Genome-resolved meta-omics analyses of microbial interactions in mining-impacted systems

By

Rose Simone Kantor

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 Banfield, Chair
Professor Mary Firestone
Professor Kara Nelson

Fall 2016

Abstract

Genome-resolved meta-omics analyses of microbial interactions in mining-impacted systems

By

Rose Simone Kantor

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor Jillian Banfield, Chair

Microbial communities play important roles in natural, engineered, and anthropogenically altered systems. Specifically, microorganisms can metabolize contaminants and contribute to nutrient cycling. These processes can be achieved by one species or by the combined effects of multiple species. Hence, communities can possess emergent properties that are not always obvious from work on isolates or taxonomic profiling. Furthermore, the majority of microorganisms have not been cultivated in the laboratory and even for cultivated species, the full metabolic capacity may not be known. Examining genomes, the blueprints for life, and proteomes, the parts assembled from these blueprints, can provide insight into microbial physiologies and their roles in systems of interest. A more detailed understanding of which organisms are responsible for key processes could improve monitoring of *in situ* bioremediation, allow better targeted biostimulation, and even direct the manipulation of applied microbial systems to be more efficient.

Genome-resolved metagenomics and metaproteomics ("meta-omics") techniques are approaches that sample biomolecules (DNA and protein) from an intact microbial community, sequence or identify these molecules, and assign the sequences to specific populations. Analysis of the resulting data can yield a species-resolved view of the metabolic potential present in a community. These methods were used to investigate the structure and functioning of microbial communities in mining-contaminated systems. Bioinformatic analyses across three different systems sought to elucidate ecological roles for members of novel bacterial phyla and identify organisms that contributed to contaminant transformation. Contaminant studies focused on removal of common mining-related compounds including thiocyanate, cyanide, and reduced sulfur species. Metagenomes taken in series were used to examine the stability of consortia over time and increased thiocyanate loading while metagenomes taken across a mining landscape were used to assess the diversity, metabolic potential, and seasonal variation of microbial communities.

Most microbial communities include bacteria from major branches of the tree of life with no cultivated representatives. These lineages are referred to as Candidate Phyla (CP), and in the absence of complete genomes or cultivated representatives, many aspects of their biology and ecological roles remained unclear. Extensive characterization of sediment- and groundwater-associated microbial communities in Rifle, Colorado, USA, provided some of the first genomic observations of these CP. The site of a former uranium and vanadium mill in Rifle has been the subject of *in situ* biostimulation experiments, most notably, acetate addition to increase uranium reduction by the native microbial community. A series of metagenomes from acetate-amended

aquifer sediment yielded three complete and one near-complete bacterial genomes from CP, some of the first ever reported. Subsequent exploration of microbial communities involved in thiocyanate remediation and acid mine drainage also recovered genomes from the CP. Metabolic analyses based on the four Rifle genomes revealed the lack of an electron transport chain and pointed to energy generation based on fermentation of organic substrates including sugars, organic acids, amino acids, and DNA. A significant portion of genes in the unusually small genomes were involved in attachment, motility, and cell surface modification. Perhaps most importantly, none of the four genomes contained genes required for the complete biosynthesis of nucleic acids and amino acids. Taken together, all evidence suggests an obligately symbiotic or parasitic lifestyle for all four organisms.

Thiocyanate (SCN^-) is a common industrial contaminant produced at high quantities in gold mining. Chemical degradation of this compound is expensive and can produce other toxic byproducts, whereas biological treatment produces sulfate, ammonium, and carbon dioxide. Thiocyanate bioremediation has been successful at the pilot and industrial scale, but the biological underpinnings of the process were not well understood. In order to identify key pathways and organisms involved in thiocyanate degradation, microbial communities of two laboratory-scale continuous flow bioreactors were studied. The first reactor was a long-running system fed at high thiocyanate loadings whereas the second, inoculated with mixed culture from the first, was fed both thiocyanate and cyanide. Metagenomic sequencing and analysis of the two reactor communities resulted in a total of 93 bacterial and two eukaryotic genome bins. Based on coverage, the most abundant organisms in both reactors belonged to the genus *Thiobacillus*. Importantly, the genomes for these organisms encoded the enzyme thiocyanate hydrolase, located in an operon with cyanase and a predicted thiocyanate transporter. Other organisms in the reactor were predicted to oxidize sulfur or ammonium produced during thiocyanate degradation, and some possessed genes encoding denitrification. Whereas prior culture-based approaches had suggested heterotrophic organisms were responsible for thiocyanate degradation and that this process requires oxygen, the *Thiobacillus* spp. genomes encoded autotrophic metabolism and anaerobic respiration using nitrate. Altogether, a qualitative model of the community suggested that further optimization of the bioreactor system to encourage the growth of autotrophs could improve thiocyanate removal and that some ammonium removal may be possible via conversion to nitrogen gas.

A second study of thiocyanate degradation examined the stability and proteome of the microbial community across increasing loadings of thiocyanate. Biomass from the same long-running thiocyanate bioreactor was used to inoculate two new laboratory-scale reactors. One was fed thiocyanate at increasing loadings, while the other was fed ammonium sulfate at parallel nitrogen loadings to mimic the breakdown products of thiocyanate. Metagenomes from the ammonium sulfate reactor showed enrichment of community members involved in nitrogen cycling and heterotrophic metabolism compared to the inoculum. Meanwhile the thiocyanate reactor showed an increase in the relative abundance of three thiocyanate-degrading *Thiobacillus* species. Two of these species were composed of substrains whose differential abundances shifted across the time series, suggestive of within-species competition or niche specialization. Proteomic data collected at the final time point, when thiocyanate loading was highest, showed that the thiocyanate hydrolase enzyme was highly expressed by all three *Thiobacillus* spp. in whose genomes it was found. Proteins involved in sulfide oxidation, ammonium oxidation, nitrite oxidation, and carbon fixation were also detected, consistent with the model for thiocyanate degradation proposed in the prior study. Supporting the prediction of anaerobic

zones in biofilm, proteins involved in denitrification were also detected in the proteomics. The biofilm and planktonic community compositions were similar, but more biomass appeared to be present in the biofilm. The importance of biofilm for uncoupling hydraulic retention time from bacterial growth rates and improving denitrification suggests consideration of biofilm-based rather than sludge-based bioreactor designs in the future.

While laboratory-scale reactors allow detailed investigations of simplified communities, microbial communities in the environment can be much more complex. Mining waste sites exhibit a range of geochemical conditions that can shape the local microbial community, and in turn this impacted community may contribute to *in situ* remediation. Metagenomics was applied to four high-sulfur mining wastewaters and to a reservoir used to treat those wastes in order to construct a picture of the biological processes in play. Elemental sulfur was the most abundant form of sulfur at all sites, and specific clades of sulfur oxidizing organisms were enriched at each. Acidic waste rock sites and wastewater piped to the reservoir were enriched in acidophilic iron- and sulfur-oxidizing organisms. In contrast, drainage from a tailings-dewatering site high in organic carbon contained mainly methylotrophic bacteria, suggesting less sulfur oxidation occurs there. Metagenomes from the reservoir showed that it contained common freshwater bacteria as would be expected due to inputs from higher in the watershed, but it also contained numerous sulfur oxidizing bacteria. Some of these bacteria were likely capable of sulfur oxidation coupled to nitrate reduction under anoxic summer conditions. Importantly, sulfur oxidizers were present in late summer and early winter at multiple depths, despite a seasonal shift in the reservoir community. Mining operations rely on unmanaged *in situ* remediation in the reservoir to convert all sulfur compounds to sulfate, preventing acidification of receiving water bodies. This improved understanding of microbial metabolism present at various sites could lead to the development of active management strategies to achieve more complete, reliable remediation.

Metagenomic approaches provide foundational information about microbial communities and can be used to monitor these communities during remediation. Further biochemical validation and controlled experiments are needed to enable quantitative modeling of reaction rates required for engineering new remediation solutions. Mining represents perhaps the largest anthropogenic manipulation of the surface and subsurface of the planet, but it is also tied to the economic development of many countries. Ultimately, deeper knowledge of mining waste microbiology may lead to safer, more effective handling of mining wastes, benefiting environmental and public health.

Table of Contents

Dedication

This work is dedicated to the memory of my grandparents, Pearl Greenwald, Victor Greenwald, Mildred Kantor, and Robert Kantor. All were first-generation Americans who understood the value of education and were my strongest supporters and biggest admirers. This is for you.

This work is also dedicated to the farmers of Santa Rosa de Pacto, Ecuador, whose land is now a mining concession. Living with you in 2008 was an experience I will never forget, and it prompted me to learn everything I could about the environmental impacts of mining and biological mitigation strategies.

Introduction

Bioremediation, mining, and metagenomics

Microorganisms play an integral part in biogeochemical cycles, which occur in natural, disturbed, and human-engineered systems. In particular, they are involved in cycling nitrogen, carbon, sulfur, and metals via oxidation/reduction (redox) chemistry (Madsen, 2011). Complex interactions and a stunning diversity of enzymes make possible many different reactions including breakdown of toxic compounds that may be recalcitrant to photo- or chemical degradation. Biological degradation offers an alternative to chemical treatment processes that can be costly and can produce toxic byproducts. Additionally, biological treatment can often remove multiple contaminants at once. Bioremediation as a whole emcompasses not only biodegradation but also biosorption, whereby contaminants physically interact with biomass and are retained, and biomineralization, in which metals are oxidized or reduced, resulting in their precipitation out of solution. In contaminated sites, native microbial communities can sometimes provide remediation over time, in conjunction with chemical and physical processes occurring in water, soils, and sediments, a phenomenon termed natural attenuation. In cases where more rapid remediation is desired or the native community does not include key active members, *in situ* attenuation can be enhanced via biostimulation (addition of nutrients or other bioactive compounds) or bioaugmentation (direct addition of new microbial community members) (Scow and Hicks, 2005). In cases where waste generation is ongoing, bioreactors can retain and remediate contaminated water via activated sludge or biofilm processes.

The importance of microbial communities can be seen clearly in the area of mining and mine waste management. Nutrient cycling and redox reactions occurring in mining waste sites are important in both perpetuating and remediating waste. Biological iron and sulfur oxidation catalyze the production of acid mine drainage (AMD), and the microbial communities involved in this process have been studied extensively (Denef et al., 2010). These communities have also been harnessed in bioleaching, a technology in which microbial metabolism drives mineral dissolution and metal release. In addition to producing AMD, mine waste can have other environmental effects. Tailings can contain radioactive metals, cyanide or other compounds used to extract metals, and semi-stable sulfur compounds. These tailings and associated wastewater are often stored for long time periods and pose a risk to local freshwater bodies. While some bioremediation practices exist for these contaminants, knowledge of the biological processes involved is limited. The quantity and rate of mine waste production make expensive remediation practices infeasible, motivating the development of simple, cost effective methods that can be implemented on-site. This need is especially acute in mining regions with limited available freshwater, such as South Africa and Australia.

