fragments based on tetranucleotide sequence composition. For validation, 20 draft genomes from a sample with a high proportion of CPR genomes (GWA2) were chosen at random. Each data point represents a 5-10 Kbp genome fragment. The ESOM was trained for 100 epochs with normalized tetranucleotide frequencies. Dark lines between data points indicate strong separation between regions. Data points are colored based on the genome the fragment originated from. The ESOM shows well-delineated clusters for most of the 20 draft genomes, with few sequence fragments falling outside of these clusters. Two genomes from the same Microgenomates (OP11) phylum were not well delineated in the tetranucleotide-based ESOM (genomes 18 and 19). This shows how the method we used for binning, which takes into account abundance patterns in addition to sequence signatures, provides more accurate genome reconstructions. The white box distinguishes a single period on the repeating map. Genomes split into multiple clusters are labeled in red.

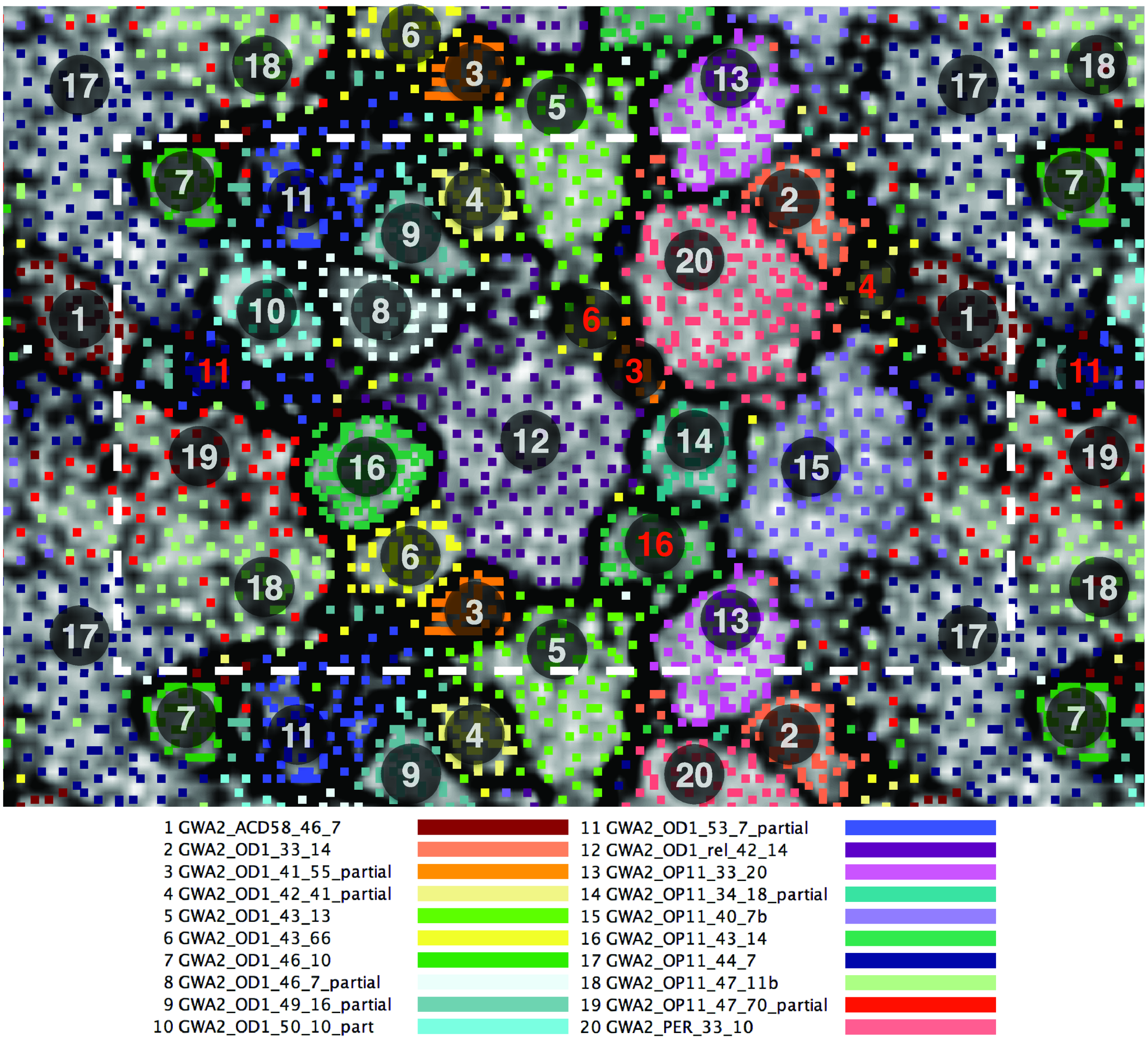


Figure 2.4 | Relative abundance of bacterial community members during acetate amendment. Relative abundance was calculated based on stringent mapping of paired-read sequences from each sample to 16S rRNA gene sequences assembled from all samples. **(a)** Relative abundance of cells from 0.2 µm filters and **(b)** from 0.1 µm filters. Enrichment of CPR organisms in the 0.2 µm filtrate indicates that these organisms have ultra-small cell sizes.

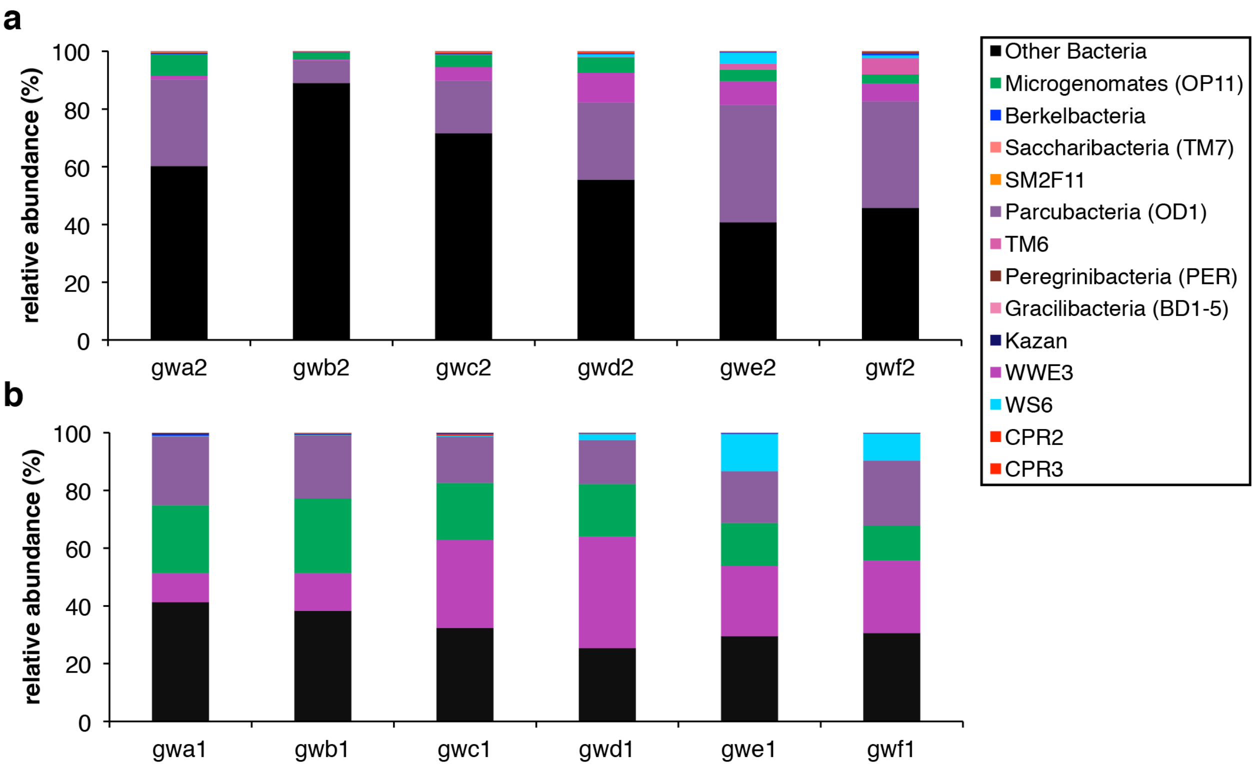


Figure 2.5 | Features of insertions encoded within CPR 16S rRNA genes. Insertions identified in assembled, unique bacterial 16S rRNA genes occur in conserved and variable (red bars) regions (**Supplementary Table 2.5**). Histograms show the frequency of insertions. Insertions are of several types distinguishable by catalytic RNA introns and/or ORFs. (IVP = intervening sequence protein).

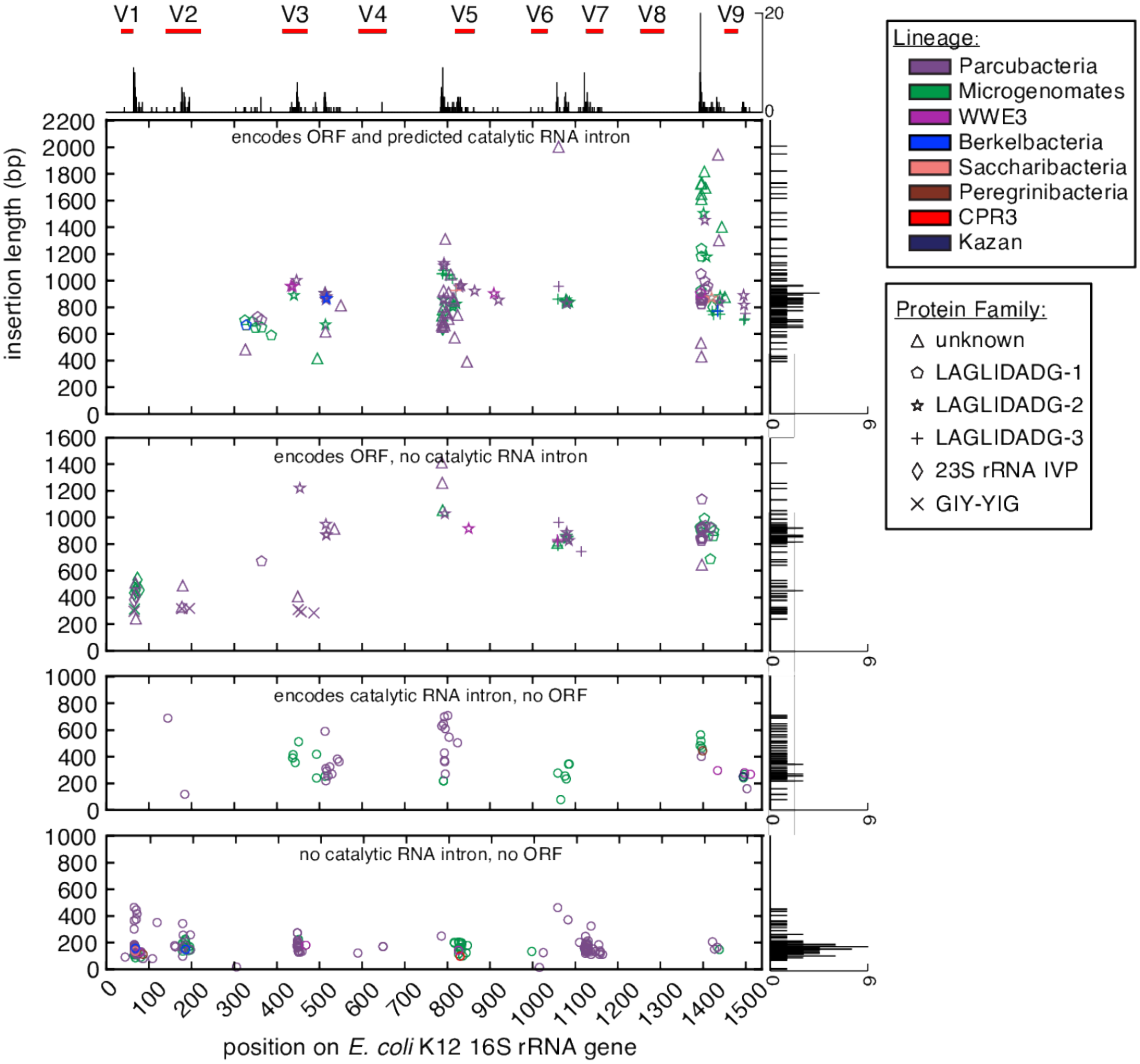


Figure 2.6 | Features of insertion sequences encoded within 16S rRNA genes from the Silva database. The non-redundant Silva 16S rRNA gene database (v. 115) was analyzed in order to assess the prevalence of insertions. Only 761 of the 418,498 16S rRNA gene sequences from bacteria encode insertions. While many small insertions were identified, unlike the 16S rRNA gene sequences from CPR bacteria, these sequences i) rarely encode large insertions, ii) do not contain both ORFs and introns, iii) do not encode ORFs that could be assigned to Pfam families, and iv) may be found in one of multiple copies of the 16S rRNA gene.

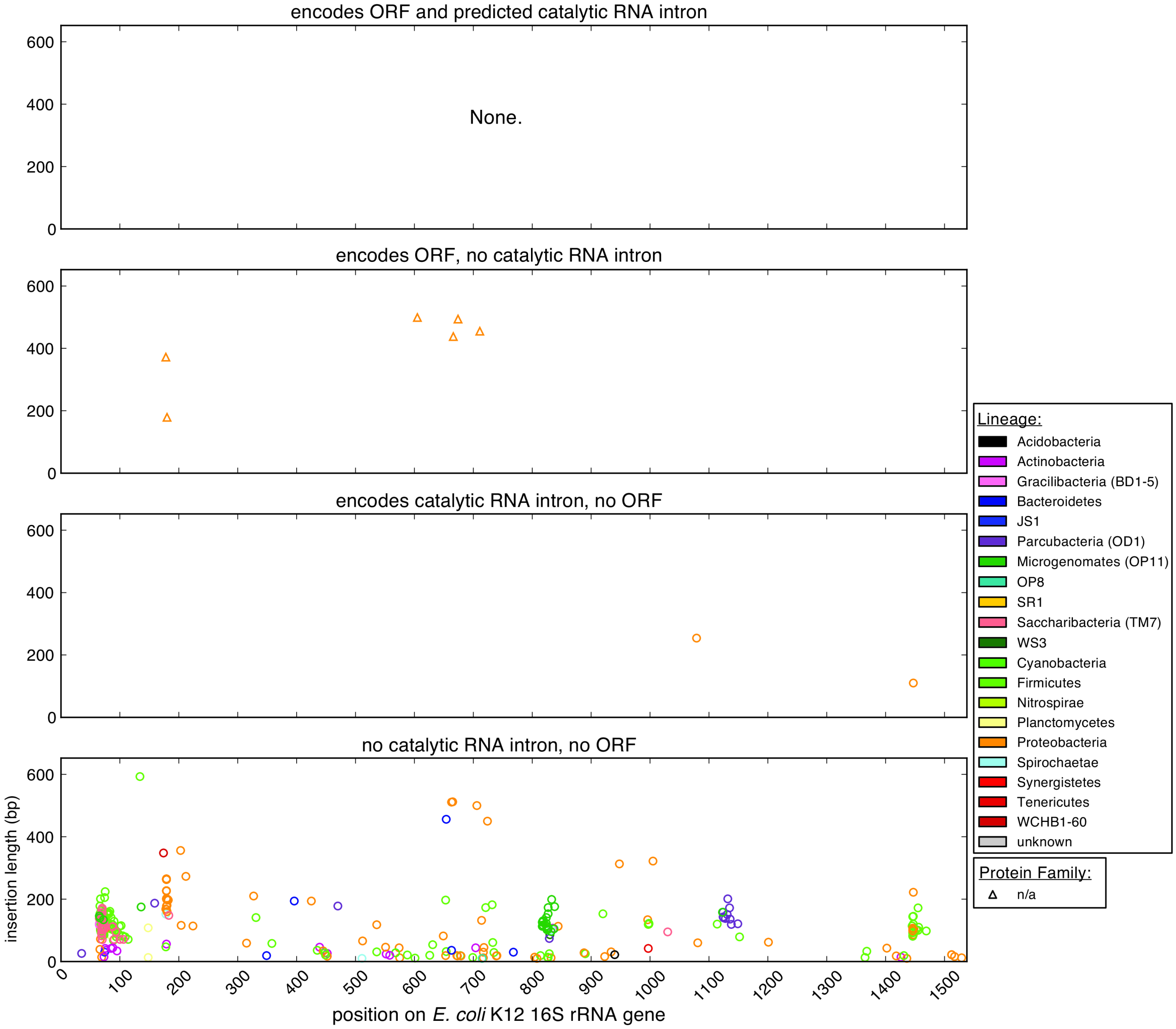


Figure 2.7 | 16S rRNA gene copy number estimations for genomes reconstructed from groundwater metagenomics. 16S rRNA gene copy number was estimated for all draft CPR genomes and genome bins for organisms outside the CPR. This was achieved by comparing the coverage of 16S rRNA gene regions to the coverage of the rest of the genome. Importantly, coverage was determined with stringently mapped reads (no mismatches were allowed) in order to improve the accuracy of coverage calculations. **(a)** Histogram of the number of 16S rRNA gene sequence copies estimated for each genome by calculating (16S rRNA gene coverage)/(genome coverage). Several WWE3 genomes were estimated to have high 16S rRNA gene copy number (**Supplementary Table 2.7**), but it was later determined that these estimates were skewed by the presence of a highly abundant closely related strain. The complete WWE3 genome assembled previously (Kantor et al., 2013) has an identical 16S rRNA gene and confirms that it is found in only one copy for this genotype. Thus, we removed these estimates from subsequent copy number analysis. **(b)** Density plot comparing estimated copy number of genomes for organisms found within and outside the CPR, where the longer tail for non-CPR genomes depicts the propensity for multiple 16S rRNA copies, a trait absent from the CPR.

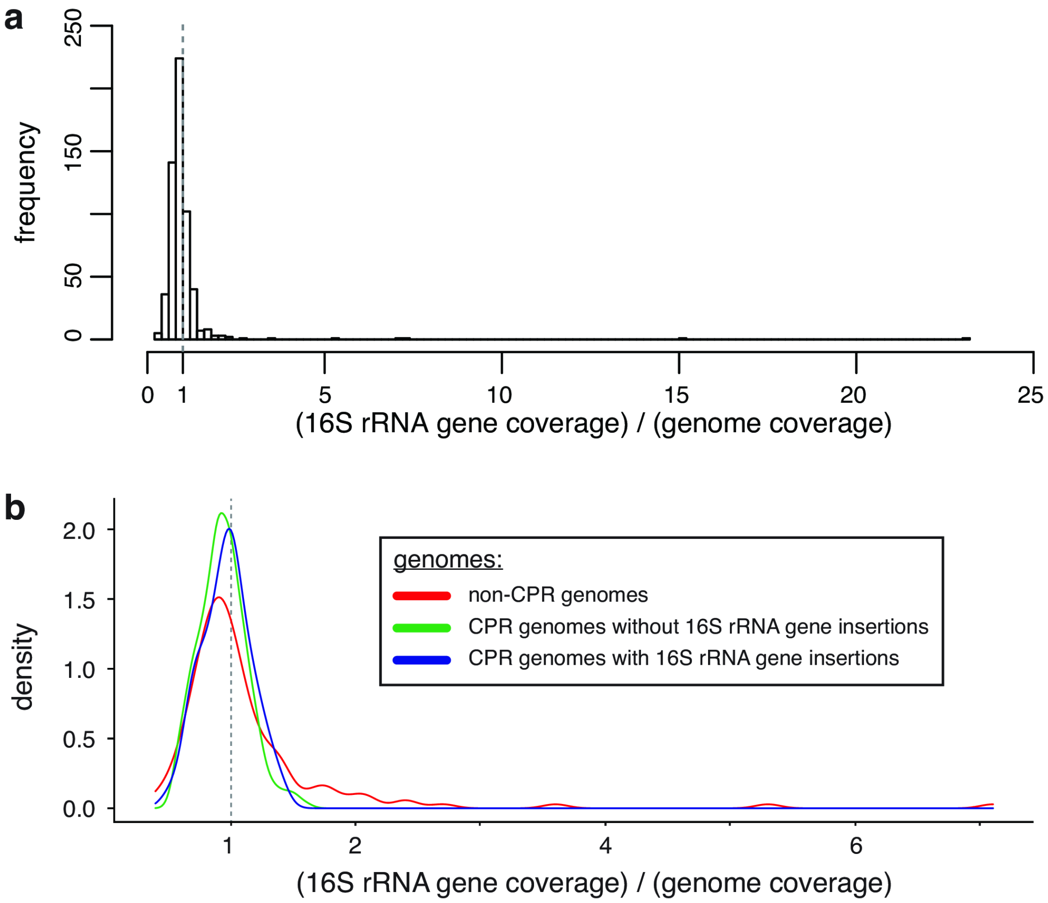


Figure 2.8 | Intron-encoding 16S rRNA gene from complete Microgenomates genome. (a) Stringent mapping of paired-read metagenome sequences confirms the assembly. **(b)** 16S rRNA encoding regions, but not insertions, are covered by perfectly-matched metatranscriptome sequences. Absence of RNA sequences for insertions indicates they are introns. Shown are regions corresponding to *E. coli* K12 gene positions, RNA catalytic introns, ORFs, and insertions. **(c)** Structural models of encoded proteins (1, 2, and 4: colored by rainbow – N to C terminus) and predicted structure for a catalytic RNA intron (3: colored by base-pairing probability – red is high, green is moderate, and blue is low). Protein Data Bank (PDB) structures were used as templates for structural modeling (1: 1R7M, 2: 1B24, 4: 1B24).

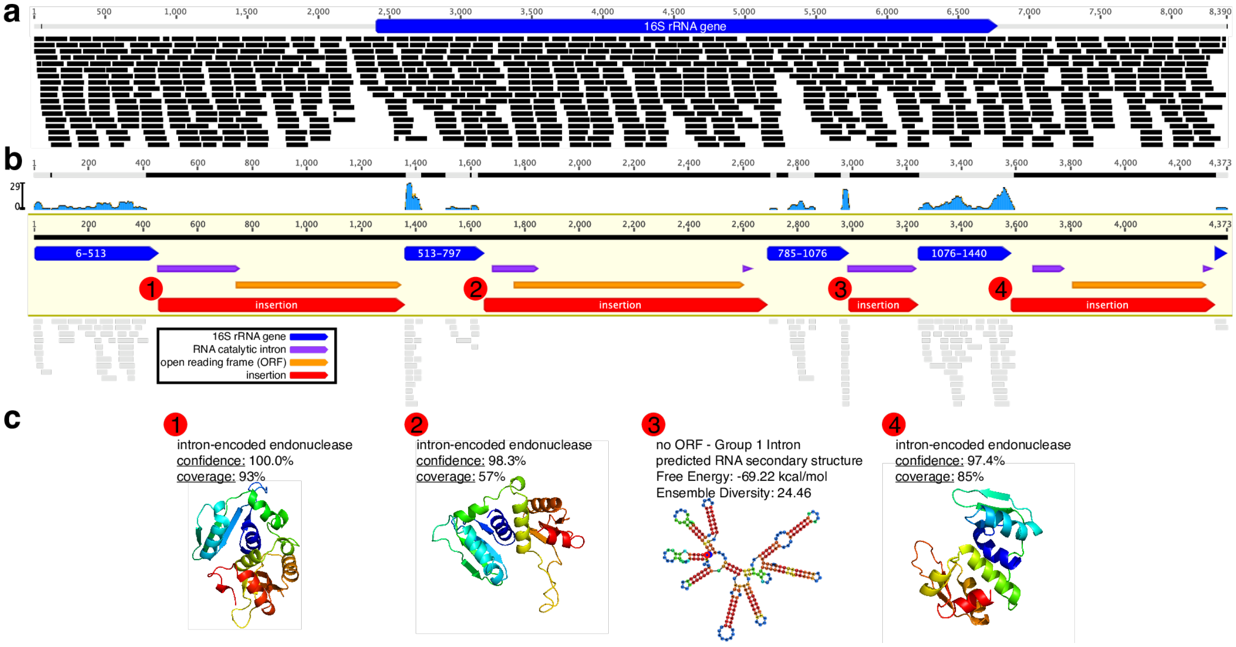


Figure 2.9 | Features of insertion sequences encoded within 23S rRNA genes recovered from groundwater-associated bacteria. Bacteria associated with the CPR encode insertions within their 23S rRNA genes (**Supplementary Table 2.5**). These insertions share many features with those identified in 16S rRNA gene sequences from CPR bacteria. Taxonomy was determined by inclusion in a genome with an established phylogeny.

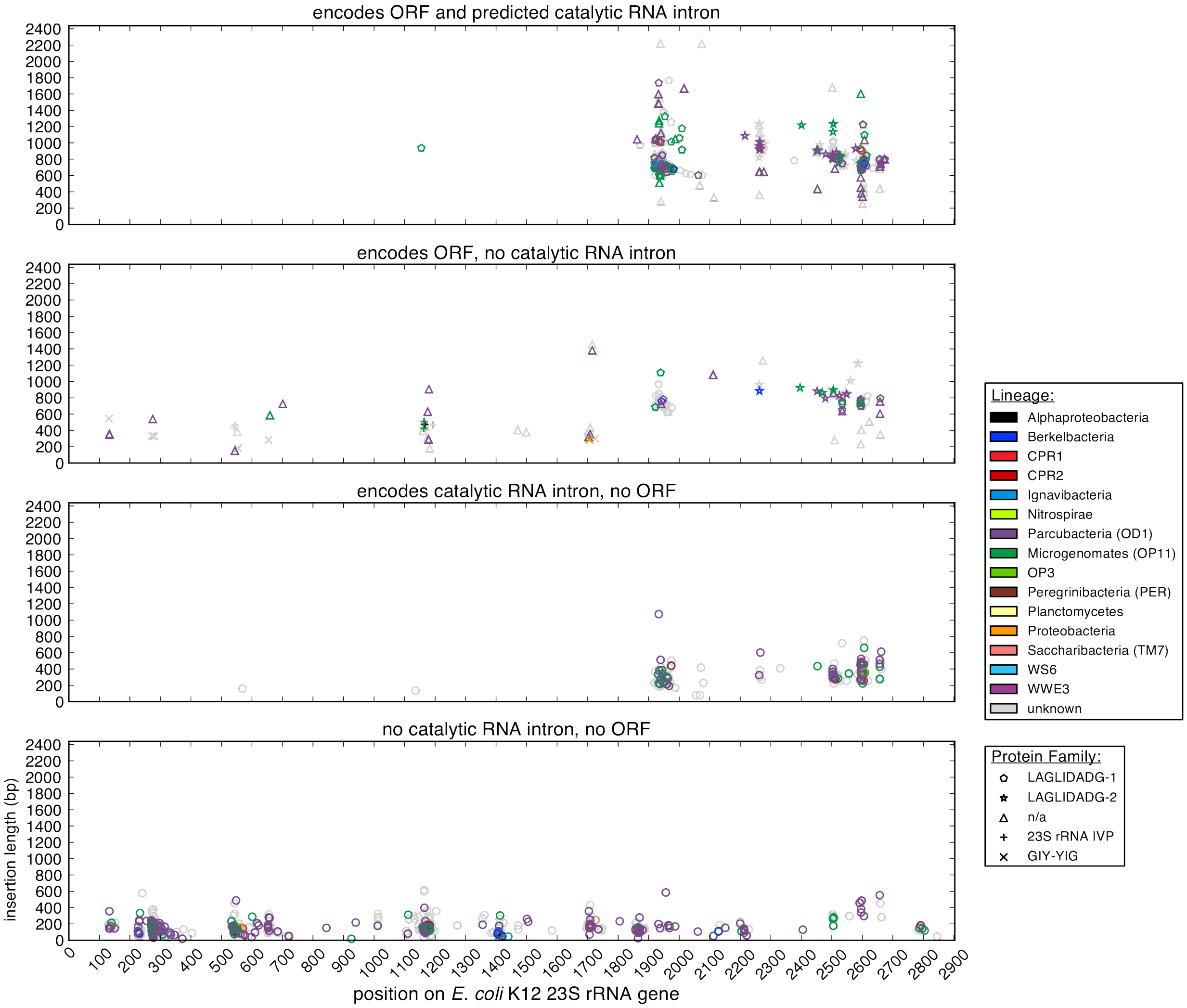


Figure 2.10 | Analysis of the ability of PCR primers 515F and 806R to bind to recovered groundwater-associated 16S rRNA gene sequences. PrimerProspector was used to assess the ability of primers 515F and 806R to bind a non-redundant set of assembled near-complete 16S rRNA gene sequences (clustered at 97% sequence identity). The percent of sequences that would be amplified by these primers is shown on the left axis and the total number of sequences analyzed is on the top of each bar The number of sequences these primers would not bind to is indicated by the shading. Many assembled groundwater-associated 16S rRNA gene sequences would evade amplification by PCR primers 515F and 806R. Results of the analysis are shown at **(a)** the domain and **(b)** superphylum or phylum levels.

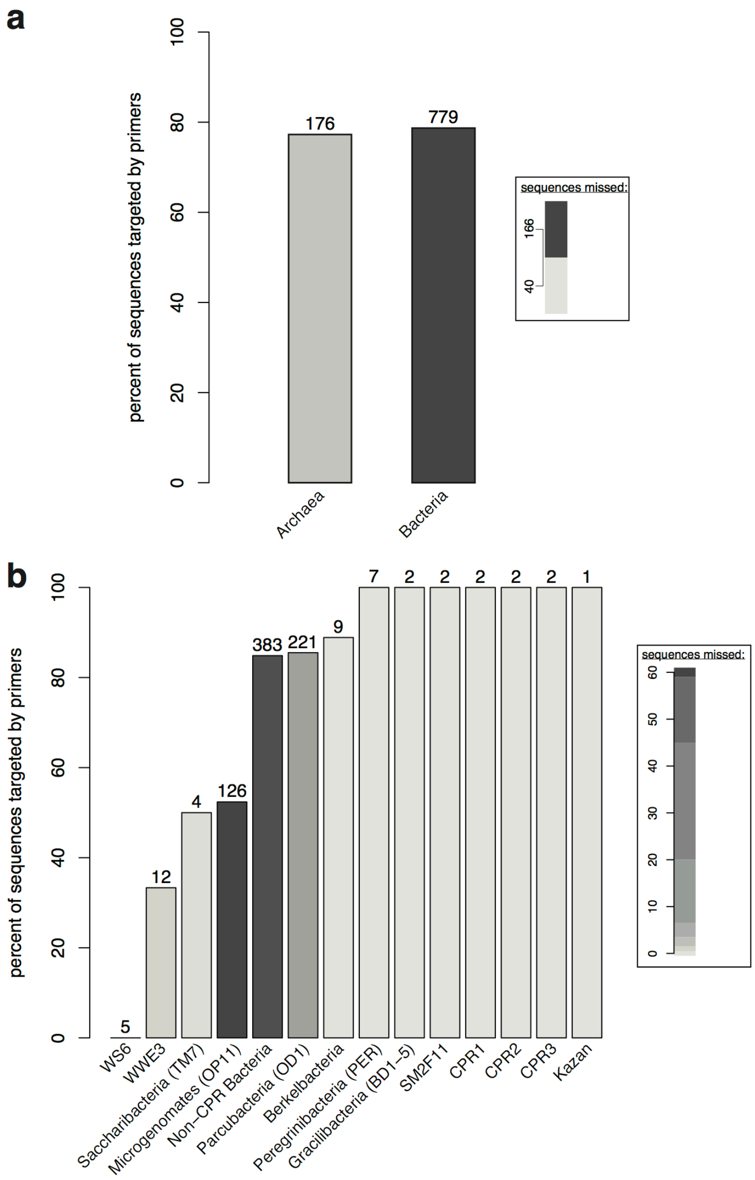


Figure 2.11 | Metabolic potential and ribosomal protein analysis of genomes from CPR and TM6 organisms. Assembled genomes were analyzed using ggKbase (**Supplementary File 2.5**). Shown here is a non-redundant set of complete and near-complete genomes (≥75% of single copy genes, ≤1.125 copies) organized based on a subset of a maximum-likelihood 16S rRNA gene phylogeny (**Supplementary File 2.2**). CPR organisms have partial tricarboxylic acid (TCA) cycles and lack electron transport chain (ETC) complexes. In addition, they have incomplete biosynthetic pathways for nucleotides and amino acids (AA Syn. is short for amino acid synthesis). The Peregrinibacteria are a notable exception to some of these limitations. Several Parcubacteria exhibit a complete ubiquinol (cytochrome b_o) oxidase operon, as previously seen in Saccharibacteria (Kantor et al., 2013). However, lack of NADH dehydrogenase and other ETC components suggests this enzyme is involved in oxygen scavenging/detoxification rather than energy production. (PP is short for the pentose phosphate pathway).

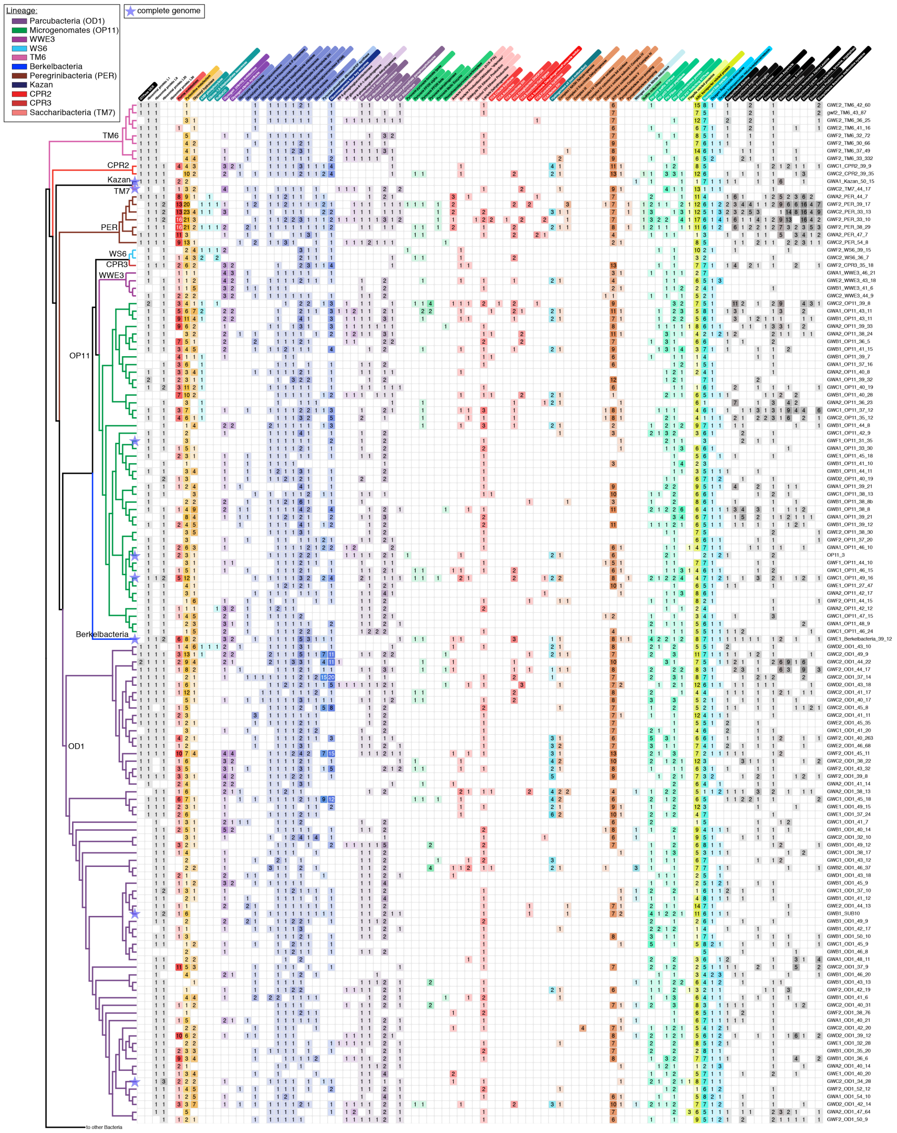


Table 2.1 | Genomes from candidate phyla (CP) bacteria. The percentage of 43 single copy genes (SCGs) identified in each genome was used to estimate completeness. CPR1 through 3 are novel CPR lineages.

Lineage	**Complete genomes**	**Draft genomes**	**Median SCGs**	**Average genome size in bp (stdev)**	**Average %GC (min/max)**	**Missing ribosomal protein(s)**
Parcubacteria	2	427	91%	707,464 (295,862)	43 (31/60)	L30, OD1-L1 missing L1
Microgenomates	3	252	91%	788,693 (261,196)	41 (31/50)	L30, L9*
WWE3	0	41	93%	719,830 (344,415)	43 (41/46)	L30, L9
WS6	0	16	91%	584,741 (167,526)	34 (33/39)	L30, L9
Peregrinibacteria	0	15	91%	1,183,124 (344,415)	42 (33/54)	L30
TM6	0	15	98%	1,060,264 (167,526)	36 (28/43)	L30, L25
Berkelbacteria	1	6	88%	581,936 (243,398)	39 (34/46)	L30
Kazan	1	5	95%	657,191 (214,462)	49 (45/52)	L30
CPR2	0	6	100%	1,032,375 (183,809)	39 (38/39)	L30
Saccharibacteria	1	2	99%	971,756 (157,794)	47 (46/48)	L30, L9
CPR1	0	2	72%	578,470 (266,611)	46 (42/49)	L30
CPR3	0	2	86%	945,288 (153,931)	35 (34/35)	L30
all	**8**	**789**	**91%**	**749,453 (263,507)**	**42 (28/60)**	

*One genotype has rpL9.

Table 2.2 | Proposed names for CPR phyla based on microbiology lifetime achievement award recipients.

Award	Year Awarded	Recipient	Proposed Name	Superphylum
ASM Lifetime Achievement Award	2014	Roy Curtiss	Curtissbacteria	Microgenomates (OP11)
ASM Lifetime Achievement Award	2013	Julian Davies	Daviesbacteria	Microgenomates (OP11)
ASM Lifetime Achievement Award	2012	Stuart B. Levy	Levybacteria	Microgenomates (OP11)
ASM Lifetime Achievement Award	2011	Susan Gottesman	Gottesmanbacteria	Microgenomates (OP11)
ASM Lifetime Achievement Award	2010	Lucy Shapiro	Shapirobacteria	Microgenomates (OP11)
ASM Lifetime Achievement Award	2009	Carl Woese	Woesebacteria	Microgenomates (OP11)
ASM Lifetime Achievement Award	2008	Bernard Roizman	Roizmanbacteria	Microgenomates (OP11)
ASM Lifetime Achievement Award	2007	Norman R. Pace	Pacebacteria	Microgenomates (OP11)
ASM Lifetime Achievement Award	2006	R. John Collier	Collierbacteria	Microgenomates (OP11)
ASM Lifetime Achievement Award	2005	Jonathan Beckwith	Beckwithbacteria	Microgenomates (OP11)
ASM Lifetime Achievement Award	2004	Alan Campbell	Campbellbacteria	Parcubacteria (OD1)
ASM Lifetime Achievement Award	2003	Stanley Falkow	Falkowbacteria	Parcubacteria (OD1)
ASM Lifetime Achievement Award	2002	Masayasu Nomura	Nomurabacteria	Parcubacteria (OD1)
ASM Lifetime Achievement Award	2001	Bruce N. Ames	Amesbacteria	Parcubacteria (OD1)
ASM Lifetime Achievement Award	2000	Boris Magasanik	Magasanikbacteria	Parcubacteria (OD1)
ASM Lifetime Achievement Award	1999	Jonathan W. Uhr	Uhrbacteria	Parcubacteria (OD1)
ASM Lifetime Achievement Award	1998	Charles Yanofsky	Yanofskybacteria	Parcubacteria (OD1)
ASM Lifetime Achievement Award	1997	A. Dale Kaiser	Kaiserbacteria	Parcubacteria (OD1)
ASM Lifetime Achievement Award	1996	Ralph S. Wolfe	Wolfebacteria	Parcubacteria (OD1)
ASM Lifetime Achievement Award	1995	Julius Adler	Adlerbacteria	Parcubacteria (OD1)
ISME Jim Tiedje Award	2014	Nancy Moran	Moranbacteria	Parcubacteria (OD1)
ISME Jim Tiedje Award	2012	Stephen Giovannoni	Giovannonibacteria	Parcubacteria (OD1)
ISME Jim Tiedje Award	2010	Bo Barker Jorgensen	Jorgensenbacteria	Parcubacteria (OD1)
ISME Jim Tiedje Award	2008	Norman R. Pace	Pacebacteria	Microgenomates (OP11)
ISME Jim Tiedje Award	2006	Gijs Kuenen	Kuenenbacteria	Parcubacteria (OD1)
ISME Jim Tiedje Award	2004	Farooq Azam	Azambacteria	Parcubacteria (OD1)

Chapter 3

Measurement of bacterial replication rates in microbial communities

C. T. Brown, M. R. Olm, B. C. Thomas, J. F. Banfield

Published in *Nature Biotechnology*, November 2016.

Abstract

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 in all genome-resolved microbiome studies to track organism responses to varying conditions, identify actively growing populations and measure replication rates for use in modeling studies.

Introduction

Dividing cells in a natural population contain, on average, more than one copy of their genome (**Figure 3.1**). 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 replication (Bremer and Churchward, 1977). This decrease can be detected by measuring changes in DNA sequencing coverage across complete genomes (Skovgaard et al., 2011). Bacterial genome replication proceeds bi-directionally from a single origin of replication (Prescott and Kuempel, 1972; Wake, 1972), therefore the origin and terminus of replication can be deduced based on this coverage pattern (Skovgaard et al., 2011). GC skew (Anantharaman et al., 2016a; Gao et al., 2013; Sernova and Gelfand, 2008) and genome coverage (Korem et al., 2015) 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 replication (Cooper and Helmstetter, 1968), which results in an average of more than two genome copies in rapidly growing cells.

Korem *et al.* used the ratio of sequencing coverage at the origin compared to the terminus of replication to measure replication rates for bacteria (Korem et al., 2015). 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 genomes (Baker et al., 2010; Brown et al., 2015; Castelle et al., 2015; Iverson et al., 2012; Nielsen et al., 2014; Seitz et al., 2016; Sharon et al., 2012; Tyson et al., 2004) (**Figure 3.1 and Figure 3.2**). Often these organisms are from little known microbial phyla, and are vastly different from organisms for which there are complete genomes in databases (Brown et al., 2015; Castelle et al., 2013; 2015; Di Rienzi et al., 2013; Eloe-Fadrosh et al., 2016; Seitz et al., 2016; Wrighton et al., 2012). 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, despite the fact 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 Index of Replication (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 (**Figure 3.1 and Figure 3.2**; **Methods and Code Availability**). 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 windows are calculated (window slide length 100 bp; **see Figure 3.3, Supplementary Table 3.1, and Methods** for evaluation of sliding window methods). Then, a sequencing GC bias correction is applied (**Figure 3.3; Methods**). The average coverage values for each window are then ordered from lowest to highest to assess the coverage trend across the genome. Because coverage values for each window are re-arranged, the order of the fragments in the complete genome need not be known. Extreme high and low coverage windows are excluded (>8-fold difference compared to the median), as they are well known to correlate with highly conserved regions, strain variation, or integrated phage. Finally, the overall slope of coverage across the genome is used to calculate iRep, a measure of the average genome copy number across a population of cells. In a population in which most cells are replicating (making a single copy of their chromosome), iRep would be two. Since iRep is an average across the population, some organisms may not be replicating, but for that to be the case others would have to be in the process of conducting two, or more, simultaneous rounds of genome replication. An iRep value of 1.25 would indicate that, on average, only one quarter of the cells are replicating.

iRep is accurate for complete or draft genomes

In order to evaluate the ability of iRep to measure replication rates, we compared iRep to PTR using 17 samples sequenced to sufficient depth from the growth rate experiments reported by Korem *et al.* as part of their validation of the PTR method. As there is no open-source version of the PTR software, we re-implemented the PTR method, with some improvements that include an option to determine the origin and terminus positions based on GC skew (Lobry, 1996) (**Methods**). PTRs generated using the Korem *et al.* software (kPTRs) use a genome database of unknown composition that can be neither viewed nor modified, and no metrics for evaluating measurement reliability are provided. These limitations are addressed in our PTR implementation (named bPTR). kPTR and bPTR values for this dataset were highly correlated, and each was correlated with iRep (**Figure 3.4a and Supplementary Table 3.2**). We used growth rates calculated using counts of colony forming units (CFU), as reported by Korem *et al.*, to verify that iRep values correlate as well as PTRs (**Figure 3.4b**). It should be noted that growth rates derived from CFU data are based on total population size, which includes effects of cell death and can be negative. iRep and PTR methods only measure replication, and thus represent the physiological state of the cells independent of death rates.