Many industrial processes (e.g. wastewater treatment) rely on mixed microbial consortia whose activities were not fully understood at the time of development of the process. In attempts to identify which organisms were responsible for key chemical transformations, characterization of community members was historically performed using culture-based techniques. Given that the vast majority of microorganisms cannot be cultivated by standard laboratory methods, this approach left much to be discovered. More recently, DNA fingerprinting methods have made use of the 16S rRNA gene and other marker genes to identify bacteria and archaea in industrial systems. Fingerprinting relies on knowledge of previously characterized organisms, but even closely related organisms can have different metabolisms. Assembly-based metagenomics followed by binning to separate different microbial genomes from one another can provide a

more complete picture of the identity, functional capacity, and relative abundances of distinct populations present in microbial communities (**Figures i.1** and **i.2**). Community genomic information can suggest how microbes are interacting with their environment and each other. Metagenomics of mining-impacted and otherwise disturbed systems has led to the discovery of members of new branches on the tree of life, which may be important parts of the biosphere enriched only under certain conditions. Further, genome-resolved approaches can suggest mechanisms underpinning applied processes that occur in bioreactors and *in situ* bioremediation processes. In this thesis, three types of mine tailings-related systems were studied using metagenomics, with the goal of understanding microbial diversity and interactions at the population level.

Radioactive metals, sediment communities, and candidate phyla

Rifle, Colorado, USA was the site of a uranium and vanadium mill, operated intermittently until 1958 near the banks of the Colorado River. Tailings were removed from the site by 1996, along with contaminated surface material, but alluvial groundwater remains contaminated with uranium, vanadium, and selenium (US Dept of Energy, 2015). Bioremediation studies were undertaken with the goal of converting uranium from a soluble valence state (U^{6+}) to an insoluble state (U(IV)), causing it to precipitate out of groundwater and become trapped in sediments. Iron-reducing bacteria are known to mediate this transformation by using uranium instead of iron as an electron acceptor. Because acetate can be respired under anaerobic conditions but cannot be fermented, this was chosen as an electron donor to stimulate uranium reduction. Groundwater acetate amendment experiments were conducted at the Rifle site, monitored at first with 16S rRNA gene sequencing (Anderson et al., 2003), followed by chip-based functional gene detection (Liang et al., 2012) and later, with metagenomics (Wrighton et al., 2014; 2012). Sediment acetate amendment studies were also undertaken, initially monitored with PhyloChip (Handley et al., 2012), and later, amendment of sediment-packed columns were studied with reconstructed 16S rRNA genes (Handley et al., 2015) and metagenomics (presented here).

Groundwater studies at Rifle provided some of the first genomes for organisms representing new branches of the tree of life (Wrighton et al., 2012), and acetate amendment increased the abundance of these organisms. Despite their detection in diverse environments (Borrel et al., 2010; Harris et al., 2004; Hugenholtz et al., 1998b; Peura et al., 2012), these organisms had been hitherto unstudied because they could not be cultivated in the laboratory. Thus they were referred to as candidate divisions or candidate phyla (CP), and some of the first complete genomes, collected from Rifle, are presented here. Subsequently, this clade has been shown to be a phylogenetically coherent group within the bacterial domain, and has been subdivided into more than 35 phyla, now known as the Candidate Phyla Radiation (CPR) (Brown et al., 2015; Hug et al., 2016). Genomes from members of the CPR were found in each study presented here and have also been detected across an ever wider variety of other environments.

Thiocyanate and laboratory-scale bioreactors

In addition to remediation practices implemented after mine closure, microbial degradation can be harnessed to treat continuous waste streams generated by active mining processes. In the metals extraction process, ore is finely milled and mixed with water to make a slurry. Bioleaching can help further dissolve the crushed rock to release gold and other metals. After

milling, cyanide (CN^-) is often used as a lixiviant to solubilize gold, and the resulting solution enters the carbon in leach or carbon in pulp process, where metals are recovered. If the ore is sulfidic, excess cyanide can react with sulfur in the tailings to produce thiocyanate (SCN^-). In some cases CN^- is deliberately reacted with sulfur to make it less dangerous, but free CN^- and heavy metals may be co-contaminants (Gould et al., 2012). Although less dangerous than CN^-, SCN^- is more stable in the environment. SCN^- is toxic at low levels to aquatic organisms and can affect human health by interfering with thyroid function (Erdogan, 2003; Speyer and Raymond, 1988; Watson and Maly, 1987). Aside from environmental impacts, SCN^- is also toxic to organisms involved in bioleaching (e.g. the BIOX™ process), limiting direct reuse of wastewater in upstream mining processes (van Hille et al., 2015). Additionally, thiocyanate is present in other industrial wastewaters, especially those produced during coal and steel processing.

SCN^- can be biologically degraded by several different groups of bacteria that hydrolyze it into distinct molecules for each of its component atoms, S, C, and N (**Figure i.1**), and many technologies have been employed for bioremediation (Gould et al., 2012). The molecular mechanisms for SCN^- decomposition have been probed in *Thiobacillus thioparus* TH115 (Katayama et al., 1998; 1992) as well as *Thiohalophilus thiocyanoxidans* (Bezsudnova et al., 2007). However, most biotechnologies for thiocyanate remediation utilize microbial communities rather than single-strain isolates. One of the first cyanide and thiocyanate remediation processes based on a microbial community was developed at the Homestake Mine (ND, USA), reaching full-scale operation by 1985 (Whitlock, 1990), and several others followed in subsequent decades (Jeong and Chung, 2006; Stott et al., 2001).

In South Africa, Outotec (formerly Biomin Ltd) developed the Activated Sludge Tailings Effluent Remediation (ASTER™) process to treat SCN^- containing wastewater from gold mines. The inoculum was a mixture of sludges from an SCN^- laden tailings pond and from a domestic wastewater treatment plant. Thiocyanate degradation was stably observed, but microbiological analysis was limited to culture-based methods with heterotroph media (van Buuren et al., 2011). A demonstration-scale plant was constructed for the ASTER™ process and sludge from this plant was given to the Centre for Bioprocess Engineering Research at the University of Cape Town for optimization studies. Microbial characterization at this stage included cultivation on multiple media types as well as 16S rRNA gene fingerprinting. This approach revealed a more diverse community than previously described (Huddy et al., 2015; van Zyl et al., 2011). While full-scale facilities have been constructed at mining sites on multiple continents, the biological mechanisms behind SCN^- degradation in these mixed microbial communities remained unknown. A "stock culture" from ASTER™ was maintained in a laboratory-scale reactor at the University of Cape Town and could be used as a model system for further investigations. Studies described herein made use of this "stock reactor" to provide inoculum for new experimental laboratory-scale reactors, and these studies provided the first metagenomic data for thiocyanate-degrading communities.

Thiosalts and landscape-scale analysis

At the field-scale, mining waste creates a variety of microenvironments, but fluid flow across a landscape can connect them and can transport contaminants. If the ore being mined is sulfidic, then thiosalts such as thiosulfate ($S_2O_3^{2-}$), trithionate ($S_3O_6^{2-}$), and tetrathionate ($S_4O_6^{2-}$) can be abundant in run-off (Miranda-Trevino et al., 2013). These salts are toxic to aquatic organisms only at high levels, but they contribute to acidification by oxidation to sulfuric acid (Kuyucak and Yaschyshyn, 2007). Therefore, wastewaters are processed by either chemical or biological

oxidation and neutralized before being released. Measurement and treatment methods for thiosalts have been a long-term focus of the Canadian mining industry, in order to increase awareness of this source of contamination and to lessen environmental impacts (Wasserlauf and Dutrizac, 1984).

Natural attenuation of sulfur compounds is known to occur, but it can be very slow under cold conditions. Bioreactors present an efficient but more costly solution, and development of this technology is ongoing but has not reached full-scale (Iglesias et al., 2016; Liljeqvist et al., 2011). Thus, *in situ* remediation is still preferred where possible. The Glencore mining district (Ontario, Canada) contains a freshwater lake that is in long-term use as an oxidation reservoir, converting thiosalts to sulfate. The reservoir has the advantage of natural watershed flows, which dilute mining inputs. Geochemical investigations of the reservoir and source inputs have recently revealed low thiosalts but very high levels of elemental sulfur and have shown that thiosalts oxidation is slower in winter (Warren et al. in prep). To date, the microbial communities of these sites have been studied using enrichment cultures only (unpublished), and sequencing efforts, beginning with the work described here, are now in progress.

Looking forward: mining, microbiology and biotechnology

At their most severe, environmental impacts of mining may entail a centuries-long legacy of soil and water contamination or the complete erasure of a landscape. Mine wastes affect human health and livelihoods as well as biodiversity. There is an increasing call for environmentally and socially responsible mining practices (Miranda and Sauer, 2010; Miranda et al., 2003), and microorganisms are critical to these practices. Metagenomics provides the opportunity to learn about organisms involved in *in situ* remediation and large-scale biotechnologies. Cross-disciplinary collaborations between biologists, geochemists, and engineers will apply knowledge gained to process design with the goal of producing economical and thorough waste treatment biotechnologies.

Figures

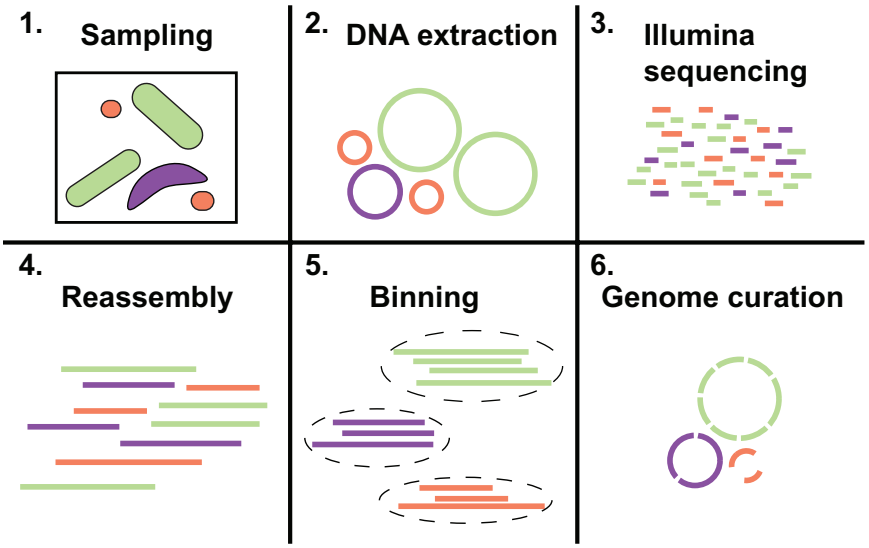

Figure i.1. Metagenomics methods overview.

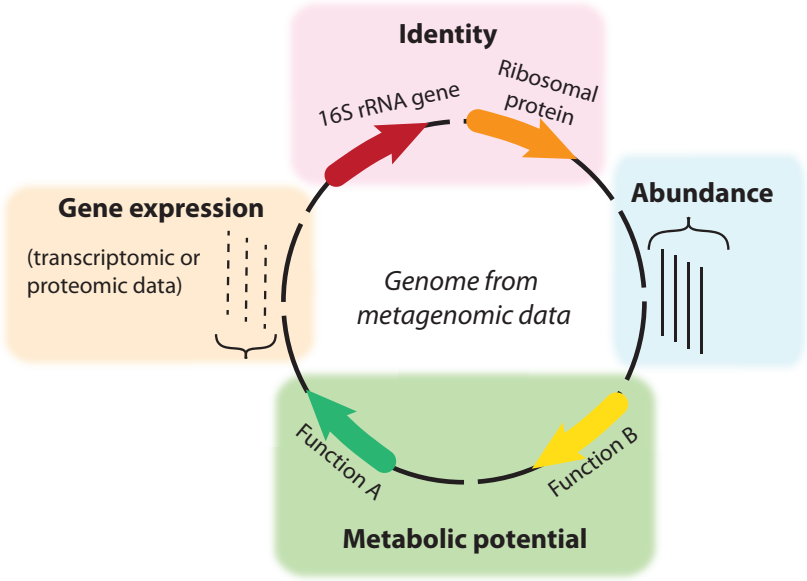

Figure i.2. Metagenomics for microbial ecology.

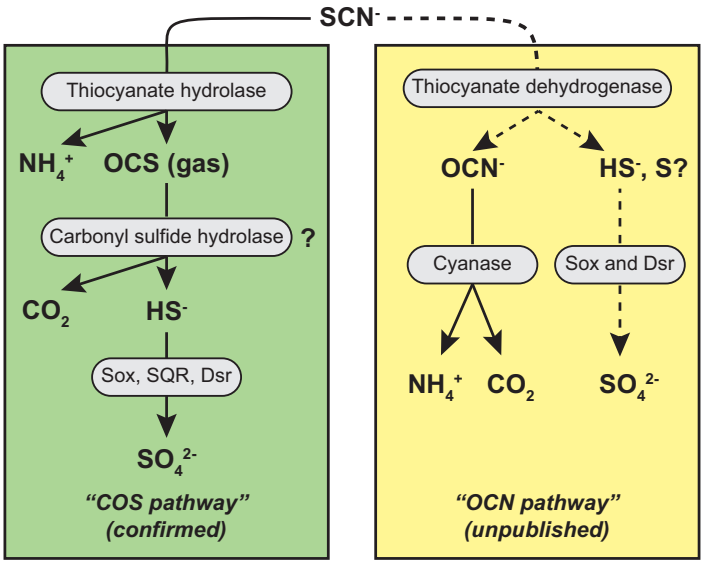

Figure i.3 Microbial thiocyanate degradation pathways. Dashed lines indicate unknown portions of the pathway that have been suggested in the literature.

Acknowledgements

I would like to thank my advisor, Jill Banfield, for her tremendous support and for fostering my development as a scientist. I also gratefully acknowledge my committee members, Mary Firestone and Kara Nelson for their feedback. Every member of the Banfield group during my time here has contributed to this thesis and helped me build skills and confidence. A special thanks to Chris Brown for his camaraderie and aid with all things coding. Thank you to the talented post doctoral mentors in the lab, to Brian Thomas and Andrea Singh for their assistance with programming and ggkbase, and to everyone else for the opportunity to learn from you and with you.

I am deeply indebted to my collaborators and co-authors at the Rifle, CO, USA site (Department of Energy and Lawrence Berkeley National Lab), the Centre for Bioprocess Engineering Research at the University of Cape Town, South Africa, the Warren group at the University of Toronto, Canada, and the Hettich group at Oak Ridge National Laboratory. I am grateful for the staff, faculty, and students in the Plant and Microbial Biology Department at UC Berkeley, especially Rocío Sanchez and Dana Jantz. I acknowledge funding provided by the NSF-GRFP, NSF-GROW with USAID, and the Berkeley Fellowship.

Most importantly, this thesis would not have been possible without the support and encouragement of my wonderful family and friends.

Chapter 1

Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla

Published in mBio, 2013

Abstract

Cultivation-independent surveys of microbial diversity have revealed many Bacterial phyla that lack cultured representatives. These lineages, referred to as candidate phyla, have been detected across many environments. Here, we deeply sequenced microbial communities from acetate-stimulated aquifer sediment to recover complete and essentially complete genomes for single representatives of the candidate phyla SR1, WWE3, TM7, and OD1. All four of these genomes are very small, 0.7 – 1.2 Mbp in size, and have large inventories of novel proteins. Additionally, all lack identifiable biosynthetic pathways for several key metabolites. The SR1 genome uses the UGA codon to encode glycine, and the same codon is very rare in the OD1 genome, suggesting that the OD1 organism could also transition to alternate coding. Interestingly, the relative abundance of the SR1 increased with the appearance of sulfide in groundwater, a pattern mirrored by a member of the phylum Tenericutes. All four genomes encode type IV pili, possibly involved in inter-organism interaction. Based on these results and other recently published research, metabolic dependence on other organisms may be widely distributed across multiple Bacterial candidate phyla.

Importance

Few or no genomic sequences exist for members of the numerous bacterial phyla lacking cultivated representatives, making it difficult to assess their roles in the environment. This paper presents two complete and two essentially complete genomes for members of four candidate phyla, documents consistently small genome size, and predicts metabolic capabilities based on gene content. These metagenomic analyses expand our view of a lifestyle apparently common across these candidate phyla.

Introduction

The current number of bacterial phyla recognized by rRNA databases is between 63 and 84 (Silva and Greengenes, accessed August, 2013), although the true count is almost certainly higher, with some estimates as high as 100 phyla (Harris et al., 2013). A careful examination of naming conventions across databases and phylogeny allowed for dereplication of the list of phyla and suggested that there are at least 38 without cultivated representatives (McDonald et al., 2012); these are referred to as candidate divisions (CD) (Pace, 2009; Rappe and Giovannoni, 2003) or candidate phyla (CP). The increase in of the number of CP over the last ten years can be attributed in part to the subdivision of one major clade, initially referred to as OP11 (Hugenholtz et al., 1998a), that is now recognized to comprise several phyla including OP11, OD1, and SR1 (Harris et al., 2004). Some CP, such as the TM7, are relatively well defined (Hugenholtz et al., 1998a), while others, including the WWE3 (Guermazi et al., 2008), and the PER (Wrighton et al., 2012) have only recently been proposed.

CP organisms have been detected by 16S rRNA gene sequencing surveys spanning a wide array of environment types, including in the human oral microbiome (Campbell et al., 2013; Dewhirst et al., 2010), mammalian gut (Ley et al., 2008), bioreactors (Albertsen et al., 2013a; Guermazi et al., 2008; Hugenholtz et al., 2001), fresh water lakes (Borrel et al., 2010; Peura et al., 2012), hypersaline microbial mats (Harris et al., 2013), and deep-sea vents (Perner et al., 2007). Targeted 16S rRNA gene primer assays have documented diversity within the CP and assessed the abundance of these organisms in certain environments, especially under anoxic and sulfidic conditions (Borrel et al., 2010; Davis et al., 2009; Dinis et al., 2011; Guermazi et al., 2008; Harris et al., 2004; Hugenholtz et al., 2001; Peura et al., 2012). However, the full diversity and roles of CP organisms in the environment remain unclear. These questions have motivated the use of two cultivation-independent approaches for sequencing CP genomes: single-cell flow systems coupled to multiple displacement amplification (MDA) and metagenomics. Single-cell sequencing has generated genomic information for representatives of several CP (Campbell et al., 2013; Marcy et al., 2007; McLean et al., 2013; Podar et al., 2007; Rinke et al., 2013; Youssef et al., 2011), but genomes are typically highly incomplete and amplification bias affects sequenced gene copy number (Lasken, 2012). Metagenomic methods have yielded near-complete and complete genomic sequences for uncultivated groups in natural environments
(Baker et al., 2010; Castelle et al., 2013; Iverson et al., 2012; Tyson et al., 2004; Woyke et al., 2006; Wrighton et al., 2012). However, an impediment to metagenomic reconstruction of genomes is the relatively low abundance of CP cells in environmental samples. Biostimulation is one means to alter the profile of a microbial community, enriching for certain metabolic capabilities, e.g. (Anderson et al., 2003). For example, acetate amendment of groundwater combined with filtering (enriching for cells < 1.2 μm in diameter) at an aquifer in Rifle, CO, USA recently led to the successful reconstruction of 49 partial and near-complete genomes (Wrighton et al., 2012) from members of several CP including OD1, OP11, BD1-5, and PER. Such strategies have the additional advantage that they can potentially illuminate organism response to specific stimuli as well as correlations in organism abundance patterns.

Here, we applied improved metagenomic methods to Illumina sequence datasets recovered from a series of biostimulated sediment communities and assembled complete genomes corresponding to four CP: SR1, WWE3, TM7, and OD1. We report genome characteristics, metabolic potential, and abundance across a range of geochemical conditions and community compositions Analysis of these new genomes and several other recently published

genomes from the same CP indicates a surprisingly consistent life strategy and provides insight into why members of these phyla remain uncultivated.

Results

We conducted a time-course biostimulation experiment using flow-through sediment columns suspended within an aquifer at Rifle, CO. Thirteen columns were pumped with acetate-amended groundwater for between 13 and 63 days, and individual columns were sacrificed at points of geochemical interest (Handley et al., 2015). Metagenomic sequencing of DNA from acetate-amended column sediment and an unamended background sediment sample revealed the presence of a variety of bacteria, including diverse members of the CP.