We tested the minimum sequencing coverage requirements for iRep, kPTR and bPTR using sequencing data of cultured *Lactobacillus gasseri* from the Korem *et al.* study. We first subsampled reads to achieve 25x coverage of the genome and then calculated replication rates to use as reference values. Then, the dataset was subsampled to lower coverage values and the replication rates re-calculated. Comparing these rates to the reference values enabled evaluation of the amount of noise introduced by increasingly lower coverage. Results show that all three methods are affected by coverage, and that although kPTR has the least amount of variation at 1x coverage, all methods are reliable when the coverage is ≥5x (**Figure 3.4c and Supplementary Table 3.3**).

Because iRep does not require knowledge of the order of genome fragments, it can be used to obtain replication rates when only draft quality genomes are available. Therefore, we evaluated the minimum percentage of a genome that is required to obtain accurate results by conducting a random genome subsampling experiment (**Figure 3.4d, Figure 3.3, and Supplementary Table 3.1**). iRep values were determined for *L. gasseri* cells sampled when growing at different rates (Korem et al., 2015), and then compared with values determined from genomes at various decreasing levels of completeness. Our analysis revealed that ≥75% of the genome sequence is required for iRep to be accurate (difference from known value <0.15). Although extensive genome fragmentation will introduce noise into iRep calculations, values are accurate for genomes with less than 175 scaffolds per Mbp of sequence (**Figure 3.3 and Supplementary Table 3.1**). Genome completeness and contamination can be estimated based on the presence and copy number of expected single copy genes (SCGs). Based on these findings, we selected genomes for iRep analysis if they were estimated to be ≥75% complete based on inventory of 51 expected single copy genes (SCGs), if they also had fewer than two duplicate SCGs and less than 175 scaffolds per Mbp of sequence. Lack of additional SCG copies indicates that a genome is free of substantial contamination. As shown below, these standards can be met for a substantial number of genomes recovered from metagenomic data sets.

The human microbiome includes some bacteria with genomes that are sufficiently similar to reference genomes to enable ordering and orienting of draft genome fragments, making it possible to calculate both iRep and bPTR for comparison. We carried out an analysis using five genomes reconstructed in a metagenomics study of premature infants (GC range: 28-56%) (Raveh-Sadka et al., 2015). Importantly, unlike when using kPTR, the reads were mapped to the genome that was reconstructed from the infant gut metagenomes in order to achieve more robust results than would be achieved using a public database-derived reference genome, due to the fact that differences in gene content and gene order will perturb coverage trends. The correct ordering of the scaffolds in the reconstructed genome was confirmed based on both coverage patterns and cumulative GC skew (**Figure 3.5**). For all 24 comparisons involving populations with iRep values of 1.8-1.9, there was a strong correlation between iRep and bPTR values (Pearson's r = 0.83, p-value = 5.9 x 10^{-7}; **Figure 3.4e**).

Although a few complete reference genomes were similar enough to reconstructed draft genomes to facilitate scaffold ordering, these reference genomes were from organisms relatively distantly related to those present in samples of interest. Specifically, for the five genomes with available similar reference genomes (average nucleotide identity 91-99%), as much as 19.5% of reference genomes was not represented by metagenome reads (min. = 1.6%, average = 13.5%), compared with essentially perfect mapping to reconstructed genomes (**Figure 3.6 and Supplementary Table 3.4)**. This level of genome deviation compared to reference genomes would preclude accurate replication rate calculations due to perturbation of coverage trends, as noted above, and emphasizes the need to reconstruct genomes for organisms of interest. We also compared iRep and bPTR replication rate metrics for a large, manually curated genome scaffold ~2.5 Mbp in length that was reconstructed from a complex groundwater metagenome. Because the scaffold contains both the origin and terminus of replication, as identified both by coverage and cumulative GC skew (**Figure 3.7**), it was possible to calculate both bPTR and iRep. For this single time point measurement, the bPTR value of 1.20 agrees with the iRep value of 1.25. Importantly, it would not have been possible to obtain this information based on mapping to complete reference genomes because this is the first sequence for an organism affiliated with a novel genus within the Deltaproteobacteria (Sharon et al., 2015). This finding demonstrates the iRep method in the context of a very complex natural environment.

Replication rates in environmental and human microbiomes

We obtained 241 iRep measurements using 152 genomes reconstructed as part of a study of premature human infant gut microbiomes (Raveh-Sadka et al., 2015), and 51 draft genomes that we reconstructed from an adult human microbiome dataset (Di Rienzi et al., 2013) (**Figure 3.8a, Supplementary Table 3.5, Supplementary Table 3.6, Supplementary Table 3.7 and see Data Availability**). In infant microbiomes, members of the Firmicutes had the highest replication rates and Proteobacteria had the highest median replication rates (**Figure 3.8b**). In the premature infant dataset, 63 iRep measurements were obtained for 8 species that could be matched to results from the kPTR program; however, there was no strong correlation between the values

(Pearson's r = 0.52, **Figure 3.9, Supplementary Table 3.5 and Supplementary Table 3.8**). Because of the strong correlation between these methods when the organisms were represented by reference genomes (**Figure 3.4a-b**), we attribute this to measurement errors due to differences between the database reference genomes used by kPTR and the genomes of the organisms sampled **(Figure 3.6**).

Using iRep, we obtained replication rates for 51 of the 54 organisms for which we had draft genomes (≥75% complete) from an adult human microbiome sample (**see Methods**; **Figure 3.8, Supplementary Table 3.6, and Supplementary Table 3.7**). Due to a lack of overlap with reference genomes, the kPTR method returned only three values, none of which were credible because all were <1 (**Supplementary Table 3.9**). Similarly, we attempted to select complete reference genomes for bPTR, but were only able to do so in five cases (**Figure 3.10**). Even for these five cases, on average only 94% (min. = 88%, max. = 98%) of each complete reference genome was covered by metagenome sequences.

The Candidate Phyla Radiation (CPR) is a major subdivision within domain Bacteria known almost exclusively from genome sequencing (Brown et al., 2015). Almost nothing is known about the growth rates of these enigmatic organisms. We measured 378 replication rates from CPR organisms using a time series of samples collected from an acetate amended aquifer near the Colorado River, and 99 different draft genome sequences reconstructed from those datasets (Brown et al., 2015) (**Supplementary Table 3.10**). Only 33 of 378 iRep values were calculated using complete genome sequences. One member of the CPR superphylum Microgenomates (OP11) had iRep values amongst the highest observed across CPR and human gut associated microorganisms (**Figure 3.8b**). However, only 16.1% of iRep values from CPR organisms were >1.5, compared with 35.8% of premature infant and 19.6% of adult human microbiome measurements. Median iRep values from CPR bacteria were significantly lower compared with those from premature infant microbiomes (**Figure 3.8a**; CPR = 1.34, premature infant = 1.42, and adult = 1.37). Overall, the results show that CPR bacteria only rarely replicate quickly, and that iRep can be applied in communities with different levels of complexity.

Microbiome responses to antibiotic administration

Twelve samples were collected during periods following antibiotic therapy for five of the ten infants (Raveh-Sadka et al., 2015) (**Figure 3.11**). To measure microbial responses to antibiotics, we compared iRep values from samples collected within five days after antibiotic administration to values from other time points. This showed that the median replication rate for organisms present after administration of antibiotics is higher compared to those present during periods without antibiotic treatment (**Figure 3.12a**). Fast replicating organisms were from the genera *Klebsiella, Lactobacillus, Escherichia, Enterobacter, Staphylococcus,* and *Enterococcus* (iRep >1.5; **Supplementary Table 3.5**).

iRep values for bacteria associated with premature infants

The premature infant dataset consisted of 55 metagenomes collected from ten co-hospitalized premature infants, half of whom developed necrotizing enterocolitis (NEC). There was no statistically significant difference between iRep values from NEC and control infant microbiomes (**Figure 3.12b**), nor was there a statistically significant difference between values determined for the same species found in both infant groups (**Figure 3.12c**). However, organisms from the genus *Clostridium* were replicating significantly faster in microbial communities associated with NEC versus control infants (Mann-Whitney p-value = 5.1 x 10^{-3}; **Figure 3.12d**). Although *Klebsiella pneumoniae* was found to replicate rapidly in control infant microbiomes, it was only infrequently detected in infants that developed NEC, and no iRep values could be determined. Intriguingly, high iRep values for *Clostridium* species were detected in two infants prior to development of NEC (**Figure 3.12e and Figure 3.11**).

iRep documentation of community dynamics

Raveh-Sadka *et al.* measured absolute cell counts per gram of feces collected using droplet digital PCR (ddPCR) as part of a premature infant microbiome study (Raveh-Sadka et al., 2015). Using these measurements and metagenome-derived relative abundance calculations we were able to track absolute changes in the population sizes of 51 genotypes (**Supplementary Table 3.5 and Figure 3.11**). For nine of the ten infants in the study, iRep and both relative and absolute abundance values could be determined for the bacterial populations. Interestingly, despite fast replication rates of *Clostridium* species in two infants before NEC diagnosis, total observed cell counts were either very low or decreasing, emphasizing that populations of active organisms may not necessarily undergo large changes in population size (**Figure 3.11**).

Doubling times are usually calculated for organisms growing in pure culture without resource limitation or host suppression. We used the absolute abundance of *Klebsiella oxytoca* following antibiotic administration to calculate an *in situ* doubling time of 19.7 hours across a four-day period starting three days after the infant was treated with antibiotics (**Figure 3.13a)**. iRep values for *K. oxytoca* during this period were consistently high (1.74–1.80), as required for the population growth that was well described by an exponential equation (r^2 = 0.97). Notably, *K. oxytoca* was essentially the only organism present during this time.

In one infant, iRep values for *Clostridium difficile* and *Enterobacter cloacae* prior to the first NEC diagnosis were unusually high compared to values for organisms found in other infants. However, these organisms remained at low absolute abundance (**Figure 3.13b**). Total cell counts were low following antibiotic treatment; however, this period was associated with high *E. cloacae* replication rates and a subsequent 2.7-fold increase in population size, as determined by ddPCR, prior to the second NEC diagnosis. Interestingly, low-abundance *Clostridium paraputrificum* and *C. difficile* were also replicating quickly before the second diagnosis.

A clear finding from analysis of replication rates for bacteria in multi-species consortia in the premature infant gut is the general lack of correlation between high iRep values and increased population size in the subsequently collected sample (**Figure 3.11)**. Notably, iRep measures the instantaneous population-average replication rate, which provides insights into population dynamics at a physiological level and time scale that cannot be determined by abundance measurements, especially when more than a day separates sampling time points. Using cell counts alone as a metric for replication would miss key features of the ecosystem because the approach measures the cumulative effect of both cell replication and death rates over a specific time period.

Discussion

We developed a method named iRep that uses metagenome sequences and draft-quality genomes, which are routinely assembled in metagenomics analyses, to determine bacterial replication rates *in situ*. As long as accurate genome bins are obtained from the metagenomes of interest (**see below**), bacterial replication rates derived using iRep are more accurate than those obtained using PTR with complete reference genomes. Even when complete genomes are available, superior results can be obtained using iRep rather than PTR, owing to the potential for error when identifying the origin and terminus of replication (**Methods**). The combination of obtaining draft genomes from metagenomes and iRep measurements from read data from multiple samples from the same environment can provide a comprehensive view of microbiome membership, metabolic potential, and *in situ* activity.

Despite the premature infant gut microbiome having relatively consistent community composition over time, iRep analyses indicate that brief periods of rapid replication are common during colonization, possibly due to varying conditions in the infant gut. Even transitory levels of increased replication, especially for potential pathogens, could have phenotypic outcomes that affect clinical presentation since bacteria are known to produce different metabolites concordant with different growth rates (Paczia et al., 2012). An important finding relates to the faster bacterial replication rates after antibiotic treatment, an observation that we attribute to high resource availability following elimination of antibiotic sensitive strains. Interestingly, rapid replication rates of several different but potentially pathogenic organisms from the genus *Clostridium*, including *C. difficile*, precede some NEC diagnoses, consistent with NEC being a multi-faceted disease. Further studies that include more samples and infants may establish a link between rapid cell division and NEC.

iRep measurements provide information about activity around the time of sampling. The approach could be used to probe the responses of specific bacteria to environmental stimuli. However, periods of fast bacterial replication may not lead to increased population size because other processes exert controls on absolute abundances (e.g., predation and immune responses). In a few cases where community complexity was low, fast replication rates did predict an increase in absolute cell numbers in subsequent samples (**Figure 3.13 and Figure 3.11**). The fact that high replication rates do not necessarily predict increases in population size of bacteria growing in community context is unsurprising since iRep directly measures replication, which represents the physiological state of the organisms, but does not account for cell death rates. Replication rates and population size are distinct measurements, and both are important for studying microbial community dynamics.

An interesting question relates to how quickly organisms proliferate in the premature infant gut compared to the adult gut environment. Measurements in such environments are very challenging using alternative approaches such as isotope tracing (Kopf et al., 2015). These studies typically target specific organisms, and such measurements have only recently been implemented in the human lung microbiome (Kopf et al., 2015). Large-scale comparisons using PTR are not possible due to a lack of complete reference genomes. Using iRep, we found that bacteria from premature infant gut microbiomes had higher replication rates compared with those from a more complex adult gut consortium. If future studies confirm this finding, it might reflect greater levels of competition for resources or other factors related to gut development in adults compared to premature infants.

Candidate Phyla Radiation (CPR) organisms have been detected in a wide range of environments (Luef et al., 2015). Together, they make up considerably more than 15% of bacterial diversity (Brown et al., 2015; Hug et al., 2016), yet they are known almost exclusively from genomic sampling (Albertsen et al., 2013; Anantharaman et al., 2016a; Brown et al., 2015; Kantor et al., 2013; Nelson and Stegen, 2015; Podar et al., 2007; Rinke et al., 2013; Wrighton et al., 2012). Based on having small cells and genomes with only a few tens of ribosomes, it was inferred that these organisms grow slowly (Burstein et al., 2016; Luef et al., 2015). Our analysis of CPR organisms sampled across a range of geochemical gradients (Brown et al., 2015) directly demonstrated their slow replication rates. However, the analysis also showed that some CPR bacteria grow rapidly under certain conditions (**Figure 3.8**). Symbiosis has been inferred as a general life strategy for these organisms (Albertsen et al., 2013; Anantharaman et al., 2016a; Brown et al., 2015; Kantor et al., 2013; Nelson and Stegen, 2015; Podar et al., 2007; Rinke et al., 2013; Wrighton et al., 2012), and has been demonstrated in a few cases (Gong et al., 2014; He et al., 2015; Luo et al., 2016; Soro et al., 2014). Rapid growth of CPR bacteria may require rapid growth of host cells. If CPR cells typically depend on a specific bacterial host, as is the case for some Saccharibacteria (TM7) (He et al., 2015), replication rate measurements may provide insights into possible host-symbiont relationships, paving the way for co-cultivation studies.

It is important to consider factors that could lead to erroneous results. For example, the presence of multiple strains similar enough that their conserved genes co-assemble, could introduce error. This usually results in draft genomes that are so fragmented that they do not meet the genome quality requirements for iRep. However, error can also be introduced if a user maps reads from a sample containing multiple closely related strains to a high-quality genome reconstructed from a different sample. If the latter approach is used, we recommend checking for evidence of strain variation by analysis of polymorphism frequencies in mapped reads.

Conclusion

An important objective for microbial community studies is the establishment of models that can accurately predict microbial community dynamics and functions under changing environmental conditions. Prior to the current study, these models could include growth rate information derived from laboratory experiments involving isolates, inferred from fixed genomic features such as 16S rRNA gene copy number or codon usage bias (Vieira-Silva and Rocha, 2010), or from *in situ* measurements such as PTR (Korem et al., 2015). Further complicating matters, relative abundance measurements commonly determined from DNA sequencing can obscure understanding of population dynamics, and overall measurements of community composition can be confounded by the presence of DNA derived from dead cells (Carini et al., 2016). We used iRep to quantify replication rates for most bacteria in infant gut microbial communities and found that the rates can be highly variable (**Figure 3.11, Figure 3.12, and Figure 3.13**). Such measurements could be used in models that seek to understand microbial ecosystem functioning, allowing incorporation of organism-specific behavior throughout the study period. Importantly, iRep can be applied to identify actively growing bacterial populations in any ecosystem, regardless of how distantly related they are to cultivated bacteria, and to track bacterial replication in response to changing conditions. The ability to make these measurements has the potential to improve our understanding of relationships between bacterial functions and biogeochemical processes or health and disease.

Methods

Calculating bPTR for complete genomes

Our implementation of the PTR method (**see Code Availability**) differs from the method described by Korem *et al.* in several key respects (Korem et al., 2015). To distinguish between these two methods, we refer to our method as bPTR and the Korem *et al.* method as kPTR. Both methods involve mapping DNA sequencing reads to complete (or near-complete, in the case of bPTR) genome sequences in order to measure differences in sequencing coverage at the origin (Ori_{cov}) and terminus (Ter_{cov}) of replication.

$$PTR = \frac{Ori_{cov}}{Ter_{cov}}$$

kPTR makes use of a database of reference genome sequences, whereas bPTR is designed to be more flexible and can use mapping of reads to any genome sequence. For our bPTR analyses, we used Bowtie2 (Langmead and Salzberg, 2012) with default parameters for read mapping.

Both bPTR and kPTR can determine the location of the origin and terminus of replication of growing cells by identifying coverage "peaks" and "troughs" associated with these positions. Identification of the origin and terminus of replication requires measuring changes in coverage along the genome sequence. This is accomplished by calculating the average coverage over 10 Kbp windows at positions along the genome separated by 100 bp. To increase the accuracy of results, a mapping quality threshold can be used in which both reads in a set of paired reads are required to map to the genome sequence with no more than a specified number of mismatches (this option is unique to bPTR). Since highly conserved regions, strain variation, or integrated phage can result in highly variable coverage, high and low coverage windows are filtered out of the analysis. Coverage windows are excluded if the values differ from the median by a factor greater than eight (threshold also used by kPTR), or if the values differ from the average of 1,000 neighboring coverage windows by a factor greater than 1.5 (threshold unique to bPTR). If more than 40% of the windows are excluded, no bPTR value will be calculated (threshold also used by kPTR). The origin and terminus are identified by fitting a piecewise linear function to the filtered, $\log_2$-transformed coverage values. Coverage values are $\log_2$-transformed to improve fitting, but the transformation is reversed prior to calculating bPTR. Fitting is conducted as described by Korem *et al.* by non-linear least squares minimization using the Levenberg-Marquardt algorithm implemented by lmfit (Newville et al., 2014).

Piecewise linear function modified from Korem *et al.*:

$$f(x) = \begin{cases} -ax + y_1 + ax_1, & x \leq x_1 \\ ax + y_1 - ax_1, & x_1 < x < x_2 \\ -ax + y_2 + ax_2, & x \geq x_2 \end{cases}$$

$$a = \frac{Ter_{cov} - Ori_{cov}}{Ter_{loc} - Ori_{loc}}$$

$$x_1 = \min\left(Ter_{loc}, Ori_{loc}\right) \qquad y_1 = \begin{cases} Ter_{cov} \; if \; x_1 = Ter_{loc} \\ Ori_{cov} \; if \; x_1 = Ori_{loc} \end{cases}$$

$$x_2 = \max\left(Ter_{loc}, Ori_{loc}\right) \qquad y_2 = \begin{cases} Ter_{cov} \; if \; x_2 = Ter_{loc} \\ Ori_{cov} \; if \; x_2 = Ori_{loc} \end{cases}$$

Ori_{loc}

Ori_{loc} and Ter_{loc} refer to the locations of the origin and terminus of replication, respectively, and Ori_{cov} and Ter_{cov} refer to $\log_2$-transformed coverage at those positions. All *x* values refer to positions on the genome, and *y* values to $\log_2$-transformed coverage values. The fitting is constrained such that Ori_{loc} and Ter_{loc} are separated by 45-55% of the genome length (Korem et al., 2015). In order to reduce the amount of noise introduced by fluctuations in sequencing coverage, a median filter is applied to the coverage data before calculating bPTR. This smoothing operation replaces the coverage value at each position with the median of values sampled from the 1,000 neighboring windows. The $\log_2$-transformed, median-filtered values corresponding with Ori_{loc} and Ter_{loc} ($Ori_{cov\text{-}med}$ and $Ter_{cov\text{-}med}$, respectively) are used to calculate bPTR.

Since the values have been $\log_2$-transformed, the final value is calculated as:

$$bPTR = \frac{2^{Ori_{cov-med}}}{2^{Ter_{cov-med}}}$$

Ori_{loc} and Ter_{loc} are determined based on sequencing from each available sample. In order to calculate bPTR using the same positions for all samples, consensus Ori_{loc} and Ter_{loc} positions are determined by finding the circular median of the positions determined from each individual sample (all Ori_{loc} and Ter_{loc} positions with bPTRs ≥1.1 are considered), as is done for kPTR (Korem et al., 2015). Once these values are determined, all bPTR values are re-calculated using the coverage at the consensus positions. It is important to note that Ori_{loc} and Ter_{loc} may vary depending on what samples are analyzed, and that with bPTR this can be avoided by using GC skew to identify Ori_{loc} and Ter_{loc} (**see below**).

For bPTR, we added the option to find Ori_{loc} and Ter_{loc} based on GC skew. GC skew is calculated over 1 Kbp windows at positions along the genome separated by 10 bp. Since Ori_{loc} and Ter_{loc} coincide with a transition in the sign (+/-) of GC skew, these positions can be identified as the transition point in a plot of the cumulative GC skew (Grigoriev, 1998) (for examples see **Figure 3.7, Figure 3.5, and Figure 3.10**). These transition points are identified by finding extreme values in the cumulative GC skew data separated by 45-55% of the genome length. Once Ori_{loc} and Ter_{loc} are identified, bPTR is calculated from median-filtered $\log_2$-transformed coverage values calculated over sliding windows as described above. bPTR provides visual representation of both coverage and GC skew patterns across genome sequences that enable verification of genome assemblies and predicted Ori_{loc} and Ter_{loc} positions (this visualization is not provided by kPTR).

Calculating the Index of Replication (iRep) for complete and draft-quality genomes

iRep analyses are conducted by first mapping DNA sequencing reads to genome sequences with Bowtie2 (default parameters). For genomes in multiple pieces, the coverage values determined at each position along the fragments are combined, and then average coverage is calculated over 5 Kbp windows at positions along the concatenated genome that are separated by 100 bp (**Figure 3.2; see Figure 3.3 and below** for accuracy metrics related to sliding window calculations). As with bPTR, a mapping quality threshold can be used to increase the accuracy of results by ensuring that both reads in a set of paired reads mapped to the genome sequence with no more than a specified number of mismatches. Coverage values from the first and last 100 bp of each scaffold are excluded due to possible edge effects. Coverage windows are filtered out of the analysis if the values differ from the median by a factor greater than eight, and then GC sequencing bias is measured and corrected (**see below**). Coverage values are $\log_2$-transformed and then sorted from lowest to highest coverage. Because the coverage windows are re-ordered in this step, it does not matter if the correct order of genome fragments is unknown. The lowest and highest 5% of sequences are excluded, and then the slope of the remaining coverage values is determined by linear regression. As with bPTR, $\log_2$-transformations are conducted to improve regression analysis, but are removed before comparing coverage values. iRep, which is a measure of the ratio between Ori_{cov} and Ter_{cov}, can be determined based on the slope (m) and y-intercept (which is synonymous with Ter_{cov}, **see Figure 3.2**) of the regression line, and the total length of the genome sequence (l):

$$iRep = \frac{m \text{ x } l + Ter_{cov}}{Ter_{cov}}$$

However, since the values have been $\log_2$-transformed, the final value is calculated as:

$$iRep = 2^{m \text{ x } l}$$

Since partial genome sequences will include a random assortment of genome fragments, the coverage trend determined from the available sequence will be representative of the coverage trend across the complete genome. Several quality thresholds are used to ensure the accuracy of iRep measurements: i) coverage depth must be ≥5x, ii) ≥98% of the genome sequence must be included after filtering coverage windows, and iii) r^2 values calculated between the coverage trend and the linear regression must be ≥0.90. These criteria are important because they ensure that enough sequencing data is present to achieve accurate measurements, and that the genome sequence is appropriate for the analysis. The 98% genome sequence coverage threshold differs from the genome completeness requirement in that this is not a measure of the quality of the genome assembly, but rather a measure of the overlap between a genome sequence and the sequencing data. Low values would indicate that the genome used for mapping is not appropriately matched with an organism present in the system. Likewise, having a strong fit of the linear regression to the coverage data indicates that sequencing coverage calculations are not influenced by strain variation, choice of an inappropriate genome sequence, or other factors that may skew replication rate measurements.

Both PTR methods involve calculations based on only two data points (Ori_{cov} and Ter_{cov}). In contrast, iRep uses coverage trends determined across an entire genome sequence, and thus is less susceptible to noise in sequencing coverage or errors in the prediction of *Ori_{loc}* and *Ter_{loc}*. Further, since both PTR methods involve predicting *Ori_{loc}* and *Ter_{loc}* based on data from multiple samples, the same positions may not be chosen for different analyses. This makes it difficult to reproduce and compare results (an issue that can be avoided by predicting *Ori_{loc}* and *Ter_{loc}* using cumulative GC skew and bPTR). iRep calculations do not depend on analysis of multiple samples, and thus results will not change based on what samples are included in an analysis. Since the order of genome fragments need not be known when calculating iRep, the method is not affected by genome assembly errors, which are present even in some genome sequences reported to be complete (**Figure 3.10**).

Determining the minimum sequencing coverage required for iRep analysis

Lactobacillus gasseri data from the Korem *et al.* study was used to determine the minimum coverage required for iRep, bPTR, and kPTR. Reads from each sample were first mapped to the complete genome sequence, and then subsampled to 25x before calculating iRep, bPTR, and kPTR. Then, each mapping was further subsampled to lower coverage levels (20x, 15x, 10x, 5x, and 1x) and replication rates were re-calculated using each method. Comparison of these values to those determined at 25x coverage enable quantification of the amount of noise introduced by increasingly lower coverage (**Fig. 2c and Supplementary Table 3**).

Determining genome quality requirements for iRep analysis

The *L. gasseri* data from Korem *et al.* subsampled to 25x coverage was also used to test the minimum fraction of a genome required for obtaining accurate iRep measurements. Four samples representing iRep values between 1.50 and 2.01 were selected in order to test the effect of missing genomic information across a range of replication rates. Genome subsampling experiments were conducted on each sample in order to evaluate the amount of noise introduced by missing genomic information. For each tested genome fraction (90%, 75%, 50%, and 25%), iRep was calculated for 100 random genome subsamples. For each subsample, the genome was fragmented into pieces with lengths determined by selecting from a gamma distribution modeled after the size of genome fragments expected for draft-quality genome sequences (alpha = 0.1, beta = 21,000, minimum length = 5 Kbp, maximum length = 200 Kbp; **Figure 3.3**). Once fragmented, the pieces were randomly sampled until the desired genome fraction was achieved. Partial fragments were included in order to prevent the desired genome fraction size from being exceeded. In order to ensure that the results were accurate even when sequencing coverage is low, iRep calculations were conducted after subsampling reads to 5x coverage. iRep values calculated after subsampling were compared to values determined at 25x coverage with the complete genome sequence in order to measure the combined affect of lower coverage and missing genome sequence information (**Figure 3.4d and Supplementary Table 3.1**). In order to determine the effect of increased genome fragmentation on iRep calculations, additional genome fragmentation experiments were conducted in which the minimum and maximum allowed fragment lengths were varied in order to determine the effects of higher than normal levels of genome fragmentation (**Figure 3.3b and Supplementary Table 3.1**).

Evaluation of iRep sliding window calculation methods

The accuracy of iRep when implemented using different sliding window coverage calculation methods was determined based on additional random genome fragmentation experiments using the *L. gasseri* data from Korem *et al.* **Figure 3.3c-e and Supplementary Table 3.1**). Three sliding window methods were tested: i) the method implemented in iRep (**described above** and referred to as the "iRep" method), ii) as implemented in iRep, except for that the iRep value is taken as the median of ten iRep values each obtained after concatenating available genome fragments in different arrangements (referred to as the "median iRep" method), and iii) obtained after calculating coverage sliding windows for each fragment individually, and then combining the sliding window data (referred to as the "scaffold windows" method). The amount of noise in the iRep calculation using each method was determined based on comparing iRep values achieved with 5x sequencing coverage and varying levels of genome completeness (**see above**) to values determined based on the standard iRep method and the complete genome sequence with 25x sequencing coverage (**Figure 3.3c**). This was repeated using different sliding window sizes in order to determine the optimal method. Furthermore, the range of iRep values obtained for tests using the "median iRep" method was used to determine the amount of noise introduced when scaffold coverage data is concatenated in a random order prior to conducting sliding window calculations (**Figure 3.3d**). Because the standard iRep method with 5 Kbp windows was determined to be the best, a final test of this method was conducted in order to compare different window slide lengths (**Figure 3.3e)**.

Correcting for GC sequencing bias

DNA sequencing platforms are biased towards sequences based on their GC content (Ross et al., 2013). Because this bias can result in a difference in the sequencing coverage across a genome sequence, it could influence iRep results. To account for this, GC sequencing bias is measured and corrected independently for each genome and metagenome. This is accomplished by first determining the GC content of sliding windows across the genome sequence that correspond with the coverage measurements used for calculating iRep. Then, linear regression is conducted between the coverage and GC values determined for each sliding window. In order to get an accurate measurement, linear regression is conducted in two steps: first with the complete data set and then after removing the 1% of data points with the largest deviation from the initial regression analysis. Then, the results of the filtered regression analysis are used to correct the coverage values for each sliding window. This method was used in the analyses of all metagenome data in this study, and is part of the iRep code (**Code Availability**). The GC sequencing bias correction resulted in better agreement between iRep and bPTR values determined using ordered and oriented genomes reconstructed from the premature infant dataset (**Figure 3.3f; see below**).

Comparative analyses of replication rate methods

iRep, bPTR, and kPTR were calculated for all samples from the Korem *et al. L. gasseri* experiments (these were the only samples sequenced to a high enough depth to enable comparison with iRep; **Supplementary Table 3.3**). For a subset of these data, replication rates could also be calculated based on counts of colony forming units (CFU/ml) (Korem et al., 2015) (**Figure 3.4b and Supplementary Table 3.2**). Pearson's correlations were calculated between replication rates based on CFU/ml data and iRep, bPTR, and kPTR, after first accounting for the time delay between start of genome replication and observable change in population size (as previously noted(Korem et al., 2015)). The time delay was determined independently for each method as the delay that resulted in the highest correlation.

iRep and bPTR values were compared for a novel Deltaproteobacterium after manually curating a recently reported draft genome sequence (Sharon et al., 2015) (**see below**). Reads from the GWC2 sample from Brown *et al.* were used to conduct the analysis (**Figure 3.7**). For this comparison, and all subsequent iRep and bPTR calculations, coverage was calculated based on reads that mapped to the genome fragment with no more than two mismatches (**see above for details**). Although enough of the genome sequence was assembled in order to calculate bPTR, the results could not be compared with kPTR because a complete reference genome sequence was not available.

In order to further compare iRep and bPTR in the context of microbial community sequencing data, bPTR values were calculated using genomes reconstructed from the premature infant dataset (Raveh-Sadka et al., 2015) that were ordered and oriented based on complete reference genome sequences (**see below; Figure 3.4e and Supplementary Table 3.4**). Although these genomes were similar enough to reference genomes to facilitate ordering and orienting the sequences, the reference genomes themselves were too divergent to facilitate replication rate calculations (**see Results; Figure 3.6**), which prevented inclusion of kPTR in this analysis.

Manual curation of a Deltaproteobacterium genome

The genome sequence of a previously reported Deltaproteobacterium was manually curated. Unplaced or misplaced paired-read sequences were used to fill scaffolding gaps, correct local assembly errors, and extend scaffolds. Overlapping scaffolds were combined when the join was supported by paired read placements. The final assembled sequence was visualized to confirm that all errors had been corrected.

Ordering and orienting draft genomes based on complete reference genomes

Reference genomes similar to draft genomes were obtained from NCBI GenBank. Genomes with aberrant GC skew patterns were not used for ordering draft genomes as they likely contain assembly errors. The average nucleotide identities (ANI) between each draft genome and associated reference genomes were calculated using the ANIm method(Richter and Rossello-Mora, 2009), 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 less than 20% alignment coverage was discarded. The remaining sequence was then aligned to the reference genome using progressive Mauve (Rissman et al., 2009), resulting in an ordered and oriented genome to be used for calculating bPTR. These genomes were manually inspected and curated based on cumulative GC skew and genome coverage patterns based on graphs generated by the bPTR script (**Figure 3.5**).

iRep measurements for premature infant metagenomes

Previously reconstructed genomes from the premature infant gut microbiome study (Raveh-Sadka et al., 2015) were included in the iRep analysis if they were estimated to be ≥75% complete based on analysis of universal single copy genes (SCGs), had no more than two duplicate SCGs, and had less than 175 fragments/Mbp of sequence. In order to maximize the number of iRep values that could be determined, custom read mapping databases were used for each metagenome. Each database was constructed by first including genomes reconstructed from the metagenome that passed the above thresholds, and then by adding additional draft-quality genomes reconstructed from other metagenomes from the same infant. This prioritizes genomes reconstructed from the metagenome used for mapping, but also attempts to include genomes from organisms that may have been present, but for which a genome sequence was not assembled.

Overlap in community membership across time-series studies results in the same genome sequence being reconstructed in multiple samples. Including highly similar or identical genome sequences in databases used for read mapping would lead to aberrant coverage calculations. This becomes a concern when including genomes reconstructed from additional samples in read mapping databases for iRep calculations. To prevent adding highly similar genomes to the databases, only the representatives of 98% ANI genome clusters (**see below**) were added to mapping databases, and only if a representative of the cluster was not already included. Consistent with clustering genomes based on sharing 98% ANI, iRep calculations were conducted based on coverage calculations determined from reads mapping to genomes with no more than two mismatches (**see above for details; Supplementary Table 3.5**).

Clustering genomes based on average nucleotide identity (ANI)

Average nucleotide identity was determined between all pairs of genome sequences using the Mash algorithm (Ondov et al., 2016) (kmer set to 21). Clusters were defined by selecting groups of genomes connected by ≥98% ANI. Representatives of each cluster were chosen by selecting the longest genome with less than 175 fragments/Mbp that had the most SCGs and the fewest SCG duplicates.

Comparison of iRep and kPTR measurements for premature infant gut metagenomes

The kPTR software from Korem *et al.* was run on the premature infant metagenomes (Raveh-Sadka et al., 2015) (**Supplementary Table 3.8**). Comparisons between iRep and kPTR were made when it was possible to link the name of the genome provided by kPTR with the taxonomy given to reconstructed genome sequences (**Supplementary Table 3.5**).

Genome binning and iRep measurements for adult human metagenomes

Genomes were binned from the adult human metagenome (Di Rienzi et al., 2013) based on coverage, GC content, and taxonomic affiliation using ggKbase tools (ggkbase.berkeley.edu), as previously described (Brown et al., 2015; Raveh-Sadka et al., 2015). Genome completeness was evaluated based on the fraction of universal single copy genes (Raes et al., 2007; Raveh-Sadka et al., 2015) that could be identified (**Supplementary Table 3.6**). Genomes estimated to be ≥75% complete, with no more than two additional single copy genes, and no more than 175 fragments per Mbp of sequence, were used in the analysis. iRep was conducted using reads mapped to genomes with no more than two mismatches (**Supplementary Table 3.7**).

bPTR and kPTR measurements from the adult human metagenome

The kPTR software from Korem *et al.* was run on the adult human metagenome (Di Rienzi et al., 2013) (**Supplementary Table 3.9**). bPTR calculations were conducted based on mapping metagenome reads to selected complete reference genomes (≤2 mismatches; **Figure 3.10**). Reference genomes for bPTR analysis were selected by searching scaffolds from reconstructed genome sequences against complete genomes from NCBI GenBank. The complete genome with the best BLAST hit to each reconstructed genome was selected for bPTR analysis.

iRep measurements for Candidate Phyla Radiation (CPR) organisms

CPR genomes identified by Brown *et al.* to be ≥75% complete, with no more than two additional single copy genes, and no more than 175 fragments per Mbp of sequence, were selected for iRep analysis. These genomes were reconstructed previously from multiple metagenomes spanning an acetate amendment time-series field experiment. Reads from each of 12 metagenomes sequenced from groundwater filtrates, collected from serial 0.2 and 0.1 µm filters at six time points, were mapped to the genome sequences for iRep calculations (≤2 mismatches; **Supplementary Table 3.10**).

Absolute abundance and doubling time determinations

Raveh-Sadka *et al.* determined the concentration of cells in each collected fecal sample using droplet-digital PCR (Raveh-Sadka et al., 2015). In this study, the population size of each species was determined by multiplying total cell counts by the fractional (relative) abundance calculated based on genome sequencing (**Figure 3.11 and Supplementary Table 3.5**). These values were used to calculate the doubling time for *Klebsiella oxytoca* (**Figure 3.13**).

Code Availability

iRep and bPTR software are maintained under github.com/christophertbrown/iRep (v1.10 used in this analysis: github.com/christophertbrown/iRep/releases/tag/v1.10).

Data Availability

DNA sequencing reads are available from the NCBI Sequence Read Archive for the groundwater (Brown et al., 2015) (SRP050083), premature human infant (Raveh-Sadka et al., 2015) (SRP052967), and adult human (Di Rienzi et al., 2013) (SRR3496379) microbiome projects. Genomes analyzed as part of this study are available from ggKbase for the groundwater (Brown et al., 2015) (ggkbase.berkeley.edu/CPR-complete-draft/organisms), premature human infant (ggkbase.berkeley.edu/project_groups/necevent_samples), and adult human (ggkbase.berkeley.edu/LEY3/organisms) datasets, as well as for the curated novel Deltaproteobacterium (ggkbase.berkeley.edu/novel_delta_irep/organisms). CPR genomes (BioProject PRJNA273161) and adult human microbiome genomes (BioProject PRJNA321218) are available from NCBI GenBank, and the Deltaproteobacterium genome from DDBJ/ENA/GenBank under the accession LVEI00000000 (version LVEI02000000 described here; **see Supplementary Tables** for additional accession numbers).

Author Contributions

CTB and JFB developed the iRep and bPTR methods. MRO ordered and oriented draft genome sequences for bPTR calculations and conducted kPTR analyses. CTB conducted the iRep, bPTR, and kPTR comparisons, and determined the accuracy of the iRep method. JFB binned the adult human metagenome and curated the Deltaproteobacterium genome, with input from CTB. CTB implemented the iRep method. BCT provided bioinformatics support. CTB and JFB drafted the manuscript. All authors contributed to iRep development, reviewed results, and approved the manuscript. The authors declare no competing financial interests.

Acknowledgements

Funding was provided by NIH grant R01AI092531, Sloan Foundation grant APSF-2012-10-05, and by the US Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research under award number DE-AC02-05CH11231 (Sustainable Systems Scientific Focus Area and DOE-JGI) and award number DE-SC0004918 (Systems Biology Knowledge Base Focus Area). We thank T. Raveh-Sadka, B. Brooks, and D. Burstein for helpful discussions, and M. Albertsen for comments regarding GC sequencing bias.

Supplementary Tables

Supplementary Table 3.1 | Analysis of the impact of genome completeness on iRep replication rate measurements. iRep was first calculated using sequencing data from *Lactobacillus gasseri* (NC_008530) experiments (Korem et al., 2015) with the complete genome at 25x coverage, and then compared with values calculated at 5x coverage after subsampling the genome 100 times for each targeted percent of the genome sequence ("delta"). This was conducted using different coverage sliding window methods (**see Methods** for descriptions), sliding window sizes, and window slide lengths. In order to test different levels of genome fragmentation, the minimum and maximum allowed fragment size was also varied. iRep range is the difference between the minimum and maximum iRep values determined when using the "median iRep" sliding window calculation method.

Supplementary Table 3.2 | Comparison of iRep, bPTR, and kPTR measurements. The Korem *et al. Lactobacillus gasseri* (NC_008530) data was used to measure replication rates using iRep, bPTR, and kPTR. iRep % windows refers to the percent of coverage windows that passed the iRep filters, and iRep r^2 is the r^2 value calculated between the sequencing coverage trend and regression used for calculating iRep. Coverage is the average sequencing depth calculated across the genome sequence. Colony forming units per ml of culture (CFU/ml) was obtained from the Korem *et al.* study.

Supplementary Table 3.3 | iRep, bPTR, and kPTR measurements for minimum genome sequencing coverage analyses. Genome coverage tests were conducted using *Lactobacillus gasseri* (NC_008530) data from previously published experiments (Korem et al., 2015). Target coverage is the level of coverage achieved after sub-sampling sequencing reads.

Supplementary Table 3.4 | Comparison of iRep and bPTR measurements for draft-quality genomes ordered and oriented based on complete genome sequences. Genomes reconstructed for organisms sampled as part of the Raveh-Sadka *et al.* premature infant dataset were ordered based on complete reference genome sequences. iRep and bPTR values were calculated for all pairs of genomes and samples by mapping reads from the samples to the genome sequences. GC r^2 is the r^2 value from the linear regression between sequencing coverage and GC content that is used for correcting GC sequencing bias. GC bias is the GC r^2 value multiplied by the slope of the regression line, and is a measure of the magnitude and direction of GC sequencing bias. Un-filtered iRep values include iRep values that may not have passed the quality thresholds. Raw iRep values are iRep values determined without the GC sequencing bias correction. Fragments/Mbp is the number of genome fragments per Mbp of genome sequence. Coverage breadth is the percent of the genome covered by sequencing reads.