Genomic sequences for four CP organisms of interest were present in multiple samples. The fragmentation patterns of each genome differed across the metagenomic datasets, making it possible to identify overlaps in scaffolds from different samples and to generate high quality draft genomes. Differential organism abundance patterns across the samples were used to confirm that all fragments > 3 kb in length were correctly assigned to the four genomes (**Figure 1.1**). Draft genomes were subsequently curated to complete and near-complete genomes using paired-end read information (see methods). The reconstructed genomes are presented here as RAAC1-RAAC4 for "Rifle Acetate Amendment Columns". Based on phylogenetic analysis, the RAAC1 genome falls within SR1, RAAC2 within WWE3, a clade sister to the OP11 (Guermazi et al., 2008), RAAC3 into TM7 group 3, and RAAC4 into OD1 (**Figures 1.1** and **1.2**). Notably, all four genomes are very small, 0.7 to 1.17 Mb in length (**Table 1.1**), within the range typically seen only in obligate symbionts (**Figure 1.3**). Only the RAAC4 (OD1) genome is not circularized. Attempts to circularize this genome using PCR amplification failed, despite control reactions demonstrating that other segments of the OD1 genome could be amplified from the sediment column extraction. Closely related genomes of similar sizes have been recovered from the same site, suggesting that the genome is likely near-complete (unpublished data). While each genome represents the composite sequence of a population, very little strain variation was observed.

A previous study of planktonic cells from the same site reconstructed genomes related to those reported here (Wrighton et al., 2012), but lack of associated 16S rRNA gene sequences complicated classification efforts. The current genomes allow phylogenetic resolution via 16S rRNA and protein trees (**Figures 1.1** and **1.4**). One partial genome, ACD25, previously classified as OP11 (Wrighton et al., 2012), matches the WWE3 genome (RAAC2) at the species level and has an identical protein sequence for the DNA-directed RNA polymerase beta subunit (**Figure 1.4**). Based on analyses reported here, ACD22 and ACD24 are also reassigned to the phylum WWE3. Another genome, ACD80, previously classified as a distant BD1-5 relative, is reclassified as belonging to the phylum SR1, in agreement with Campbell *et al.* (Campbell et al., 2013), and a tentatively binned scaffold containing a 16S rRNA gene was confirmed as belonging to that genome (**Figures 1.1** and **1.2**) (Wrighton et al., 2012).

Time series relative abundance

The four CP genomes exhibit distinct enrichment patterns across the fourteen samples (**Figure 1.5**). Acetate stimulation resulted in the early increase and then decline of Betaproteobacteria, followed by an increase in Clostridia (**Figure 1.5**), a pattern that largely paralleled the shift in terminal electron accepting processes during acetate amendment (Handley et al., 2015; Williams

et al., 2011). The change from iron reduction to sulfate reduction is reflected in the exponential variation in groundwater sulfide concentration across samples (**Figure 1.6A**). The SR1 genome rose in relative abundance by as much as 500 fold as sulfide concentrations increased; its abundance also correlated strongly with that of a member of the Tenericutes (**Figure 1.6B**). The log relative abundance of the SR1 genome is directly correlated with log sulfide concentration until peak sulfate reduction (**Figure 1.6A**, $R^2 = 0.78$, $p < .0005$), after which abundance continued to rise and sulfide levels began to decline. This correlation could be due to a variety of factors including metabolic relationships between the SR1 and organisms performing sulfate reduction. Of the CP genomes, the WWE3 genome appears the most stably abundant (at 0.13-0.58%) from mid-iron reduction through sulfate reduction. The TM7 and OD1 genomes had the highest relative abundance in samples 3 and 7, before notable sulfate reduction occurred (**Figure 1.5**).

Novel proteins and metabolic characterization

Based on annotations, each of the four CP genomes contains a large fraction of proteins lacking functional predictions and proteins whose only homologs are hypothetical proteins from CP genomes (**Table 1.1**) (Albertsen et al., 2013a; Campbell et al., 2013; Marcy et al., 2007; Podar et al., 2007; Wrighton et al., 2012). When the four genomes were searched against Hidden Markov Models (HMMs) representing all known protein families (called "sifting families" or "SFams," see ref. (Sharpton et al., 2012), between 39% (OD1) and 53% (SR1) of the CP sequences were not matched to any family (**Figure 1.7**). Attempts to cluster novel proteins in order to identify previously unknown protein families were largely unsuccessful, likely due to the wide phylogenetic distances between the four genomes (data not shown).

Analyses using KAAS (KEGG Automatic Annotation Server, (Moriya et al., 2007), independent BLAST searches against the CP genomes, and gene-by-gene manual analysis for all genomes were performed to assess completeness of central metabolic pathways. Consistent with previous analyses of other bacteria from these CP (Albertsen et al., 2013a; Campbell et al., 2013; Podar et al., 2007; Wrighton et al., 2012), all genomes support fermentative metabolisms (**Figure 1.8**) and lack TCA cycle genes. Superoxide dismutase and alkyl hydroperoxide reductase genes are present, presumably for scavenging oxygen and free-radicals. The four genomes encode distinct fermentation pathways, different abilities to use and store complex carbon, different electron-carrying proteins and some key unique pathways, although more consistency was observed within the candidate phyla (**Figure 1.9**).

RAAC1 (SR1)

The SR1 genome lacks genes involved in the initial steps of Embden-Meyerhof-Parnas (EMP), pentose phosphate, and Entner-Doudoroff pathways, consistent with our analysis of ACD80 and SR1_OR1 (Campbell et al., 2013; Wrighton et al., 2012). In RAAC1, genes were identified encoding triose phosphate isomerase and the lower portion of the EMP pathway, from the conversion of glyceraldehyde-3-phosphate to 3-phosphoglycerate by a possible non-phosphorylating GAP dehydrogenase (GapN) through formation of pyruvate (**Figure 1.8A and Figure 1.9**). The other two SR1 genomes, ACD80 and SR1-OR1, do not appear to contain these genes. The potential for gluconeogenesis is indicated by the presence of pyruvate phosphate dikinase, but the other gene involved in this pathway, fructose 1,6-bisphosphatase, was not identified in any SR1 genome. The RAAC1 genome possesses a gene cluster responsible for fermentation of pyruvate to acetate and formate via pyruvate:formate lyase (PFL), the alternative

phosphotransacetylase (pduL) (Pierce et al., 2008), acetate kinase, and a putative formate transporter. This pathway is also found in SR1-OR1, though not in a gene cluster. The organism may also degrade complex carbon using a dockerin-like protein, a cellulosome anchoring protein and several cell-surface glucanases and pectin lyases, as observed in the other SR1 genomes.

Also as reported previously for the two partially reconstructed SR1 genomes (Campbell et al., 2013; Wrighton et al., 2012), the SR1 genome presented here harbors a type II/III intermediate form of ribulose 1,5-bisphosphate carboxylase-oxygenase (RuBisCO). This type of RuBisCO is implicated in the AMP salvage pathway (Sato et al., 2007), identified in RAAC1 (**Figure 1.8A**) and both other SR1 genomes sequenced to date. The RAAC1 RuBisCO possesses conserved catalytic residues and an additional 29 amino acid sequence unique to this form of RuBisCO. The sequence identity across all three SR1 RuBisCOs is high (72%), suggesting they play a similar role in the metabolism of these organisms. As suggested by Sato *et al.* (Sato et al., 2007), the 3-phosphoglycerate produced from the AMP salvage pathway may enter glycolysis for energy generation.

Lastly, the RAAC1 SR1 genome encodes several electron-carrying proteins including a cytochrome B_5, three Fe-S cluster proteins of unknown function, a ferredoxin-reductase-like protein, three flavodoxins, and a rubrerythrin. Some of these may be important for response to oxidative stress and/or reoxidation of reduced ferredoxin or NADH.

RAAC2 (WWE3)

The WWE3 genome possesses most genes for the EMP pathway, pyruvate:ferredoxin oxidoreductase (PFOR, or possibly 2-oxoglutarate ferredoxin oxidoreducase; subunits alpha and beta), which converts pyruvate to acetyl-CoA, and an acetyl-CoA synthetase (ADP-forming, EC: 6.2.1.13) that produces acetate and ATP (**Figure 1.8B**). There is also a gene cluster for the lower portion of the pentose phosphate pathway, converting ribulose-5-phosphate to glyceraldehyde-3-phosphate. Located in the same gene cluster is a protein identified as belonging to the D-isomer-specific 2-hydroxyacid dehydrogenases, a superfamily that contains glyoxylate reductase and D-lactate dehydrogenase (Fauvart et al., 2007). The presence of phosphoenolpyruvate (PEP) synthase suggests the WWE3 organism may engage in gluconeogenesis, but it lacks fructose 1,6-bisphosphatase. These observations are consistent with analysis of the newly reclassified WWE3, ACD22, ACD24, and ACD25 (**Figure 1.9**), although these genomes are fragmented and incomplete, and may be subject to binning error. The RAAC2 WWE3 organism may also be capable of synthesizing and utilizing glycogen, a common energy storage polysaccharide, via an operon containing a predicted galactose-1-P uridylyltransferase, 2 glycogen synthase genes, an alpha-amylase-like gene, and a glycosyl hydrolase of family 57, the family containing the branching enzyme required to form glycogen.

The ATP synthase operon in the WWE3 genome contains alpha, beta, gamma, and delta/epsilon subunits, however, the adjacently located a, b, and c subunits lack significant homology to known ATPase subunits. Instead, the putative c-subunit, responsible for ion translocation, bears homology to a similar transmembrane protein found in the ATPase operons of the OP11 and WWE3 genomes, ACD38 and ACD24 (Wrighton et al., 2012). Given the lack of homology to characterized ATPases (Müller and Grüber, 2003), it remains unclear whether the ATPase, if functional, operates primarily in the forward or reverse direction and whether it pumps protons or sodium ions. This genome also contains a membrane-bound proton/sodium pumping pyrophosphatase, which could be used to generate membrane potential. Electron-carriers in the WWE3 organism include a ferredoxin oxidoreductase-like protein, a cytochrome

b_5, and a plastocyanin-like blue copper protein encoded in the genome adjacent to a predicted membrane-associated ferric reductase-like protein (**Figure 1.8B**). Additional electron flow may proceed through a type 3B-like cytoplasmic nickel-dependent hydrogenase homologous to those found in another WWE3 genome, ACD22, and several OD1 genomes (Wrighton et al., 2012).

RAAC3 (TM7)

The pentose phosphate pathway and most of the EMP pathway are present in the TM7 genome (**Figure 1.8C)** but enolase was not detected, nor was a means for converting pyruvate to acetyl-CoA (e.g. PFOR, PFL, or pyruvate dehydrogenase) or for using acetyl-CoA. We confirmed that enolase is not annotated or identifiable in any of the currently available TM7 genome sequences (Albertsen et al., 2013a; Marcy et al., 2007; Podar et al., 2007) (**Figure 1.9**). RAAC3 contains phosphoketolase but not acetate kinase, needed for the generation of ATP from this branch off the pentose phosphate pathway (**Figure 1.8C**, box 19), although both genes were identified in a gene cluster in a related genome (Albertsen et al., 2013a) and synteny is lacking between the two genomes in this region. Other possible routes for fermentation and regeneration of NAD^+ by the RAAC3 TM7 organism include two dehydrogenases related to D-lactate dehydrogenase, which are conserved in other TM7 genomes (Albertsen et al., 2013a). Malate/lactate dehydrogenase, reported in another TM7 genome (Albertsen et al., 2013a), was not identified in RAAC3. As an alternative means for producing ATP, RAAC3 encodes the arginine deiminase pathway (**Figure 1.8C**). We identified genes in this pathway in other TM7 genome sequences (Albertsen et al., 2013a; Marcy et al., 2007), suggesting it may be common to members of the TM7 phylum.

The TM7 (RAAC3) genome has several genes involved in complex carbon degradation, including beta-glucosidase, a predicted secreted endo-1,3(4)beta-glucanase, alpha-amylase, and a glycogen phosphorylase-like protein. The presence of a bifunctional trehalose synthase/phosphatase indicates usage of trehalose, which is synthesized from glucose-6-P and UDP-glucose. Alpha-alpha trehalase is also present, providing for the subsequent degradation of trehalose and release of glucose for cellular consumption.

The RAAC3 TM7 genome contains a complete ubiquinol oxidase (cytochrome b_o) operon, with intact functional residues as well as residues known to distinguish cytochrome O ubiquinol oxidases from their closely related counterparts, cytochrome *c* oxidases (Abramson et al., 2000). This complex could be used for oxygen scavenging, as all other information points to a fermentation-based metabolism and we did not find the complex in other TM7 genomes. The electron source for ubiquinol oxidase may be a single-subunit form of NADH:ubiquinone dehydrogenase (NDH) most similar to type II NDH in structural searches (Feng et al., 2012; Kelley and Sternberg, 2009). This NDH is located adjacent to the ubiquinol oxidase operon in the RAAC3 genome and has homologs in other TM7 genomes (Albertsen et al., 2013a; Marcy et al., 2007).

RAAC4 (OD1)

The OD1 is a very diverse CP, so it is unsurprising that variation exists in metabolic capabilities. Like the TM7, the RAAC4 OD1 genome contains genes for the pentose phosphate pathway, as does the recently sequenced AAA011-A08 genome (Rinke et al., 2013). However, the AAA255-P19 OD1 genome from the same work does not contain this pathway, and two other partial OD1 genomes examined support only the latter half of the pathway (**Figure 1.9**). RAAC4 possesses a modified EMP pathway from mannose-6-P isomerase to pyruvate (**Figure 1.8D**). Also similar to the TM7, and consistent with other OD1 analyzed here, it does not contain any identifiable PFL,

PFOR, or pyruvate dehydrogenase genes, nor genes for utilizing acetyl-CoA. Like the WWE3 organism, the OD1 may be capable of fermentation to lactate, indicated by a 2-hydroxyacid dehydrogenase family gene found adjacent to enolase and pyruvate kinase in the genome. Synteny of the genes in this cluster is not conserved for the other OD1 genomes examined. A predicted lactate transporter is encoded elsewhere in the RAAC4 genome. Presence of a putative cellulosome-related gene cluster, 2 glycosyl hydrolases and an end-specific cellobiohydrolase suggest that the OD1 is capable of complex carbon degradation. The RAAC4 OD1 genome also contains a gene cluster involved in alternative polyamine biosynthesis as described by Lee *et al*. (Lee et al., 2009), which may be important in biofilm formation.

The RAAC4 OD1 genome may use a membrane-bound sodium/proton pumping pyrophosphatase to generate a proton motive force. Electron transport proteins include a rubredoxin and flavoprotein of unknown function, perhaps involved in oxygenic stress tolerance. While other OD1 genomes previously described were found to contain putative nickel-iron hydrogenases involved in uptake and hydrogen production, no hydrogenases could be identified in RAAC4.

Essential biosynthetic pathways

While the RAAC genomes have easily identifiable enzymes for the inter-conversion of amino acids and nucleotides (e.g. serine hydroxymethyltransferase with the exception of the WWE3; dCTP- or dCMP- deaminase), complete biosynthesis pathways for nucleotides (**Table S1.2 and Figure 1.10**), lipids, and most amino acids (White, 2000) could not be identified in any of the four genomes based on annotations or using KAAS (Moriya et al., 2007). Based on this preliminary analysis, we concluded that these organisms may be auxotrophic for many essential metabolites or may contain novel biosynthetic pathways.

Further analysis was performed to assess completeness of nucleotide biosynthesis using reference sets of genes from diverse genomes to query RAAC genomes (see methods). The SR1 appear to be missing most of this pathway. If novel genes present in this phylum could perform parts of this pathway, we might expect them to be conserved and possibly located near identifiable genes involved in nucleotide biosynthesis. However, few biosynthesis genes are present all in the three SR1 genomes, and the regions containing these genes lack synteny, offering no clues to novel conserved genes that may perform the missing functions. The possibility remains that such genes are found elsewhere in the genomes. Similarly, the WWE3 and TM7 genomes show few genes involved in nucleotide biosynthesis and lack synteny surrounding these genes.

Some sequenced representatives of the OD1 may have functional pathways for nucleotide biosynthesis (e.g. single cell genome AAA255-P19) (Rinke et al., 2013), but RAAC4 (OD1) does not (**Figure 1.10**). Gene loss or horizontal gene transfer could explain these differences in metabolic potential between members of the OD1. Whereas the AAA255-P19 genome contains pyrimidine and purine biosynthesis genes in distinct operons, genes that correspond to these pathways in the other CP genomes are not found in gene clusters (**Table S1.2**).

Given the lack of complete pathways for biosynthesis of some essential metabolites, we examined the possibility that the RAAC CP organisms could scavenge these compounds from their surroundings. The RAAC genomes contain numerous nucleases (**Figure 1.10**) and proteases, as well as several transporters whose substrates are unknown.

Cell-surface and environmental interactions

None of the RAAC1-4 CP organisms appear to make lipid A or lipopolysaccharide, as indicated by the absence of genes for biosynthesis, including LpxC and KdsA (Sutcliffe, 2010), and as suggested by previous work on another TM7 genome (Albertsen et al., 2013a). All genomes except for the WWE3 contain complete, identifiable pathways for peptidoglycan synthesis (**Figure 1.11A**). The abundance of glycosyl transferases in all four genomes, particularly the WWE3, suggests the organisms devote significant energy to production of polysaccharides, glycoproteins and/or a glycosylated S-layer. Additionally, the SR1 and WWE3 genomes contain genes for the synthesis of dTDP-rhamnose (**Figure 1.11A**). The genomes encode proteins containing one or more of the following domains: concanavalins/lectins, pectin lyases, fibronectin III, beta propeller, and polycystic kidney disease (PKD), some of which are predicted to be cell-surface domains in Bacteria and Archaea (**Figure 1.11B**) (Jing et al., 2002). Many of these predicted proteins are large, up to 5900 amino acids in size, and some have signal peptides or sortase motifs suggesting possible cell wall localization.

Sortases, which covalently attach surface proteins to the cell wall of gram-positive bacteria, are present in the SR1, WWE3, and TM7 genomes and predicted sorted proteins were found in the WWE3 and TM7. Each of the four genomes encodes the required components for type IV pili biosynthesis, including pilT (Aukema et al., 2005), for twitching motility and several predicted pilins (**Figure 1.11B**) (Proft and Baker, 2009). The TM7 genome has multiple type II and IV pili-related gene clusters and additional pilins, totaling 60 genes, a full 6% of the TM7 genome. Importantly, the pili present in these CP genomes are not related to the sortase-associated pili more commonly found gram-positive bacteria (Kang and Baker, 2012). Rather, they are homologous to type IV pili sometimes involved in the uptake of environmental DNA (Chen and Dubnau, 2004). Consistent with this possible function, each genome contains at least one copy of ComEC, the DNA-specific pore-forming protein required for competence, and DNA protection protein, DprA (**Figure 1.11B**).

Translation and coding

Codon usage in both the SR1 and OD1 genomes is skewed toward low-GC codons, as expected given the low GC content across these genomes (**Figure 1.12**). Additionally, the SR1 genome uses alternate coding, as reported for another SR1 genome (Campbell et al., 2013). This genome, RAAC1, and the previously reported SR1 genomes (Campbell et al., 2013; Wrighton et al., 2012) contain near-identical genes for $tRNA_{UCA}$, suggesting that the corresponding codon, UGA, is not read as termination but rather as an amino acid. Concordantly, ORF detection using code 11 (bacterial) yielded extremely short sequences and an unreasonably high frequency of split genes (2288 total predicted ORFs with average length of 289 bp), while translation with code 4 (UGA read as tryptophan) gave typical ORF sizes and complete genes (**Table 1.1**). However, unlike code 4, where UGA encodes tryptophan, conserved positions in protein alignments indicate UGA most likely encodes glycine, as described for the oral SR1 genotype by Campbell *et al*. (Campbell et al., 2013). Interestingly, the SR1 genome also harbors duplicated and interrupted tRNA synthetases. Specifically, there is an extra, fragmented copy of both valine- and alanine-tRNA synthetase, and an unusual isoleucine tRNA-synthetase, which appears split into four regions by a nudix hydrolase domain, and two phosphoglycerate mutase domains. The same tRNA synthetases were also found in the ACD80 genome.

The OD1 genome has an extremely small number of UGA stop codons (5% of all ORFs, compared to 22% in the WWE3 and 15% in the TM7) suggesting this codon may be approaching extinction and possible reassignment, leading to alternate coding. The OD1 genome also

contains a predicted suppressor tRNA that reads the codon UAA. However, the majority of predicted genes in this genome use UAA for termination, which suggests that this may be a pseudo-tRNA or may recognize a different codon.

Discussion

Metabolic predictions for all four genomes point to a primarily fermentation-based lifestyle, and an inability to synthesize essential metabolites. However, many predicted ORFs were annotated only at the protein domain level or not at all (**Table 1.1**), and some unannotated proteins may complete metabolic pathways that appear broken or absent in the CP genomes. As more sequence data from the CPs become available, conserved protein families will likely emerge, preparing the way for further exploration at the phylogenetic, biochemical, and structural levels.