Supplementary Table 3.5 | iRep measurements for organisms associated with premature infant microbiomes. iRep measurements were determined using genomes and metagenomes from the Raveh-Sadka *et al.* premature infant dataset. DOL = day of life and NEC = necrotizing enterocolitis. The DOL – sample column indicates whether additional samples were collected on a particular day, the DOL – NEC diagnosis column includes day of life relative to NEC diagnosis, and condition indicates whether or not the infant was diagnosed with NEC. The antibiotics column indicates whether or not antibiotics were administered at, or within five days prior to, the time of sample collection. Relative abundance was calculated for each organism based on the number of sequencing reads mapped to the genome sequence as a percent of sequences mapped to all draft-quality genomes. Absolute abundance (cells/g) was determined for each organism based on relative abundance and previously published ddPCR measurements of total cells/g of feces (Raveh-Sadka et al., 2015). kPTR values are provided for cases where there was a clear match with results from the kPTR software (**Supplementary Table 3.8**).

Supplementary Table 3.6 | Single copy gene inventory for genomes reconstructed from an adult human gut metagenome. Genomes were binned as part of this study from a previously published metagenome dataset from Di Rienzi *et al.* (SAMN04978193). The number of single copy marker genes, which can be used as a proxy for genome completeness, was determined for each genome.

Supplementary Table 3.7 | iRep measurements for organisms associated with an adult human microbiome. iRep measurements were determined using the metagenome from Di Rienzi *et al.* (SAMN04978193) and the genomes reconstructed as part of this study.

Supplementary Table 3.8 | kPTR values determined from the premature infant metagenomes. The kPTR software was used to measure replication rates for organisms represented in the Raveh-Sadka *et al.* premature infant metagenomes.

Supplementary Table 3.9 | kPTR values determined from the adult human metagenome. The kPTR software was used to measure replication rates for organisms represented in the metagenome from Di Rienzi *et al.* (SAMN04978193).

Supplementary Table 3.10 | iRep measurements for Candidate Phyla Radiation (CPR) organisms. iRep measurements were determined using genomes and metagenomes from the Brown *et al.* CPR dataset.

Figure 3.1 | iRep determines replication rates for bacteria using genome-resolved metagenomics. (**a**) Populations of bacteria undergoing rapid cell division differ from slowly growing populations in that the individual cells of a growing population are more actively in the process of replicating their genomes (purple circles). (**b**) Differences in genome copy number across a population of replicating cells can be determined based on sequencing read coverage over complete genome sequences. The ratio between the coverage at the origin ("peak") and terminus ("trough") of replication (PTR) relates to the replication rate of the population. The origin and terminus can be determined based on cumulative GC skew. (**c-d**) If no complete genome sequence is available, it is possible to calculate the replication rate based on the distribution of coverage values across a draft-quality genome using the iRep method. Coverage is first calculated across overlapping segments of genome fragments. Growing populations will have a wider distribution of coverage values compared with stable populations (histograms). These values are ordered from lowest to highest, and linear regression is used to evaluate the coverage distribution across the genome in order to determine the coverage values associated with the origin and terminus of replication. iRep is calculated as the ratio of these values. (**e**) Genome-resolved metagenomics involves DNA extraction from a microbiome sample followed by DNA sequencing, assembly, and genome binning. Binning is the grouping together of assembled genome fragments that originated from the same genome. This can be done based on shared characteristics of each fragment, such as sequence composition, taxonomic affiliation, or abundance.

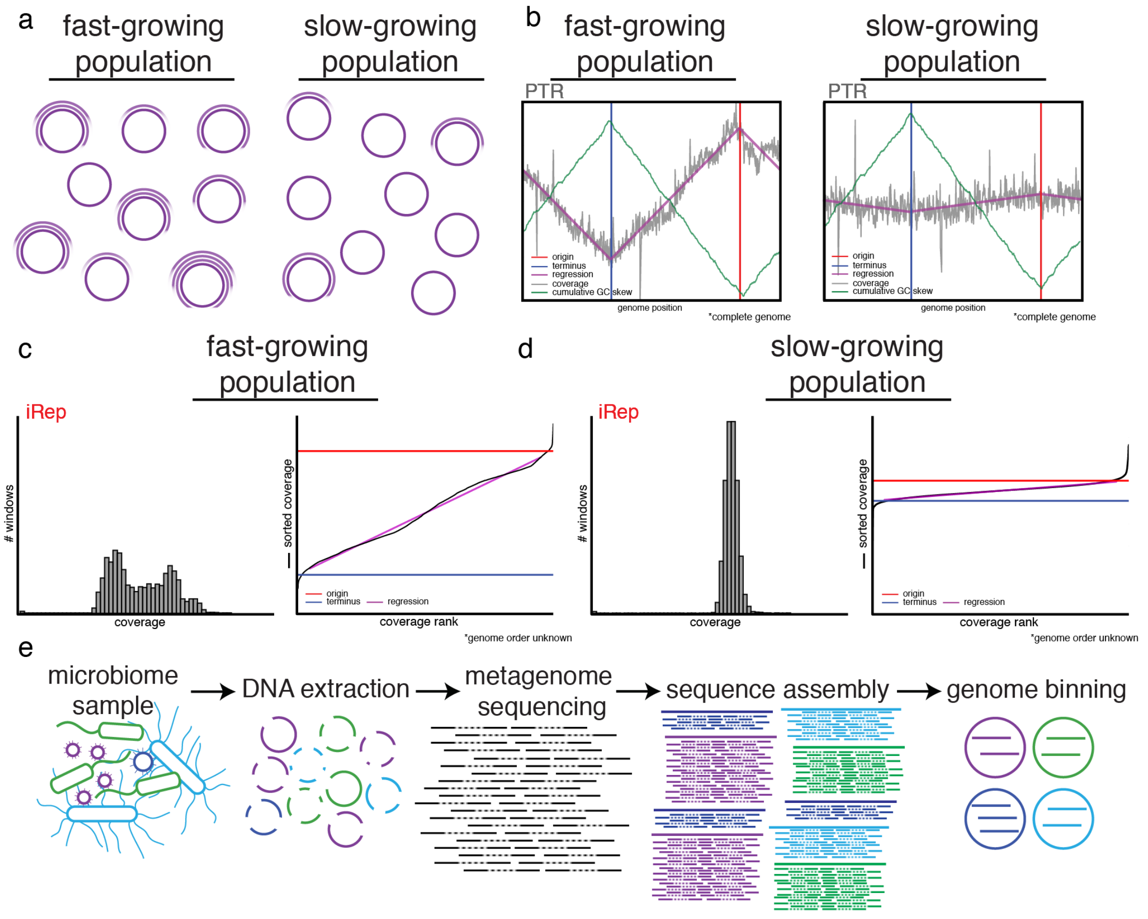
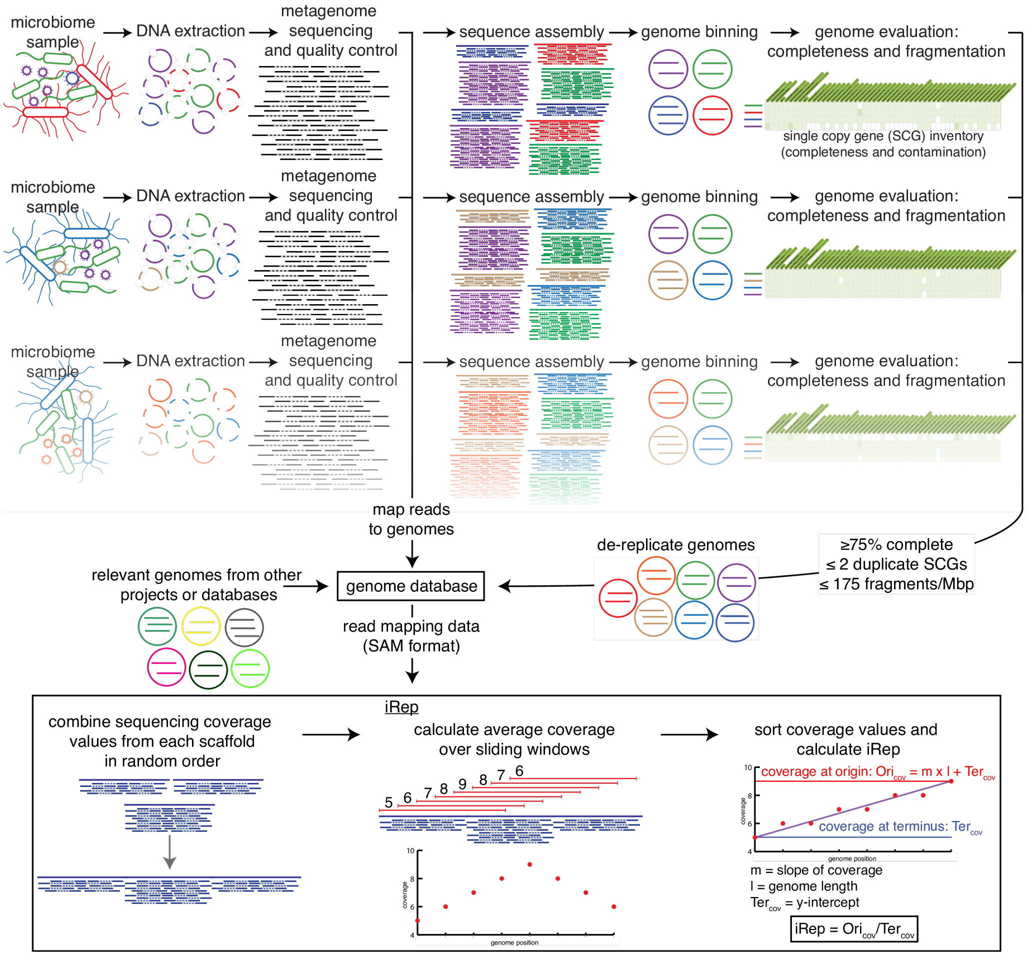


Figure 3.2 | Schematic showing steps involved in a genome-resolved metagenomics study that includes iRep analysis. Microbiome sample collection and DNA extraction methods should be determined on a per-project basis, and metagenome sequencing can be conducted on the Illumina, PacBio, or another sequencing platform. Sequencing reads are trimmed based on quality scores (e.g. using Sickle (Joshi)) and filtered for contamination (e.g. removal of human genome sequences). High-quality reads are then assembled (e.g. using IDBA_UD (Peng et al., 2012)), and the resulting scaffolds are binned either manually (e.g. based on GC content, taxonomic affiliation, coverage), and/or using a clustering algorithm such as ESOM (Dick et al., 2009; Raveh-Sadka et al., 2015; Sharon et al., 2012)) or using an automated binning program (e.g. MaxBin (Wu et al., 2015), CONCOCT (Alneberg et al., 2014), or ABAWACA (Brown et al., 2015)). Genome bins can then be assessed for completion and contamination based on inventory of expected single copy genes (SCGs), either based on identification of these genes from genome annotations (see (Brown et al., 2015; Raes et al., 2007; Raveh-Sadka et al., 2015)), or using software such as CheckM (Parks et al., 2015). High-quality genomes are then compared with one another and grouped into clusters based on average nucleotide identity (ANI; e.g., based on sharing 98% ANI determined using Mash (Ondov et al., 2016)). A representative of each cluster should be included in a genome database that will be used for iRep analysis, along with genomes from other projects that may be appropriate for the analysis. Reads from each metagenome are then mapped to the genome database (e.g. using Bowtie2 (Langmead and Salzberg, 2012)), and iRep is calculated from the read mapping data (**see Methods**).



microbiome sample
DNA extraction
metagenome sequencing and quality control
sequence assembly
genome binning
genome evaluation: completeness and fragmentation
single copy gene (SCG) inventory (completeness and contamination)
microbiome sample
DNA extraction
metagenome sequencing and quality control
sequence assembly
genome binning
genome evaluation: completeness and fragmentation
microbiome sample
DNA extraction
metagenome sequencing and quality control
sequence assembly
genome binning
genome evaluation: completeness and fragmentation
map reads to genomes
de-replicate genomes
≥75% complete
≤ 2 duplicate SCGs
≤ 175 fragments/Mbp
relevant genomes from other projects or databases
genome database
read mapping data (SAM format)
iRep
combine sequencing coverage values from each scaffold in random order
calculate average coverage over sliding windows
5 6 7 8 9 8 7 6
10
8
6
4
coverage
genome position
sort coverage values and calculate iRep
coverage at origin: $Ori_{cov} = m \times l + Ter_{cov}$
coverage at terminus: Ter_{cov}
10
8
6
4
coverage
genome position
m = slope of coverage
l = genome length
Ter_{cov} = y-intercept
$iRep = Ori_{cov}/Ter_{cov}$

Figure 3.3 | Evaluation of iRep method parameters. (**a**) Gamma distribution used to simulate genome fragmentation for genome completeness analyses. The frequency of genome fragment sizes from all genomes analyzed in this study are compared with genome fragment sizes simulated using a gamma distribution with parameters: alpha = 0.1, beta = 21,000, min. = 5,000, max. = 200,000. These parameters were first estimated by fitting to the genome data, and then manually adjusted. Similarity between the two distributions shows that this gamma distribution can be used to approximate the level of genome fragmentation expected for draft-quality genome sequences. (**b**) iRep was calculated from random genome fragmentation simulations in order to survey a range of fragmentation levels (**Supplementary Table 3.1**). The analysis was conducted for an *L. gasseri* sample from the Korem *et al.* study in which iRep was determined to be 2.01 using the complete genome with 25x sequencing coverage. This known iRep value was then compared with iRep values determined from each genome fragmentation simulation after subsampling to 75% of the genome and using only 5x sequencing coverage. This enabled analysis of the influence of fragmentation on iRep calculations at the completeness and coverage limits of the method. Results show that 91.8% of iRep values are within the expected range of 0.15 when genomes have fewer than 175 fragments/Mbp of genome sequence. (**c**) Four *L. gasseri* samples from the Korem *et al.* study that represent iRep values between 1.50 and 2.01 were selected in order to test different coverage sliding window calculation methods (**see Methods** for description of each method) and window sizes. For each sample, 100 random genome fragmentations and subsets were conducted in order to assess each method based on various levels of genome completion. The results show that the "iRep" and "median iRep" methods using 5 Kbp windows exhibited the least amount of variation. (**d**) Because the iRep method involves randomly combining coverage data from different genome fragments prior to calculating coverage sliding windows, some sliding windows will include coverage values from different locations on the complete genome sequence. In order to evaluate the variation introduced by the (random) order in which scaffolds are combined, iRep calculations were conducted for ten random orderings of 100 random genome fragmentations conducted using the sample set described in (c). Results show a very minimal amount of variation in iRep values as described by the difference between the lowest and highest values determined from each of the ten orderings ("iRep range"). Because of this, we chose not to implement the "median iRep" strategy. (**e**) Using the sample set described in (c), the iRep method was implemented using 5 Kbp windows using different window slide values in order to test whether or not the slide value would change the results. Because both 10 and 100 bp window slides produced similar results, we implemented the iRep method using a 100 bp window slide. (**f**) iRep is not as strongly correlated with bPTR without the GC sequencing bias correction for five genome sequences assembled from premature infant metagenomes (**Supplementary Table 3.4**; compare with GC corrected data in **Figure 3.4e**).

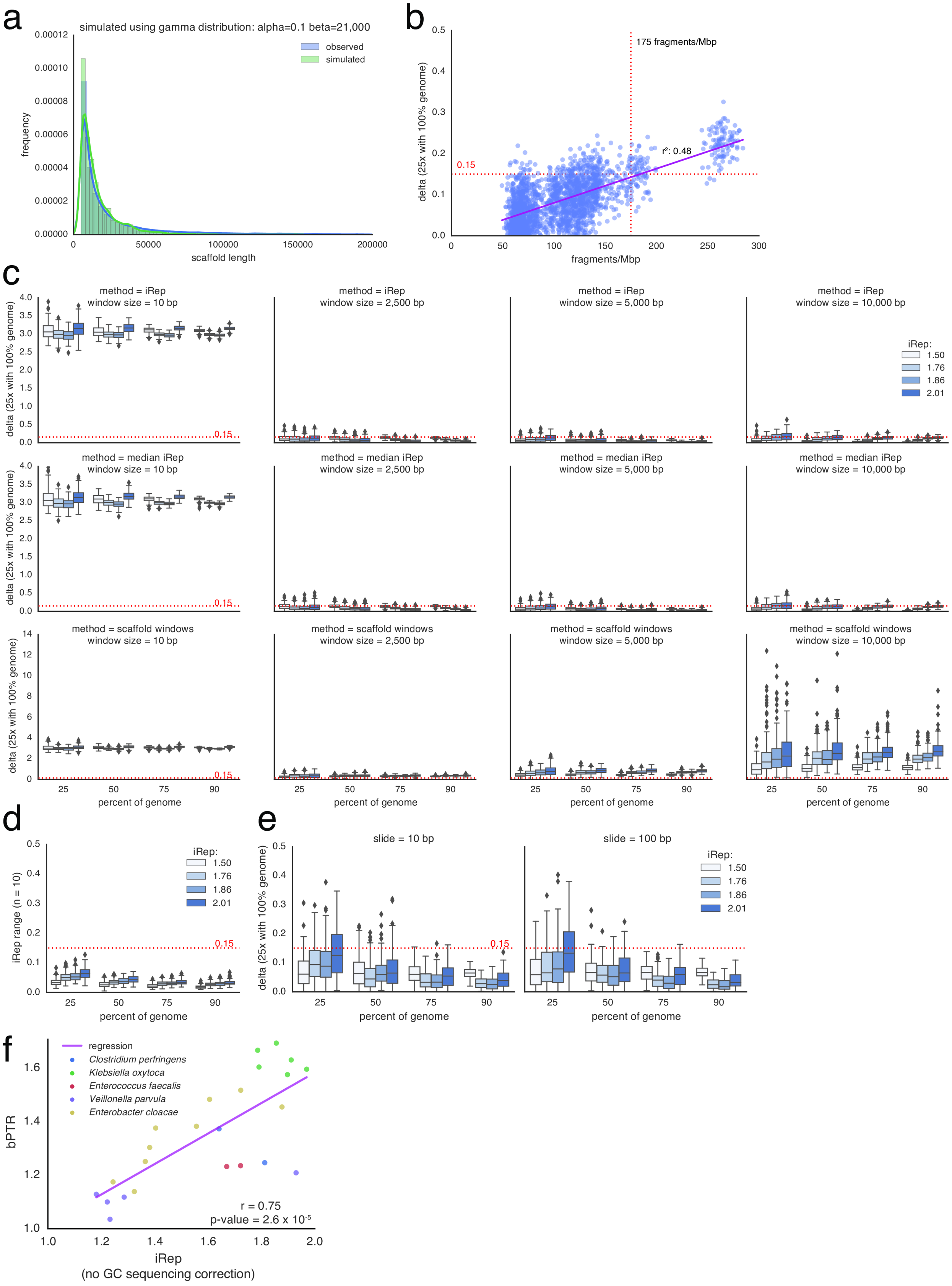

a
simulated using gamma distribution: alpha=0.1 beta=21,000
observed
simulated
frequency
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
50000
100000
150000
200000
scaffold length
b
175 fragments/Mbp
r²: 0.48
0.15
delta (25x with 100% genome)
0.5
0.4
0.3
0.2
0.1
0.0
0
50
100
150
200
250
300
fragments/Mbp
c
method = iRep
window size = 10 bp
method = iRep
window size = 2,500 bp
method = iRep
window size = 5,000 bp
method = iRep
window size = 10,000 bp
iRep:
1.50
1.76
1.86
2.01
method = median iRep
window size = 10 bp
method = median iRep
window size = 2,500 bp
method = median iRep
window size = 5,000 bp
method = median iRep
window size = 10,000 bp
method = scaffold windows
window size = 10 bp
method = scaffold windows
window size = 2,500 bp
method = scaffold windows
window size = 5,000 bp
method = scaffold windows
window size = 10,000 bp
delta (25x with 100% genome)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.15
14
12
10
8
6
4
2
0
25
50
75
90
percent of genome
d
iRep:
1.50
1.76
1.86
2.01
0.15
iRep range (n = 10)
0.5
0.4
0.3
0.2
0.1
0.0
25
50
75
90
percent of genome
e
slide = 10 bp
slide = 100 bp
iRep:
1.50
1.76
1.86
2.01
0.15
delta (25x with 100% genome)
0.5
0.4
0.3
0.2
0.1
0.0
25
50
75
90
percent of genome
f
regression
Clostridium perfringens
Klebsiella oxytoca
Enterococcus faecalis
Veillonella parvula
Enterobacter cloacae
r = 0.75
p-value = 2.6 x 10⁻⁵
bPTR
1.6
1.4
1.2
1.0
1.0
1.2
1.4
1.6
1.8
2.0
iRep
(no GC sequencing correction)

Figure 3.4 | iRep is an accurate measure of *in situ* replication rates. (**a**) iRep, bPTR, and kPTR measurements made for cultured *Lactobacillus gasseri* (Korem et al., 2015) were compared (r = Pearson's r value), showing strong agreement between all methods. (**b**) Colony forming unit (CFU) counts were available for a subset of these samples (Korem et al., 2015), and used to calculate growth rates (n = 2). All methods were highly correlated with CFU-derived rates after first accounting for the delay between start of genome replication and observable change in population size (as noted previously (Korem et al., 2015)). Replication rates from CFU data were adjusted by variable amounts before calculating correlations with sequencing-based rates (best correlation shown; d = time adjustment). CFU data are plotted with a -90 minute offset. (**c**) Using the *L. gasseri* data, minimum coverage requirements were determined for each method by first measuring the replication rate at 25x coverage, and then comparing to values calculated after simulating lower coverage. This shows that ≥5x coverage is required. (**d**) The minimum required genome fraction for iRep was determined by conducting 100 random fragmentations and subsets of the *L. gasseri* genome. Sequencing was subset to 5x coverage before calculating iRep to show the combined affect of low coverage and missing genomic information. With ≥75% of a genome sequence, most iRep measurements are accurate ±0.15. (**e**) iRep and bPTR measurements were calculated using five genome sequences assembled from premature infant metagenomes, showing that these methods are in agreement in the context of microbiome sequencing data.

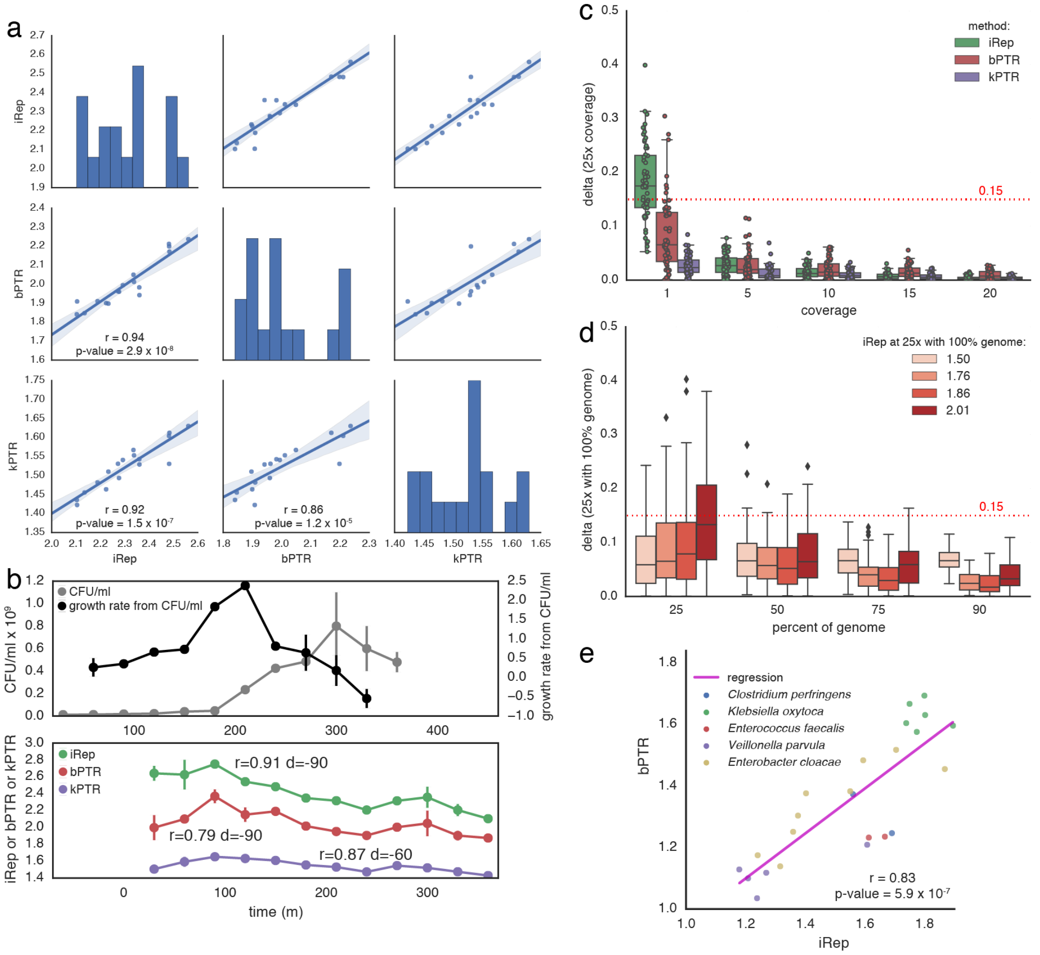


Figure 3.5 | Coverage, GC skew patterns, and bPTR measurements for reconstructed genomes oriented and ordered based on complete reference genome sequences. (a-e) Read mapping was conducted using sequences from the sample used for genome recovery. bPTR was calculated after determining the origin and terminus of replication based on cumulative GC skew. Coverage was calculated for 10 Kbp windows calculated every 100 bp (extremely low and high coverage windows were filtered out; see Methods). bPTR was calculated as the ratio between the coverage at the origin and terminus after applying a median filter. Cumulative GC skew and coverage patterns confirm the ordering of genome fragments.

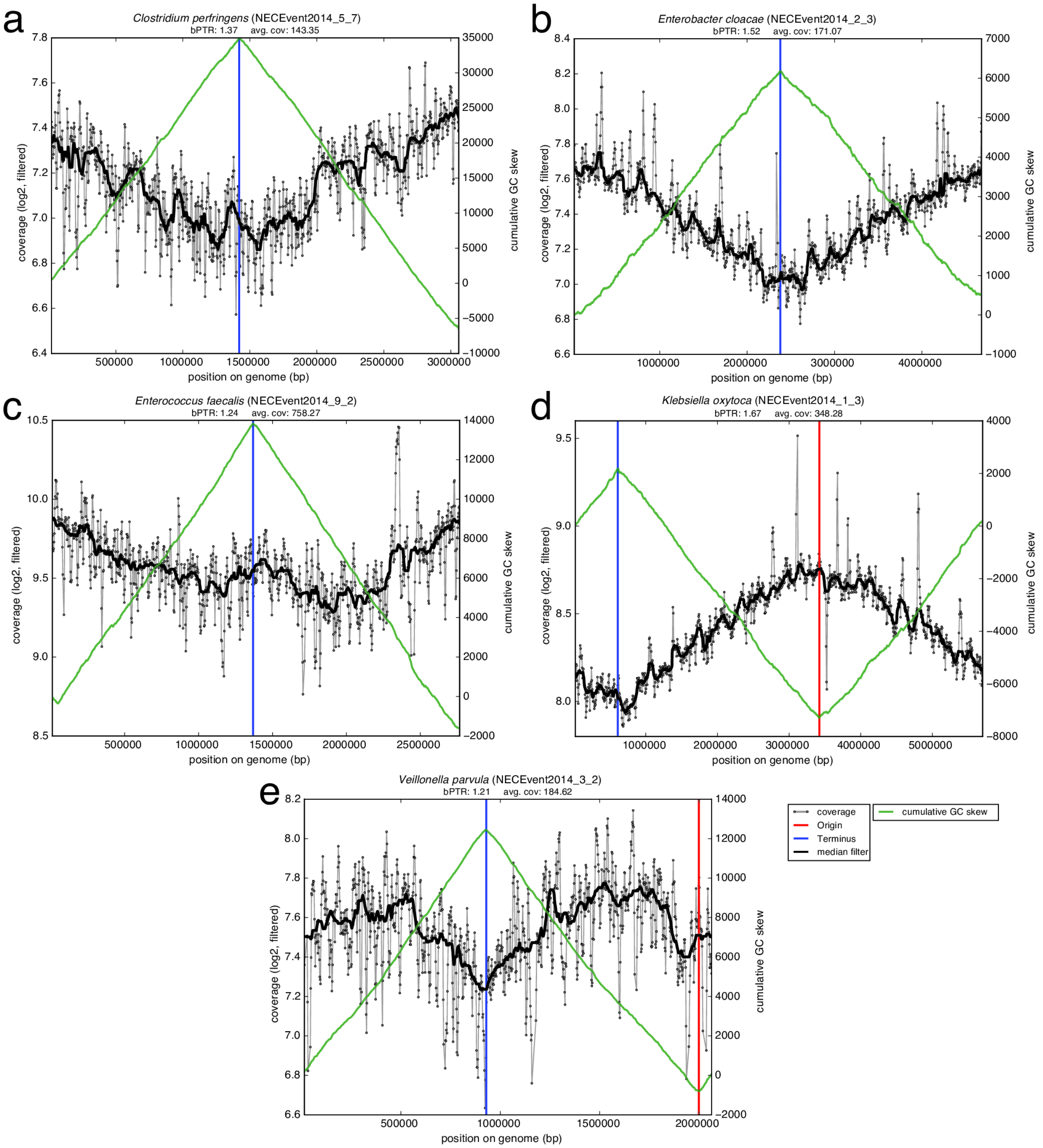


Figure 3.6 | Reference genomes are not representative of organisms surveyed in the premature infant microbiome study. Reads were mapped to both reconstructed genomes and closely related reference genomes (**Supplementary Table 3.4**), and the percent of each genome covered by sequencing reads is reported. Average nucleotide identity (ANI) is reported between each reconstructed genome and the paired reference genome. The large fractions of reference genomes not represented by metagenome sequencing show that extensive genomic variation is present between surveyed and reference genomes, despite high ANI values in some cases.

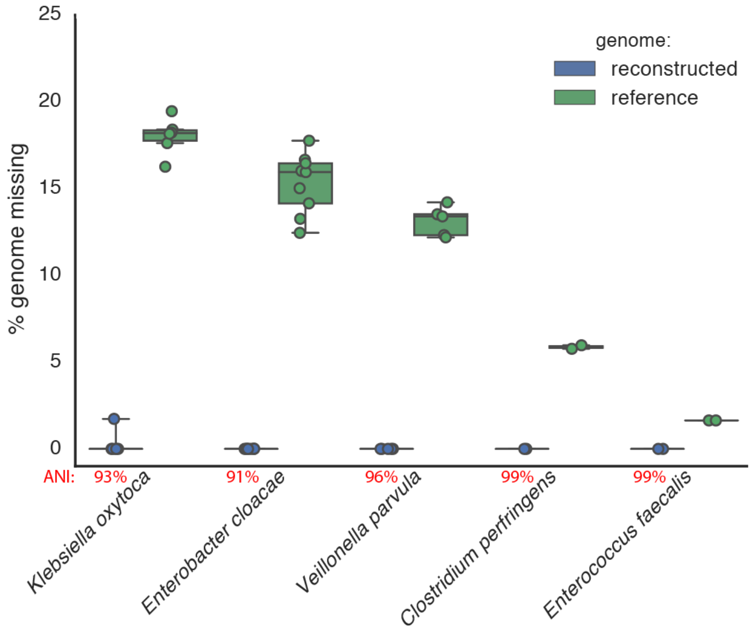


Figure 3.7 | iRep and bPTR calculations agree for a novel Deltaproteobacterium sampled from groundwater. (**a**) bPTR was calculated after determining the origin and terminus of replication based on regression to coverage calculated across the genome. Coverage was calculated for 10 Kbp windows sampled every 100 bp (**see Methods**). The ratio between the coverage at the origin and terminus was determined after applying a median filter. The cumulative GC skew pattern confirms the genome assembly and locations of the origin and terminus of replication. (**b**) iRep was determined by first calculating coverage over 5 Kbp windows sampled every 100 bp, and then sorting the resulting values. High and low coverage windows were removed, and then the slope of the remaining (trimmed) values was determined and used to evaluate the coverage at the origin and terminus of replication: iRep was calculated as the ratio of these values. (r^2 was calculated between trimmed data and the linear regression).

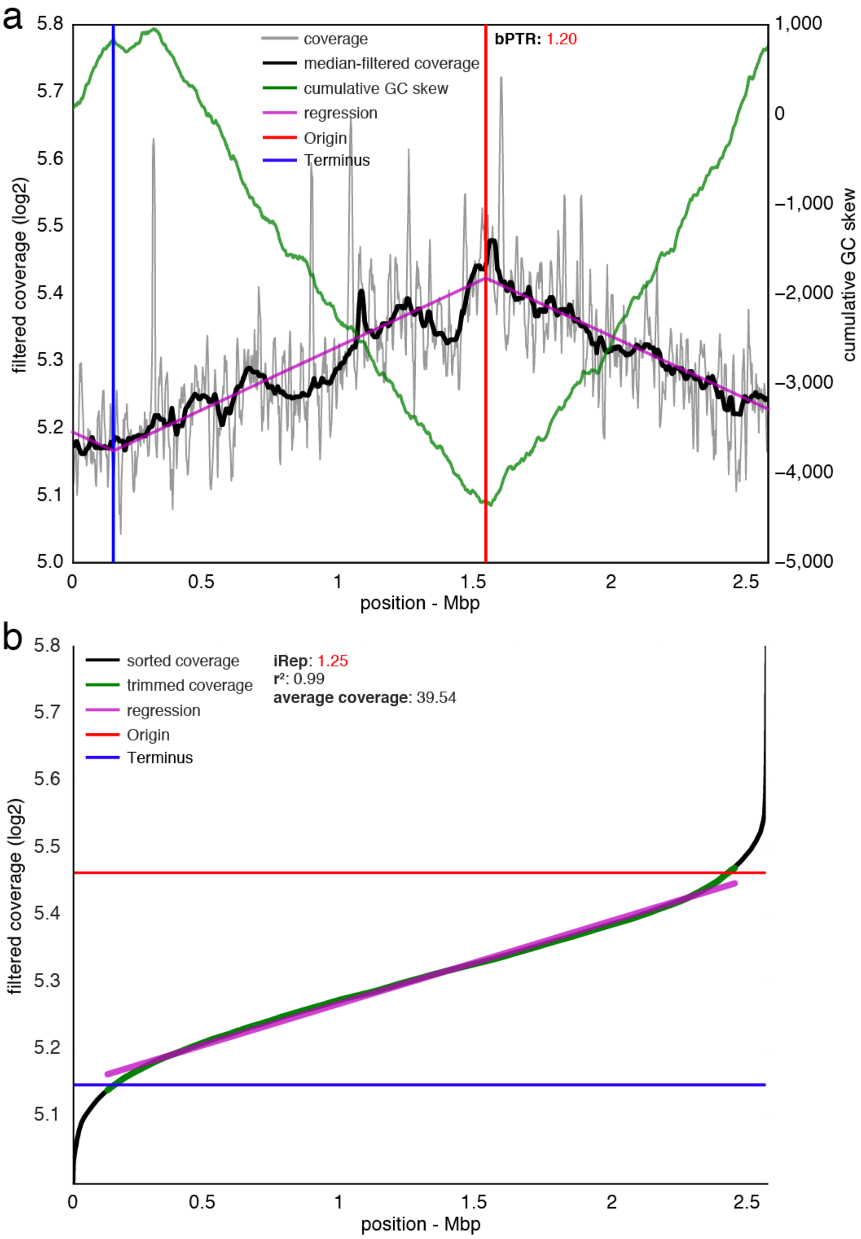


Figure 3.8 | Replication rates were determined for Candidate Phyla Radiation (CPR) and human microbiome-associated organisms. iRep values were measured and compared across studies (**a**; MW = Mann-Whitney, n = number of measured replication rates), and compared based on taxonomic affiliation (**b**).

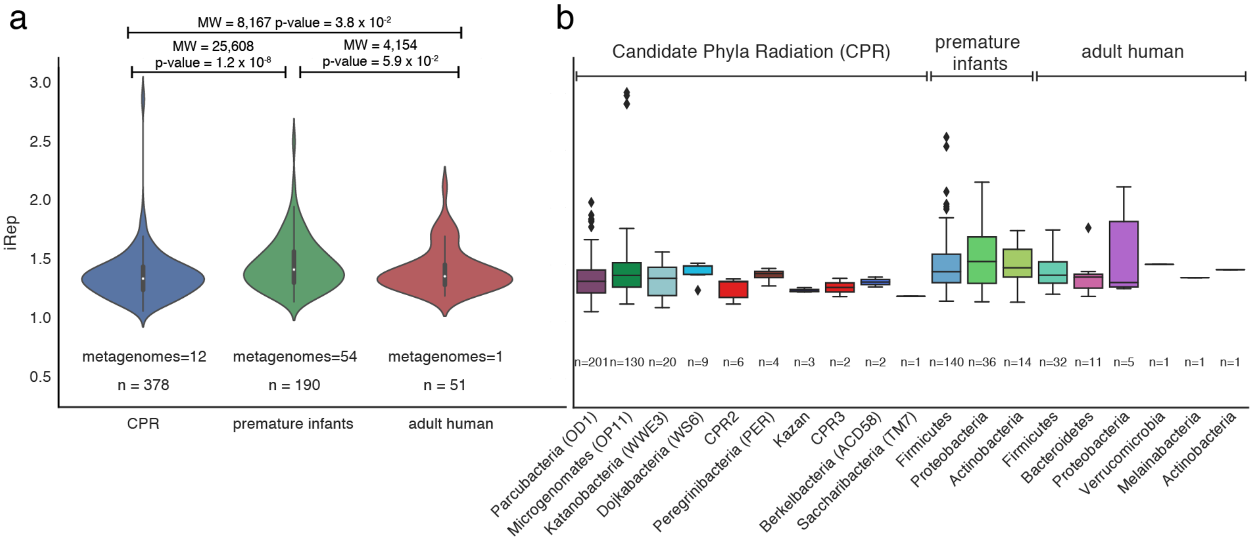


Figure 3.9 | Replication rates determined by iRep and kPTR are not in strong agreement for the premature infant study. iRep values were determined based on reconstructed genomes and kPTR values based on complete reference genomes (r = Pearson's r value).

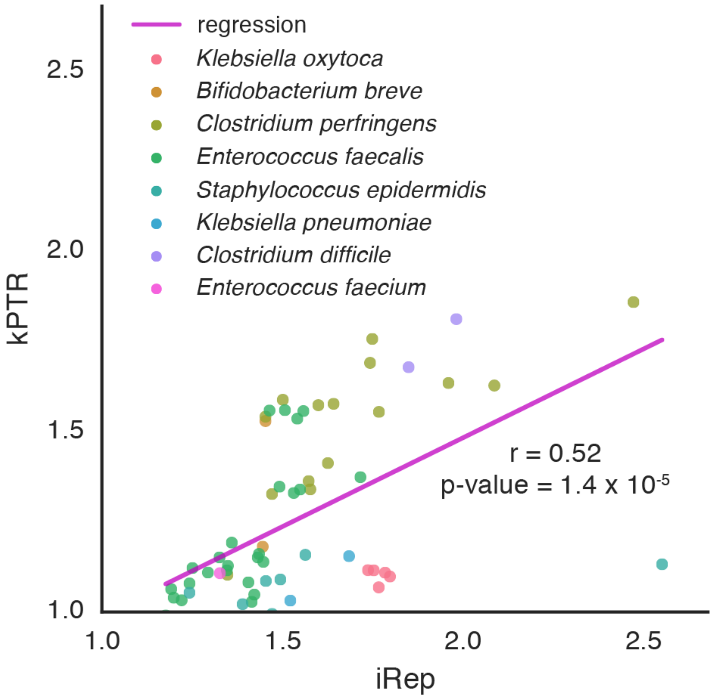


Figure 3.10 | Coverage, cumulative GC skew, and bPTR measurements for complete reference genomes with similarity to genomes from the adult human microbiome sample. (a-e) Reads from the adult human microbiome were mapped to complete reference genome sequences. Coverage was calculated for 10 Kbp windows every 100 bp (extremely low and high coverage windows were filtered out; see Methods). The origin and terminus of replication were determined based on coverage. bPTR was calculated as the ratio between the coverage at the origin and terminus after applying a median filter. Cumulative GC skew and coverage patterns suggest the presence of genomic variation or assembly errors for some genomes (b-c, e).

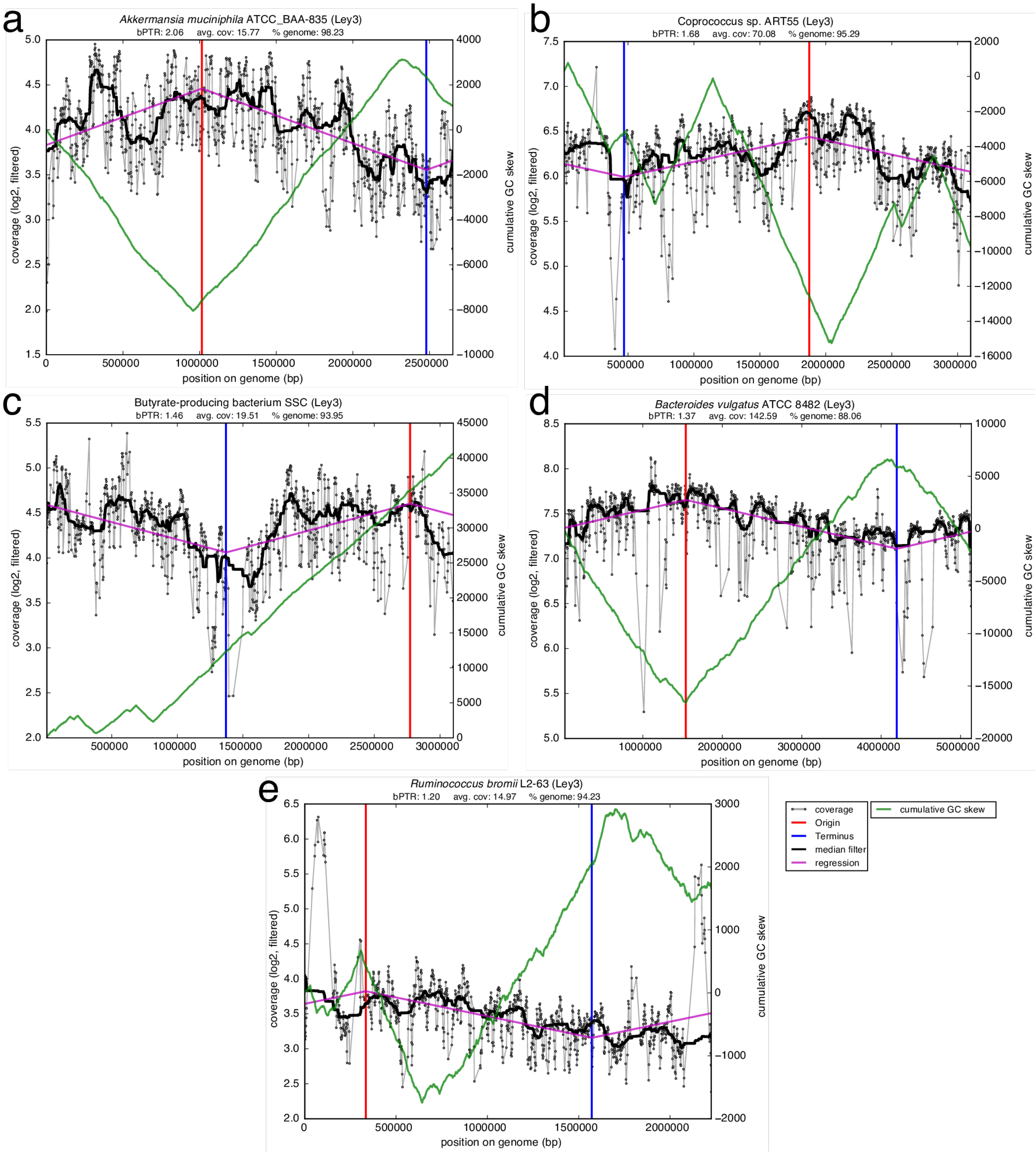


Figure 3.11 | Absolute abundance (bars, left axis) and iRep (scatter plot, right axis) for bacteria associated with premature infants. The five days following antibiotic administration are indicated using a color gradient (DOL = day of life).