The most remarkable feature of all four complete or essentially complete CP genomes is their small size. The first estimate of genome size for a member of the TM7, based on fragmentary single cell sequencing data, was substantially larger (Podar et al., 2007). Our result, a genome size of 0.85 Mbp for RAAC3, parallels the recent finding of a 1.01 Mbp genome for another TM7 organism (Albertsen et al., 2013a). Campbell *et al.* suggested the genome for a member of the SR1 was less than 2 Mbp (Campbell et al., 2013), somewhat larger than the genome size of 1.17 Mbp for RAAC1. The expectation of small genome size for OP11 and OD1 genomes was also noted by Wrighton *et al*. (Wrighton et al., 2012). Together, data presented here and recently published results suggest that small genome size is common across multiple phyla.

Genomes of the small sizes reported here are found in some free-living marine bacteria and in obligate symbionts (**Figure 1.5**), both of which may be descended from organisms with larger genomes. Mechanisms for genome reduction in free-living bacteria include streamlining (e.g., decreased intergenic distance and loss of non-essential genes and pathways) or metabolic specialization, and in obligate symbionts, directed loss of genes whose functions are provided by the host (Giovannoni et al., 2005; Grote et al., 2012; McCutcheon and Moran, 2012). No close relatives with larger genomes have been sequenced to date and small genome size may indeed be an ancestral trait for these CP.

Examination of pseudogenes can sometimes reveal the evolutionary trajectory of bacteria with reduced genome size. Genome erosion and an accumulation of pseudogenes is characteristic of the early stages of evolving symbiosis in bacteria (McCutcheon and Moran, 2012), whereas an elimination of pseudogenes is suggestive of genomic streamlining (Grote et al., 2012) or later stages of symbiosis (McCutcheon and Moran, 2012). However, until more closely-related CP genomes are sequenced, it will be difficult to determine which unannotated genes are truly pseudogenes and which serve novel functions in certain lineages. The coding density for the four CP genomes (around 0.90; **Table 1.1**) is lower than that reported for some organisms that have undergone genomic streamlining (e.g. 0.97 in *Prochlorococcus marinus* and *Pelagibacter ubique*) (Dufresne et al., 2003; Giovannoni et al., 2005) but higher than the value for an obligate symbiont, where streamlining was not a mechanism for genome reduction (0.81 in *Candidatus* Atelocyanobacterium thalassa) (Thompson et al., 2012; Tripp et al., 2010).

Genome reduction has been suggested as a driving factor in the switch to alternate coding in some symbiotic Alphaproteobacteria and mitochondria (Knight et al., 2001; McCutcheon et al., 2009). In these groups, which use code 4, a single $tRNA^{Trp}$ has mutated to accommodate both UGG and UGA via wobble-pairing, allowing elimination of tRNA and termination factor genes

(Knight et al., 2001; McCutcheon et al., 2009). While there is currently no consistent bioinformatic method for defining tRNA-to-amino acid specificity (Perona and Hadd, 2012), protein alignments and biochemical evidence (Campbell et al., 2013) are convincing arguments for the recoding of UGA to glycine in all SR1 genomes reported thus far, and this coding in RAAC1 supports the suggestion by Campbell *et al*. that this may be a phylum-level trait (Campbell et al., 2013). In contrast to the code 4 organisms mentioned above, the SR1 genomes do not appear to use wobble-pairing but rather an additional $tRNA^{Gly}_{UCA}$ to achieve alternate coding (Campbell et al., 2013). Use of UGA for glycine could be a mechanism for reducing genomic GC content, as all other glycine codons are more GC-rich (**Figure 1.12**).

The CP genomes appear smaller than those of typical free-living bacteria at least in part due to missing metabolic functions. Recently, McLean *et al.* noted a similarly small genome size for a member of the TM6 (McLean et al., 2013), and suggested it may be a symbiont. Our analysis suggests this may also be true for some or all of the organisms reported here. If several core biosynthetic pathways are in fact absent in the CP bacteria described here, the organisms likely rely heavily on one (possibly a member of the Tenericutes in the case of the SR1) or more community members in a manner similar to symbiosis. Type IV pili, encoded by all four genomes, may aid the cells in interacting with the environment and with other organisms via adhesion to extracellular surfaces, DNA-uptake, and biofilm formation (Proft and Baker, 2009). Other adhesion- or biofilm-related proteins may also be important to the life-strategies of these organisms. Transporters, nucleases and proteases could allow the organisms to make use of metabolites provided by biomass in their environment or by a host. A potential dependence on other organisms may explain why these CP bacteria remain uncultured.

Methods

Field experiment

The experimental conditions and sample collection in the field have been described in detail by Handley *et al*., (2015). The present study focuses on acetate-amended sediment, collected between August and November of 2010. Thirteen flow-through columns packed with sieved (< 2 mm) sediment were placed into one of three wells at the US Department of Energy's (DOE) Integrated Field Research Challenge (IFRC) aquifer in Rifle, CO, USA. Columns were equilibrated to conditions in the subsurface for one week. Subsequently, groundwater amended with 10 mM sodium acetate was pumped upward through the columns at an approximate rate of 52 ml day^{-1}. Individual columns were sacrificed for sampling between 13 and 63 days of amendment such that a range of geochemical conditions from iron reduction to sulfate reduction was sampled. Un-amended sieved sediment was taken as a background sample. Geochemical measurements of filtered column effluent were made on or the day prior to column sacrifice. Aqueous ferrous iron and sulfide were quantified immediately following collection of effluent using the colorimetric assays 1,10-Phenanthroline and Methylene Blue (Hach Company, Loveland, CO). Sulfate was measured with ion chromatography (IC, ICS-2100, Dionex, Sunnyvale, CA, fitted with an AS-18 guard and analytical column).

DNA extraction and sequencing

DNA was extracted from an average of 25 ± 7g of acetate-amended sediment (samples 2-14) and 42.1g of sieved background sediment using PowerMax Soil DNA Isolation Kits (MoBio Laboratories, Inc., Carlsbad, CA, USA) with the following modification to the manufacturer's

instructions. Sediment was vortexed for an additional 3 minutes in SDS at maximum speed, and then incubated for 30 minutes at 65°C in place of bead beating. Extracted DNA was precipitated with cold ethanol, Na-Acetate (0.3 M, pH 5.2) and glycogen (50 µg/ml), and re-eluted in 50 µl EB. Illumina sequencing was conducted at UC Davis DNA Technologies Core Facility (http://dnatech.genomecenter.ucdavis.edu) using paired-end 101 bp reads with insert size of 500 bp. Sequencing was distributed across 5 lanes to produce between 3 and 6 Gbp of sequence for each of the 13 amended samples and 15 Gbp for the background sample.

Metagenomic assembly and curation

Sequence datasets for each sample were assembled independently using idba_ud with default parameters (Peng et al., 2012). This generated 61.0 - 107.1 Mbp of sequence information per sample and 16.8 Mbp for background sediment on scaffolds > 5 kb. Genes on all scaffolds > 5 kb in length were predicted with Prodigal using the metagenome option (Hyatt et al., 2010; 2012). Scaffolds for which Prodigal chose code 4 (UGA translated as tryptophan instead of termination) were manually curated into a genome identified as belonging to the SR1. The genome was retranslated with UGA encoding glycine (see Results). For each scaffold, we determined the GC content, coverage, genetic code, and the profile of phylogenetic affiliation based on the best hit for each gene to Uniref90 (Suzek et al., 2007). Based on analyses of these data, as well as ESOM-based analyses of tetranucleotide frequencies and time-series relative abundance (Dick et al., 2009; Sharon et al., 2013), draft genomes were generated that included scaffolds from multiple samples. Scaffolds for the same genome found in different samples were aligned to yield longer fragments, leveraging the observation that fragmentation of assemblies is, to some extent, dependent on context (community composition). Read-mapping used Bowtie (Langmead et al., 2009), and paired-read information was used to extend and join contigs and to fill in gaps left by the assembler (Sharon et al., 2013). A few regions, particularly those containing short repeats (a few hundred bases or less), could not be completely resolved but connectivity of their two sides was confirmed.

Functional annotation

Predicted ORFs were run through a multi-database search pipeline for functional prediction, as described (Wrighton et al., 2012). Briefly, sequence similarity searches were performed using Usearch (Edgar, 2010) against UniRef90 (Suzek et al., 2007), KEGG (Kyoto University Bioinformatics Center). Domain-level functional annotation was done using InterproScan (Zdobnov and Apweiler, 2001). RNA was predicted using a combination of database searching and tRNAScan-SE (Lowe and Eddy, 1997) for tRNAs. Cellular localization was predicted with PSORTb (version 3.0) (Yu et al., 2010) and detection of the sortase cleavage motif, LPXTG (Comfort and Clubb, 2004), with in-house scripts.

Phylogenetic analysis

Genomes were placed into phylogenetic context based on analysis of the 16S rRNA gene sequences. Sequences were aligned to the SILVA database using the SINA alignment service (http://www.arb-silva.de/aligner) (Pruesse et al., 2012). Representative sequences from TM7, PER, BD1-5, SR1, OD1, WS6, WWE3, and OP11 were obtained from SILVA in aligned form (see accession numbers provided). Conserved gaps were removed from compiled aligned sequences using GapStreeze v.2.1.0, with gap tolerance set to 99% (http://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html). The alignment was

further trimmed to remove uninformative regions. The maximum likelihood tree was constructed with RAxML using the GTRCAT model with 1000 bootstraps. For the WWE3, which is not recognized as a phylum by SILVA and is called "otu_4443" in Greengenes (accessed June 2013), the SINA aligner was used to find sequences with > 80% identity to the genomic 16S rRNA gene sequence. These sequences were included on a 16S rRNA gene tree containing multiple CP and formed a phylum-level branch.

Overall community composition

Percent relative abundances of genomes in each of the 13 samples and the background sediment sample were calculated based on mapping of unpaired reads from each sample against each genome using Bowtie2 under the following settings: --phred33 and --fast with default specificity (version 2.0.4) (Langmead and Salzberg, 2012). Community composition by sample was determined with EMIRGE (Miller et al., 2011) using 80 iterations and SILVA 108 clustered at 97% identity as the reference database. Chimeras were detected using Uchime (Edgar, 2010). Reconstructed 16S rRNA gene sequences were analyzed with RDP-classifier (http://rdp.cme.msu.edu/classifier/classifier.jsp) and relative abundances were calculated by EMIRGE based on read mapping normalized for sequence length. OTU abundances were summed within each phylogenetic order represented (Proteobacteria were summed at the class level). All orders with abundance < 5% in all samples were included in the category "other" (**Figure 1.3**).

Novel protein analysis

An original database and an updated database of Hidden Markov Models (HMMs) representing Sifting Families (Sharpton et al., 2012) were obtained from http://edhar.genomecenter.ucdavis.edu/sifting_families. These were compiled into searchable form using hmmpress (Janelia Farm, 2010, h3.0). Amino acid sequences from each genome were searched against the database using hmmsearch with a reporting cut-off of 1E-5 and parsed with an alignment coverage threshold of 80% for both the HMM and the query gene.

Metabolic pathways analysis

Initial analysis relied upon gene annotations from an in-house pipeline (above) with functional residues confirmed in proteins of interest. Subsequently, amino acid sequences were submitted to KAAS (http://www.genome.jp/kaas-bin/kaas_main?mode=partial) (Moriya et al., 2007) using a customized search list of diverse Bacteria and Archaea (KAAS IDs: pfa, eco, son, cje, gme, sme, rsp, mtu, bsu, cac, ctr, bfr, fjo, emi, cau, tma, mja, afu, pho, tac, ape, sso, pai, tne, tko, pab, pfu, mma, aae, dra, det, cte, pma, syw, fnu, fsu, cao, sru, lil, fra). Searches were run independently in both bi-directional and single-directional best-hit modes. Additional searches for specific genes (**Figure 1.8 and Table S1.1**) were conducted by generating a diverse reference set from 75 bacterial and archaeal genomes in the IMG database (Markowitz et al., 2012), and using these as queries for BLAST (Altschul et al., 1990) to search for potential homologs within the CP genomes.

Nucleotide and amino acid sequences

All sequences and annotations can be accessed at genegrabber.berkeley.edu/aac. Sequences are also available at NCBI through bioproject numbers PRJNA217185 (RAAC1), PRJNA217183 (RAAC2), PRJNA217186 (RAAC3), and PRJNA216121 (RAAC4).

Figures and Tables

Table 1.1. Genome information for the four CP genomes.

Genome	RAAC1 (SR1)	RAAC2 (WWE3)	RAAC3 (TM7)	RAAC4 (OD1)
Completeness	Circular Closed	Circular Closed	Circular 8 gaps	Not circular 2 gaps
Length (bp)	1,177,827	878,137	845,953	693,134
GC content	0.31	0.43	0.49	0.31
Average ORF length (bp)	1011	926	825	906
Average intergenic distance (bp)*	92	66	72	86
Protein coding density	0.91	0.92	0.91	0.9
tRNA count	37	45	43	42
Protein-coding gene count	1060	874	931	686
ORFs with predicted function	596 (56%)	618 (71%)	704 (76%)	504 (73%)
Domain-only prediction	115 (11%)	49 (6%)	38 (4%)	33 (5%)
Conserved hypothetical	106 (10%)	121 (14%)	145 (16%)	96 (14%)

*Average intergenic distance was calculated including all non-zero distances between protein- and RNA-coding sequences.

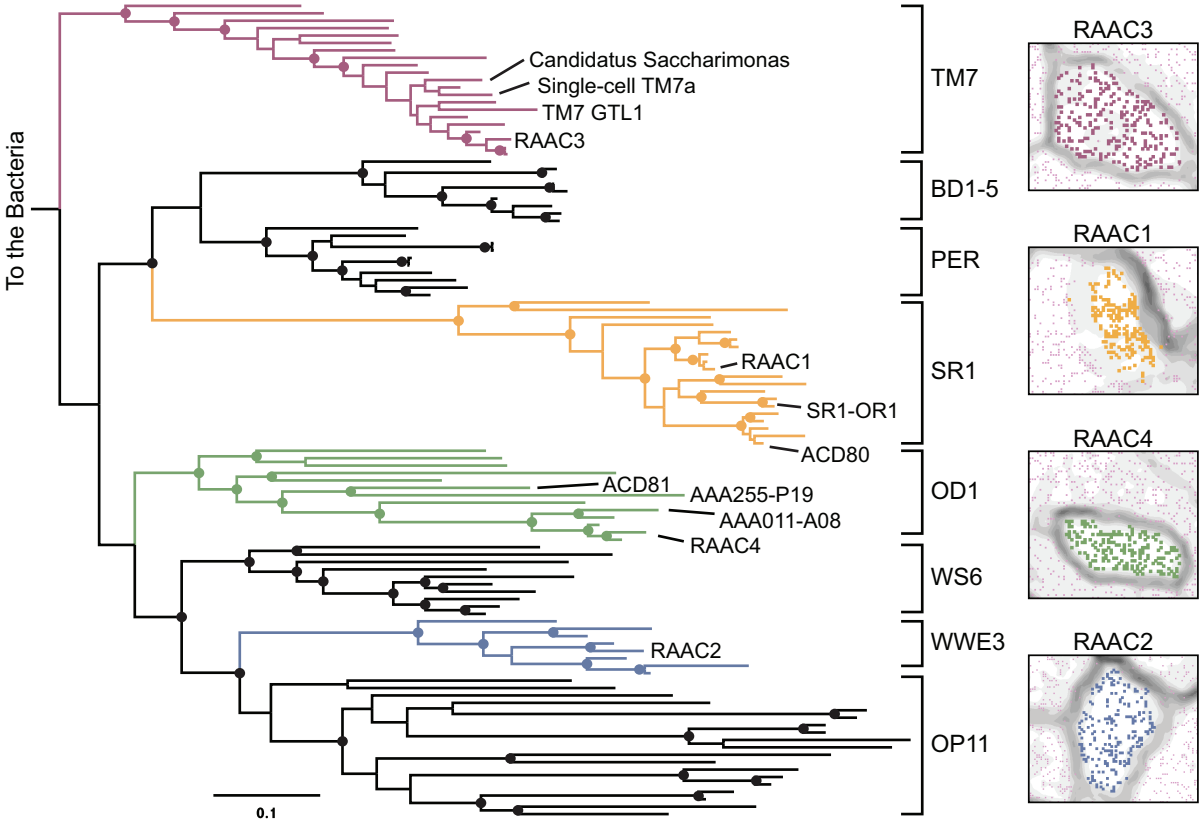

Figure 1.1. Complete and near-complete genomes from four candidate phyla (called RAAC1-4) were reconstructed. Left: Phylogeny of selected CP based on 16S rRNA genes. The maximum likelihood tree was constructed from an alignment containing 111 taxa and 1565 unambiguously aligned positions. Bootstrap values greater than 80% are displayed as filled circles (for accession numbers, see Figure S1). Also noted are previously sequenced partial and complete genomes for related members of the CP for which full-length 16S rRNA gene sequences were available

(Albertsen et al., 2013a; Campbell et al., 2013; Marcy et al., 2007; Podar et al., 2007; Rinke et al., 2013; Wrighton et al., 2012). Right: emergent self-organizing maps (ESOMs) made using differential coverage across the 14 sediment columns confirmed genome binning.

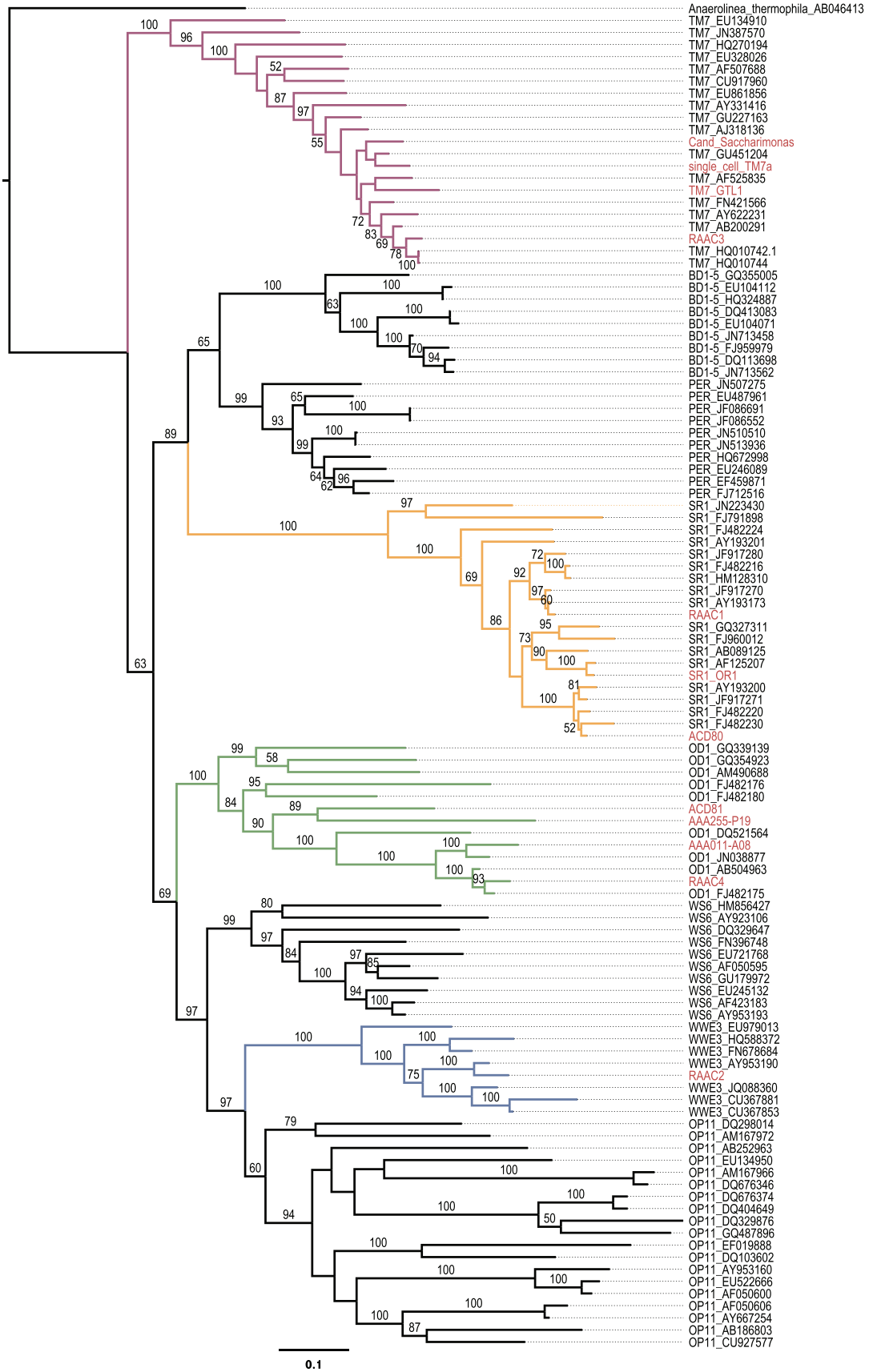
Anaerolinea_thermophila_AB046413
TM7_EU134910
TM7_JN387570
TM7_HQ270194
TM7_EU328026
TM7_AF507688
TM7_CU917960
TM7_EU861856
TM7_AY331416
TM7_GU227163
TM7_AJ318136
Cand_Saccharimonas
TM7_GU451204
single_cell_TM7a
TM7_AF525835
TM7_GTL1
TM7_FN421566
TM7_AY622231
TM7_AB200291
RAAC3
TM7_HQ010742.1
TM7_HQ010744
BD1-5_GQ355005
BD1-5_EU104112
BD1-5_HQ324887
BD1-5_DQ413083
BD1-5_EU104071
BD1-5_JN713458
BD1-5_FJ959979
BD1-5_DQ113698
BD1-5_JN713562
PER_JN507275
PER_EU487961
PER_JF086691
PER_JF086552
PER_JN510510
PER_JN513936
PER_HQ672998
PER_EU246089
PER_EF459871
PER_FJ712516
SR1_JN223430
SR1_FJ791898
SR1_FJ482224
SR1_AY193201
SR1_JF917280
SR1_FJ482216
SR1_HM128310
SR1_JF917270
SR1_AY193173
RAAC1
SR1_GQ327311
SR1_FJ960012
SR1_AB089125
SR1_AF125207
SR1_OR1
SR1_AY193200
SR1_JF917271
SR1_FJ482220
SR1_FJ482230
ACD80
OD1_GQ339139
OD1_GQ354923
OD1_AM490688
OD1_FJ482176
OD1_FJ482180
ACD81
AAA255-P19
OD1_DQ521564
AAA011-A08
OD1_JN038877
OD1_AB504963
RAAC4
OD1_FJ482175
WS6_HM856427
WS6_AY923106
WS6_DQ329647
WS6_FN396748
WS6_EU721768
WS6_AF050595
WS6_GU179972
WS6_EU245132
WS6_AF423183
WS6_AY953193
WWE3_EU979013
WWE3_HQ588372
WWE3_FN678684
WWE3_AY953190
RAAC2
WWE3_JQ088360
WWE3_CU367881
WWE3_CU367853
OP11_DQ298014
OP11_AM167972
OP11_AB252963
OP11_EU134950
OP11_AM167966
OP11_DQ676346
OP11_DQ676374
OP11_DQ404649
OP11_DQ329876
OP11_GQ487896
OP11_EF019888
OP11_DQ103602
OP11_AY953160
OP11_EU522666
OP11_AF050600
OP11_AF050606
OP11_AY667254
OP11_AB186803
OP11_CU927577
0.1