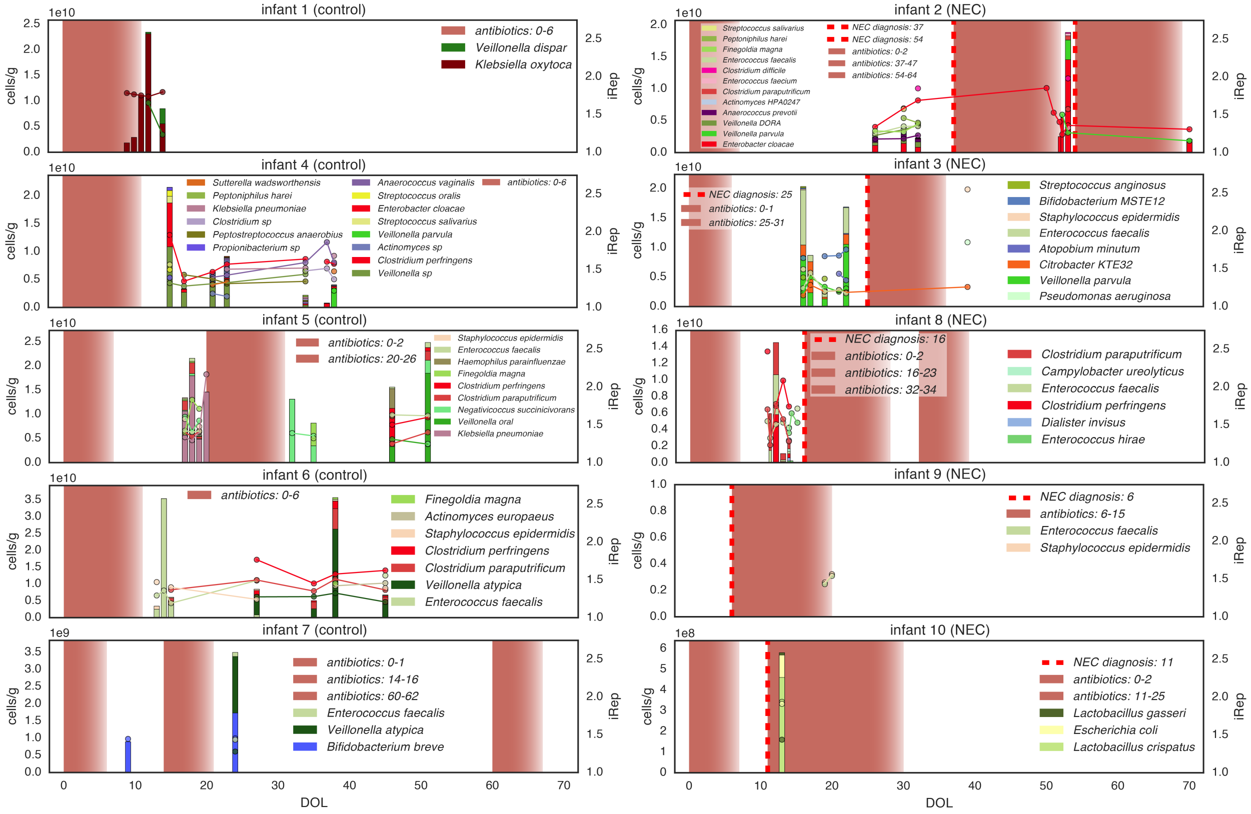


Figure 3.12 | Elevated replication rates are associated with antibiotic administration and were detected prior to onset of necrotizing enterocolitis (NEC) in premature infants. iRep distributions were compared (a) between samples collected during or within five days after antibiotic administration and samples from other time points, and (b) between samples collected from NEC and control infants. (c-d) Comparison of iRep values measured for different species (c) and genera (d) sampled from NEC and control infants (shown are taxa with ≥5 observations from either group). (e) iRep for the fastest growing organism observed for each control infant, and for the fastest growing organism from each day of life (DOL) sampled for each NEC infant, reported relative to NEC diagnosis. High replication rates for members of the genus *Clostridium* were detected in infants surveyed prior to NEC diagnosis.

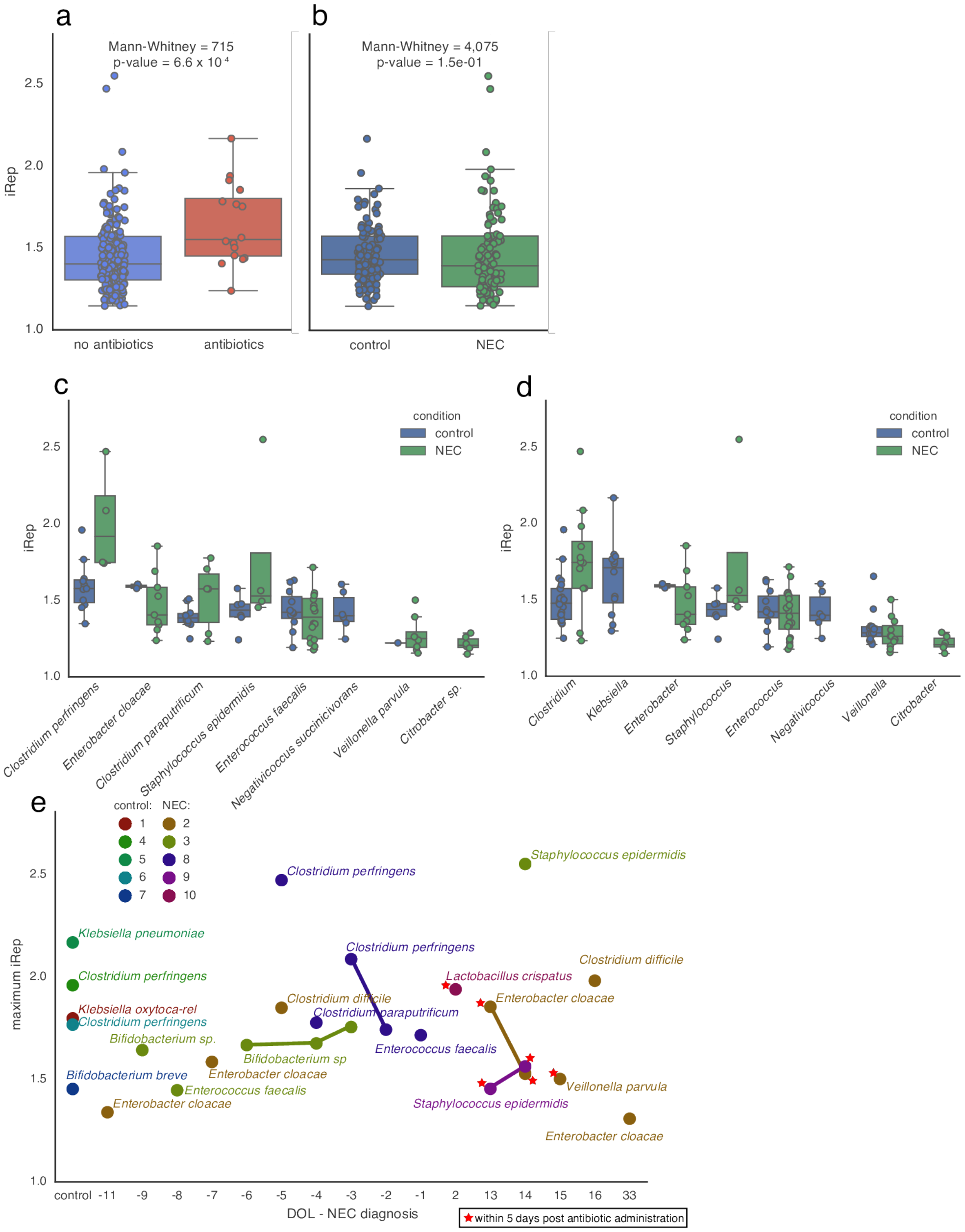
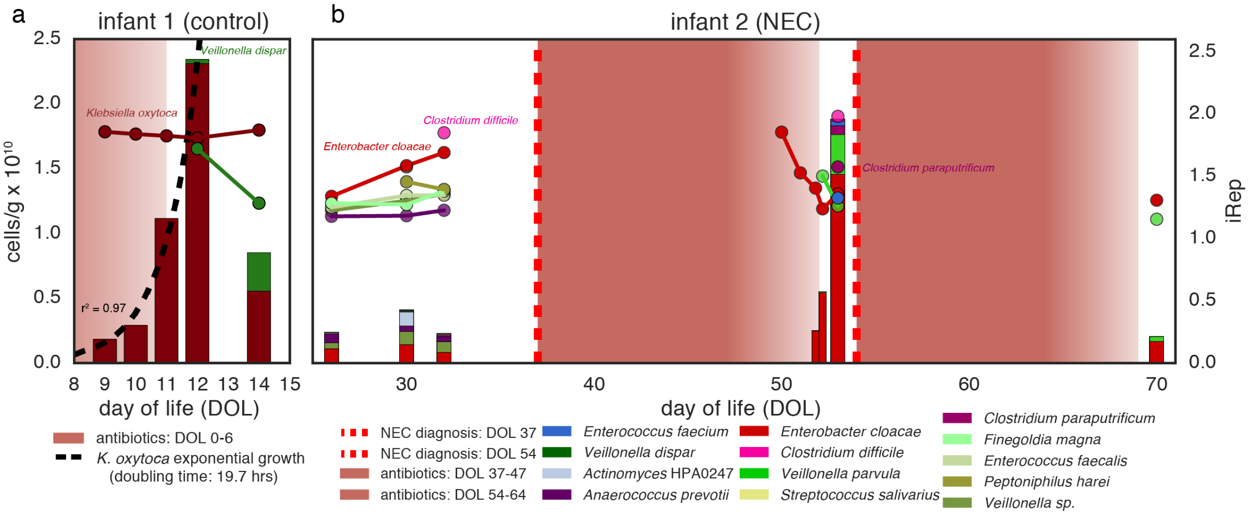


Figure 3.13 | Absolute abundance (bars, left axis) and iRep (scatter plot, right axis) values for bacterial species associated with two premature infants. The 5 d following antibiotic administration are indicated using a color gradient. (**a**) Exponential growth was determined by regression to *K. oxytoca* absolute abundance values (black dotted line). (**b**) Infant 2 was diagnosed with two cases of necrotizing enterocolitis (NEC; dotted red lines) during the study period.



Chapter 4

Linking microbial community dynamics to metabolic shifts during colonization of the premature infant gut

C. T. Brown, W. Xiong, M. R. Olm, B. C. Thomas, M. J. Morowitz, R. L. Hettich, J. F. Banfield

Unpublished.

Abstract

The first weeks of life are an important developmental period for premature infants. During this time infants are colonized by microbes, which are thought to contribute to immune system maturation and other processes. In premature infants, aberrant microbial communities have been implicated in onset of necrotizing enterocolitis (NEC), a life-threatening intestinal disease. Currently, little is known about microbial community dynamics during this time from the perspective of composition, activity, and metabolism. Of particular interest is how genetically similar microbes may modulate their activity and metabolic characteristics in different community contexts. In order to study this process, gut microbiome samples collected during the first three months of life from 11 premature infants, four that developed NEC, were selected for detailed metagenome and metaproteome analyses. Samples were selected in part based on the presence of members of the same microbial species. In total, 711 draft-quality genomes representing 98 different species groups were reconstructed from 144 metagenomes. These genomes were used to measure *in situ* replication rates, and for proteomic analysis of microbial metabolic profiles. Members of the species *Enterococcus faecalis*, *Klebsiella pneumoniae*, and *Staphylococcus epidermidis* colonized essentially all infants, but many other organisms were present. Communities were classified into six types based on community composition. Infant health status and development did not determine microbial community type. Interestingly, community type switched within individual infants, sometimes multiple times, and communities sampled from the same infant at subsequent time points were sometimes more similar to those from other infants than to earlier communities. In some cases, switches preceded onset of NEC, but no community type was associated with NEC. However, members of several species were found to be replicating at different rates in healthy infants compared with those that went on to develop NEC. Differences in the abundance of proteins involved in specific transporter and sugar degradation systems, as well as proteins involved in other processes, distinguished community types from one another. Community-specific protein abundances were in part driven by shifts in protein expression of members of the same species living in different community contexts. These analyses characterized the early infant microbiome as a highly variable system and uncovered microbial dynamics not apparent from community composition alone.

Introduction

The infant microbiome has been characterized as having high levels of between-individual variation compared with adult human microbiomes (Costello et al., 2009; Palmer et al., 2007). During the first one to two years of life the gut microbiomes of infants begins to converge upon an adult-like state (Bokulich et al., 2016; Palmer et al., 2007). However, aberrations in this process may contribute to diseases such as type 1 and 2 diabetes, irritable bowl syndrome, and in necrotizing enterocolitis (NEC) in premature infants (Brown et al., 2011; Mai et al., 2011; Morrow et al., 2013; Mshvildadze et al., 2010; Qin et al., 2012; Xavier and Podolsky, 2007). Because establishment of the microbiome is a key driver of immune system development, changes in the process of colonization may have life-long implications even if they do not result in a drastically different microbiome composition later in life (Lathrop et al., 2011; Maslowski et al., 2009).

Infants born prematurely have low-diversity microbial communities compared with full term infants (Brown et al., 2013; Raveh-Sadka et al., 2016; Sharon et al., 2012), and are susceptible to life-threatening diseases such as NEC (Neu and Walker, 2011). While it has long been thought that NEC is caused by bacterial infections, strain-resolved microbial community analysis has shown that no single pathogen is responsible for the disease (Raveh-Sadka et al., 2015). However, it is still likely that microbial communities play an important role, whereby the metabolism of organisms may be critical to infant health and disease. In order to better understand how microbes modulate their replication rates and metabolism during the colonization process, we conducted a combined metagenomics and metaproteomics study of the microbiome of both healthy premature infants and infants that went on to develop NEC. Microbiome samples were collected during the first three months of life with the goal of measuring the physiological changes of dominant and ubiquitous bacterial species.

Results and discussion

Metagenome sequencing and genome binning

In order to study the developing gut microbiome, stool samples were collected during the first three months of life for 11 infants born prematurely. One of the infants in the study cohort (N1_019) developed a case of sepsis and four infants (N1_021, N2_039, N2_069, and N2_071) developed necrotizing enterocolitis (NEC) (**Table 4.1**). To study the gut microbiome, we sequenced 474.5 Gbp of DNA across 144 metagenomes with an average of 3.4 Gbp of sequencing per sample (**Figure 4.1 and Supplementary Table 4.1**). Metagenomes were assembled into 3.15 Gbp of scaffolds ≥1,000 bp that represented 92.6% of all sequenced DNA.

Genomes were binned based on Emergent Self Organizing Map (ESOM) clustering of scaffold time-series abundance profiles (**Figure 4.2**), and manually based on GC content, single time point coverage, and taxonomic profiles. This resulted in 1,697 bins, 711 of which were draft-quality (≥75% complete). These genomes were clustered into 98 groups approximating different bacterial species based on sharing ≥98% average nucleotide identity (ANI), each of which was represented by a draft-quality genome (**see below and Supplementary Table 4.2**). These genomes account for 89% of the total sequences.

Premature infants are colonized by members of the same species

Most prior studies of the infant gut microbiome have depended on marker gene sequences in order to track the presence and abundance of microbial community members. Consequently, these studies have not been able to address the question of whether or not members of the same species reside in the microbiomes of different premature infants. This lack of resolution has also obscured time-series studies. We approximated species as groups of organisms with genome sequences that shared ≥98% ANI, and found that at least eight of the eleven infants were colonized by members of the species *Enterococcus faecalis*, *Staphylococcus epidermidis*, and *Klebsiella pneumoniae* (**Figure 4.3**). Clustering of infants based on species membership showed that infants that developed NEC have similar species inventories compared with those that did not (**Figure 4.4**). Infants were largely distinguished based on the presence of either *Staphylococcus epidermidis* and/or *Negativicoccus succinicivorans* (Pearson's correlation p-value ≤0.05).

Bacterial species replicate at different rates during colonization

iRep is a newly-developed method for determining microbial replication rates based on measurements of DNA sequencing coverage trends that result from genome replication (Brown et al., 2016). We applied the iRep method using draft-quality genomes recovered from metagenomes sequenced for each infant in the study. Results show that populations of several species of bacteria were replicating more quickly in either infants that developed NEC or healthy controls (**Figure 4.5**). However, overall iRep values collected from infants that did and did not go on to develop NEC were not statistically different. Across all infants, *Staphylococcus aureus*, members of the genus *Veillonella*, and *Klebsiella pneumoniae* exhibited some of the highest replication rates.

Proteome sequencing and identification of predicted proteins

Metaproteomics was conducted on the same samples that were used for metagenomics analysis, or, in cases where enough sample was not available, on samples that were collected at a similar time. Conducting metagenomics and metaproteomics sequencing of samples from the same infant is critical for obtaining an appropriate database for matching peptides to proteins. Metaproteomics resulted in measurement of 37,590,440 spectra counts across 61 collected samples (most have two technical replicates), with an average of 63,549 spectra per sample that could uniquely be assigned to microbial proteins (**Figure 4.1 and Supplementary Table 4.3**). The 3,083,935 open reading frames (ORFs) predicted from the metagenome sequencing data were represented by 552,375 non-redundant sequences determined based on their inclusion in a representative genome sequence.

ORFs were grouped into 85,437 putative protein families (2.8% of total and 15.5% of non-redundant protein sets). Of the 51,629 protein families represented by draft-quality genomes, only 4,621 were detected by proteomics. Of the undetected protein families, 95% were each found in 12 species or less, and are thus somewhat rarely encoded in the infant gut microbiome. However, some undetected protein families were fairly commonly encoded in recovered genome sequences. While some of these may not have been detected due to difficulties associated with protein extraction, such as the membrane-bound cell division protein FtsW and several transporters, others are likely not frequently expressed in the gut microbiome, such as the mutagenic DNA polymerase IV (**Supplementary Table 4.4**). Of the 44 families identified in 75 or more genomes, only six were hypothetical proteins. Amongst the most abundant detected proteins were those involved in glycolysis, translation, transport, and protein maturation.

Different species express varying amount of their proteome in the infant gut

Microbes present in the gut environment are not expected to express their complete complement of proteins at all times. In order to investigate this we compared the average proteome sequencing depth for each organism to the percent of the predicted proteome that could be detected (**Figure 4.6**). The median proteome detection across all samples was 10.2%, but this was largely due to a lack of proteome sequencing depth. Higher protein sequencing depth corresponded with more proteins being detected. The median percent of the proteome detected for organisms with the best detection in each sample was 30%. For several frequently detected colonists, including *Klebsiella pneumoniae*, *Klebsiella oxytoca*, and members of the genus *Enterobacter,* maximum proteome expression was determined to be around 50%. Members of the species *Bifidobacterium bifidum*, *Propionibacterium sp*, and *Anaerococcus vaginalis*, expressed a greater proportion of their encoded genes, suggesting that these organisms may be more specialized to the infant gut.

Activity of bacterial community members based on iRep and proteomics

Comparison of DNA and proteome abundance levels for specific organisms enables determination of whether or not an organism contributes as much to the proteome as would be expected based on abundance determined based on DNA sequencing (**Figure 4.7**). The difference between these values is related to the relative proteome expression level for a particular organism. *Veillonella sp.*, *Enterococcus faecium*, *Clostridium perfringens*, *Peptostreptococcus sp.*, *Bifidobacterium bifidum*, and *Clostridium sp.* were all found to frequently have higher than estimated protein expression levels (**Figure 4.8**). Many of these were also found to have high iRep replication rates at certain times, although the two values are not correlated. The fact that activity measures based on iRep and proteomics are not correlated is not surprising since they represent distinct measurements of microbial physiology.

Studied microbial communities cluster into six types

Microbial communities were clustered based on species membership and abundance in order to identify microbial consortia common during the colonization process. Six distinct community types were identified (**Figure 4.9**), each of which is characterized based on dominance of different community members (**Figure 4.10 and Figure 4.11**). Members of the bacterial species *Citrobacter freundii, Escherichia coli, Enterobacter sp.* and *Klebsiella pneumoniae, Enterococcus faecalis, Haemophilus parainfluenzae,* and *Enterobacter sp.* and *Klebsiella oxytoca* dominated individual microbial community types. Microbiomes from different infants clustered into the same community type, and the microbiome of individual infants was found to switch types, sometimes multiple times, during the colonization process (**Figure 4.11**). Microbiomes samples from the same infant at different time points may be more similar to those from other infants than to microbiomes collected at other time points. Furthermore, there was no strong correlation between microbial community type and infant health status, developmental age, or antibiotic usage (**Figure 4.9**). Microbiomes associated with infants that did and did not go on to develop NEC were often classified in the same community type, indicating that organism physiology, rather than simple microbial community membership and abundance, may be important to infant health.

Abundant microbial proteins correlate with microbial community type

The finding that microbial communities associated with premature infants group into distinct types raises the question of their functional similarity. Analysis of the most abundant protein families identified in samples from each infant showed that they clustered primarily based on microbial community type (**Figure 4.12**), indicating that they are functionally distinct. However, some samples from the same community type were not clustered together, indicating that differences in protein expression exist even when community composition is similar. Similar to what was found for clustering based on community composition alone, clusters based on abundant proteins did not correlate with infant, infant health or development, or antibiotic usage.

Protein expression profiles associated with microbial community types

In order to identify which proteins best distinguish microbial community types, protein expression profiles were compared between sets of samples from different microbial community types. Groups of co-varying proteins with statistically different abundance levels between community types were identified (**Figure 4.13**). Notably, the community type dominanted by *Citrobacter freundii* was distinguished from other community types based on having an abundance of proteins involved in bacterial chemotaxis, microcompartment formation, propanediol utilization, and respiration. Several other communities were distinguished based on specific types of sugar transporters and associated processing machinery.

Species-specific metabolic profiles shift in association with microbial community context and replication rate

Interestingly, *Klebsiella pneumoniae*, *Enterococcus faecalis*, and *Citrobacter freundii* were found to be present in different microbial community types. This raises the question of whether these organisms are maintaining the same metabolic strategies, or shifting expression in accordance with microbial community context, and whether or not these shifts are contributing to overall proteome variation between microbial community types. Expression levels of the proteins found to best distinguish between microbial community types (**see above and Figure 4.13**) were evaluated for each of the organisms. In order to compare organism-specific changes in proteome profile, protein abundance measurements were normalized for each taxon, instead of across all the sample data. This allowed for comparison of relative expression levels of identified proteins. Each of these organisms exhibited proteome expression profiles that correlated with microbial community type, indicating that the organisms are using different metabolic strategies in each context (**Figure 4.14**).

When abundant, *K. pneumoniae* expressed a diverse set of metabolic proteins that were not expressed when the organism was less abundant. Lower abundance coincided with a shift to the production of several TCA cycle proteins and a multidrug transporter. Likewise, when *E. faecalis* was less abundant it was producing fewer proteins that make up phosphotransferase systems, and suspended production of a glycerol dehydrogenase that was one of the most expressed proteins when the organism was more abundant. *C. freundii* was found to be expressing several ABC transporters and a microcompartment protein that were not identified when the organisms was found in a different microbial community context where it was less abundant. However, when less abundant, a *C. freundii* dimethyl sulfoxide reductase was identified, suggesting a switch in the electron acceptor being used for respiration. Notably, proteomes from *E. faecalis* and *C. freundii* clustered primarily based on community context, and subsequently based on iRep replication rate. Thus, proteome expression in each community is not driven by replication rate alone, but rather by modulation of the expression of environment-specific proteins. These findings illustrate how organisms can behave differently depending on their environment, and can drive microbiome function through expression of different proteins in a context-dependent manner.

Conclusion

Microbial colonization is a dynamic process only partially described by changes in microbial community membership and abundance. We used genome-resolved time-series metagenomics in conjunction with iRep replication rate and metaproteomics measurements to further probe the colonization process, and found that even within microbial communities that appear similar based on community composition, differences in microbial metabolism exist. Furthermore, because genetically similar organisms were found in multiple community contexts, we were able to identify changes in their proteome that were related to community context and replication rate.

Methods

Sample collection and metagenome sequencing

Samples were collected and processed for metagenome sequencing as previously described (Raveh-Sadka et al., 2016). Briefly, stool samples were collect from infants and stored at −80°C. DNA was extracted from frozen fecal samples using the QIAamp DNA Stool mini-Kit (Qiagen) with modifications (Zoetendal et al., 2006). DNA libraries were sequenced on an Illumina HiSeq for 100 or 150 cycles (Illumina, San Diego, CA). The protocol for sample collection and processing was approved by the University of Pittsburgh Institutional Review Board (IRB PRO10090089). All samples were collected with parental consent.

Genome binning and clustering into species groups

Samples from infants N1_003, N1_019, N1_021, N1_023 were analyzed previously (Raveh-Sadka et al., 2016). However, the sequencing data were re-assembled and analyzed for the current study. All raw sequencing reads were trimmed using Sickle (https://github.com/najoshi/sickle). Each metagenome was assembled separately using IDBA_UD (Peng et al., 2012). Open reading frames were predicted using Prodigal (Hyatt et al., 2010) with the option to run in metagenome mode. Predicted protein sequences were annotated based on USEARCH (–ublast) (Edgar, 2010) searches against UniProt (The UniProt Consortium, 2015), UniRef100 (Suzek et al., 2007), and KEGG (Kanehisa et al., 2012; Minoru Kanehisa, 2000). Scaffold coverage was calculated by mapping reads to the assembly using Bowtie2 (Langmead and Salzberg, 2012) with default parameters for paired reads.

Scaffolds from infants N1_003, N1_019, N1_021, N1_023 were binned into genome sequences using Emergent Self-Organizing Maps (ESOMs), as previously described (Dick et al., 2009) but with several modifications. Reads from every sample were mapped independently to every assembly using SNAP (Zaharia et al., 2011), and the resulting coverage data were combined. Coverage was calculated over non-overlapping 3 Kbp windows. Coverage values were normalized first by sample, and then the values for each scaffold fragment were normalized from 0-1. Combining coverage data from scaffolds assembled from different samples prior to normalization made it possible to generate a single ESOM map to bin genomes assembled independently from each sample. ESOMs were trained for 10 epochs using the Somoclu algorithm (Wittek et al., 2013) with the option to initialize the codebook using Principal Component Analysis (PCA). Genomes were binned by manually selecting data points on the ESOM map using Databionics ESOM Tools (Ultsch, 2005). Binning was aided by coloring scaffold fragments on the map based on BLAST (Altschul et al., 1990) hits to the genomes assembled in the prior study (Raveh-Sadka et al., 2016).

Scaffolds from other assembled metagenomes were binned based on their GC content, DNA sequence coverage, and taxonomic affiliation using ggKbase tools (ggkbase.berkeley.edu). Genome bins from all datasets were classified based on the lowest possible consensus of taxonomic assignments for predicted protein sequences. Genome completeness and contamination were estimated using CheckM (Parks et al., 2015). Genome bins were clustered into species groups based on sharing ≥98% average nucleotide identity (ANI) as estimated by MASH (Ondov et al., 2016). Representative genomes were selected for each cluster as the largest genome with the highest expected completeness and smallest amount of contamination. Representative genomes estimated to be ≥75%, but that had duplicate single copy genes were manually curated by removing contaminating sequences based on identifying scaffolds with extreme GC and/or coverage values. Genomes were classified as draft-quality based on the requirements for iRep analysis: ≥75% complete, ≤2.5% estimated contamination, and ≤175 scaffolds per Mbp of sequence (Brown et al., 2016).

iRep analysis

Accurate iRep (https://github.com/christophertbrown/iRep) analyses require precise read mapping, which can be in part be achieved by compiling appropriate genome databases. Individual mapping databases comprised of representatives of each genome cluster were created for each metagenome. Genomes reconstructed from the same sample were given highest priority for inclusion in the database. Genomes were selected to represent genome clusters using the following priority scheme: 1) draft-quality genomes assembled from the same sample, 2) draft-quality genomes assembled from another metagenome from the same infant, and 3) the highest quality genome sequence from the sample. In some case no representative was included for a genome cluster. Genomes that did not pass the draft-quality genome requirements were included in the database for mapping and abundance calculation purposes, but not for iRep. iRep was conducted using reads that mapped to genome sequences with ≤1 mismatch per read sequence.

Metaproteomics sequencing

Metaproteomics sequencing was conducted on 0.3 g of stool suspended in 10 mL cold phosphate buffered saline. Samples were filtered through a 20 μm size filter to enrich for microbial cells and proteins. Microbial cells were collected by centrifugation, boiled in 4% sodium dodecyl sulfate for 5 minutes, and sonicated to lyse cells. The resulting protein extract was precipitated with 20% trichloroacetic acid at -80°C overnight. The protein pellet was washed with ice-cold acetone, solubilized in 8 M urea, and cysteines were blocked with 20 mM iodoacetamide. Then sequencing grade trypsin was used to digest the proteins into peptides. Proteolyzed peptides were then salted and acidified by adjusting the sample to 200 mM NaCl, 0.1% formic acid, followed by filtering through a 10 kDa cutoff spin column filter to collect tryptic peptides.

Peptides were quantified by BCA assay and 50 μg peptides of each sample were analyzed via two-dimensional nanospray LC-MS/MS system on an LTQ-Orbitrap Elite mass spectrometer (Thermo Scientific). Each peptide mixture was loaded onto a biphasic back column containing both strong-cation exchange and reverse phase resins (C18). As previously described, loaded peptides were separated and analyzed using a 11-salt-pusle MudPIT protocol over a 22-h period (Xiong et al., 2015). Mass spectra were acquired in a data-dependent mode with following parameters: full scans were acquired at 30 k resolution (1 microscan) in the Orbitrap, followed by CID fragmentation of the 20 most abundant ions (1 microscan). Charge state screening and monoisotopic precursor selection were enabled. Unassigned charge and charge state +1 were rejected. Dynamic exclusion was enabled with a mass exclusion width of 10 ppm and exclusion duration of 30 seconds. Two technical replicates were conducted for each sample.

Protein databases were generated for each infant from protein sequences predicted from assembled metagenomes (**see above**). The database also included human protein sequences (NCBI Refseq_2011), common contaminants, and reverse protein sequences, which are used to control the false discovery rate (FDR). Collected MS/MS spectra were matched to peptides using MyriMatch v2.1 (Tabb et al., 2007), filtered, and assembled into proteins using IDPicker v3.0 (Ma et al., 2009). All searches included the following peptide modifications: a static cysteine modification (+57.02 Da), an N-terminal dynamic carbamylation modification (+43.00 Da), and a dynamic oxidation modification (+15.99). A maximum 2% peptide spectrum match level FDR and a minimum of two distinct peptides per protein were applied to achieve confident peptide identifications (FDR <1%). To alleviate the ambiguity associated with shared peptides, proteins were clustered into protein groups by 100% identity for microbial proteins and 90% amino acid sequence identity for human proteins using USEARCH (Edgar, 2010). Spectral counts were balanced between shared proteins.

Identification of putative protein families

Putative protein families were identified in order to track the presence and abundance of different protein types across samples. ORFs were first pre-clustered at 95% identity (usearch -cluster_smallmem -target_cov 0.50 -query_cov 0.95 -id 0.95), and then all-versus-all protein searches were conducted (usearch –ublast -evalue 10e-10 -strand both). Protein families were delineated from within the all-versus-all network graph using the MCL clustering algorithm (-I 2 -te 10) (Enright et al., 2002). The most common annotation observed across all protein sequences in the group was selected as the annotation for the putative protein family.

Identification of proteins with statistically significant differences in abundance

EdgeR (Robinson et al., 2009) was used to calculate statistically significant differences in balanced spectral counts between conditions using quasi-likelihood linear modeling (glmQLFTest).

Author Contributions

CTB and MRO assembled and annotated the metagenome data. CTB and JFB carried out the genome binning and curation, functional, time series abundance, protein expression, and iRep analyses, and drafted the manuscript. MRO and BCT provided bioinformatics support. MJM oversaw sample collection. WX and RLH generated the proteomics data. The authors declare that they have no competing interests.

Acknowledgements

Funding was provided by NIH grant R01AI092531.

Supplementary Tables

Supplementary Table 4.1 | DNA sequencing statistics.

Supplementary Table 4.2 | Genomes reconstructed from metagenomes. The representative of each genome is the best genome sequence within the same 98% ANI cluster. The draft-rep. column indicates whether or not the representative genome sequence is a draft-quality genome. The sample rep. genome is the best genome sequence within the same 98% ANI cluster that was assembled either from the same sample or, if no draft-quality genome was available, from a different metagenome from the same infant. The sample rep. genome was used for calculating relative abundance (see sample rep. coverage) and iRep, if it was a draft-quality genome. DOL stands for day of life.

Supplementary Table 4.3 | Proteomics sequencing statistics.

Supplementary Table 4.4 | Abundance of protein families predicted from draft-quality genome sequences.

Figure 4.1 | Metagenome and metaproteome sequencing conducted on microbiome samples collected from premature infants. a, Metagenome sequencing, and **b**, the percentage of each metagenome represented by assembled draft-quality genome sequences. **c**, The number of proteomics spectra counts that could be uniquely assigned to bacteria for each analyzed microbiome sample.

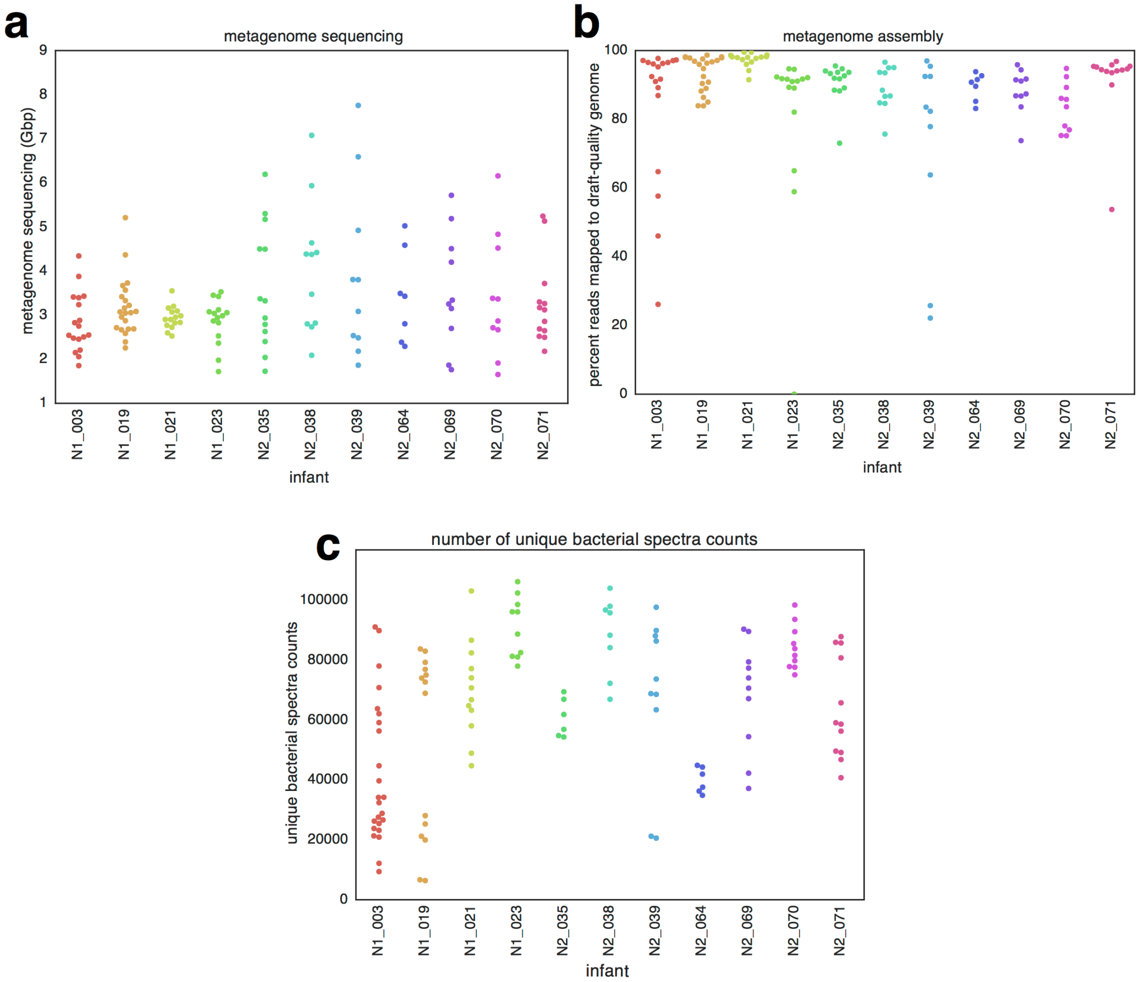


Figure 4.2 | ESOM genome binning. Genome binning was conducted based on Emergent Self-Organizing Map (ESOM) clustering of scaffolds assembled from individual metagenomes. Data points represent 3 Kbp fragments of assembled scaffolds. Coloring is based on the species-level assignment of reconstructed draft-quality genomes. The map is periodic, and red boxes indicate a single period.

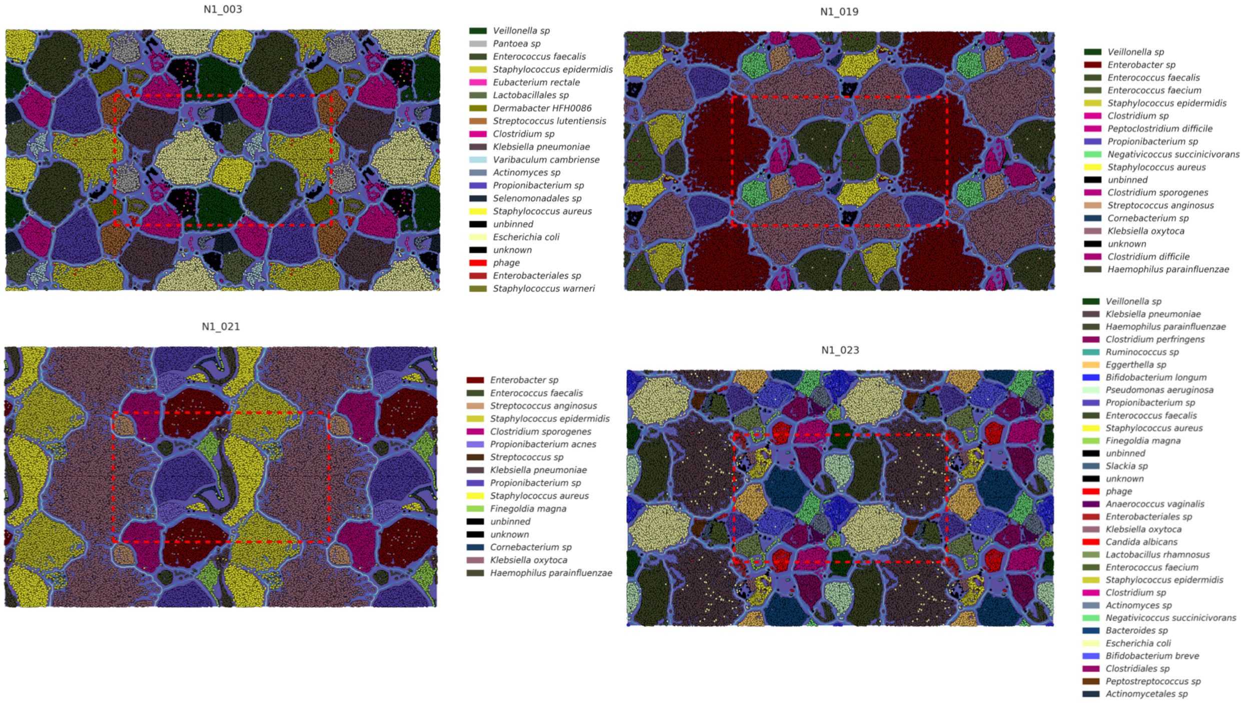


Figure 4.3 | Clustering of genomes reconstructed from metagenomes. Reconstructed genomes were clustered based on sharing 98% average nucleotide identity (ANI). Genomes were classified based on the lowest possible consensus of taxonomic assignments for predicted protein sequences. **a**, The number of genomes assigned to each genome cluster and **b**, the number of infants in the study with a reconstructed genome assigned to each cluster. Shown are clusters comprised of five or more genomes.

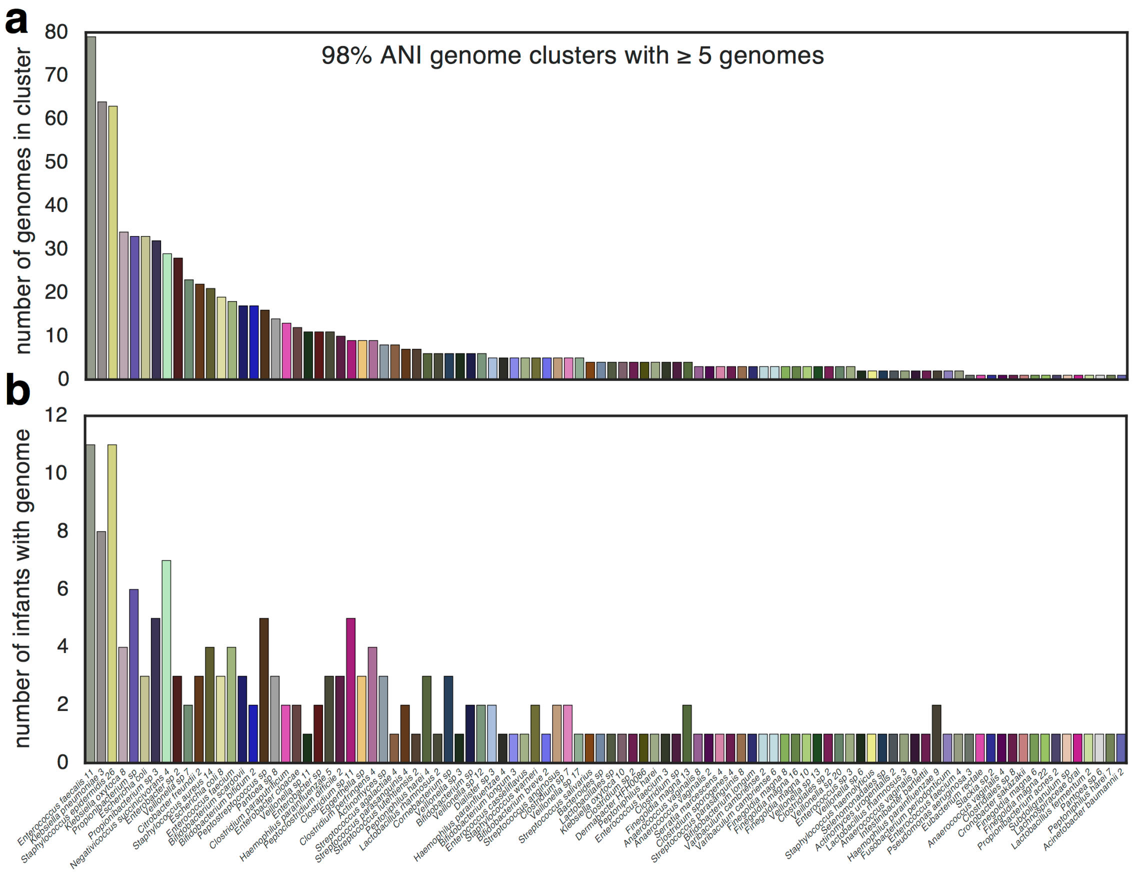


Figure 4.4 | Infants that developed NEC and healthy controls are colonized by some of the same species of bacteria. **a**, Presence (dark boxes) and absence (white boxes) of species identified in microbial communities from different infants. Species were identified based on sharing ≥98% genome ANI. **b**, Principal component analysis (PCA) clustering of infants based on the presence and absence of microbial species.

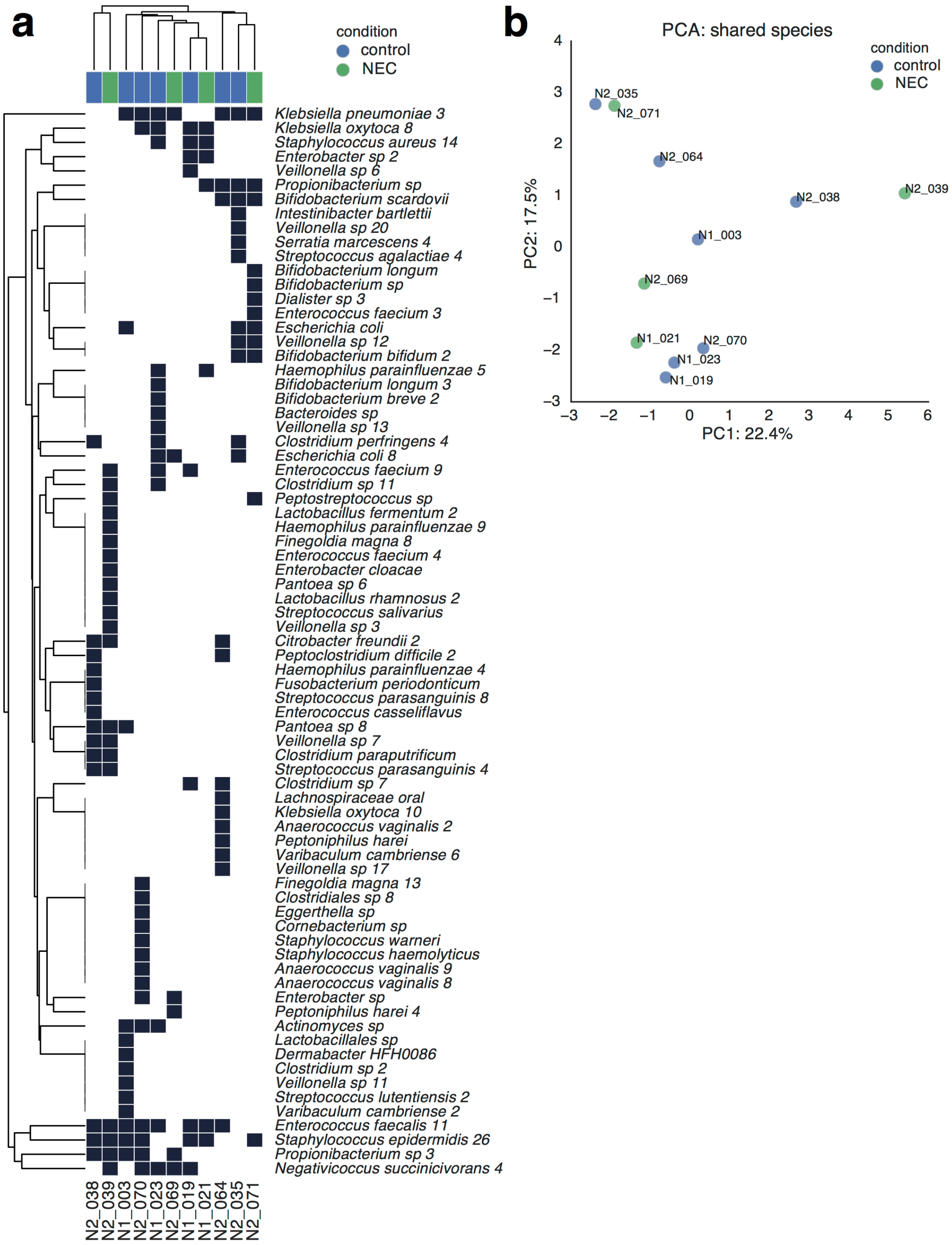


Figure 4.5 | Bacteria from the same species exhibited different replication rates over time and between infants that developed NEC and healthy controls. **a**, Species with significantly different replication rates between infants that did and did not go on to develop NEC are indicated with an asterisk (Mann-Whitney U Test p-value ≤0.01). **b**, Comparison of iRep replication rates from control infant microbiomes compared with those associated with infants that developed NEC. **c**, Comparison of iRep values from samples collected within five days of NEC diagnosis (NEC samples) to all other samples from both NEC and healthy control infants (non-NEC samples). Overall community replication rates were not statistically different between NEC and control samples (**b, c**).

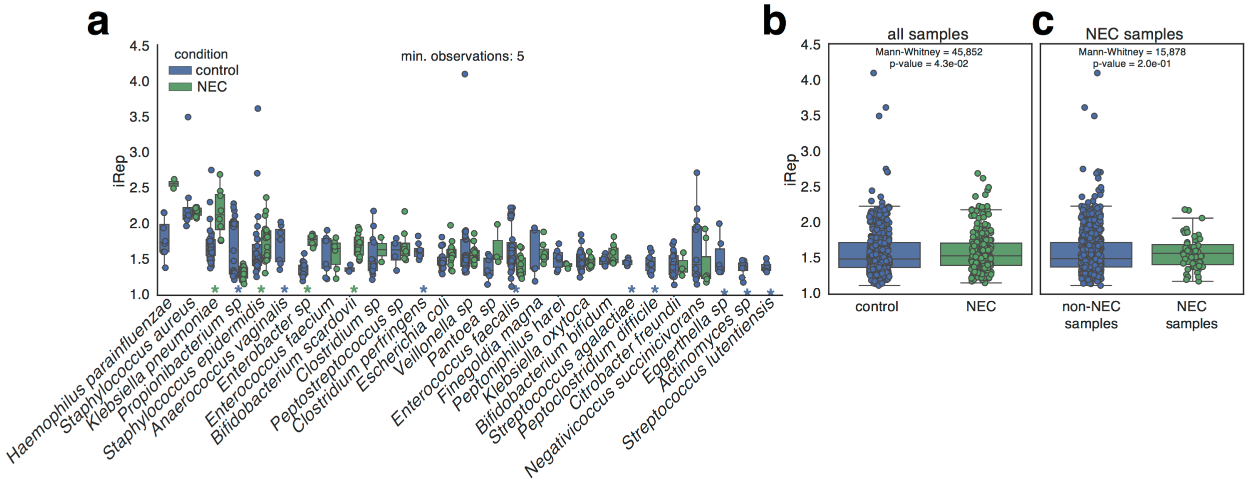


Figure 4.6 | Proteome detection for species colonizing premature infants. a, The proteome sequencing depth achieved for organisms in each sample is compared against the percent of predicted proteins that could be detected. Data point sizes and histograms are scaled based on organism abundance as determined by metagenome sequencing. **b**, Histogram showing the distribution of the maximum percent of the proteome detected for all organisms present in each sample.

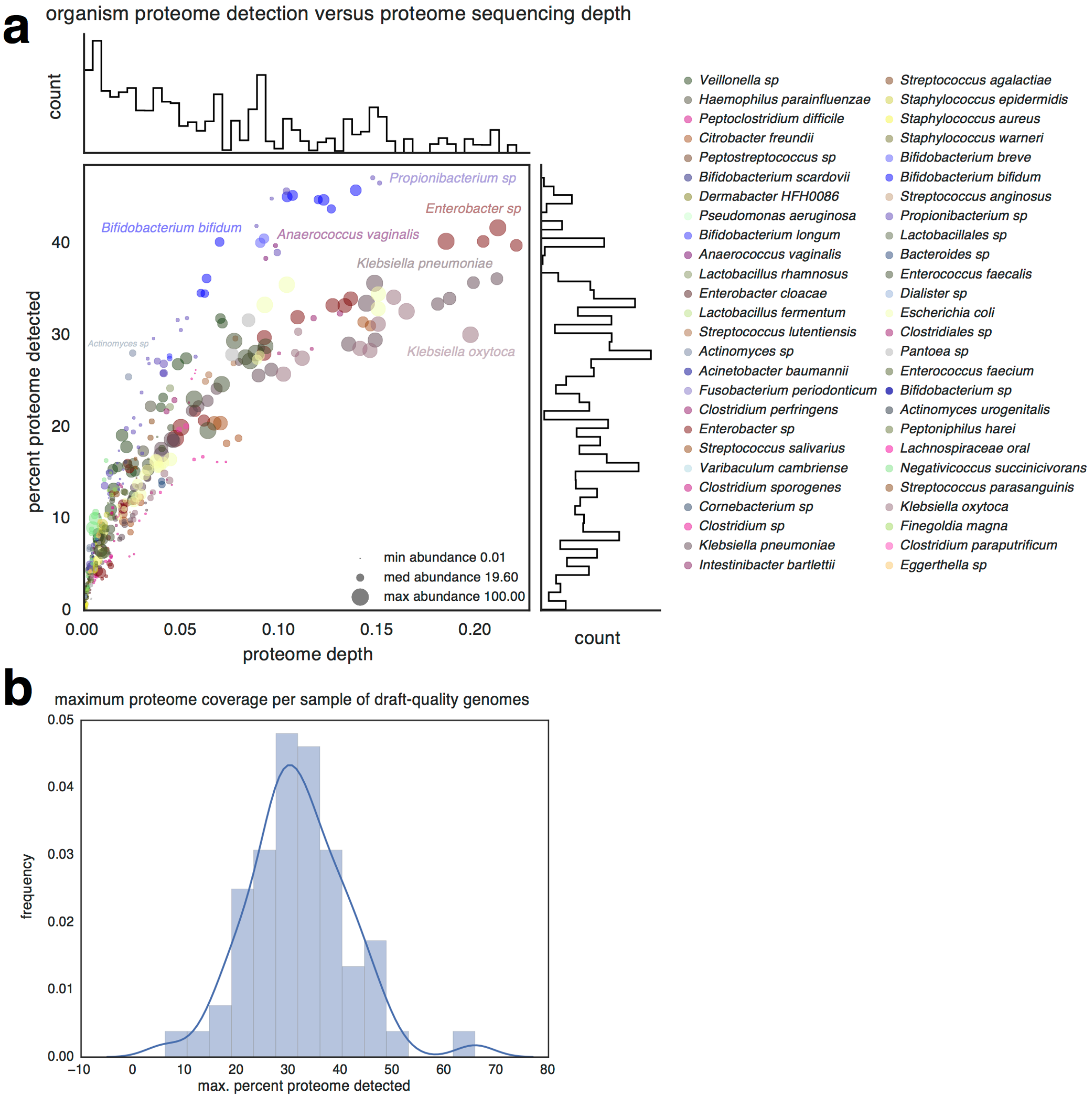


Figure 4.7 | Comparison of DNA and protein abundances determined for organisms colonizing premature infants. a, The abundance of each organism was determined based on both DNA and proteomics sequencing, and compared. Shown is the r^2 for the linear regression between the measurements. **b**, Histogram showing the difference between the measured and expected proteome abundances for each organism. The expected abundance was calculated from the linear regression conducted in b. Values below zero are from organisms where less of the proteome was detected than would be expected based on the abundance of the organism as calculated based on DNA sequencing, and values above zero are from organisms where more of the proteome was detected.

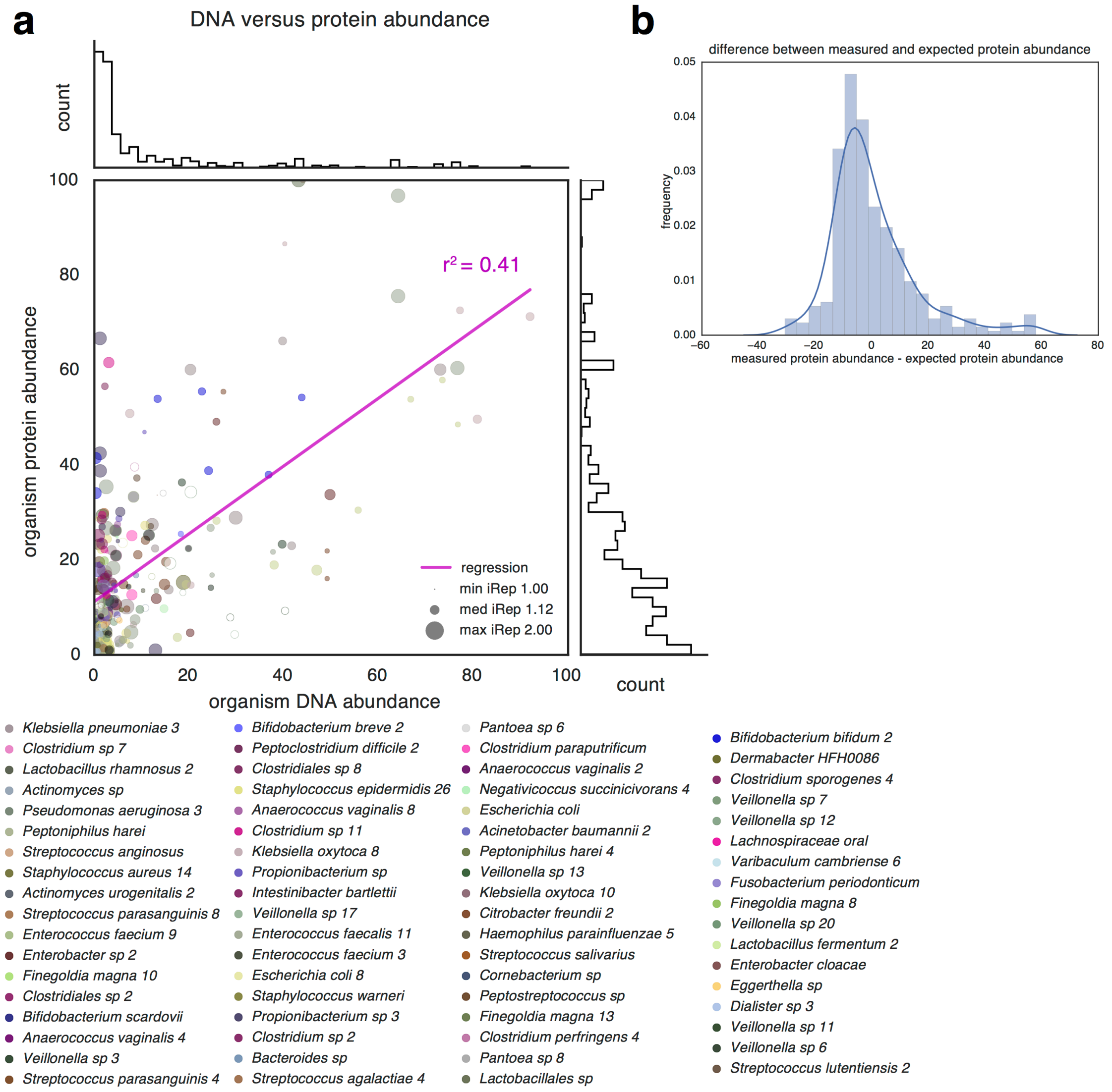


Figure 4.8 | Bacteria colonizing premature infants exhibit varying activity levels. Organism activity levels were measured based on proteomics (**a**) and iRep (**b**). Protein expression was measured as the difference between the measured proteome abundance and the expected proteome abundance determined based on DNA sequencing abundance (**see Figure 4.7**). iRep and proteomics measure different aspects of an organisms physiology and are not correlated. Shown are organisms with at least three measurements.

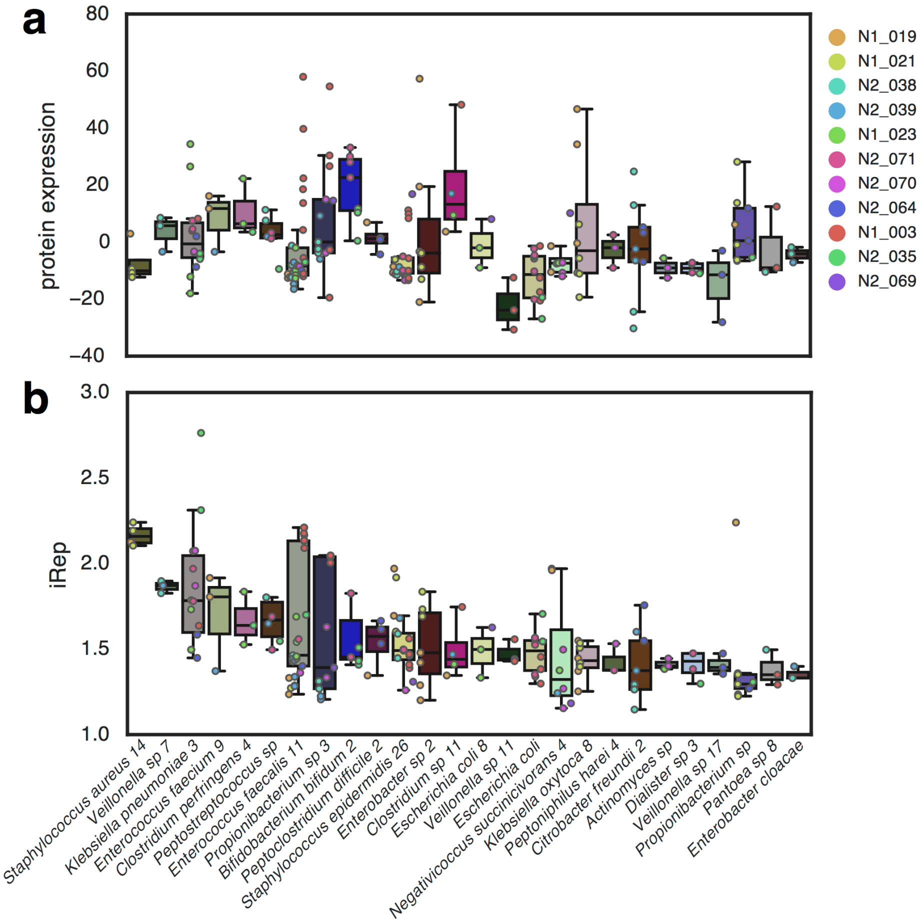


Figure 4.9 | Studied infant gut microbial communities associate into six distinct community types. a, Clustering was based on the presence and abundance of bacterial species (defined based on inclusion in 98% ANI genome cluster; hierarchical clustering was conducted based on a Euclidean distance matrix). Microbial community types are numbered and identified by colored boxes. Abundant species driving clustering of communities into types are shown in smaller boxes and the species names are colored. **b-g**, PCA clustering of microbial communities with associated metadata: infant the sample was collected from (**b**), infant health (**c**), number of days before NEC diagnosis (**d**), antibiotics usage (**e**), developmental age as measured by the number of days since conception (gestational age, GA + day of life, DOL; **f**), and clusters defined based on hierarchical clustering (**g**).

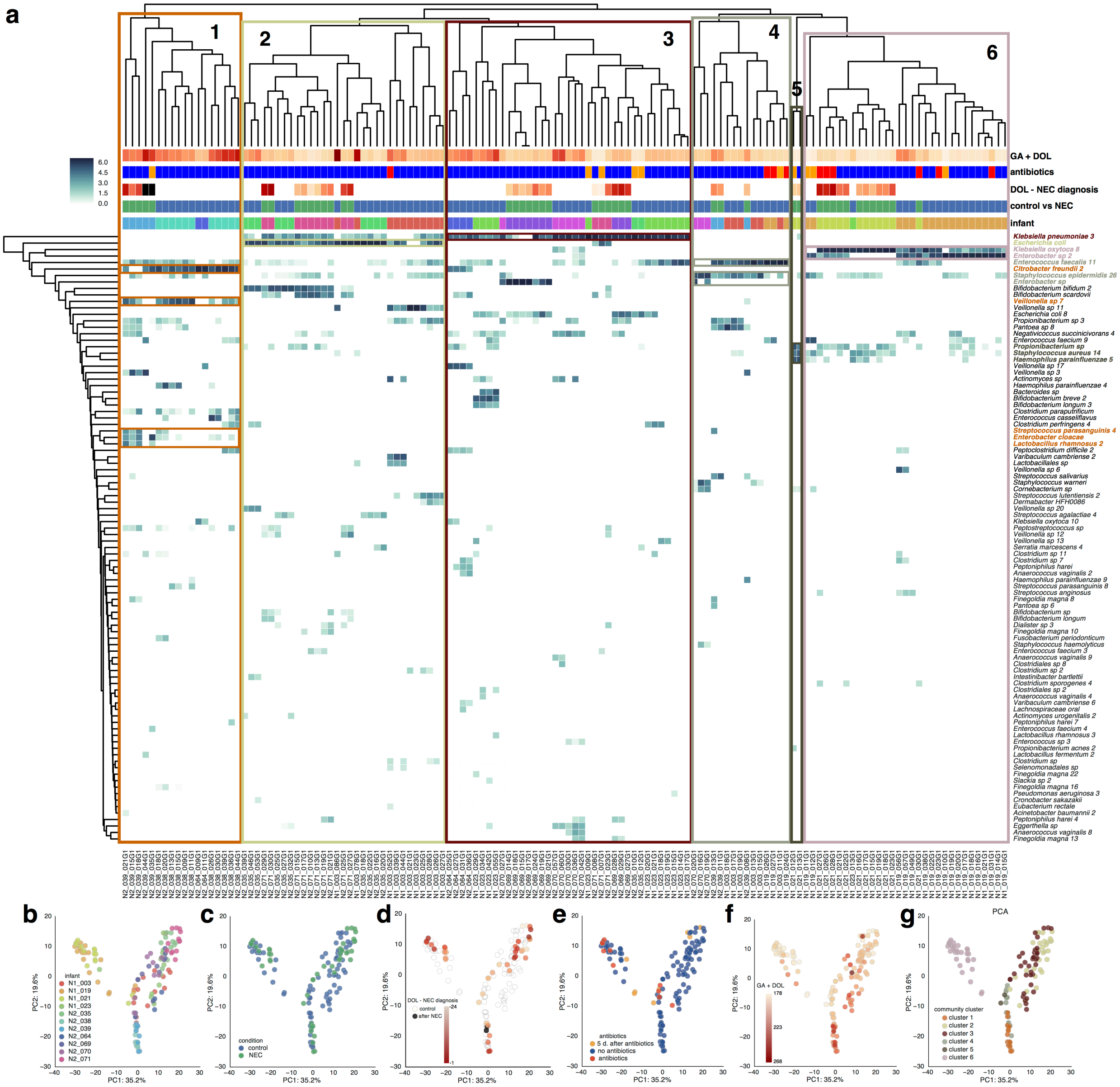


Figure 4.10 | Microbial community types are distinguished by their abundant members. a-f, Rank abundance curves showing the average and range of organism abundances associated with each community type.

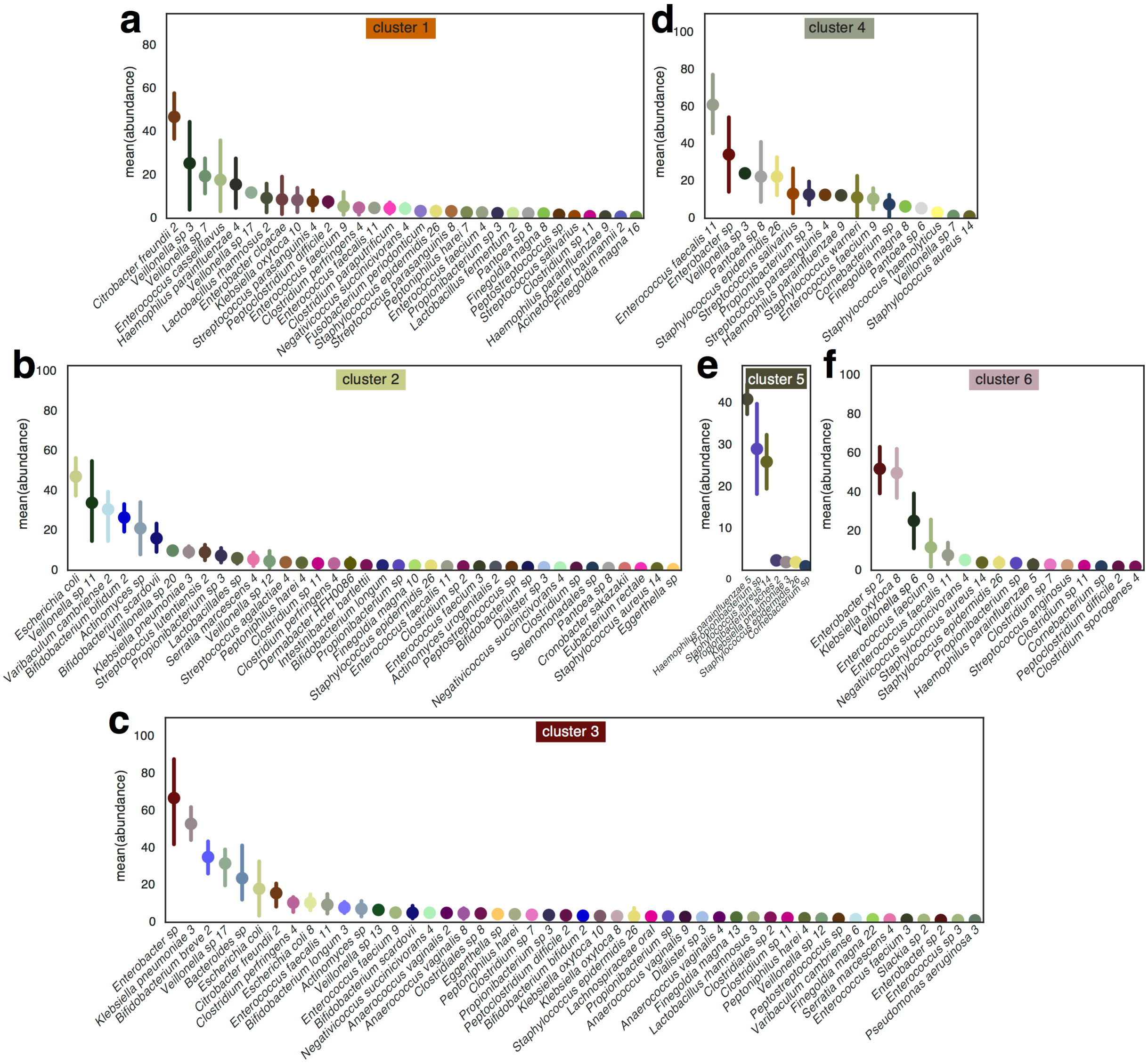


Figure 4.11 | Infants are colonized by several microbial community types during the first three months of life. a-d, Microbial community profiles for infant microbiomes clustered into similar community types. Community type is indicated by colored bars that match designations from **Figure 4.9**. DOL stands for day of life.

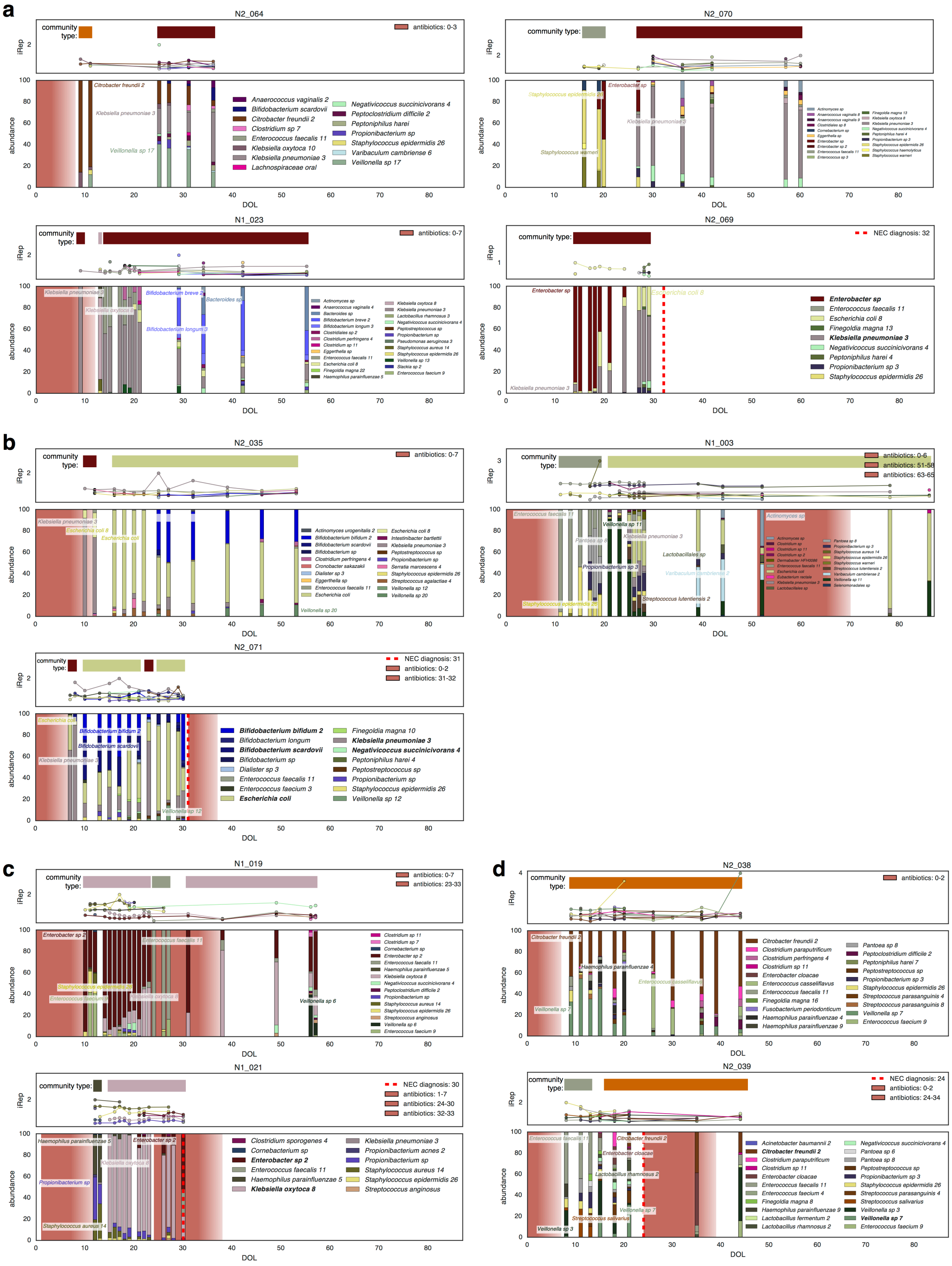


Figure 4.12 | Premature infant gut microbiome protein abundance patterns cluster by community type. Clustering of samples was conducted based on the relative abundance of abundant protein families (hierarchical clustering was conducted based on a Euclidean distance matrix). The 100 most abundant proteins for each infant were included. **b-g**, PCA clustering of proteomics data for all detected proteins with associated metadata: infant the sample was collected from (**b**), infant health (**c**), number of days before NEC diagnosis (**d**), antibiotics usage (**e**), developmental age as measured by the number of days since conception (gestational age, GA + day of life, DOL; **f**), and microbial community type (**g**).

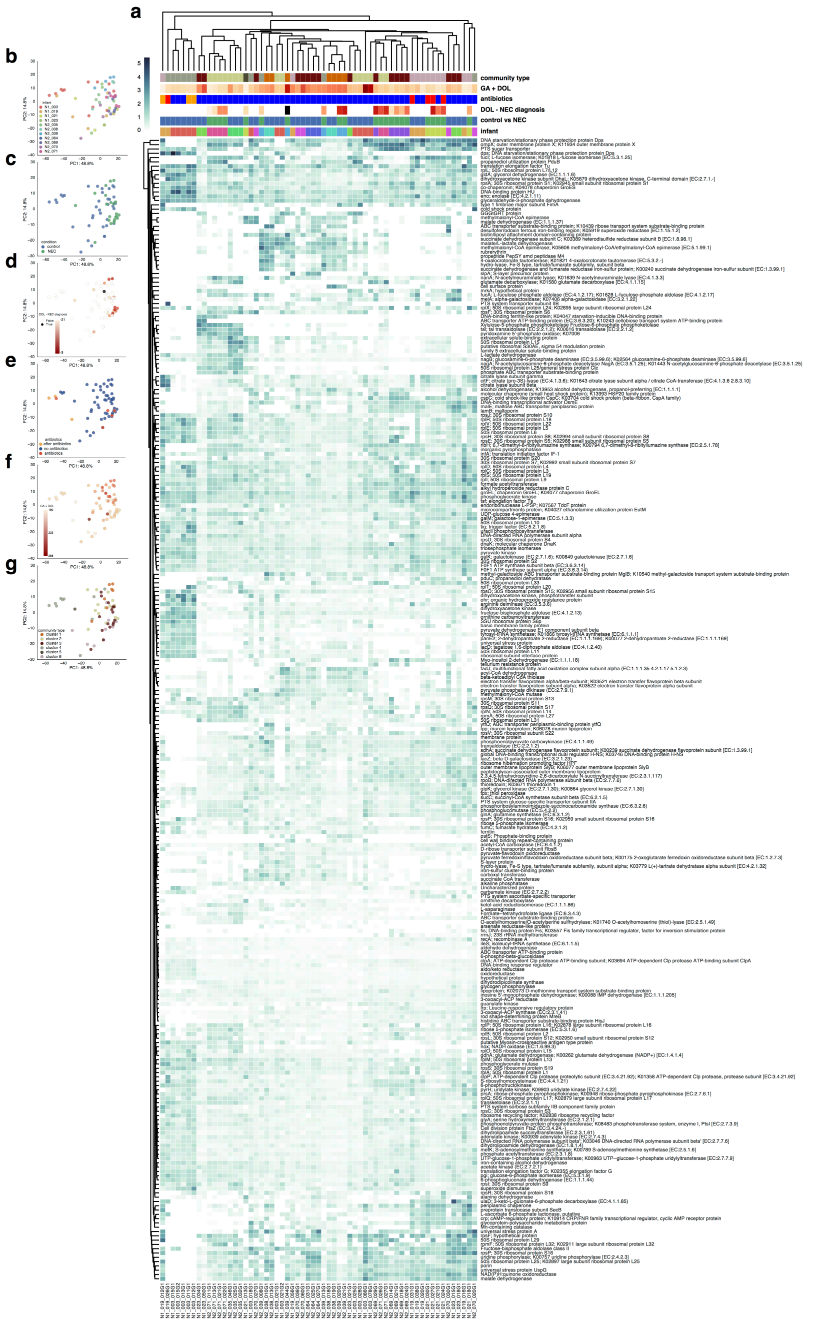


Figure 4.13 | Proteins associated with microbial community types. Abundance profiles are shown for protein families if they were among the most differentially expressed between community types. Proteome profiles were clustered based on their relative abundance within the sample (hierarchical clustering using Euclidean distance matrix). Colored boxes indicate microbial community type. Abundant proteins driving clustering are shown in smaller boxes, and the protein names are colored. Shown are protein families within the top 25 most differentially expressed between each pair of community types with a p-value ≤0.01. Cluster names listed on rows with proteins indicate which cluster had the highest protein expression.

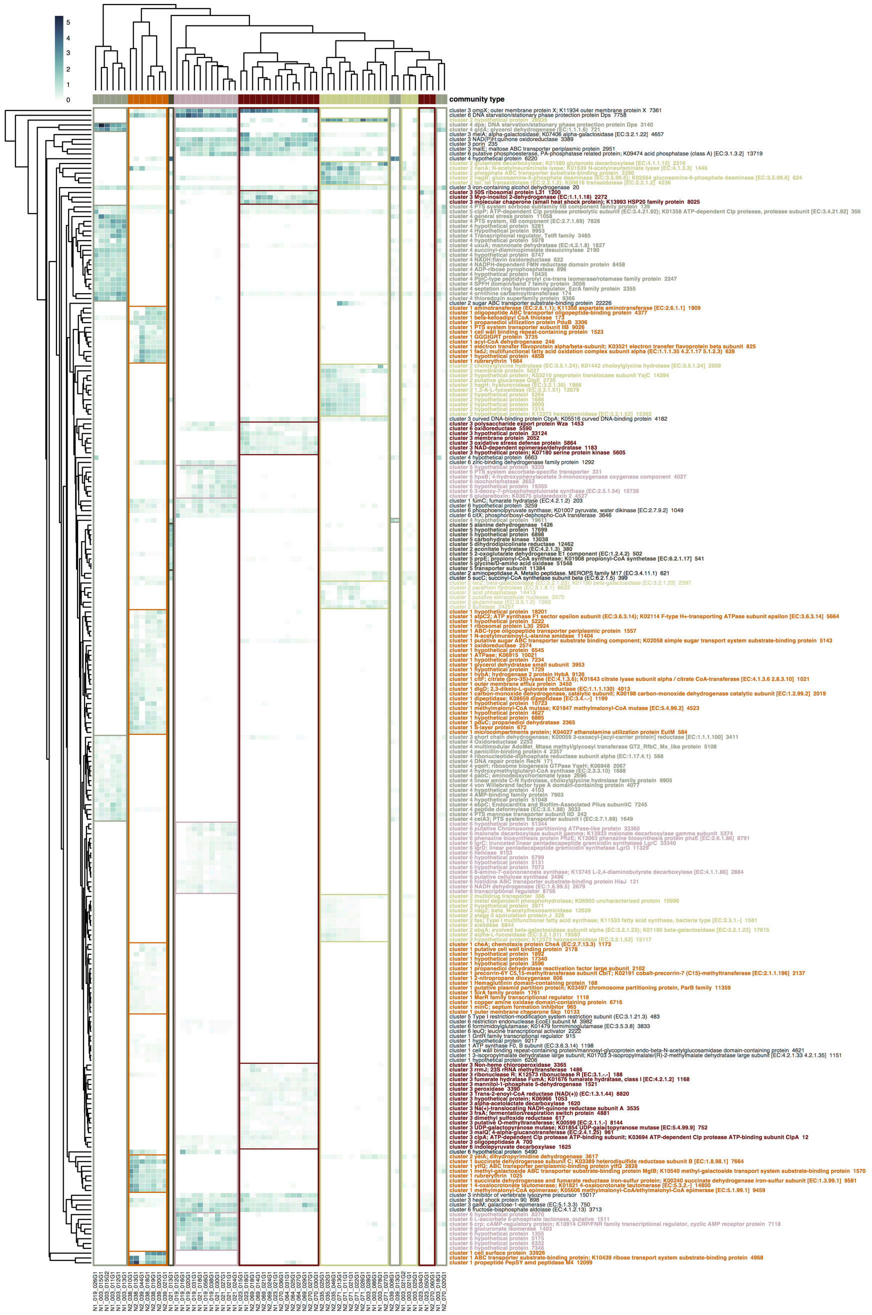
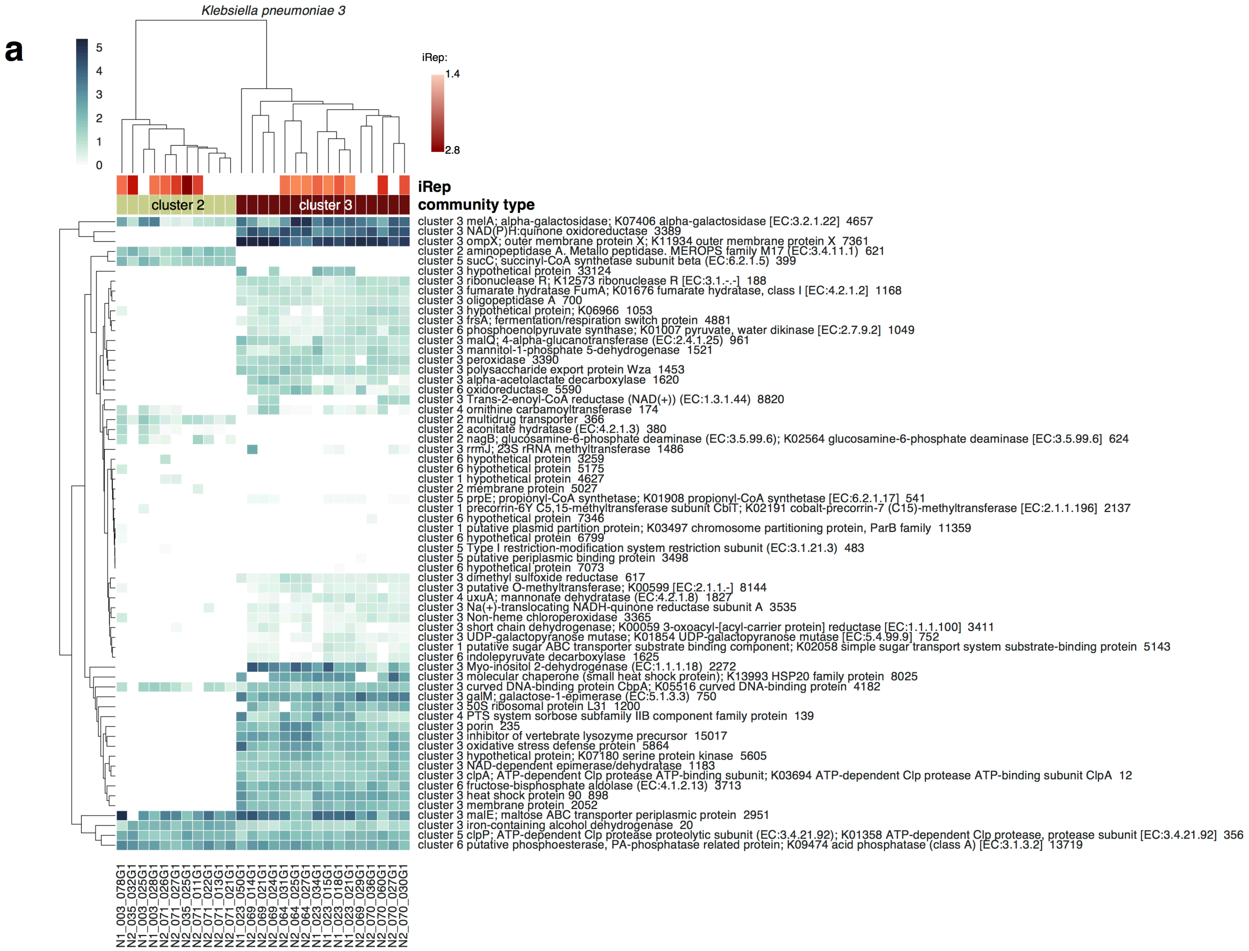
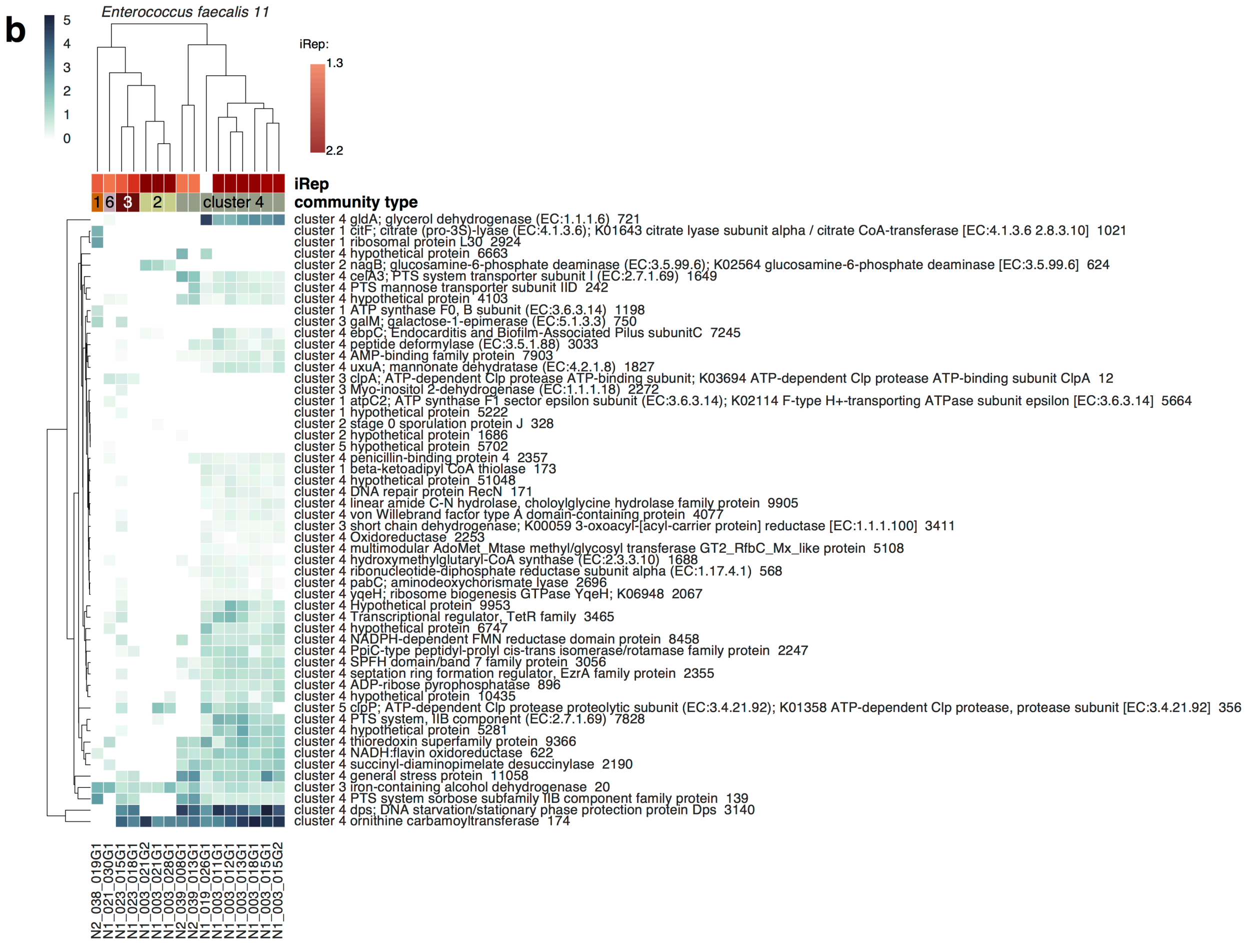


Figure 4.14 | Species-specific proteomic profiles correlate with community type and iRep replication rate. a-c, The expression levels of proteins that distinguish community types (**see Figure 4.13**) were determined for species present in multiple community types. Protein abundances were normalized for each genome in order to show changes in organism-specific relative proteome expression. Shown are proteome profiles for organisms in samples with ≥5% proteome detection. Cluster names listed on rows with proteins indicate which cluster had the highest protein expression.




b
5
4
3
2
1
0
Enterococcus faecalis 11
iRep:
1.3
2.2
iRep
community type
1
6
3
2
cluster 4
cluster 4 gldA; glycerol dehydrogenase (EC:1.1.1.6) 721
cluster 1 citF; citrate (pro-3S)-lyase (EC:4.1.3.6); K01643 citrate lyase subunit alpha / citrate CoA-transferase [EC:4.1.3.6 2.8.3.10] 1021
cluster 1 ribosomal protein L30 2924
cluster 4 hypothetical protein 6663
cluster 2 nagB; glucosamine-6-phosphate deaminase (EC:3.5.99.6); K02564 glucosamine-6-phosphate deaminase [EC:3.5.99.6] 624
cluster 4 celA3; PTS system transporter subunit I (EC:2.7.1.69) 1649
cluster 4 PTS mannose transporter subunit IID 242
cluster 4 hypothetical protein 4103
cluster 1 ATP synthase F0, B subunit (EC:3.6.3.14) 1198
cluster 3 galM; galactose-1-epimerase (EC:5.1.3.3) 750
cluster 4 ebpC; Endocarditis and Biofilm-Associated Pilus subunitC 7245
cluster 4 peptide deformylase (EC:3.5.1.88) 3033
cluster 4 AMP-binding family protein 7903
cluster 4 uxuA; mannonate dehydratase (EC:4.2.1.8) 1827
cluster 3 clpA; ATP-dependent Clp protease ATP-binding subunit; K03694 ATP-dependent Clp protease ATP-binding subunit ClpA 12
cluster 3 Myo-inositol 2-dehydrogenase (EC:1.1.1.18) 2272
cluster 1 atpC2; ATP synthase F1 sector epsilon subunit (EC:3.6.3.14); K02114 F-type H+-transporting ATPase subunit epsilon [EC:3.6.3.14] 5664
cluster 1 hypothetical protein 5222
cluster 2 stage 0 sporulation protein J 328
cluster 2 hypothetical protein 1686
cluster 5 hypothetical protein 5702
cluster 4 penicillin-binding protein 4 2357
cluster 1 beta-ketoadipyl CoA thiolase 173
cluster 4 hypothetical protein 51048
cluster 4 DNA repair protein RecN 171
cluster 4 linear amide C-N hydrolase, choloylglycine hydrolase family protein 9905
cluster 4 von Willebrand factor type A domain-containing protein 4077
cluster 3 short chain dehydrogenase; K00059 3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100] 3411
cluster 4 Oxidoreductase 2253
cluster 4 multimodular AdoMet_Mtase methyl/glycosyl transferase GT2_RfbC_Mx_like protein 5108
cluster 4 hydroxymethylglutaryl-CoA synthase (EC:2.3.3.10) 1688
cluster 4 ribonucleotide-diphosphate reductase subunit alpha (EC:1.17.4.1) 568
cluster 4 pabC; aminodeoxychorismate lyase 2696
cluster 4 yqeH; ribosome biogenesis GTPase YqeH; K06948 2067
cluster 4 Hypothetical protein 9953
cluster 4 Transcriptional regulator, TetR family 3465
cluster 4 hypothetical protein 6747
cluster 4 NADPH-dependent FMN reductase domain protein 8458
cluster 4 PpiC-type peptidyl-prolyl cis-trans isomerase/rotamase family protein 2247
cluster 4 SPFH domain/band 7 family protein 3056
cluster 4 septation ring formation regulator, EzrA family protein 2355
cluster 4 ADP-ribose pyrophosphatase 896
cluster 4 hypothetical protein 10435
cluster 5 clpP; ATP-dependent Clp protease proteolytic subunit (EC:3.4.21.92); K01358 ATP-dependent Clp protease, protease subunit [EC:3.4.21.92] 356
cluster 4 PTS system, IIB component (EC:2.7.1.69) 7828
cluster 4 hypothetical protein 5281
cluster 4 thioredoxin superfamily protein 9366
cluster 4 NADH:flavin oxidoreductase 622
cluster 4 succinyl-diaminopimelate desuccinylase 2190
cluster 4 general stress protein 11058
cluster 3 iron-containing alcohol dehydrogenase 20
cluster 4 PTS system sorbose subfamily IIB component family protein 139
cluster 4 dps; DNA starvation/stationary phase protection protein Dps 3140
cluster 4 ornithine carbamoyltransferase 174
N2_038_019G1
N1_021_030G1
N1_023_015G1
N1_023_018G1
N1_003_021G2
N1_003_021G1
N1_003_028G1
N2_039_008G1
N2_039_013G1
N1_019_026G1
N1_003_011G1
N1_003_012G1
N1_003_013G1
N1_003_018G1
N1_003_015G1
N1_003_015G2

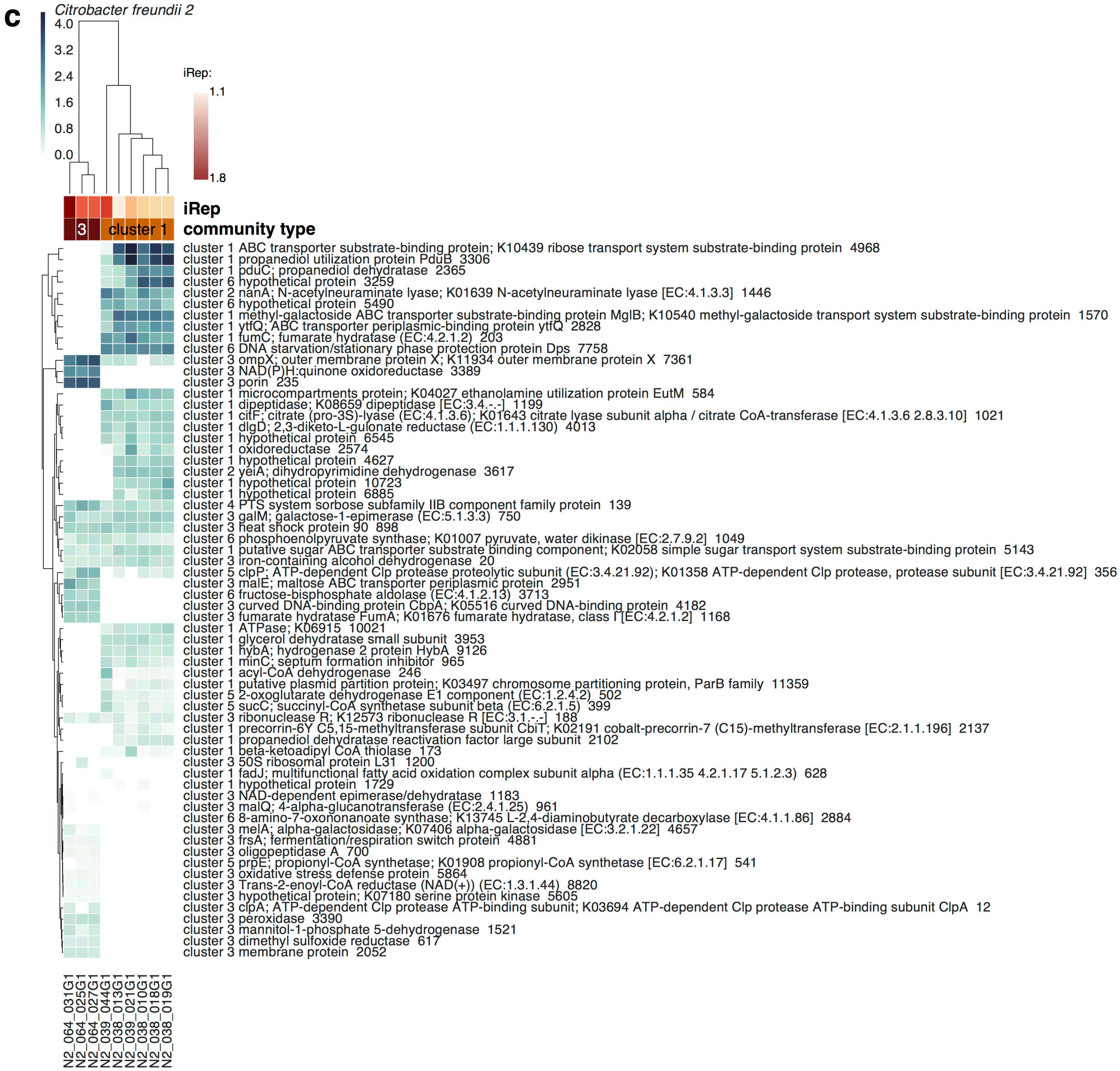

c
Citrobacter freundii 2
4.0
3.2
2.4
1.6
0.8
0.0
iRep:
1.1
1.8
iRep
community type
3
cluster 1
cluster 1 ABC transporter substrate-binding protein; K10439 ribose transport system substrate-binding protein 4968
cluster 1 propanediol utilization protein PduB 3306
cluster 1 pduC; propanediol dehydratase 2365
cluster 6 hypothetical protein 3259
cluster 2 nanA; N-acetylneuraminate lyase; K01639 N-acetylneuraminate lyase [EC:4.1.3.3] 1446
cluster 6 hypothetical protein 5490
cluster 1 methyl-galactoside ABC transporter substrate-binding protein MglB; K10540 methyl-galactoside transport system substrate-binding protein 1570
cluster 1 ytfQ; ABC transporter periplasmic-binding protein ytfQ 2828
cluster 1 fumC; fumarate hydratase (EC:4.2.1.2) 203
cluster 6 DNA starvation/stationary phase protection protein Dps 7758
cluster 3 ompX; outer membrane protein X; K11934 outer membrane protein X 7361
cluster 3 NAD(P)H:quinone oxidoreductase 3389
cluster 3 porin 235
cluster 1 microcompartments protein; K04027 ethanolamine utilization protein EutM 584
cluster 1 dipeptidase; K08659 dipeptidase [EC:3.4.-.-] 1199
cluster 1 citF; citrate (pro-3S)-lyase (EC:4.1.3.6); K01643 citrate lyase subunit alpha / citrate CoA-transferase [EC:4.1.3.6 2.8.3.10] 1021
cluster 1 dlgD; 2,3-diketo-L-gulonate reductase (EC:1.1.1.130) 4013
cluster 1 hypothetical protein 6545
cluster 1 oxidoreductase 2574
cluster 1 hypothetical protein 4627
cluster 2 yeiA; dihydropyrimidine dehydrogenase 3617
cluster 1 hypothetical protein 10723
cluster 1 hypothetical protein 6885
cluster 4 PTS system sorbose subfamily IIB component family protein 139
cluster 3 galM; galactose-1-epimerase (EC:5.1.3.3) 750
cluster 3 heat shock protein 90 898
cluster 6 phosphoenolpyruvate synthase; K01007 pyruvate, water dikinase [EC:2.7.9.2] 1049
cluster 1 putative sugar ABC transporter substrate binding component; K02058 simple sugar transport system substrate-binding protein 5143
cluster 3 iron-containing alcohol dehydrogenase 20
cluster 5 clpP; ATP-dependent Clp protease proteolytic subunit (EC:3.4.21.92); K01358 ATP-dependent Clp protease, protease subunit [EC:3.4.21.92] 356
cluster 3 malE; maltose ABC transporter periplasmic protein 2951
cluster 6 fructose-bisphosphate aldolase (EC:4.1.2.13) 3713
cluster 3 curved DNA-binding protein CbpA; K05516 curved DNA-binding protein 4182
cluster 3 fumarate hydratase FumA; K01676 fumarate hydratase, class I [EC:4.2.1.2] 1168
cluster 1 ATPase; K06915 10021
cluster 1 glycerol dehydratase small subunit 3953
cluster 1 hybA; hydrogenase 2 protein HybA 9126
cluster 1 minC; septum formation inhibitor 965
cluster 1 acyl-CoA dehydrogenase 246
cluster 1 putative plasmid partition protein; K03497 chromosome partitioning protein, ParB family 11359
cluster 5 2-oxoglutarate dehydrogenase E1 component (EC:1.2.4.2) 502
cluster 5 sucC; succinyl-CoA synthetase subunit beta (EC:6.2.1.5) 399
cluster 3 ribonuclease R; K12573 ribonuclease R [EC:3.1.-.-] 188
cluster 1 precorrin-6Y C5,15-methyltransferase subunit CbiT; K02191 cobalt-precorrin-7 (C15)-methyltransferase [EC:2.1.1.196] 2137
cluster 1 propanediol dehydratase reactivation factor large subunit 2102
cluster 1 beta-ketoadipyl CoA thiolase 173
cluster 3 50S ribosomal protein L31 1200
cluster 1 fadJ; multifunctional fatty acid oxidation complex subunit alpha (EC:1.1.1.35 4.2.1.17 5.1.2.3) 628
cluster 1 hypothetical protein 1729
cluster 3 NAD-dependent epimerase/dehydratase 1183
cluster 3 malQ; 4-alpha-glucanotransferase (EC:2.4.1.25) 961
cluster 6 8-amino-7-oxononanoate synthase; K13745 L-2,4-diaminobutyrate decarboxylase [EC:4.1.1.86] 2884
cluster 3 melA; alpha-galactosidase; K07406 alpha-galactosidase [EC:3.2.1.22] 4657
cluster 3 frsA; fermentation/respiration switch protein 4881
cluster 3 oligopeptidase A 700
cluster 5 prpE; propionyl-CoA synthetase; K01908 propionyl-CoA synthetase [EC:6.2.1.17] 541
cluster 3 oxidative stress defense protein 5864
cluster 3 Trans-2-enoyl-CoA reductase (NAD(+)) (EC:1.3.1.44) 8820
cluster 3 hypothetical protein; K07180 serine protein kinase 5605
cluster 3 clpA; ATP-dependent Clp protease ATP-binding subunit; K03694 ATP-dependent Clp protease ATP-binding subunit ClpA 12
cluster 3 peroxidase 3390
cluster 3 mannitol-1-phosphate 5-dehydrogenase 1521
cluster 3 dimethyl sulfoxide reductase 617
cluster 3 membrane protein 2052
N2_064_031G1
N2_064_025G1
N2_064_027G1
N2_039_044G1
N2_038_013G1
N2_039_021G1
N2_038_010G1
N2_038_018G1
N2_038_019G1

Table 4.1 | Infant medical information.

infant	sex	delivery	gestational age (weeks)	birth weight (g)	feeding	condition	NEC diagnosis (DOL)
N1_003	F	c-section	26	822	breast	control	n/a
N1_019	F	c-section	24	731	breast + formula	sepsis	n/a
N1_021	F	c-section	24	697	breast	NEC	30
N1_023	F	vaginal	27	875	breast	control	n/a
N2_035	M	vaginal	25	795	breast	control	n/a
N2_038	F	c-section	30	1381	breast + formula	control	n/a
N2_039	F	c-section	30	1470	breast + formula	NEC	24
N2_064	M	vaginal	28	1100	breast + formula	control	n/a
N2_069	M	c-section	26	637	breast	NEC	32
N2_070	F	c-section	26	633	breast + formula	control	n/a
N2_071	M	c-section	25	754	breast + formula	NEC	31

Chapter 5

Summary and future directions

The Candidate Phyla Radiation

The Candidate Phyla Radiation (CPR) encompasses a diverse collection of related phylum-level lineages almost completely devoid of cultured representatives. Culture-independent 16S rRNA gene sequencing had suggested the existence of this group during the last decades, and has shown that these organisms are essentially ubiquitous (Harris et al., 2004). However, only recently have we begun to recognize this as a coherent group. Our genomic sampling covered a substantial portion of previously defined lineages, and several that had never been recognized before (**Figure 3.1**). All CPR studied to date have limited metabolic potential, and organisms from most lineages have incomplete nucleotide, amino acid, and fatty acid biosynthesis pathways, indicating that they depend on other organisms for survival (Anantharaman et al., 2016a; 2016b; Brown et al., 2015; Kantor et al., 2013; Wrighton et al., 2012). The diversity and prevalence of CPR in natural environments highlights the prominence of currently little known organisms and inter-organism interactions in microbial communities.

Genome-resolved metagenomics was critical to identifying this radiation as a distinct feature of the tree of life. Complete and high-quality draft genomes enabled multiple phylogenetic analyses of the same organisms, and identification of genomic features that distinguish CPR organisms from other bacteria (**Chapter 2**). The CPR is monophyletic based on both 16S rRNA gene and concatenated ribosomal protein phylogenies, but exhibited different placement within the tree in each analysis (Hug et al., 2016). In the concatenated ribosomal protein tree the CPR emerges as a deep-branching, early evolving clade (**Appendix 1.1**), suggesting that their consistently small genomes and limited metabolism may not be the result of genome reduction, but rather reflects a metabolic platform from early life where sharing of resources may have been commonplace. While the high level of diversity throughout the CPR is consistent with the hypothesis that they evolved early on, we cannot rule out the possibility that the diversity seen within the CPR is due to accelerated evolutionary rates. Future studies of evolutionary rates across the tree of life may shed additional light on their evolutionary history. Regardless, the CPR represents a considerable amount of genetic diversity.

Based on conservative 16S rRNA gene sequence identity cutoffs, the CPR is estimated to be comprised of hundreds of phyla (Brown et al., 2015; Yarza et al., 2014). Additional comparative phylogenetic analyses were conducted in order to better understand the phylogenetic structure of the CPR. We used 16S rRNA and concatenated ribosomal protein phylogenies to delineate previously unrecognized phyla within the CPR, and found that two of the largest groups, the Microgenomates (OP11) and Parcubacteria (OD1) are superphyla. In support of this observation, a comprehensive phylogenetic analysis suggested that approximately 50% of bacterial genetic diversity is represented by the CPR (Hug et al., 2016). An outstanding question from these analyses was the placement of the Absconditabacteria (SR1), which was not consistently placed within the CPR. In follow up analysis, we have shown that the Absconditabacteria do belong with the CPR, sister to the Gracilibacteria (BD1-5) (**Appendix 3**). The grouping of the Absconditabacteria and Gracilibacteria is consistent with the finding that organisms from both of these groups have alternatively coded genomes in which UGA codes for glycine instead of functioning as a stop codon (Campbell et al., 2013; Kantor et al., 2013; Wrighton et al., 2012).

An open question related to the CPR is whether or not protein evolution is driven largely by positive selection or genetic drift. The frequency of rRNA gene sequence introns suggests that genetic drift may play a large role in CPR evolution (**Chapter 2**), but this has not been demonstrated. Although CPR bacteria have small genomes, they encode an unusually high number of proteins with no known function. Identification of proteins experiencing different selective pressures across a variety of natural systems would identify rapidly evolving pathways and cellular components. This approach would enable measurement of protein evolution, even for proteins with no known function, and provide a foundation for future protein characterization studies. Preliminary results of CPR ortholog analyses indicate that, as would be expected for organisms from different phyla, few orthologs are widely distributed. Across draft-quality genomes from the OD1-L1 group (Parcubacteria missing ribosomal protein L1), only ~150 proteins were encoded by most genomes. Approximately 100 of these proteins are also encoded by non-CPR genomes, indicating that the CPR has a small "core" genome, and a large set of novel auxiliary proteins.

Several features of CPR bacteria indicate that they typically replicate infrequently. For example, their limited metabolic potential, single copy of the rRNA gene operon, the prevalence of rRNA gene insertions (~46% of 16S and 23S rRNA genes have an insertion ≥5 bp), unusual ribosomes and ribosome biogenesis mechanisms, and an almost complete lack of CRISPR-Cas virus defense systems (Burstein et al., 2016). Furthermore, cryogenic transmission electron microscopy of ultra-small CPR bacteria showed that they have few ribosomes per cell (42 +/- 9.5) (Luef et al., 2015). iRep, the method we developed for determining replication rates based on genome sequencing coverage trends, requires that the organism replicate their genome from a single origin of replication. Analysis of GC skew patterns across complete CPR genomes confirms that these organisms, like essentially all bacteria, also replicate their genomes bi-directionally from a single origin (Anantharaman et al., 2016b; Gao et al., 2013; Sernova and Gelfand, 2008) (**Appendix 4**). We used iRep to directly measure replication for CPR bacteria sampled during an acetate amendment field experiment and found that most, but not all exhibited slow replication rates (**Chapter 3**).

Strikingly, most CPR bacteria are not able to produce their own nucleotides, amino acids, or fatty acids (Brown et al., 2015; Burstein et al., 2016). These findings strongly suggest that CPR bacteria are symbionts (Albertsen et al., 2013; Anantharaman et al., 2016a; Brown et al., 2015; Kantor et al., 2013; Nelson and Stegen, 2015; Podar et al., 2007; Rinke et al., 2013; Wrighton et al., 2012), a notion that has been confirmed in a few cases (Gong et al., 2014; He et al., 2015; Luo et al., 2016; Soro et al., 2014). Notable among these studies is a detailed determination that a Saccharibacteria (TM7) phylotype associated with the human oral cavity is an obligate epibiont of an *Actinomyces odontolyticus* strain (He et al., 2015). Microscopy showed Saccharibacteria associated with the *A. odontolyticus* membrane, indicating a possible method for acquiring lipids. Their analyses showed that Saccharibacteria have a specific symbiotic association, whereby growth was only facilitated by a particular *A. odontolyticus* strain. This high level of specificity suggests that analysis of co-abundance and/or co-replication (iRep) could identity symbiont pairs for a large number of CPR bacteria. Comparative analysis of CPR and symbiont genomes could uncover additional roles of CPR bacteria in their environment, which could have impacts on large-scale biogeochemical processes. Furthermore, this may provide necessary information for establishing co-culture systems, which would provide opportunities for genetic and biochemical analysis of these enigmatic organisms. This would enable investigation of the mechanisms behind rRNA gene intron splicing for the many cases where none could be predicted (**Figure 2.5**), and evaluation of the functionality of ribosomes that are missing proteins thought to be required.

Microbial colonization of the premature infant gut

The developing infant microbiome is of great interest due to the potential for life-long impacts on health and development (Groer et al., 2014). Basic questions exit related to the colonization process, especially during early life where little is known about the dynamics and metabolism of the microbiome, or how life events such as changes in feeding or antibiotic use can alter this process. Premature infants are an interesting study group to address these questions because they are routinely given antibiotics at birth, are kept in the relatively controlled hospital environment during early life, and have low diversity microbial communities that are tractable for high-resolution genome-resolved microbiome studies (Brown et al., 2013; Raveh-Sadka et al., 2015; 2016; Sharon et al., 2012). Furthermore, they represent an at risk population in which microbiome interventions could have significant impacts on health, especially if they were to mitigate the incidence of necrotizing enterocolitis (NEC) (Neu and Walker, 2011).

Our detailed metagenomic and metaproteomics analyses of the colonization process uncovered microbial community dynamics and metabolic shifts that would not be apparent in other types of analyses (**Chapter 1**, **Chapter 3, and Chapter 4**). These techniques enable measurement of microbial community membership, metabolic potential, replication rates, and metabolic activity, and showed that members of even the same species exhibit different replication rates and metabolism over time, which may be important to human health. Additional detailed studies of more infants will be required in order to determine whether or not the trends observed thus far are general features of the colonization process. Identification of core colonization patterns across large sets of infants may identify metabolisms associated with various colonization trajectories, some of which may lead to disease.

In situ replication rates for bacteria in microbial communities

Cells in a population undergoing division contain, on average, more than one copy of their genome (**Figure 3.1**). In an unsynchronized population, cells have replicated their genomes to different extents, resulting in a gradual decrease in average genome copy number from the origin to terminus of replication. We developed a novel method for calculating replication rates by measuring changes in genome copy number by mapping metagenome-sequencing reads to draft-quality genome sequences (**Chapter 3**). This method provides a quantitative measurement of replication, which we call the Index of Replication, or iRep. Given that metagenome studies routinely generate hundreds of genomes of suitable quality to conduct an iRep analysis, this method should continue to provide novel insights into microbial community dynamics.

Little is known about *in situ* bacterial replication strategies, and much could be learned by i) measuring replication rates for bacteria across multiple environments (e.g. human microbiome, soil, sediments, and groundwater), ii) characterizing organisms based on replication rate patterns, and iii) linking replication strategies to genomic and metabolic features. Replication rates could be determined for organisms from different environments using genome-resolved metagenomics and the newly developed iRep method. Because genome sequences would be available for all studied organisms, it would be possible to correlate genomic features such as rRNA gene copy number and codon usage bias, and metabolic potential, such as carbon and energy sources, with replication patterns. rRNA gene copy number, codon usage bias, and specific energy generating strategies, such as respiration, have been suggested to influence replication rates, but the extent to which this applies to organisms in community context is not known. Overlap in community membership across environments will enable comparison of organism replication rates across environment types. Proteomic and transcriptomic data can then be used to further profile expression patterns for specific organisms under different conditions, and provide an understanding of organism responses to changing environmental conditions (for example see **Chapter 4**).

It has been proposed that organisms can be classified as *r* or *k* strategists based on whether they replicate rapidly when conditions are favorable, versus slow, steady replication. These classifications are important because they can be incorporated into predictive models of microbial communities and ecosystems. However, until now it has not been possible to obtain *in situ* replication rates that could confirm this hypothesis. Even if not supported, replication rate measurements can be used to develop new models for how microbes persist and respond to changing conditions. Further, once the relevant groups are established, it will be possible to correlate phylogenetic and metabolic information with organisms exhibiting specific replication strategies across time and environment type.

Software Development

The work presented here depended on the development of a number of software tools (**Appendix 5**), which have also been used in several co-authored studies (**Appendix 1**). Development of several of these tools is ongoing. Future work on the iRep algorithm (**Chapter 3)** will focus on adding additional statistical tests to provide confidence levels for replication rates determined from genomes with low sequencing coverage, and inclusion of a built in method for detecting levels of strain variation that may skew results. Several other tools were developed in order to improve genome sequence assemblies. The primary advance here being the development of an automated method for identifying and correcting scaffolding errors (**ra2.py**). These small, localized errors occur during the assembly process when contigs are joined based on paired read sequences. While these rarely result in chimeric assemblies, they can result in incomplete or fragmented gene predictions. Future work on this software will focus on memory handling, which would enable automatic curation of large metagenome assemblies, and improvements in the algorithm used for re-assembling errors. Other notable tools were developed for clustering genomes based on their average nucleotide identity, identification of orthologs between genomes in order to conduct comparative analyses, identification and characterization of rRNA genes and insertions, and for identifying a set of ribosomal proteins frequently used in phylogenetic analyses. For more information see **Appendix 5**.

References

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. Bioinformatics *21*, 2104–2105.

Akanuma, G., Nanamiya, H., Natori, Y., Yano, K., Suzuki, S., Omata, S., Ishizuka, M., Sekine, Y., and Kawamura, F. (2012). Inactivation of Ribosomal Protein Genes in Bacillus subtilis Reveals Importance of Each Ribosomal Protein for Cell Proliferation and Cell Differentiation. J. Bacteriol. *194*, 6282–6291.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., and Nielsen, P.H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat. Biotechnol. *31*, 533–538.

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. Nat Meth *11*, 1144–1146.

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol. *215*, 403–410.

Anantharaman, K., Brown, C.T., Burstein, D., Castelle, C.J., Probst, A.J., Thomas, B.C., Williams, K.H., and Banfield, J.F. (2016a). Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum. PeerJ *4*, e1607–e1607.

Anantharaman, K., Brown, C.T., Hug, L.A., Sharon, I., Castelle, C.J., Probst, A.J., Thomas, B.C., Singh, A., Wilkins, M.J., Karaoz, U., et al. (2016b). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat. Commun. *7*, 13219.

Andronescu, M., Condon, A., Hoos, H.H., Mathews, D.H., and Murphy, K.P. (2007). Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics *23*, i19–i28.

Atkins, J.F., and Björk, G.R. (2009). A gripping tale of ribosomal frameshifting: extragenic suppressors of frameshift mutations spotlight P-site realignment. Microbiol. Mol. Biol. Rev. *73*, 178–210.

Baker, B.J., and Dick, G.J. Omic Approaches in Microbial Ecology: Charting the Unknown. Microbe *8*, 353–360.

Baker, B.J., Hugenholtz, P., Dawson, S.C., and Banfield, J.F. (2003). Extremely acidophilic protists from acid mine drainage host Rickettsiales-lineage endosymbionts that have intervening sequences in their 16S rRNA genes. Appl. Environ. Microbiol. *69*, 5512–5518.

Baker, B.J., Comolli, L.R., Dick, G.J., Hauser, L.J., Hyatt, D., Dill, B.D., Land, M.L., Verberkmoes, N.C., Hettich, R.L., and Banfield, J.F. (2010). Enigmatic, ultrasmall, uncultivated Archaea. Pnas *107*, 8806–8811.

Boch, J., Nau-Wagner, G., Kneip, S., and Bremer, E. (1997). Glycine betaine aldehyde dehydrogenase from Bacillus subtilis: characterization of an enzyme required for the synthesis of the osmoprotectant glycine betaine. Arch. Microbiol. *168*, 282–289.

Bokulich, N.A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., Lieber, A.D., Wu, F., Perez-Perez, G.I., Chen, Y., et al. (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. Sci. Transl. Med. *8*.

Bremer, H., and Churchward, G. (1977). An examination of the Cooper-Helmstetter theory of DNA replication in bacteria and its underlying assumptions. J. Theor. Biol. *69*, 645–654.

Brown, C.T., Davis-Richardson, A.G., Giongo, A., Gano, K.A., Crabb, D.B., Mukherjee, N., Casella, G., Drew, J.C., Ilonen, J., Knip, M., et al. (2011). Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. PLoS ONE *6*, e25792.

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., and Banfield, J.F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. Nature *523*, 208–211.

Brown, C.T., Olm, M.R., Thomas, B.C., and Banfield, J.F. (2016). Measurement of bacterial replication rates in microbial communities. Nat. Biotechnol. *34*, 1256–1263.

Brown, C.T., Sharon, I., Thomas, B.C., Castelle, C.J., Morowitz, M.J., and Banfield, J.F. (2013). Genome resolved analysis of a premature infant gut microbial community reveals a *Varibaculum cambriense* genome and a shift towards fermentation-based metabolism during the third week of life. Microbiome *1*, 30.

Burge, S.W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E.P., Eddy, S.R., Gardner, P.P., and Bateman, A. (2012). Rfam 11.0: 10 years of RNA families. Nucleic Acids Res *41*, D226–D232.

Burstein, D., Sun, C.L., Brown, C.T., Sharon, I., Anantharaman, K., Probst, A.J., Thomas, B.C., and Banfield, J.F. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat. Commun. *7*, 10613.

Burt, A., and Koufopanou, V. (2004). Homing endonuclease genes: the rise and fall and rise again of a selfish element. Current Opinion in Genetics & Development *14*, 609–615.

Campbell, J.H., O'Donoghue, P., Campbell, A.G., Schwientek, P., Sczyrba, A., Woyke, T., Soll, D., and Podar, M. (2013). UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. Pnas *110*, 5540–5545.

Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Müller, K.M., et al. (2002). The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics *3*, 2.

Caplan, M.S. (2009). Probiotic and prebiotic supplementation for the prevention of neonatal necrotizing enterocolitis. Journal of Perinatology *29 Suppl 2*, S2–S6.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L., and Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics *26*, 266–267.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010b). QIIME allows analysis of high-throughput community sequencing data. Nat Meth *7*, 335–336.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. Isme J *6*, 1621–1624.

Carini, P., Marsden, P.J., Leff, J.W., Morgan, E.E., Strickland, M.S., and Fierer, N. (2016). Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. bioRxiv 043372.

Castelle, C.J., Hug, L.A., Wrighton, K.C., Thomas, B.C., Williams, K.H., Wu, D., Tringe, S.G., W, S.S., Eisen, J.A., and Banfield, J.F. (2013). Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. Nat. Commun. *4*, 2120.

Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J., Frischkorn, K.R., Tringe, S.G., Singh, A., Markillie, L.M., et al. (2015). Genomic expansion of domain Archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr. Biol. *25*, 690–701.

Chivian, D., Brodie, E.L., Alm, E.J., Culley, D.E., Dehal, P.S., DeSantis, T.Z., Gihring, T.M., Lapidus, A., Lin, L.H., Lowry, S.R., et al. (2008). Environmental Genomics Reveals a Single-Species Ecosystem Deep Within Earth. Science *322*, 275–278.

Chu, Y.W., Wong, C.H., Chu, M.Y., Cheung, C.P.F., Cheung, T.K.M., Tse, C., Luk, W.K., and Lo, J.Y.C. (2009). *Varibaculum cambriense* infections in Hong Kong, China, 2006. Emerging Infect. Dis. *15*, 1137–1139.

Cooper, S., and Helmstetter, C.E. (1968). Chromosome replication and the division cycle of *Escherichia coli* B/r. J. Mol. Biol. *31*, 519–540.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. Science *326*, 1694–1697.

Das, A., Silaghi-Dumitrescu, R., Ljungdahl, L.G., and Kurtz, D.M. (2005). Cytochrome bd Oxidase, Oxidative Stress, and Dioxygen Tolerance of the Strictly Anaerobic Bacterium Moorella thermoacetica. J. Bacteriol. *187*, 2020–2029.

David, C. (2012). Protein-linked glycan degradation in infants fed human milk. J Glycomics Lipidomics *S1*, 002.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. *72*, 5069–5072.

Di Rienzi, S.C., Sharon, I., Wrighton, K.C., Koren, O., Hug, L.A., Thomas, B.C., Goodrich, J.K., Bell, J.T., Spector, T.D., Banfield, J.F., et al. (2013). The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. Elife *2*, e01102–e01102.

Dick, G.J., Andersson, A.F., Baker, B.J., Simmons, S.L., Thomas, B.C., Yelton, A.P., and Banfield, J.F. (2009). Community-wide analysis of microbial genome sequence signatures. Genome Biol *10*, R85–.

Eckburg, P.B.P., Bik, E.M.E., Bernstein, C.N.C., Purdom, E.E., L Dethlefsen, Les, Sargent, M.M., Gill, S.R.S., Nelson, K.E.K., and Relman, D.A.D. (2005). Diversity of the human intestinal microbial flora. Science *308*, 1635–1638.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol *7*, e1002195.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792–1797.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Edgar, R.C. UBLAST. Drive5.com.

Eloe-Fadrosh, E.A., Páez-Espino, D., Jarett, J., Dunfield, P.F., Hedlund, B.P., Dekas, A.E., Grasby, S.E., Brady, A.L., Dong, H., Briggs, B.R., et al. (2016). Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. Nat. Commun. *7*, 10476.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res *30*, 1575–1584.

Evguenieva-Hackenberg, E. (2005). Bacterial ribosomal RNA in pieces. Mol. Microbiol. *57*, 318–325.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2013). Pfam: the protein families database. Nucleic Acids Res *42*, D222–D230.

Gao, F., Luo, H., and Zhang, C.-T. (2013). DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. Nucleic Acids Res *41*, D90–D93.

Gilbert, J.A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C.T., Brown, C.T., Desai, N., Eisen, J.A., Evers, D., Field, D., et al. (2010). Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. Stand. Genomic Sci. *3*, 243–248.

Gilbert, J.A., Quinn, R.A., Debelius, J., Xu, Z.Z., Morton, J., Garg, N., Jansson, J.K., Dorrestein, P.C., and Knight, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. Nature *535*, 94–103.

Giongo, A., Davis-Richardson, A.G., Crabb, D.B., and Triplett, E.W. (2010). TaxCollector: Modifying Current 16S rRNA Databases for the Rapid Classification at Six Taxonomic Levels. Diversity *2*, 1015–1025.

Gong, J., Qing, Y., Guo, X., and Warren, A. (2014). "Candidatus Sonnebornia yantaiensis," a member of candidate division OD1, as intracellular bacteria of the ciliated protist *Paramecium bursaria* (Ciliophora, Oligohymenophorea). Syst. Appl. Microbiol. *37*, 35–41.

Grigoriev, A. (1998). Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res *26*, 2286–2290.

Groer, M.W., Luciano, A.A., Dishaw, L.J., Ashmeade, T.L., Miller, E., and Gilbert, J.A. (2014). Development of the preterm infant gut microbiome: a research priority. Microbiome *2*, 38–54.

Hall, V. (2008). Actinomyces—Gathering evidence of human colonization and infection. Anaerobe *14*, 1–7.

Hall, V.V., Collins, M.D.M., Lawson, P.A.P., Hutson, R.A.R., Falsen, E.E., Inganas, E.E., and Duerden, B.B. (2003). Characterization of some actinomyces-like isolates from human clinical sources: description of *Varibaculum cambriensis* gen nov, sp nov. J. Clin. Microbiol. *41*, 640–644.

Hallman, M., Bry, K., Hoppu, K., Lappi, M., and Pohjavuori, M. (1992). Inositol supplementation in premature infants with respiratory distress syndrome. N Engl J Med *326*, 1233–1239.

Harris, J.K., Kelley, S.T., and Pace, N.R. (2004). New perspective on uncultured bacterial phylogenetic division OP11. Appl. Environ. Microbiol. *70*, 845–849.

He, X., McLean, J.S., Edlund, A., Yooseph, S., Hall, A.P., Liu, S.-Y., Dorrestein, P.C., Esquenazi, E., Hunter, R.C., Cheng, G., et al. (2015). Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Pnas *112*, 244–249.

Heintz-Buschart, A., May, P., Laczny, C.C., Lebrun, L.A., Bellora, C., Krishna, A., Wampach, L., Schneider, J.G., Hogan, A., de Beaufort, C., et al. (2016). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat. Microbiol. *2*, 16180.

Hess, M., Sczyrba, A., Egan, R., Kim, T.W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., et al. (2011). Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. Science *331*, 463–467.

Hooper, L.V., Wong, M.H., Thelin, A., Hansson, L., Falk, P.G., and Gordon, J.I. (2001). Molecular analysis of commensal host-microbial relationships in the intestine. Science *291*, 881–884.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. Nat. Microbiol. *1*, 16048.

Hug, L.A., Castelle, C.J., Wrighton, K.C., Thomas, B.C., Sharon, I., Frischkorn, K.R., Williams, K.H., Tringe, S.G., and Banfield, J.F. (2013). Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome *1*, 22.

Hugenholtz, P., Pitulle, C., Hershberger, K.L., and Pace, N.R. (1998). Novel division level bacterial diversity in a Yellowstone hot spring. J. Bacteriol. *180*, 366–376.

Huson, D.H., and Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol *61*, 1061–1067.

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

Iverson, V., Morris, R.M., Frazar, C.D., Berthiaume, C.T., Morales, R.L., and Armbrust, E.V. (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science *335*, 587–590.

Joshi, N. Sickle. Github.com.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res *40*, D109–D114.

Kantor, R.S., Wrighton, K.C., Handley, K.M., Sharon, I., Hug, L.A., Castelle, C.J., Thomas, B.C., and Banfield, J.F. (2013). Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. mBio *4*, e00708–e00713.

Karp, P.D., Paley, S., and Romero, P. (2002). The Pathway Tools software. Bioinformatics *18*, S225–S232.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.

Kelley, L.A., and Sternberg, M.J.E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protoc. *4*, 363–371.

Koenig, J.E., Spor, A., Scalfone, N., Fricker, A.D., Stombaugh, J., Knight, R., Angenent, L.T., and Ley, R.E. (2011). Succession of microbial consortia in the developing infant gut microbiome. Pnas *108 Suppl 1*, 4578–4585.

Konstantinidis, K.T., and Tiedje, J.M. (2005). Towards a Genome-Based Taxonomy for Prokaryotes. J. Bacteriol. *187*, 6258–6264.

Kopf, S.H., Sessions, A.L., Cowley, E.S., Reyes, C., Van Sambeek, L., Hu, Y., Orphan, V.J., Kato, R., and Newman, D.K. (2015). Trace incorporation of heavy water reveals slow and heterogeneous pathogen growth rates in cystic fibrosis sputum. Pnas *113*, E110–E116.

Korem, T., Zeevi, D., Suez, J., Weinberger, A., Avnit-Sagi, T., Pompan-Lotan, M., Matot, E., Jona, G., Harmelin, A., Cohen, N., et al. (2015). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science *349*, 1101–1106.

Lagkouvardos, I., Jehl, M.-A., Rattei, T., and Horn, M. (2014). Signature protein of the PVC superphylum. Appl. Environ. Microbiol. *80*, 440–445.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth *9*, 357–359.

Lathrop, S.K., Bloom, S.M., Rao, S.M., Nutsch, K., Lio, C.-W., Santacruz, N., Peterson, D.A., Stappenbeck, T.S., and Hsieh, C.-S. (2011). Peripheral education of the immune system by colonic commensal microbiota. Nature *478*, 250–254.

Lecompte, O. (2002). Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res *30*, 5382–5390.

Leslie, S.B., Teter, S.A., Crowe, L.M., and Crowe, J.H. (1994). Trehalose lowers membrane phase transitions in dry yeast cells. Biochim. Biophys. Acta *1192*, 7–13.

Lewis, A.L., and Lewis, W.G. (2012). Host sialoglycans and bacterial sialidases: a mucosal perspective. Cell Microbiol *14*, 1174–1182.

Ley, R.E., Bäckhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., and Gordon, J.I. (2005). Obesity alters gut microbial ecology. Pnas *102*, 11070–11075.

Ley, R.E., Peterson, D.A., and Gordon, J.I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell *124*, 837–848.

Lobry, J.R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. *13*, 660–665.

Loman, N.J., Constantinidou, C., Christner, M., Rohde, H., Chan, J.Z.-M., Quick, J., Weir, J.C., Quince, C., Smith, G.P., Betley, J.R., et al. (2013). A Culture-Independent Sequence-Based

Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia coli O104:H4Outbreak of Shiga-toxigenic Escherichia coli. Jama *309*, 1502–1510.

Long, P.E., Williams, K.H., Hubbard, S.S., and Banfield, J.F. (2016). Microbial Metagenomics Reveals Climate-Relevant Subsurface Biogeochemical Processes. Trends Microbiol.

Luef, B., Fakra, S.C., Csencsits, R., Wrighton, K.C., H, W.K., Wilkins, M.J., Downing, K.H., Long, P.E., Comolli, L.R., and Banfield, J.F. (2012). Iron-reducing bacteria accumulate ferric oxyhydroxide nanoparticle aggregates that may support planktonic growth. Isme J.

Luef, B., Frischkorn, K.R., Wrighton, K.C., Holman, H.-Y.N., Birarda, G., Thomas, B.C., Singh, A., Williams, K.H., Siegerist, C.E., Tringe, S.G., et al. (2015). Diverse, Uncultivated Ultra-Small Bacterial Cells in Groundwater. Nat. Commun. *6*, 6372.

Luo, F., Devine, C.E., and Edwards, E.A. (2016). Cultivating microbial dark matter in benzene‑degrading methanogenic consortia. Environ Microbiol.

Ma, Z.-Q., Dasari, S., Chambers, M.C., Litton, M.D., Sobecki, S.M., Zimmerman, L.J., Halvey, P.J., Schilling, B., Drake, P.M., Gibson, B.W., et al. (2009). IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. J. Proteome Res. *8*, 3872–3881.

Mai, V., Young, C.M., Ukhanova, M., Wang, X., Sun, Y., Casella, G., Theriaque, D., Li, N., Sharma, R., Hudak, M., et al. (2011). Fecal microbiota in premature infants prior to necrotizing enterocolitis. PLoS ONE *6*, e20647–e20647.

Maidak, B.L., Olsen, G.J., Larsen, N., Overbeek, R., McCaughey, M.J., and Woese, C.R. (1997). The RDP (Ribosomal Database Project). Nucleic Acids Res *25*, 109–111.

Marcy, Y., Ouverney, C., Bik, E.M., Lösekann, T., Ivanova, N., Martin, H.G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D.A., et al. (2007). Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Pnas *104*, 11889–11894.

Martinez-Espinosa, R.M., Dridge, E.J., Bonete, M.J., Butt, J.N., Butler, C.S., Sargent, F., and Richardson, D.J. (2007). Look on the positive side! The orientation, identification and bioenergetics of Archaeal membrane-bound nitrate reductases. Fems Microbiol Lett *276*, 129–139.

Maslowski, K.M., Vieira, A.T., Ng, A., Kranich, J., Sierro, F., Yu, D., Schilter, H.C., Rolph, M.S., Mackay, F., Artis, D., et al. (2009). Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. Nature *461*, 1282–1286.

McLean, J.S., Lombardo, M.-J., Badger, J.H., Edlund, A., Novotny, M., Yee-Greenbaum, J., Vyahhi, N., Hall, A.P., Yang, Y., Dupont, C.L., et al. (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. Pnas *110*, E2390–E2399.

Miller, C.S., Baker, B.J., Thomas, B.C., W, S.S., and Banfield, J.F. (2011). EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol *12*, R44.

Minoru Kanehisa, S.G. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res *28*, 27.

Morowitz, M.J., Denef, V.J., Costello, E.K., Thomas, B.C., Poroyko, V., Relman, D.A., and Banfield, J.F. (2011). Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. Pnas *108*, 1128–1133.

Morris, R.L., and Schmidt, T.M. (2013). Shallow breathing: bacterial life at low O2. Nat. Rev. Microbiol. *11*, 205–212.

Morrow, A.L., Lagomarcino, A.J., Schibler, K.R., and Taft, D.H. (2013). Early microbial and metabolomic signatures predict later onset of necrotizing enterocolitis in preterm infants. Microbiome *1*, 13.

Mshvildadze, M., Neu, J., Shuster, J., Theriaque, D., Li, N., and Mai, V. (2010). Intestinal Microbial Ecology in Premature Infants Assessed with Non–Culture-Based Techniques. J. Pediatr. *156*, 20–25.

Nawrocki, E.P. (2009). Structural RNA Homology Search and Alignment using Covariance Models. In Structural RNA Homology Search and Alignment Using Covariance Models, S.R. Eddy, M. Brent, J. Buhler, J. Fay, J.I. Gordon, R. Mitra, and G. Stormo, eds. (Washington University in Saint Louis, School of Medicine).

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: inference of RNA alignments. Bioinformatics *25*, 1335–1337.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

Nelson, W.C., and Stegen, J.C. (2015). The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. Frontiers in Microbiology *6*, 713–713.

Neu, J., and Walker, W.A. (2011). Necrotizing enterocolitis. N Engl J Med *364*, 255–264.

Nevskaya, N. (2005). Ribosomal protein L1 recognizes the same specific structural motif in its target sites on the autoregulatory mRNA and 23S rRNA. Nucleic Acids Res *33*, 478–485.

Newville, M., Stensitzki, T., Allen, D.B., and Ingargiola, A. (2014). LMFIT: non-linear least-square minimization and curve-fitting for Python (Zenodo).

Ng, K.M., Ferreyra, J.A., Higginbottom, S.K., Lynch, J.B., Kashyap, P.C., Gopinath, S., Naidu, N., Choudhury, B., Weimer, B.C., Monack, D.M., et al. (2013). Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. Nature *502*, 96–99.

Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Le Chatelier, E., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. *32*, 822–828.

Norman, A. bantools. Github.com.

Nowotny, V., and Nierhaus, K.H. (1982). Initiator proteins for the assembly of the 50S subunit from Escherichia coli ribosomes. Pnas *79*, 7238–7242.

Olm, M.R., Brown, C.T., Brooks, B., Firek, B.A., Baker, R., Burstein, D., Soenjoyo, K., Thomas, B.C., Morowitz, M.J., and Banfield, J.F. (2016). Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different *in situ* growth rates.

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol *17*, 1.

Paczia, N., Nilgen, A., Lehmann, T., Gätgens, J., Wiechert, W., and Noack, S. (2012). Extensive exometabolome analysis reveals extended overflow metabolism in various microorganisms. Microbial Cell Factories 2012 11:1 *11*, 1.

Palmer, C., Bik, E.M., DiGiulio, D.B., Relman, D.A., and Brown, P.O. (2007). Development of the human infant intestinal microbiota. PLoS Biol. *5*, e177.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, gr.186072.114–gr.186072.1055.

Pauli, G., and Overath, P. (1972). ato Operon: a Highly Inducible System for Acetoacetate and Butyrate Degradation in Escherichia coli. Eur J Biochem *29*, 553–562.

Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics *28*, 1420–1428.

Pereira, G.R., Baker, L., Egler, J., Corcoran, L., and Chiavacci, R. (1990). Serum myoinositol concentrations in premature infants fed human milk, formula for infants, and parenteral nutrition. Am. J. Clin. Nutr. *51*, 589–593.

Perisin, M., Vetter, M., Gilbert, J.A., and Bergelson, J. (2015). 16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies. Isme J *10*, 1020–1024.

Podar, M., Abulencia, C.B., Walcher, M., Hutchison, D., Zengler, K., Garcia, J.A., Holland, T., Cotton, D., Hauser, L., and Keller, M. (2007). Targeted access to the genomes of low-abundance organisms in complex microbial communities. Appl. Environ. Microbiol. *73*, 3205–3214.

Prescott, D.M., and Kuempel, P.L. (1972). Bidirectional replication of the chromosome in *Escherichia coli*. Pnas *69*, 2842–2845.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE *5*, e9490.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res *40*, D130–D135.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res *41*, D590–D596.

Raes, J., Korbel, J.O., Lercher, M.J., Mering, von, C., and Bork, P. (2007). Prediction of effective genome size in metagenomic samples. Genome Biol *8*, R10.

Raghavan, R., Hicks, L.D., and Minnick, M.F. (2008). Toxic Introns and Parasitic Intein in Coxiella burnetii: Legacies of a Promiscuous Past. J. Bacteriol. *190*, 5934–5943.

Raveh-Sadka, T., Firek, B., Sharon, I., Baker, R., Brown, C.T., Thomas, B.C., Morowitz, M.J., and Banfield, J.F. (2016). Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants. Isme J.

Raveh-Sadka, T., Thomas, B.C., Singh, A., Firek, B., Brooks, B., Castelle, C.J., Sharon, I., Baker, R., Good, M., Morowitz, M.J., et al. (2015). Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. Elife *4*, –.

Richter, M., and Rossello-Mora, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Pnas *106*, 19126–19131.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature *499*, 431–437.

Rissman, A.I., Mau, B., Biehl, B.S., Darling, A.E., Glasner, J.D., and Perna, N.T. (2009). Reordering contigs of draft genomes using the Mauve aligner. Bioinformatics *25*, 2071–2073.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rolfe, R.D.R., Hentges, D.J.D., Campbell, B.J.B., and Barrett, J.T.J. (1978). Factors related to the oxygen tolerance of anaerobic bacteria. Appl. Environ. Microbiol. *36*, 306–313.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome Biol *14*, 1.

Salman, V., Amann, R., Shub, D.A., and Schulz-Vogt, H.N. (2012). Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. Pnas *109*, 4203–4208.

Schuwirth, B.S. (2005). Structures of the Bacterial Ribosome at 3.5 A Resolution. Science *310*, 827–834.

Segata, N., and Huttenhower, C. GraPhlAn. Huttenhower.Sph.Harvard.Edu.

Seitz, K.W., Lazar, C.S., Hinrichs, K.-U., Teske, A.P., and Baker, B.J. (2016). Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. Isme J *10*, 1696–1705.

Sernova, N.V., and Gelfand, M.S. (2008). Identification of replication origins in prokaryotic genomes. Briefings in Bioinformatics *9*, 376–391.

Severi, E., Hood, D.W., and Thomas, G.H. (2007). Sialic acid utilization by bacterial pathogens. Microbiology *153*, 2817–2822.

Shajani, Z., Sykes, M.T., and Williamson, J.R. (2011). Assembly of Bacterial Ribosomes. Annu. Rev. Biochem. *80*, 501–526.

Sharon, I., Morowitz, M.J., Thomas, B.C., Costello, E.K., Relman, D.A., and Banfield, J.F. (2012). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. *23*, 111–120.

Sharon, I., Battchikova, N., Aro, E.-M., Giglione, C., Meinnel, T., Glaser, F., Pinter, R.Y., Breitbart, M., Rohwer, F., and Béjà, O. (2011). Comparative metagenomics of microbial traits within oceanic viral communities. Isme J *5*, 1178–1190.

Sharon, I., Kertesz, M., Hug, L.A., Pushkarev, D., Blauwkamp, T.A., Castelle, C.J., Amirebrahimi, M., Thomas, B.C., Burstein, D., Tringe, S.G., et al. (2015). Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res. *25*, 534–543.

Shimodaira, H. (2001). Multiple Comparisons of Log-Likelihoods and Combining Nonnested Models with Applications to Phylogenetic Tree Selection. Communications in Statistics - Theory and Methods *30*, 1751–1772.

Skovgaard, O., Bak, M., Løbner-Olesen, A., and Tommerup, N. (2011). Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. Genome Res. *21*, 1388–1393.

Soro, V., Dutton, L.C., Sprague, S.V., Nobbs, A.H., Ireland, A.J., Sandy, J.R., Jepson, M.A., Micaroni, M., Splatt, P.R., Dymock, D., et al. (2014). Axenic culture of a candidate division TM7 bacterium from the human oral cavity and biofilm interactions with other oral bacteria. Appl. Environ. Microbiol. *80*, 6480–6489.

Stamatakis, A. (2006). Phylogenetic models of rate heterogeneity: a high performance computing perspective. (IEEE).

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

Strøm, A.R., and Kaasen, I. (1993). Trehalose metabolism in Escherichia coli: stress protection and stress regulation of gene expression. Mol. Microbiol. *8*, 205–210.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics *23*, 1282–1288.

Tabb, D.L., Fernando, C.G., and Chambers, M.C. (2007). MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis. J. Proteome Res. *6*, 654–661.

Takahashi, Y., Yoshida, A., Nagata, E., Hoshino, T., Oho, T., Awano, S., Takehara, T., and Ansai, T. (2011). Streptococcus anginosus l-cysteine desulfhydrase gene expression is associated with abscess formation in BALB/c mice. Mol Oral Microbiol *26*, 221–227.

Tamura, K.K., Peterson, D.D., Peterson, N.N., Stecher, G.G., Nei, M.M., and Kumar, S.S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. *28*, 2731–2739.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes.

The NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., et al. (2009). The NIH Human Microbiome Project. Genome Res. *19*, 2317–2323.

The UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res *43*, D204–D212.

Traxler, M.F., Summers, S.M., Nguyen, H.-T., Zacharia, V.M., Hightower, G.A., Smith, J.T., and Conway, T. (2008). The global, ppGpp-mediated stringent response to amino acid starvation in Escherichia coli. Mol. Microbiol. *68*, 1128–1148.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature *444*, 1027–1131.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., and Banfield, J.F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature *428*, 37–43.

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

Vieira-Silva, S., and Rocha, E.P.C. (2010). The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet.

Vimr, E.R. (2013). Unified Theory of Bacterial Sialometabolism: How and Why Bacteria Metabolize Host Sialic Acids. Hindawi *2013*, 1–26.

Wake, R.G. (1972). Visualization of reinitiated chromosomes in *Bacillus subtilis*. J. Mol. Biol. *68*, 501–509.

Walters, W.A., Caporaso, J.G., Lauber, C.L., Berg-Lyons, D., Fierer, N., and Knight, R. (2011). PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. Bioinformatics *27*, 1159–1161.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998). Prokaryotes: the unseen majority. Pnas *95*, 6578–6583.

Whitman, W.B., Goodfellow, M., Kämpfer, P., Busse, H.-J., Trujillo, M.E., Suzuki, K.-I., and Ludwig, W. (2012). Bergey's Manual of Systematic Bacteriology: Volume 5: The Actinobacteria (New York, NY: Springer New York).

Williams, K.H., Long, P.E., Davis, J.A., Wilkins, M.J., N'Guessan, A.L., Steefel, C.I., Yang, L., Newcomer, D., Spane, F.A., Kerkhof, L.J., et al. (2011). Acetate Availability and its Influence on Sustainable Bioremediation of Uranium-Contaminated Groundwater. Geomicrobiol J *28*, 519–539.

Winter, S.E., Lopez, C.A., and Bäumler, A.J. (2013a). The dynamics of gut-associated microbial communities during inflammation. EMBO Rep *14*, 319–327.

Winter, S.E., Winter, M.G., Xavier, M.N., Thiennimitr, P., Poon, V., Keestra, A.M., Laughlin, R.C., Gomez, G., Wu, J., Lawhon, S.D., et al. (2013b). Host-derived nitrate boosts growth of E. coli in the inflamed gut. Science *339*, 708–711.

Wissenbach, U., Kröger, A., and Unden, G. (1990). The specific functions of menaquinone and demethylmenaquinone in anaerobic respiration with fumarate, dimethylsulfoxide, trimethylamine N-oxide and nitrate by Escherichia coli. Arch. Microbiol. *154*, 60–66.

Wittek, P., Gao, S.C., Lim, I.S., and Zhao, L. (2013). Somoclu: An Efficient Parallel Library for Self-Organizing Maps.

Woo, T.D.H.T., Oka, K.K., Takahashi, M.M., Hojo, F.F., Osaki, T.T., Hanawa, T.T., Kurata, S.S., Yonezawa, H.H., and Kamiya, S.S. (2011). Inhibition of the cytotoxic effect of *Clostridium difficile* in vitro by *Clostridium butyricum* MIYAIRI 588 strain. J Med Microbiol *60*, 1617–1625.

Wrighton, K.C., Castelle, C.J., Wilkins, M.J., Hug, L.A., Sharon, I., Thomas, B.C., Handley, K.M., Mullin, S.W., Nicora, C.D., Singh, A., et al. (2014). Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. Isme J *8*, 1452–1463.

Wrighton, K.C., Thomas, B.C., Sharon, I., Miller, C.S., Castelle, C.J., Verberkmoes, N.C., Wilkins, M.J., Hettich, R.L., Lipton, M.S., Williams, K.H., et al. (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science *337*, 1661–1665.

Wu, Y.W., Simmons, B.A., and Singer, S.W. (2015). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics.

Xavier, R.J., and Podolsky, D.K. (2007). Unravelling the pathogenesis of inflammatory bowel disease. Nature *448*, 427–434.

Xiong, W., Abraham, P.E., Li, Z., Pan, C., and Hettich, R.L. (2015). Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. Proteomics *15*, 3424–3438.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H., Whitman, W.B., Euzéby, J., Amann, R., and Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. *12*, 635–645.

Yelton, A.P., Thomas, B.C., Simmons, S.L., Wilmes, P., Zemla, A., Thelen, M.P., Justice, N., and Banfield, J.F. (2011). A semi-quantitative, synteny-based method to improve functional predictions for hypothetical and poorly annotated bacterial and archaeal genes. PLoS Comput Biol *7*, e1002230.

Yutin, N., Puigbò, P., Koonin, E.V., and Wolf, Y.I. (2012). PLOS ONE: Phylogenomics of Prokaryotic Ribosomal Proteins. PLoS ONE *7*, e36972.

Zaharia, M., Bolosky, W.J., Curtis, K., Fox, A., Patterson, D., Shenker, S., Stoica, I., Karp, R.M., and Sittler, T. (2011). Faster and More Accurate Sequence Alignment with SNAP.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847–848.

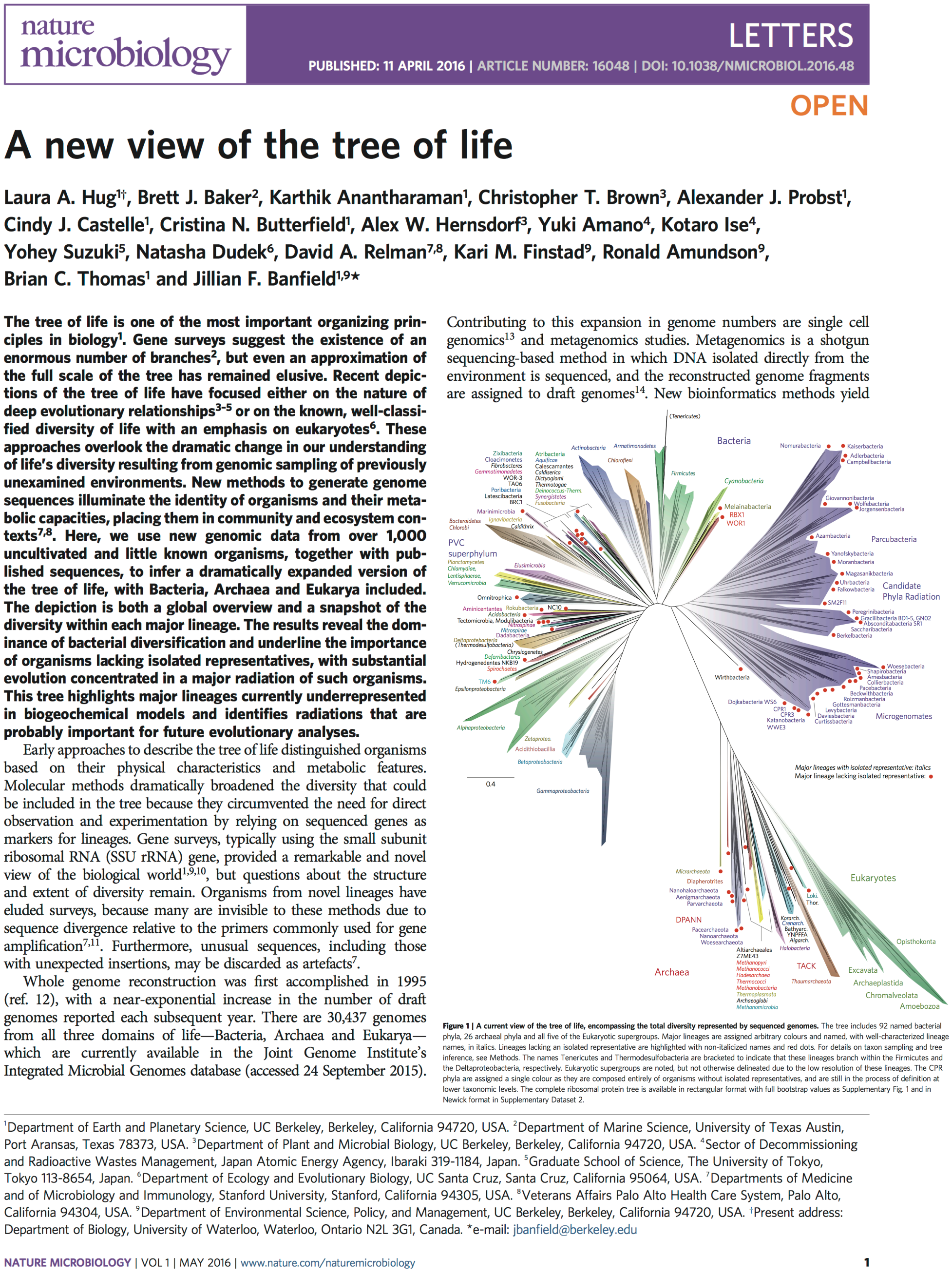
Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zoetendal, E.G., Heilig, H.G.H.J., Klaassens, E.S., Booijink, C.C.G.M., Kleerebezem, M., Smidt, H., and de Vos, W.M. (2006). Isolation of DNA from bacterial samples of the human gastrointestinal tract. Nat. Protoc. *1*, 870–873.

Appendices

Appendix 1: Select co-authored publications

Appendix 1.1 | A new view of the tree of life



nature microbiology

LETTERS

PUBLISHED: 11 APRIL 2016 | ARTICLE NUMBER: 16048 | DOI: 10.1038/NMICROBIOL.2016.48

OPEN

A new view of the tree of life

Laura A. Hug[1†], Brett J. Baker[2], Karthik Anantharaman[1], Christopher T. Brown[3], Alexander J. Probst[1], Cindy J. Castelle[1], Cristina N. Butterfield[1], Alex W. Hernsdorf[3], Yuki Amano[4], Kotaro Ise[4], Yohey Suzuki[5], Natasha Dudek[6], David A. Relman[7,8], Kari M. Finstad[9], Ronald Amundson[9], Brian C. Thomas[1] and Jillian F. Banfield[1,9*]

The tree of life is one of the most important organizing principles in biology[1]. Gene surveys suggest the existence of an enormous number of branches[2], but even an approximation of the full scale of the tree has remained elusive. Recent depictions of the tree of life have focused either on the nature of deep evolutionary relationships[3-5] or on the known, well-classified diversity of life with an emphasis on eukaryotes[6]. These approaches overlook the dramatic change in our understanding of life's diversity resulting from genomic sampling of previously unexamined environments. New methods to generate genome sequences illuminate the identity of organisms and their metabolic capacities, placing them in community and ecosystem contexts[7,8]. Here, we use new genomic data from over 1,000 uncultivated and little known organisms, together with published sequences, to infer a dramatically expanded version of the tree of life, with Bacteria, Archaea and Eukarya included. The depiction is both a global overview and a snapshot of the diversity within each major lineage. The results reveal the dominance of bacterial diversification and underline the importance of organisms lacking isolated representatives, with substantial evolution concentrated in a major radiation of such organisms. This tree highlights major lineages currently underrepresented in biogeochemical models and identifies radiations that are probably important for future evolutionary analyses.

Early approaches to describe the tree of life distinguished organisms based on their physical characteristics and metabolic features. Molecular methods dramatically broadened the diversity that could be included in the tree because they circumvented the need for direct observation and experimentation by relying on sequenced genes as markers for lineages. Gene surveys, typically using the small subunit ribosomal RNA (SSU rRNA) gene, provided a remarkable and novel view of the biological world[1,9,10], but questions about the structure and extent of diversity remain. Organisms from novel lineages have eluded surveys, because many are invisible to these methods due to sequence divergence relative to the primers commonly used for gene amplification[7,11]. Furthermore, unusual sequences, including those with unexpected insertions, may be discarded as artefacts[7].

Whole genome reconstruction was first accomplished in 1995 (ref. 12), with a near-exponential increase in the number of draft genomes reported each subsequent year. There are 30,437 genomes from all three domains of life—Bacteria, Archaea and Eukarya—which are currently available in the Joint Genome Institute's Integrated Microbial Genomes database (accessed 24 September 2015). Contributing to this expansion in genome numbers are single cell genomics[13] and metagenomics studies. Metagenomics is a shotgun sequencing-based method in which DNA isolated directly from the environment is sequenced, and the reconstructed genome fragments are assigned to draft genomes[14]. New bioinformatics methods yield

Figure 1 | A current view of the tree of life, encompassing the total diversity represented by sequenced genomes. The tree includes 92 named bacterial phyla, 26 archaeal phyla and all five of the Eukaryotic supergroups. Major lineages are assigned arbitrary colours and named, with well-characterized lineage names, in italics. Lineages lacking an isolated representative are highlighted with non-italicized names and red dots. For details on taxon sampling and tree inference, see Methods. The names Tenericutes and Thermodesulfobacteria are bracketed to indicate that these lineages branch within the Firmicutes and the Deltaproteobacteria, respectively. Eukaryotic supergroups are noted, but not otherwise delineated due to the low resolution of these lineages. The CPR phyla are assigned a single colour as they are composed entirely of organisms without isolated representatives, and are still in the process of definition at lower taxonomic levels. The complete ribosomal protein tree is available in rectangular format with full bootstrap values as Supplementary Fig. 1 and in Newick format in Supplementary Dataset 2.

[1]Department of Earth and Planetary Science, UC Berkeley, Berkeley, California 94720, USA. [2]Department of Marine Science, University of Texas Austin, Port Aransas, Texas 78373, USA. [3]Department of Plant and Microbial Biology, UC Berkeley, Berkeley, California 94720, USA. [4]Sector of Decommissioning and Radioactive Wastes Management, Japan Atomic Energy Agency, Ibaraki 319-1184, Japan. [5]Graduate School of Science, The University of Tokyo, Tokyo 113-8654, Japan. [6]Department of Ecology and Evolutionary Biology, UC Santa Cruz, Santa Cruz, California 95064, USA. [7]Departments of Medicine and of Microbiology and Immunology, Stanford University, Stanford, California 94305, USA. [8]Veterans Affairs Palo Alto Health Care System, Palo Alto, California 94304, USA. [9]Department of Environmental Science, Policy, and Management, UC Berkeley, Berkeley, California 94720, USA. †Present address: Department of Biology, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. *e-mail: jbanfield@berkeley.edu

Appendix 1.2 | Genomic expansion of domain Archaea highlights roles for organisms from new phyla in anaerobic carbon cycling

Current Biology *25*, 690–701, March 16, 2015 ©2015 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2015.01.014

Article

Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling

Cindy J. Castelle,[1] Kelly C. Wrighton,[2] Brian C. Thomas,[1] Laura A. Hug,[1] Christopher T. Brown,[3] Michael J. Wilkins,[2,4] Kyle R. Frischkorn,[5] Susannah G. Tringe,[6] Andrea Singh,[1] Lye Meng Markillie,[7] Ronald C. Taylor,[7] Kenneth H. Williams,[8] and Jillian F. Banfield[1,9,*]

[1]Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA 94720 USA
[2]Department of Microbiology, The Ohio State University, Columbus, OH 43210 USA
[3]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA 94720 USA
[4]School of Earth Sciences, The Ohio State University, Columbus, OH 43210 USA
[5]Department of Earth and Environmental Sciences and the Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York, NY 10964 USA
[6]Metagenome Program, DOE Joint Genome Institute, Walnut Creek, CA 94598 USA
[7]Environmental Molecular Sciences Laboratory, Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352 USA
[8]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[9]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA 94720 USA

Summary

Background: Archaea represent a significant fraction of Earth's biodiversity, yet they remain much less well understood than Bacteria. Gene surveys, a few metagenomic studies, and some single-cell sequencing projects have revealed numerous little-studied archaeal phyla. Certain lineages appear to branch deeply and may be part of a major phylum radiation. The structure of this radiation and the physiology of the organisms remain almost unknown.

Results: We used genome-resolved metagenomic analyses to investigate the diversity, genomes sizes, metabolic capacities, and potential roles of Archaea in terrestrial subsurface biogeochemical cycles. We sequenced DNA from complex sediment and planktonic consortia from an aquifer adjacent to the Colorado River (USA) and reconstructed the first complete genomes for Archaea using cultivation-independent methods. To provide taxonomic context, we analyzed an additional 151 newly sampled archaeal sequences. We resolved two new phyla within a major, apparently deep-branching group of phyla (a superphylum). The organisms have small genomes, and metabolic predictions indicate that their primary contributions to Earth's biogeochemical cycles involve carbon and hydrogen metabolism, probably associated with symbiotic and/or fermentation-based lifestyles.

Conclusions: The results dramatically expand genomic sampling of the domain Archaea and clarify taxonomic designations within a major superphylum. This study, in combination with recently published work on bacterial phyla lacking cultivated representatives, reveals a fascinating phenomenon of major radiations of organisms with small genomes, novel proteome composition, and strong interdependence in both domains.

Introduction

Archaea constitute a considerable fraction of the microbial biomass on Earth. The earliest archaeal phylogenetic trees contained only cultured Archaea (hyperthermophiles, halophiles, and methanogens) and included just two phyla, Crenarchaeota and Euryarchaeota [1]. New genomes and 16S rRNA gene sequences have dramatically expanded and reshaped the archaeal tree, and several new phylum-level lineages and two superphyla have been proposed. First, a new archaeal branch [2], now represented by one genome [3], was recognized as the phylum Korarchaeota. Second, Archaea [4] that belong to the recently defined Thaumarchaeota phylum were isolated [5]. This phylum comprises all known ammonia-oxidizing Archaea (AOA). Third, recent metagenomic analyses provided the genome of uncultivated crenarchaeote Candidatus (Ca.) "*Caldiarchaeum subterraneum*" that has been proposed to constitute a novel phylum, Aigarchaeota [6]. Together, the Thaumarchaeota, Aigarchaeota, Crenarchaeaota, and Korarchaeota were proposed to constitute an archaeal superphylum, referred to as TACK [7]. Fourth, a new phylum (suggested name Bathyarchaeota; [8]) was proposed

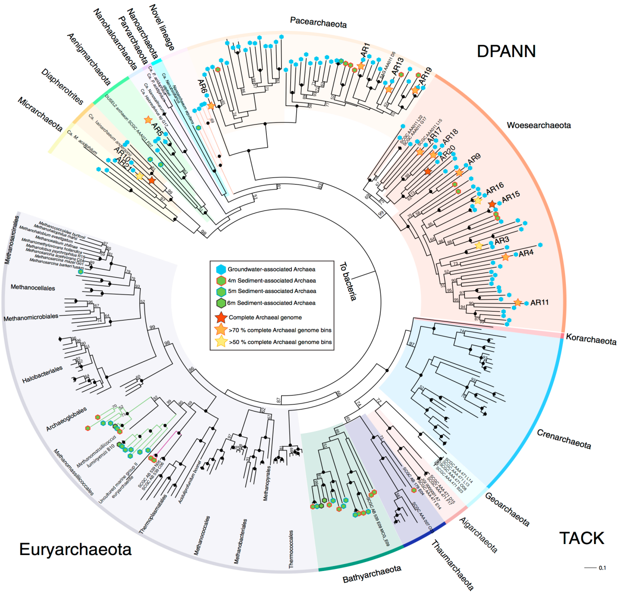
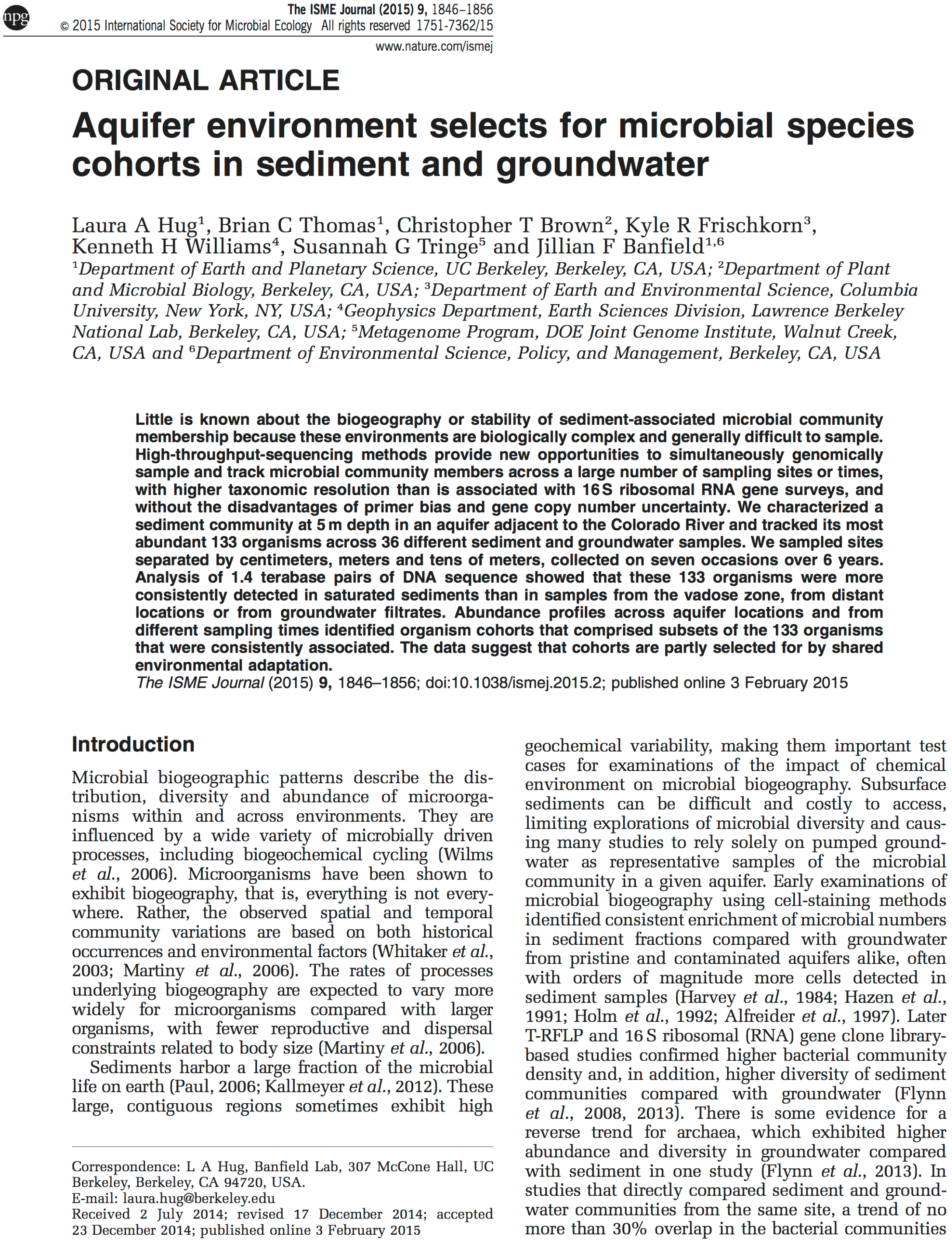


Figure 2. Phylogenetic Analyses Placing the 153 Genomically Sampled Subsurface Archaea

Maximum-likelihood phylogeny of the TACK and DPANN superphyla and Euryarchaeota phylum based on a 15-ribosomal-protein concatenated alignment. Black dots on nodes denote bootstrap support of 100%, while other support values > 50% are reported directly. Phyla are colored and named for clarity. Lineages reported in this study are marked with hexagons: blue for groundwater organisms and green for sediment-associated Archaea, with colored borders on green hexagons indicating the sediment depth from which the organism derives. Genome bins are further marked with stars, with the star color indicating genome completeness (see legend). See also Figure S1 for 16S rRNA analysis that clearly supports the definition of the two newly identified DPANN phyla in this study.

*Correspondence: jbanfield@berkeley.edu

Appendix 1.3 | Aquifer environment selects for microbial species cohorts in sediment and groundwater



The ISME Journal (2015) 9, 1846–1856
© 2015 International Society for Microbial Ecology All rights reserved 1751-7362/15
www.nature.com/ismej

ORIGINAL ARTICLE

Aquifer environment selects for microbial species cohorts in sediment and groundwater

Laura A Hug[1], Brian C Thomas[1], Christopher T Brown[2], Kyle R Frischkorn[3], Kenneth H Williams[4], Susannah G Tringe[5] and Jillian F Banfield[1,6]
[1]*Department of Earth and Planetary Science, UC Berkeley, Berkeley, CA, USA;* [2]*Department of Plant and Microbial Biology, Berkeley, CA, USA;* [3]*Department of Earth and Environmental Science, Columbia University, New York, NY, USA;* [4]*Geophysics Department, Earth Sciences Division, Lawrence Berkeley National Lab, Berkeley, CA, USA;* [5]*Metagenome Program, DOE Joint Genome Institute, Walnut Creek, CA, USA and* [6]*Department of Environmental Science, Policy, and Management, Berkeley, CA, USA*

Little is known about the biogeography or stability of sediment-associated microbial community membership because these environments are biologically complex and generally difficult to sample. High-throughput-sequencing methods provide new opportunities to simultaneously genomically sample and track microbial community members across a large number of sampling sites or times, with higher taxonomic resolution than is associated with 16 S ribosomal RNA gene surveys, and without the disadvantages of primer bias and gene copy number uncertainty. We characterized a sediment community at 5 m depth in an aquifer adjacent to the Colorado River and tracked its most abundant 133 organisms across 36 different sediment and groundwater samples. We sampled sites separated by centimeters, meters and tens of meters, collected on seven occasions over 6 years. Analysis of 1.4 terabase pairs of DNA sequence showed that these 133 organisms were more consistently detected in saturated sediments than in samples from the vadose zone, from distant locations or from groundwater filtrates. Abundance profiles across aquifer locations and from different sampling times identified organism cohorts that comprised subsets of the 133 organisms that were consistently associated. The data suggest that cohorts are partly selected for by shared environmental adaptation.

The ISME Journal (2015) **9,** 1846–1856; doi:10.1038/ismej.2015.2; published online 3 February 2015

Introduction

Microbial biogeographic patterns describe the distribution, diversity and abundance of microorganisms within and across environments. They are influenced by a wide variety of microbially driven processes, including biogeochemical cycling (Wilms *et al.*, 2006). Microorganisms have been shown to exhibit biogeography, that is, everything is not everywhere. Rather, the observed spatial and temporal community variations are based on both historical occurrences and environmental factors (Whitaker *et al.*, 2003; Martiny *et al.*, 2006). The rates of processes underlying biogeography are expected to vary more widely for microorganisms compared with larger organisms, with fewer reproductive and dispersal constraints related to body size (Martiny *et al.*, 2006).

Sediments harbor a large fraction of the microbial life on earth (Paul, 2006; Kallmeyer *et al.*, 2012). These large, contiguous regions sometimes exhibit high geochemical variability, making them important test cases for examinations of the impact of chemical environment on microbial biogeography. Subsurface sediments can be difficult and costly to access, limiting explorations of microbial diversity and causing many studies to rely solely on pumped groundwater as representative samples of the microbial community in a given aquifer. Early examinations of microbial biogeography using cell-staining methods identified consistent enrichment of microbial numbers in sediment fractions compared with groundwater from pristine and contaminated aquifers alike, often with orders of magnitude more cells detected in sediment samples (Harvey *et al.*, 1984; Hazen *et al.*, 1991; Holm *et al.*, 1992; Alfreider *et al.*, 1997). Later T-RFLP and 16 S ribosomal (RNA) gene clone library-based studies confirmed higher bacterial community density and, in addition, higher diversity of sediment communities compared with groundwater (Flynn *et al.*, 2008, 2013). There is some evidence for a reverse trend for archaea, which exhibited higher abundance and diversity in groundwater compared with sediment in one study (Flynn *et al.*, 2013). In studies that directly compared sediment and groundwater communities from the same site, a trend of no more than 30% overlap in the bacterial communities

Correspondence: L A Hug, Banfield Lab, 307 McCone Hall, UC Berkeley, Berkeley, CA 94720, USA.
E-mail: laura.hug@berkeley.edu

Received 2 July 2014; revised 17 December 2014; accepted 23 December 2014; published online 3 February 2015

Appendix 1.4 | Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages

environmental microbiology

Environmental Microbiology (2016) **18**(1), 159–173 doi:10.1111/1462-2920.12930

Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages

Laura A. Hug,[1]* Brian C. Thomas,[1] Itai Sharon,[1] Christopher T. Brown,[2] Ritin Sharma,[3†] Robert L. Hettich,[3] Michael J. Wilkins,[4,5] Kenneth H. Williams,[6] Andrea Singh[1] and Jillian F. Banfield[1,7]

Departments of [1]Earth and Planetary Science, [2]Plant and Microbial Biology and
[7]Environmental Science, Policy, and Management, UC Berkeley, Berkeley, CA, USA.
[3]Oak Ridge National Laboratory, Oak Ridge, TN, USA.
[4]Department of Microbiology and
[5]School of Earth Sciences, Ohio State University, Columbus, OH, USA.
[6]Department of Geophysics, Division of Earth Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

Summary

Nitrogen, sulfur and carbon fluxes in the terrestrial subsurface are determined by the intersecting activities of microbial community members, yet the organisms responsible are largely unknown. Metagenomic methods can identify organisms and functions, but genome recovery is often precluded by data complexity. To address this limitation, we developed subsampling assembly methods to re-construct high-quality draft genomes from complex samples. We applied these methods to evaluate the interlinked roles of the most abundant organisms in biogeochemical cycling in the aquifer sediment. Community proteomics confirmed these activities. The eight most abundant organisms belong to novel lineages, and two represent phyla with no previously sequenced genome. Four organisms are predicted to fix carbon via the Calvin–Benson–Bassham, Wood–Ljungdahl or 3-hydroxyproprionate/4-hydroxybutarate pathways. The profiled organisms are involved in the network of denitrification, dissimilatory nitrate reduction to ammonia, ammonia oxidation and sulfate reduction/oxidation, and require substrates supplied by other community members. An ammonium-oxidizing Thaumarchaeote is the most abundant community member, despite low ammonium concentrations in the groundwater. This organism likely benefits from two other relatively abundant organisms capable of producing ammonium from nitrate, which is abundant in the groundwater. Overall, dominant members of the microbial community are interconnected through exchange of geochemical resources.

Introduction

Microbial metabolism is critical to the creation, maintenance and turnover of carbon and nitrogen sinks in the subsurface. The high proportion of uncultured and little-studied organisms present in subsurface environments (Wrighton *et al.*, 2012; Castelle *et al.*, 2013), and the continued discovery of new microbial lineages implicated in major geochemical cycles (Green *et al.*, 2010; Rasigraf *et al.*, 2014) indicates that there is substantial metabolic diversity yet to be discovered within terrestrial sediments. Understanding and modelling geochemical cycling thus requires identification of, and metabolic prediction for, the microbial networks catalysing carbon, nitrogen and sulfur cycling in the subsurface.

The complexity of sediment microbial communities had prevented substantial metagenomic assembly until recently; with advances in sequencing technologies and bioinformatic methods, insight into the structure and function of diverse, low abundance sediment communities is now tractable through assembly and genome curation (Castelle *et al.*, 2013; Kantor *et al.*, 2013). Metagenomics provides simultaneous taxonomic identification and metabolic profiling for the community, or, with binning and genome curation, for specific organisms from the environment. Draft genomes from sediment-associated organisms have allowed definition of new radiations on the tree of life (Ettwig *et al.*, 2009; Wrighton *et al.*, 2012; Castelle *et al.*, 2013; Rinke *et al.*, 2013), and prediction of previously unknown roles in biogeochemical cycles, including carbon fixation in *Chloroflexi* (Hug *et al.*, 2013) and an

Received 22 April, 2015; revised 26 May, 2015; accepted 27 May, 2015. *For correspondence. E-mail laura.audrey.hug@gmail.com; Tel. +510 643 2155; Fax +510 643 9980. †Current affiliation: Molecular Oncology, Moffitt Cancer Center and Research Institute, Tampa, FL, USA.

Appendix 1.5 | Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems

nature COMMUNICATIONS

ARTICLE

Received 30 Jul 2015 | Accepted 4 Jan 2016 | Published 3 Feb 2016

DOI: 10.1038/ncomms10613

OPEN

Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems

David Burstein[1], Christine L. Sun[1,2], Christopher T. Brown[3], Itai Sharon[1], Karthik Anantharaman[1], Alexander J. Probst[1], Brian C. Thomas[1] & Jillian F. Banfield[1,4]

Current understanding of microorganism–virus interactions, which shape the evolution and functioning of Earth's ecosystems, is based primarily on cultivated organisms. Here we investigate thousands of viral and microbial genomes recovered using a cultivation-independent approach to study the frequency, variety and taxonomic distribution of viral defence mechanisms. CRISPR-Cas systems that confer microorganisms with immunity to viruses are present in only 10% of 1,724 sampled microorganisms, compared with previous reports of 40% occurrence in bacteria and 81% in archaea. We attribute this large difference to the lack of CRISPR-Cas systems across major bacterial lineages that have no cultivated representatives. We correlate absence of CRISPR-Cas with lack of nucleotide biosynthesis capacity and a symbiotic lifestyle. Restriction systems are well represented in these lineages and might provide both non-specific viral defence and access to nucleotides.

[1] Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, California 94720, USA. [2] Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, California 94305, USA. [3] Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, California 94720, USA. [4] Department of Environmental Science, Policy and Management, University of California, Berkeley, California 94720, USA. Correspondence and requests for materials should be addressed to J.F.B. (email: jbanfield@berkeley.edu).

Appendix 1.6 | Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum

Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum

Karthik Anantharaman[1], Christopher T. Brown[2], David Burstein[1], Cindy J. Castelle[1], Alexander J. Probst[1], Brian C. Thomas[1], Kenneth H. Williams[3] and Jillian F. Banfield[1,3]

[1] Department of Earth and Planetary Sciences, University of California, Berkeley, California, United States

[2] Department of Plant and Microbial Biology, University of California, Berkeley, California, United States

[3] Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States

ABSTRACT

Five closely related populations of bacteria from the Candidate Phylum (CP) Peregrinibacteria, part of the bacterial Candidate Phyla Radiation (CPR), were sampled from filtered groundwater obtained from an aquifer adjacent to the Colorado River near the town of Rifle, CO, USA. Here, we present the first complete genome sequences for organisms from this phylum. These bacteria have small genomes and, unlike most organisms from other lineages in the CPR, have the capacity for nucleotide synthesis. They invest significantly in biosynthesis of cell wall and cell envelope components, including peptidoglycan, isoprenoids via the mevalonate pathway, and a variety of amino sugars including perosamine and rhamnose. The genomes encode an intriguing set of large extracellular proteins, some of which are very cysteine-rich and may function in attachment, possibly to other cells. Strain variation in these proteins is an important source of genotypic variety. Overall, the cell envelope features, combined with the lack of biosynthesis capacities for many required cofactors, fatty acids, and most amino acids point to a symbiotic lifestyle. Phylogenetic analyses indicate that these bacteria likely represent a new class within the Peregrinibacteria phylum, although they ultimately may be recognized as members of a separate phylum. We propose the provisional taxonomic assignment as '*Candidatus* Peribacter riflensis', Genus Peribacter, Family Peribacteraceae, Order Peribacterales, Class Peribacteria in the phylum Peregrinibacteria.

Submitted 19 October 2015
Accepted 28 December 2015
Published 28 January 2016

Corresponding author
Jillian F. Banfield,
jbanfield@berkeley.edu

Academic editor
David Baltrus

Additional Information and Declarations can be found on page 11

DOI **10.7717/peerj.1607**

Subjects Bioinformatics, Environmental sciences, Genomics, Microbiology
Keywords Peregrinibacteria, Candidate phyla radiation, Metagenomics, Complete genomes, Strain variation

INTRODUCTION

Metagenomic analyses of microbial communities in natural environments have revealed numerous previously unrecognized microorganisms, including those now described to be

How to cite this article Anantharaman et al. (2016), Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum. **PeerJ 4:e1607; DOI 10.7717/peerj.1607**

Appendix 1.7 | RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria

The ISME Journal (2016) 10, 2702–2714
© 2016 International Society for Microbial Ecology All rights reserved 1751-7362/16
www.nature.com/ismej

OPEN

ORIGINAL ARTICLE

RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria

Kelly C Wrighton[1,7], Cindy J Castelle[2,7], Vanessa A Varaljay[1], Sriram Satagopan[1], Christopher T Brown[3], Michael J Wilkins[1,4], Brian C Thomas[2], Itai Sharon[2], Kenneth H Williams[5], F Robert Tabita[1] and Jillian F Banfield[2,6]

[1]Department of Microbiology, The Ohio State University, Columbus, OH, USA; [2]Department of Earth and Planetary Science, UC Berkeley, Berkeley, CA, USA; [3]Department of Plant and Microbial Biology, UC Berkeley, Berkeley, CA, USA; [4]School of Earth Sciences, The Ohio State University, Columbus, OH, USA; [5]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA and [6]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA

Metagenomic studies recently uncovered form II/III RubisCO genes, originally thought to only occur in archaea, from uncultivated bacteria of the candidate phyla radiation (CPR). There are no isolated CPR bacteria and these organisms are predicted to have limited metabolic capacities. Here we expand the known diversity of RubisCO from CPR lineages. We report a form of RubisCO, distantly similar to the archaeal form III RubisCO, in some CPR bacteria from the Parcubacteria (OD1), WS6 and Microgenomates (OP11) phyla. In addition, we significantly expand the Peregrinibacteria (PER) II/III RubisCO diversity and report the first II/III RubisCO sequences from the Microgenomates and WS6 phyla. To provide a metabolic context for these RubisCOs, we reconstructed near-complete (>93%) PER genomes and the first closed genome for a WS6 bacterium, for which we propose the phylum name Dojkabacteria. Genomic and bioinformatic analyses suggest that the CPR RubisCOs function in a nucleoside pathway similar to that proposed in Archaea. Detection of form II/III RubisCO and nucleoside metabolism gene transcripts from a PER supports the operation of this pathway *in situ*. We demonstrate that the PER form II/III RubisCO is catalytically active, fixing CO_2 to physiologically complement phototrophic growth in a bacterial photoautotrophic RubisCO deletion strain. We propose that the identification of these RubisCOs across a radiation of obligately fermentative, small-celled organisms hints at a widespread, simple metabolic platform in which ribose may be a prominent currency.

The ISME Journal (2016) **10,** 2702–2714; doi:10.1038/ismej.2016.53; published online 3 May 2016

Introduction

The vast majority of the organisms in the environment have not been cultivated, obscuring our knowledge of their physiology. Recent metagenomic investigations have revealed the presence of a large diversity of uncultivated organisms in marine and terrestrial subsurface environments (Castelle *et al*., 2013, 2015; Lloyd *et al*., 2013; Baker *et al*., 2015). For example, in a metagenomic study of a shallow alluvial aquifer, we determined that many of the uncultivated bacteria were associated with the candidate phyla radiation (CPR), a monophyletic group of at least 35 phyla that accounts for >15% of all bacterial diversity (Brown *et al*., 2015). Metagenomic sampling of almost 800 CPR bacteria suggested that members of this radiation consistently have small genomes with many metabolic limitations, including lack of an electron transport chain, no more than a partial tricarboxylic acid cycle and mostly incomplete nucleotide and amino-acid biosynthesis pathways. Using metabolic predictions from complete and near-complete genomes (Wrighton *et al*., 2012), we inferred that members of the CPR are fermenters whose primary biogeochemical impact is primarily on subsurface organic carbon and hydrogen cycling (Wrighton *et al*., 2014). However, many of the genes in these genomes remain poorly annotated and the metabolic platform of CPR bacteria remains uncertain. Further, no transcription and enzyme activities have been validated, hindering knowledge of the actual reactions catalyzed by these organisms.

Correspondence: JF Banfield, Department of Earth and Planetary Sciences, and Department of Environmental Science, Policy, and Man, University of California at Berkeley, 336 Hilgard Hall, Berkeley, CA 94720, USA.
E-mail: jbanfield@berkeley.edu

[7]These authors contributed equally to this work.

Received 25 August 2015; revised 29 February 2016; accepted 4 March 2016; published online 3 May 2016

Appendix 1.8 | Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system

nature COMMUNICATIONS

ARTICLE

Received 31 Mar 2016 | Accepted 13 Sep 2016 | Published 24 Oct 2016

DOI: 10.1038/ncomms13219 OPEN

Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system

Karthik Anantharaman[1], Christopher T. Brown[2], Laura A. Hug[1], Itai Sharon[1], Cindy J. Castelle[1], Alexander J. Probst[1], Brian C. Thomas[1], Andrea Singh[1], Michael J. Wilkins[3], Ulas Karaoz[4], Eoin L. Brodie[4], Kenneth H. Williams[4], Susan S. Hubbard[4] & Jillian F. Banfield[1,4]

The subterranean world hosts up to one-fifth of all biomass, including microbial communities that drive transformations central to Earth's biogeochemical cycles. However, little is known about how complex microbial communities in such environments are structured, and how inter-organism interactions shape ecosystem function. Here we apply terabase-scale cultivation-independent metagenomics to aquifer sediments and groundwater, and reconstruct 2,540 draft-quality, near-complete and complete strain-resolved genomes that represent the majority of known bacterial phyla as well as 47 newly discovered phylum-level lineages. Metabolic analyses spanning this vast phylogenetic diversity and representing up to 36% of organisms detected in the system are used to document the distribution of pathways in coexisting organisms. Consistent with prior findings indicating metabolic handoffs in simple consortia, we find that few organisms within the community can conduct multiple sequential redox transformations. As environmental conditions change, different assemblages of organisms are selected for, altering linkages among the major biogeochemical cycles.

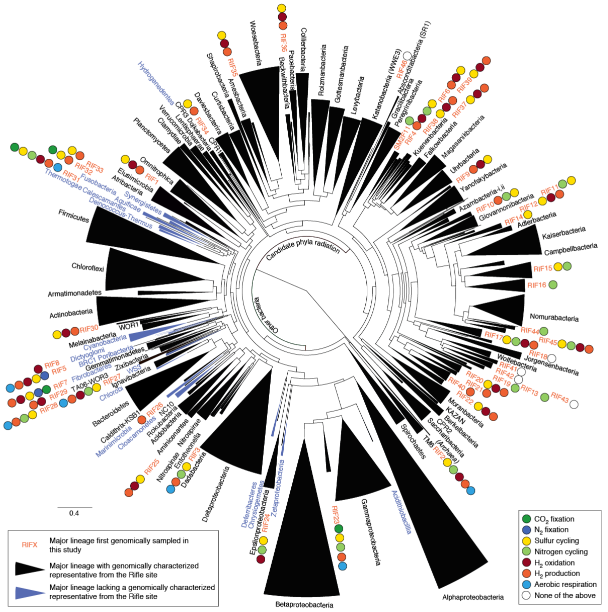


Figure 2 | Phylogeny of bacterial genomes inferred by maximum likelihood. The phylogenetic tree is based on 16 concatenated RPs and was collapsed at the phylum level. Colours of the wedges indicate the following: black: phylum-level lineage identified at Rifle; blue: phylum-level lineage not identified at Rifle. Coloured circles describe important biogeochemical roles inferred for newly described phylum-level lineages. Proposed names for newly described phylum-level lineages (RIF1-RIF46 and SM2F11) are detailed in Table 1. The phylogenetic inference configurations with detailed branch support values are provided in Supplementary Fig. 2 and Supplementary Data 12.

[1]Department of Earth and Planetary Science, University of California, Berkeley, California 94720, USA.[2] Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA. [3] School of Earth Sciences and Department of Microbiology, The Ohio State University, Columbus, Ohio 43210, USA. [4] Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. Correspondence and requests for materials should be addressed to J.F.B. (email: jbanfield@berkeley.edu).

Appendix 1.9 | Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants

The ISME Journal (2016) 10, 2817–2830
© 2016 International Society for Microbial Ecology All rights reserved 1751-7362/16
www.nature.com/ismej

ORIGINAL ARTICLE

Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants

Tali Raveh-Sadka[1], Brian Firek[2], Itai Sharon[1], Robyn Baker[3], Christopher T Brown[1], Brian C Thomas[1], Michael J Morowitz[2] and Jillian F Banfield[1]
[1]*Department of Earth and Planetary Science, UC Berkeley, Berkeley, CA, USA;* [2]*Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA and* [3]*Department of Pediatrics, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA*

The potentially critical stage of initial gut colonization in premature infants occurs in the hospital environment, where infants are exposed to a variety of hospital-associated bacteria. Because few studies of microbial communities are strain-resolved, we know little about the extent to which specific strains persist in the hospital environment and disperse among infants. To study this, we compared 304 near-complete genomes reconstructed from fecal samples of 21 infants hospitalized in the same intensive care unit in two cohorts, over 3 years apart. The genomes represent 159 distinct bacterial strains, only 14 of which occurred in multiple infants. *Enterococcus faecalis* and *Staphylococcus epidermidis,* common infant gut colonists, exhibit diversity comparable to that of reference strains, inline with introduction of strains from infant-specific sources rather than a hospital strain pool. Unlike other infants, a pair of sibling infants shared multiple strains, even after extensive antibiotic administration, suggesting overlapping strain-sources and/or genetic selection drive microbiota similarities. Interestingly, however, five strains were detected in infants hospitalized three years apart. Three of these were also detected in multiple infants in the same year. This finding of a few widely dispersed and persistent bacterial colonizers despite overall low potential for strain dispersal among infants has implications for understanding and directing healthy colonization.
The ISME Journal (2016) **10,** 2817–2830; doi:10.1038/ismej.2016.83; published online 3 June 2016

Introduction

Gut microbes have important roles in health and disease. Colonization of the gut may begin *in utero* (Funkhouser and Bordenstein, 2013; Moles *et al*., 2013; Aagaard *et al.,* 2014), but progresses rapidly after birth. According to 2012 data from the Center for Disease Control (Macdorman *et al*., 2013), 99% of US infants are born in hospitals, suggesting that very early stages of gut colonization occur in the hospital environment. We know very little about the extent to which hospital-associated microbial communities can influence gut colonization. However, given that many pathogens are found in hospital settings, it is likely that acquisition of hospital-derived organisms can lead to aberrant colonization and long-term detrimental health effects (Arrieta *et al*., 2014). This is of particular importance for premature infants that are hospitalized for long periods, especially given frequent antibiotic administration at birth, their immature gastrointestinal tracts and the fact that their immune system is more vulnerable than that of term infants (Groer *et al*., 2014). Hospitalized children and adults, especially those with compromised immune systems, are also vulnerable to acquisition of hospital-associated pathogens, particularly those that persist in the room environment.

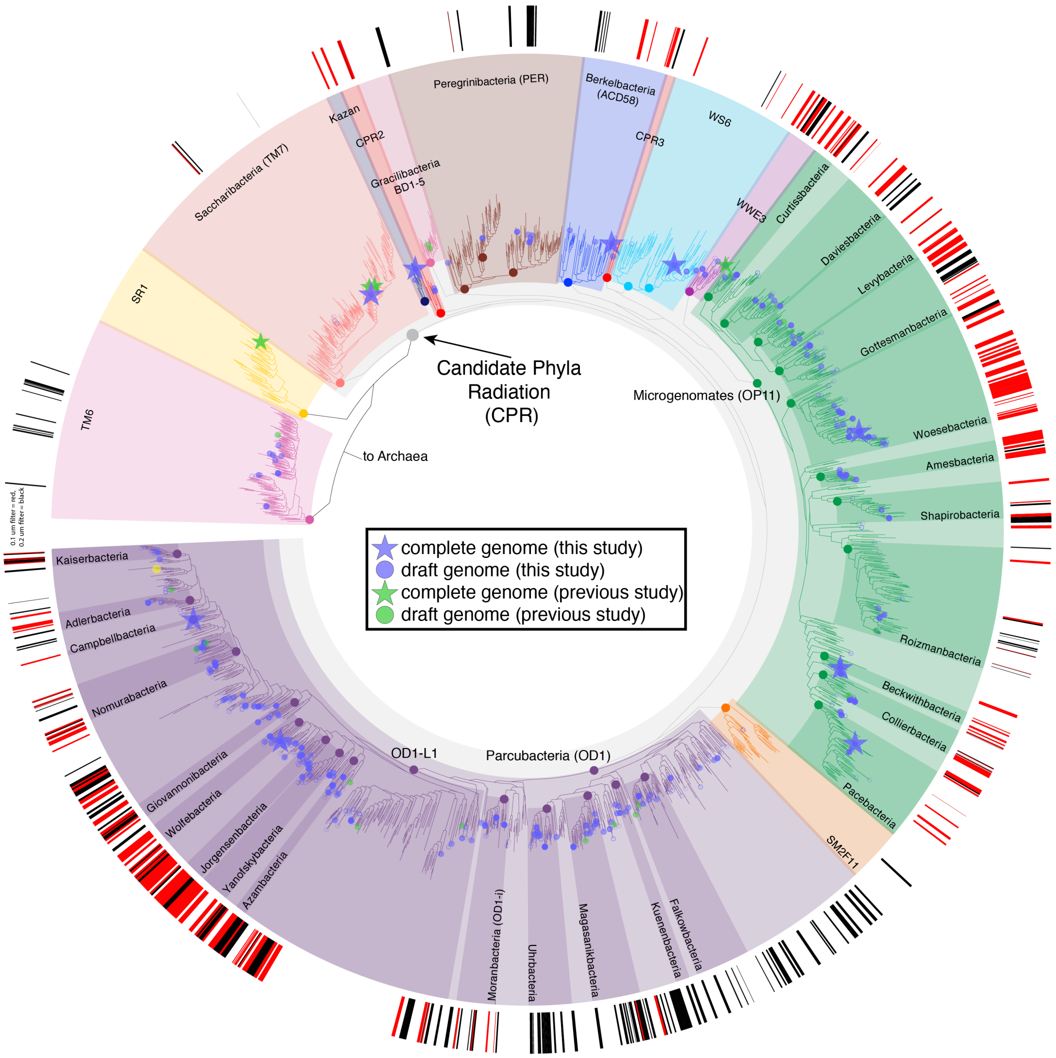
Some studies of hospital outbreaks have focused on a small number of specific strains of interest, typically bacterial pathogens (Chin *et al*., 2010; Köser *et al*., 2012; Snitkin *et al*., 2012; He *et al*., 2013; Loman *et al*., 2013), and tracked them among hospitalized individuals. However, an outbreak is a very specific phenomenon that involves both spread and infection by a single, potentially highly virulent organism. Consequently, lessons from outbreak studies provide only limited insight into the behavior of the larger consortia of hospital-associated bacteria that may be relevant for initial gut colonization. Numerous recent studies followed microbial communities colonizing hospitalized neonates over time (Morowitz *et al*., 2011; Brown *et al*., 2013; Costello *et al*., 2013; Sharon *et al*., 2013;

Correspondence: JF Banfield, Department of Earth and Planetary Sciences, and Department of Environmental Science, Policy, and Man, 369 McCone Hall, UC Berkeley, Berkeley CA 94720, USA.
E-mail: jbanfield@berkeley.edu
Received 18 October 2015; revised 6 March 2016; accepted 12 April 2016; published online 3 June 2016

Appendix 2: Candidate Phyla Radiation (CPR) lineages associated with small-cell groundwater filtrates

Ultra-small cell size has been demonstrated for some Candidate Phyla Radiation (CPR) bacteria (Luef et al., 2015); however, due to the diversity of the CPR, it is likely that some may have larger cells. In order to determine whether or not specific CPR lineages are more likely to have ultra-small cells, the abundance of each CPR organism on both 0.1 and 0.2 µm filters from the 2011 Rifle, CO groundwater study was compared. Results indicate that the Microgenomates (OP11), Parcubacteria that are missing ribosomal protein L1 (OD1-L1), and Katanobacteria (WWE3) are more likely to be detected in small-cell filtrates, while the Peregrinibacteria (PER) and other Parcubacteria (OD1) are more often found on the 0.2 µm filter, and thus likely have larger cells (**Appendix 2 Figure 1**).

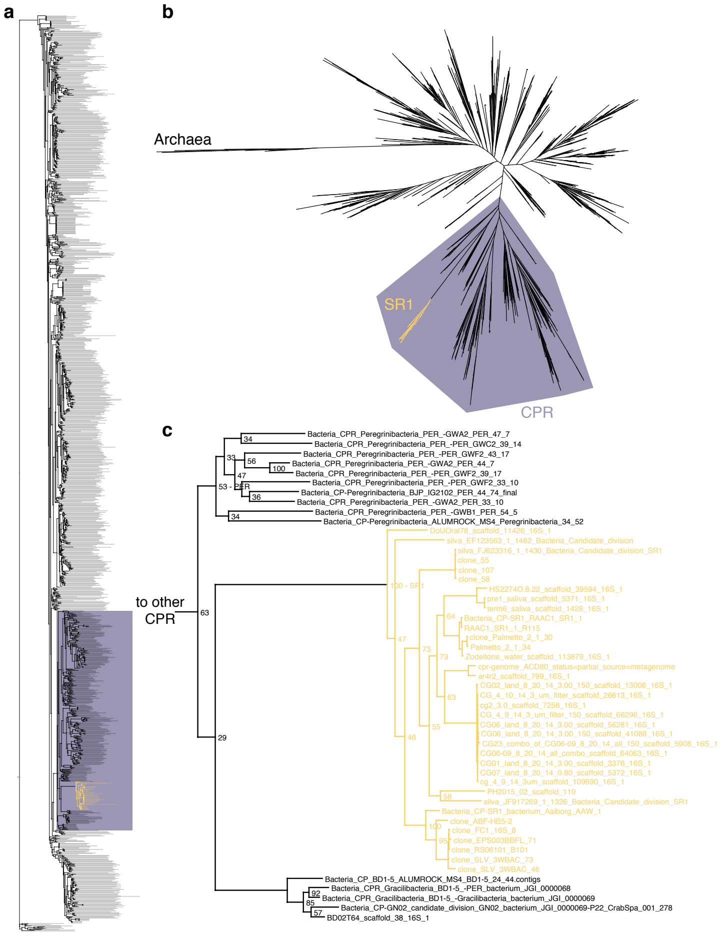
Appendix 2 Figure 1 | Candidate Phyla Radiation (CPR) bacteria associated with different groundwater filters. Subset of a maximum-likelihood 16S rRNA gene phylogeny (**Figure 2.1 and Supplementary File 2.2**) showing the CPR, and whether organisms from each lineage were more abundant on 0.2 (black) or 0.1 (red) µm filters.



Appendix 3: Absconditabacteria (SR1) phylogeny

The 16S rRNA gene sequence phylogeny presented in **Chapter 2** (**Figure 2.1 and Supplementary File 2.2**) placed the Absconditabacteria (SR1) outside of the Candidate Phyla Radiation (CPR), in contrast to placement within the concatenated ribosomal protein tree and in analyses from prior studies (Kantor et al., 2013; Rinke et al., 2013). We hypothesized that the difference in topology was due to an improved sequence alignment after removing insertion sequences. However, an updated analysis of the tree of life that took insertions into account also found the phylum to be within the CPR in multiple phylogenetic analyses (**Appendix 1.1**) (Hug et al., 2016). Due to the presence of only a few draft-quality Absconditabacteria genomes, placement remained an open question. Analysis of the sequence set from **Chapter 2,** which included 16S rRNA gene sequences from clone libraries and assembled from metagenomes, identified several chimeric sequences that were included from the Silva database. Removal of these sequences, in addition to improved sampling of several clades, confirmed the placement of the Absconditabacteria within the CPR (**Appendix 3 Figure 1**).

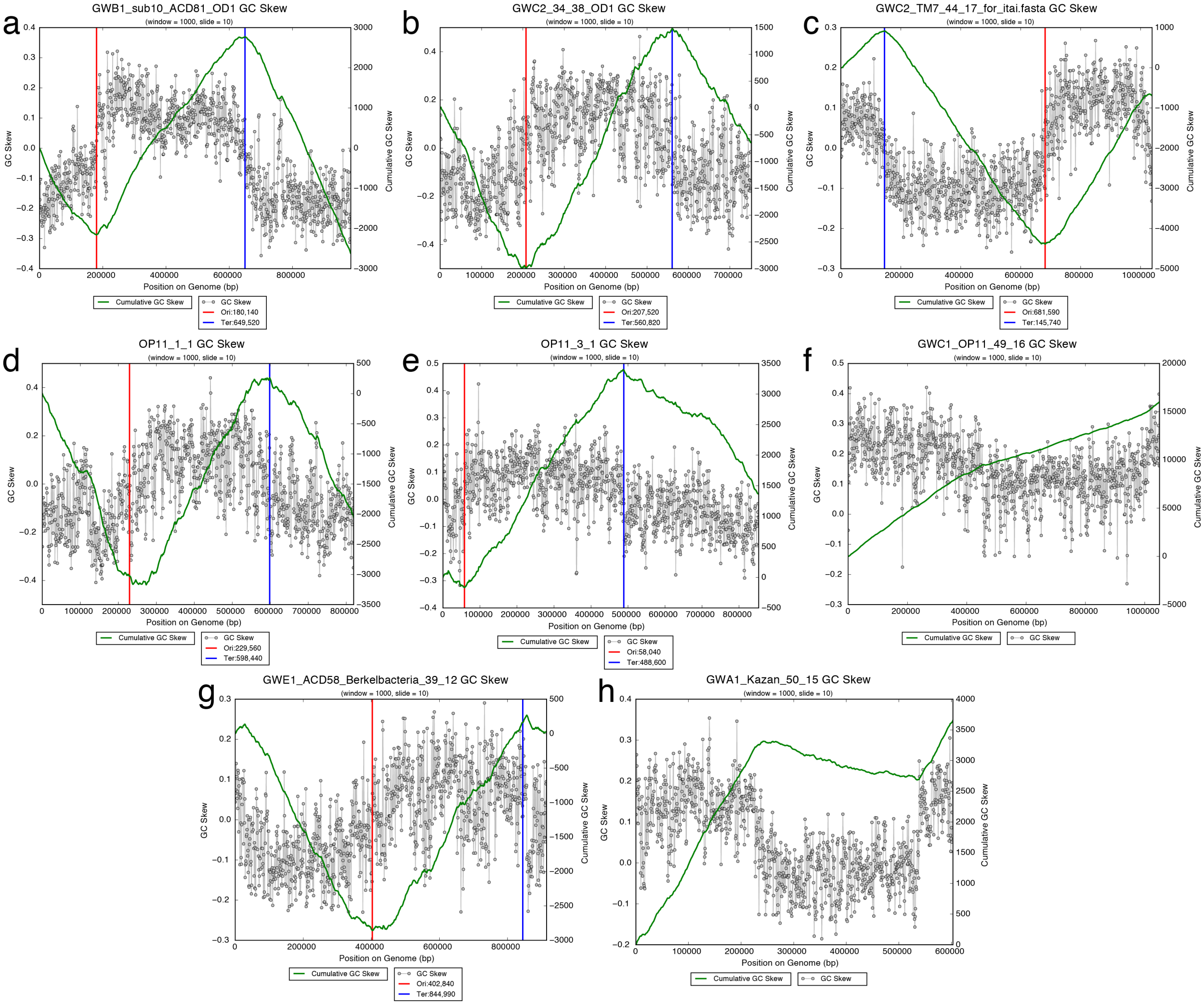
Appendix 3 Figure 1 | Updated 16S rRNA gene sequence phylogeny shows that the Absconditabacteria (SR1) are part of the Candidate Phyla Radiation (CPR). Phylogeny showing all included organisms (**a**), overall tree topology (**b**), and sampling of the Absconditabacteria (**c**).


a
b
Archaea
SR1
CPR
c
Bacteria_CPR_Peregrinibacteria_PER_-GWA2_PER_47_7
Bacteria_CPR_Peregrinibacteria_PER_-PER_GWC2_39_14
Bacteria_CPR_Peregrinibacteria_PER_-PER_GWF2_43_17
Bacteria_CPR_Peregrinibacteria_PER_-GWA2_PER_44_7
Bacteria_CPR_Peregrinibacteria_PER_-PER_GWF2_39_17
Bacteria_CPR_Peregrinibacteria_PER_-PER_GWF2_33_10
Bacteria_CP-Peregrinibacteria_BJP_IG2102_PER_44_74_final
Bacteria_CPR_Peregrinibacteria_PER_-GWA2_PER_33_10
Bacteria_CPR_Peregrinibacteria_PER_-GWB1_PER_54_5
Bacteria_CP-Peregrinibacteria_ALUMROCK_MS4_Peregrinibacteria_34_52
34
33
56
100
47
53 - PER
36
34
to other CPR
63
29
100 - SR1
DolJOral78_scaffold_11426_16S_1
silva_EF123563_1_1482_Bacteria_Candidate_division
silva_FJ623316_1_1430_Bacteria_Candidate_division_SR1
clone_55
clone_107
clone_58
HS2274O.8.22_scaffold_39594_16S_1
pre1_saliva_scaffold_5371_16S_1
term6_saliva_scaffold_1428_16S_1
Bacteria_CP-SR1_RAAC1_SR1_1
RAAC1_SR1_1_R115
clone_Palmetto_2_1_30
Palmetto_2_1_34
Zodeltone_water_scaffold_113879_16S_1
cpr-genome_ACD80_status=partial_source=metagenome
ar4r2_scaffold_799_16S_1
CG02_land_8_20_14_3.00_150_scaffold_13008_16S_1
CG_4_10_14_3_um_filter_scaffold_26613_16S_1
cg2_3.0_scaffold_7256_16S_1
CG_4_9_14_3_um_filter_150_scaffold_66296_16S_1
CG06_land_8_20_14_3.00_scaffold_56281_16S_1
CG06_land_8_20_14_3.00_150_scaffold_41088_16S_1
CG23_combo_of_CG06-09_8_20_14_all_150_scaffold_5908_16S_1
CG06-09_8_20_14_all_combo_scaffold_64063_16S_1
CG01_land_8_20_14_3.00_scaffold_3376_16S_1
CG07_land_8_20_14_0.80_scaffold_5372_16S_1
cg_4_9_14_3um_scaffold_109690_16S_1
PH2015_02_scaffold_110
silva_JF917269_1_1326_Bacteria_Candidate_division_SR1
Bacteria_CP-SR1_bacterium_Aalborg_AAW_1
clone_ABF-HB5-2
clone_FC1_16S_8
clone_EPS003BBFL_71
clone_RS06101_B101
clone_SLV_3WBAC_73
clone_SLV_3WBAC_46
47
73
64
73
63
55
46
58
100
95
Bacteria_CP_BD1-5_ALUMROCK_MS4_BD1-5_24_44.contigs
Bacteria_CPR_Gracilibacteria_BD1-5_-PER_bacterium_JGI_0000068
Bacteria_CPR_Gracilibacteria_BD1-5_-Gracilibacteria_bacterium_JGI_0000069
Bacteria_CP-GN02_candidate_division_GN02_bacterium_JGI_0000069-P22_CrabSpa_001_278
BD02T64_scaffold_38_16S_1
92
85
57

Appendix 4: GC skew for complete genomes from Candidate Phyla Radiation (CPR) bacteria

iRep analysis is based on the expectation that genome replication is occurring bi-directionally from a single origin (theta replication). We addressed the question of whether or not CPR bacteria replicate their genomes in this manner by identifying whether or not GC skew patterns match the pattern expected for genomes that undergo theta replication. All complete bacterial genome sequences from the Brown *et al.* study were analyzed (**Appendix 4 Figure 1**) (Brown et al., 2015), showing the expected pattern for all but two genome sequences. Theta replication has also been demonstrated for members of the Peregrinibacteria (PER) (Anantharaman et al., 2016a).

Appendix 4 Figure 1 | GC skew patterns indicate that CPR bacteria replicate their genomes from a single origin of replication. a-h, GC skew is calculated over 1 Kbp windows every 10 bp. The origin and terminus of replication can be identified for genomes that undergo theta replication as the transition points in a plot of cumulative GC skew.



Appendix 5: Software

Measurement of bacterial replication rates

The iRep algorithm and software (**Chapter 3**), and my implementation of the Korem *et al.* PTR algorithm, bPTR, are maintained under github.com/christophertbrown/iRep. My script for plotting GC skew for complete genome sequences is maintained under the same repository.

Genome curation

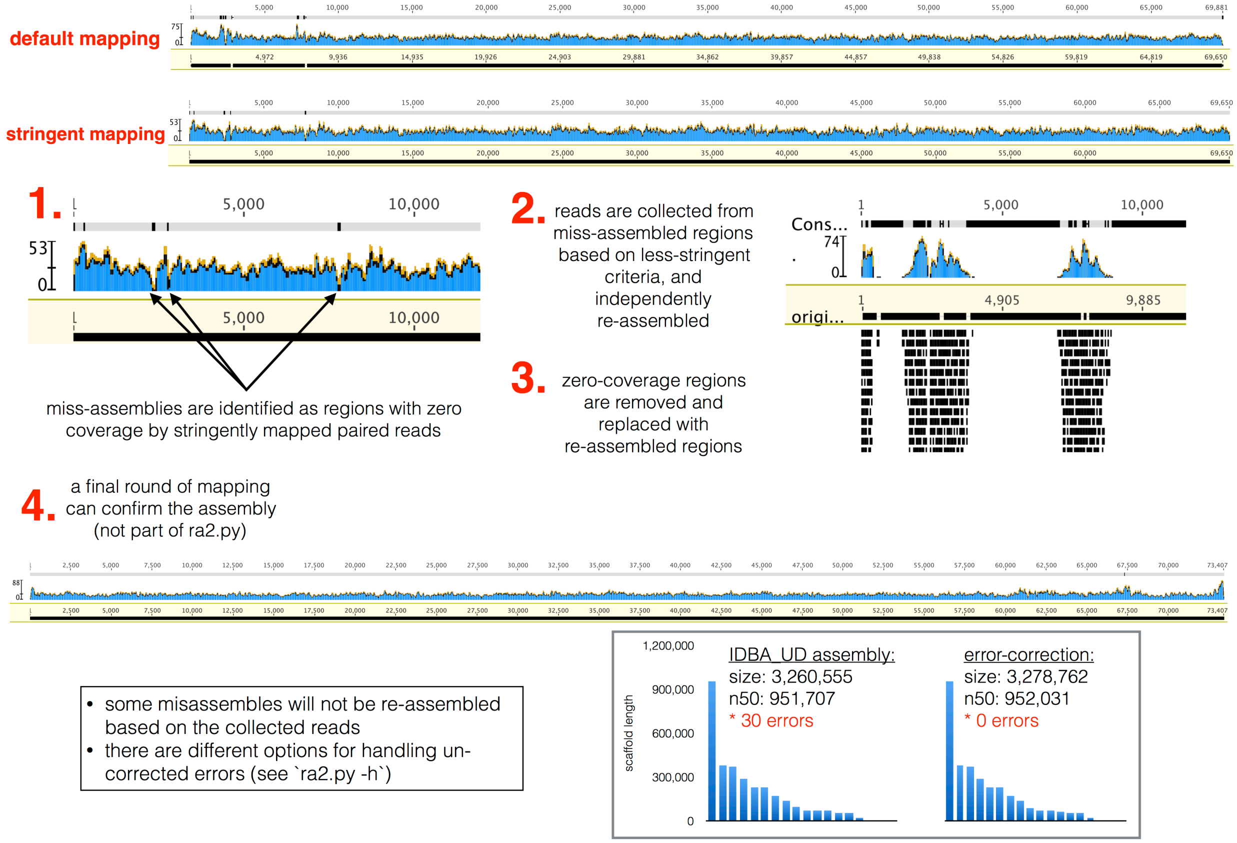
Several tools were developed in order to aid genome curation efforts. This endeavor began because we needed an automated way to accurately check and curate Candidate Phyla Radiation (CPR) rRNA genes with insertion sequences. The first version of this program, **re_assemble_errors.py**, achieved this goal by identifying scaffolding errors as regions of a genome assembly with no coverage by stringently mapped reads (**Chapter 2 and Appendix 5 Figure 1**). After identification, errors are re-assembled by collecting reads that mapped to the region, and then re-assembling the reads using Velvet (Zerbino and Birney, 2008). Re-assembled fragments are checked for scaffolding errors, and errors are replaced with the re-assembled sequence. In cases where the error could not be corrected, the sequence is broken and each broken end is extended, if possible. Several advancements on this original program were made, including the option to only break scaffolds in cases where an error cannot be corrected if there are no paired read sequences spanning the error. In cases where an error is not fixed, the error containing sequence can be replaced with Ns. This new version is referred to as **ra2.py**, which is short for "re-assemble errors version 2," and has been used in several co-authored studies (**Appendix 1.4, Appendix 1.6, and Appendix 1.8**). Because this process requires read tracking in memory, large assemblies cannot be curated. However, modification to use the compressed BAM format could improve performance and enable automatic curation of large metagenome assemblies. In the current implementation, the method is only suitable for single bacterial genomes. Another area for improvement has to do with the method used to re-assemble error containing regions. The current method runs Velvet using multiple kmer sizes. While this works in many cases, an overlap style assembler, rather than a de Bruijn graph assembler, has the potential to be much more effective. I have written a prototype assembler to accomplish this by using read overlaps to find a path between the high-quality sequences found on both sides of an error (**seq_extend.py**).

Genome curation often involves using paired read sequences to extend assembled scaffolds until overlaps can be found, a process I have automated with a script called **scaffolder.py**. This method extends scaffolds using methods implemented in **ra2.py**, identifies overlaps between scaffolds using BLAST (Altschul et al., 1990), and then validates scaffold joins based on stringent paired read mapping (based on **ra2.py**).

Another useful tool is **mapped.py**, which filters SAM read mapping files based on highly customizable criteria for how accurately paired reads map to an assembly. This can be used for manual genome curation, in which it is advantageous to visualize only high-quality read mappings, or can be used for abundance calculations, in which off target mappings need to be avoided (for example see **Appendix 1.3**).

This software is maintained under github.com/christophertbrown/fix_assembly_errors.

Appendix 5 Figure 1 | Schematic for the ra2.py automated genome curation method.


default mapping
stringent mapping
1.
53
0
5,000
10,000
miss-assemblies are identified as regions with zero coverage by stringently mapped paired reads
2.
reads are collected from miss-assembled regions based on less-stringent criteria, and independently re-assembled
Cons...
74
0
1
5,000
10,000
origi...
4,905
9,885
3.
zero-coverage regions are removed and replaced with re-assembled regions
4.
a final round of mapping can confirm the assembly (not part of ra2.py)
• some misassembles will not be re-assembled based on the collected reads
• there are different options for handling un-corrected errors (see `ra2.py -h`)
1,200,000
900,000
600,000
300,000
0
scaffold length
IDBA_UD assembly:
size: 3,260,555
n50: 951,707
* 30 errors
error-correction:
size: 3,278,762
n50: 952,031
* 0 errors

Clustering genome sequences based on average nucleotide identity

Clustering genomes based on shared sequence identity is a common and useful task. However, this can be a very time consuming process when hundreds or thousands of genomes need to be compared with one another. To accomplish this I wrote a script that uses the MASH (Ondov et al., 2016) algorithm to quickly estimate average nucleotide identity (ANI), and then group together any genomes with at least a specified amount of sequence similarity (**cluster_ani.py**). The script also chooses a representative genome for each cluster based on genome size, completeness, and contamination. This script is maintained under github.com/christophertbrown/bioscripts.

Ortholog detection for comparative genomics

Detection of orthologs between genome sequences is useful for comparative genomics. I wrote a script that conducts reciprocal best USEARCH searches between all pairs of genomes of interest in order to identify orthologs (**orthologer.py**). The script outputs a table that maintains information about gene synteny, the format of which is based on a prior method (Yelton et al., 2011). This method is distinct from prior approaches in that it allows for a comparison between any number of genomes, and can be run in either a reference or global mode. In reference mode all genomes are compared to a single reference, while in global mode all genomes are compared against all other genomes. This script is maintained under github.com/christophertbrown/bioscripts.

rRNA gene detection and characterization

The presence of insertions in CPR 16S and 23S rRNA gene sequences required the development of methods for accurately identifying these genes and their insertions (**Chapter 2**). This is accomplished using the scripts **16SfromHMM.py** and **23SfromHMM.py**, respectively. These scripts can be used in conjunction with **rRNA_insertions.py**, which analyzes insertion features, and **rRNA_copies.py**, which evaluates relative gene copy number. These scripts are maintained under github.com/christophertbrown/bioscripts.

Detection of a syntenic block of ribosomal proteins

In order to accurately identify the 16 ribosomal proteins frequently used for phylogenetic analysis (Hug et al., 2013; 2016), I developed a method for detecting them that takes into account the fact that they are frequently co-localized on a genome sequence (**rp16.py**). This is advantageous because instead of setting arbitrary cutoffs based on sequence similarity, putative annotations are validated based on their relative genome positions. This script is maintained under github.com/christophertbrown/bioscripts.