Figure 1.2. Phylogeny of selected CP using 16S rRNA genes including NCBI accession numbers for the tree shown in Figure 1. Bipartition values from 1000 bootstraps are expressed as percentages for all bootstraps greater than 50%. The 16S rRNA genes from genomes in this study and other genomic sequences are in red. Previously reported TM7 single-cell genomic sequences are represented by TM7 genomosp. GTL1 (accession AAXS01000094) (20), single cell TM7a (accession NZ_ABBV01000356) (Marcy et al., 2007) and *Candidatus* Saccharimonas aalborgensis (NC_021219) (12) and are 81% and 85% identical to RAAC3 at the 16S rRNA gene level, respectively. The SR1 sequences ACD80 and SR1-OR1 are 87% and 89% identical to RAAC1, respectively. The OD1 sequences AAA011-A08 and AAA255-P19 represent partial genomes described in (Rinke et al., 2013) and accessed through IMG (Markowitz et al., 2012). RAAC4 shares 75% sequence identity with AAA011-A08, its nearest relative with a genomic sequence.

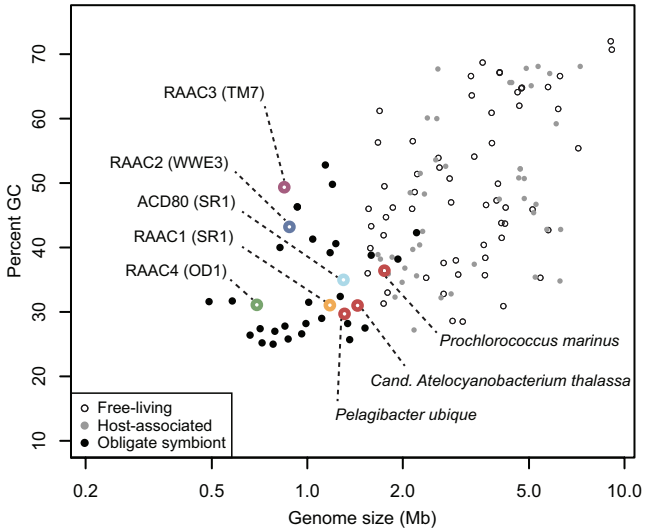

Figure 1.3. GC content versus genome size for the RAAC1-4 genomes with reference data from NCBI. Biotic relationship categories are as in Giovannoni *et al*. (Giovannoni et al., 2005). Open red circles represent marine bacteria with small genomes (left to right): *Pelagibacter ubique* (free-living), Candidatus Atelocyanobacterium thalassa (UCYN-A, a symbiont), and *Prochlorococcus marinus* (free-living).

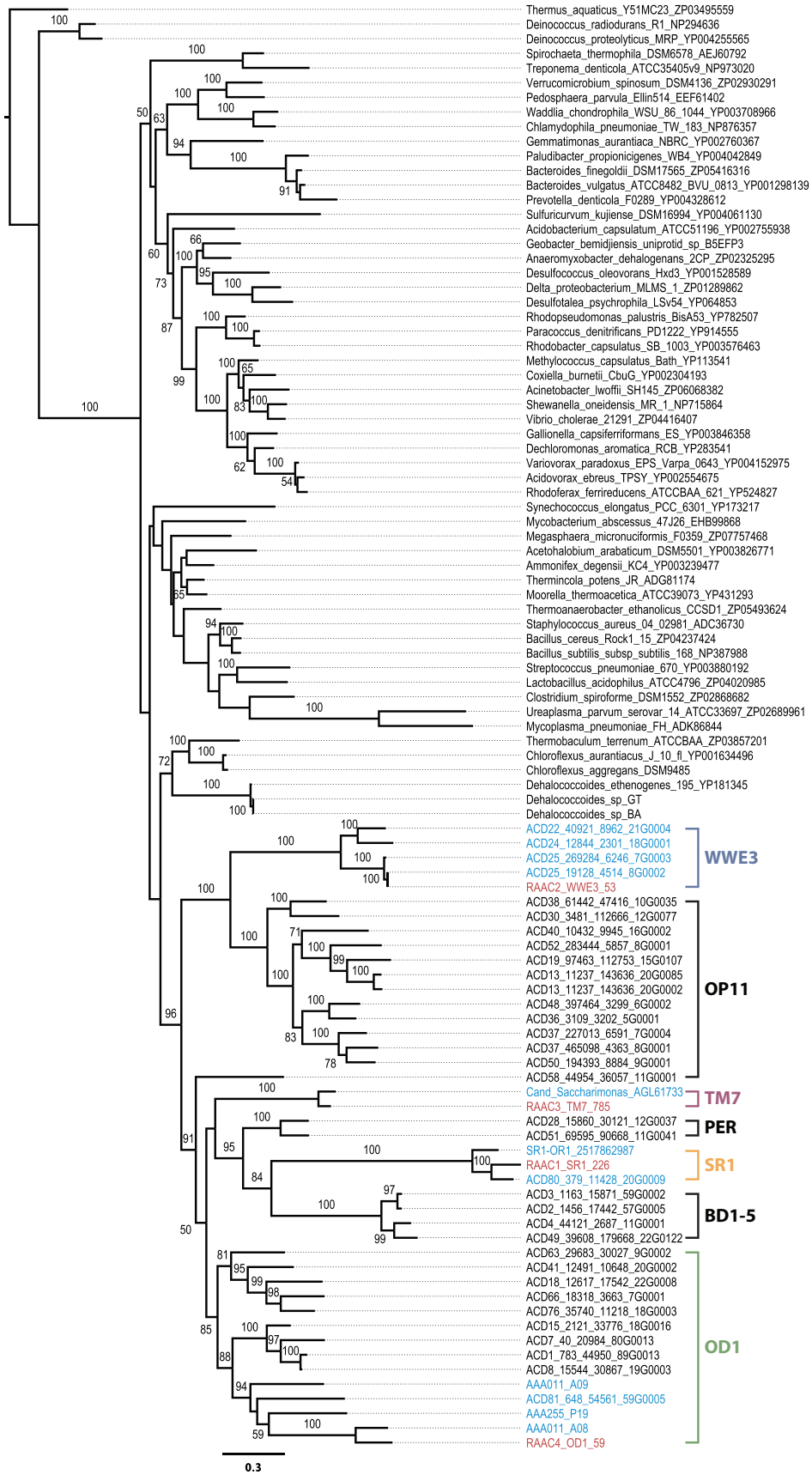
Thermus_aquaticus_Y51MC23_ZP03495559
Deinococcus_radiodurans_R1_NP294636
Deinococcus_proteolyticus_MRP_YP004255565
Spirochaeta_thermophila_DSM6578_AEJ60792
Treponema_denticola_ATCC35405v9_NP973020
Verrucomicrobium_spinosum_DSM4136_ZP02930291
Pedosphaera_parvula_Ellin514_EEF61402
Waddlia_chondrophila_WSU_86_1044_YP003708966
Chlamydophila_pneumoniae_TW_183_NP876357
Gemmatimonas_aurantiaca_NBRC_YP002760367
Paludibacter_propionicigenes_WB4_YP004042849
Bacteroides_finegoldii_DSM17565_ZP05416316
Bacteroides_vulgatus_ATCC8482_BVU_0813_YP001298139
Prevotella_denticola_F0289_YP004328612
Sulfuricurvum_kujiense_DSM16994_YP004061130
Acidobacterium_capsulatum_ATCC51196_YP002755938
Geobacter_bemidjiensis_uniprotid_sp_B5EFP3
Anaeromyxobacter_dehalogenans_2CP_ZP02325295
Desulfococcus_oleovorans_Hxd3_YP001528589
Delta_proteobacterium_MLMS_1_ZP01289862
Desulfotalea_psychrophila_LSv54_YP064853
Rhodopseudomonas_palustris_BisA53_YP782507
Paracoccus_denitrificans_PD1222_YP914555
Rhodobacter_capsulatus_SB_1003_YP003576463
Methylococcus_capsulatus_Bath_YP113541
Coxiella_burnetii_CbuG_YP002304193
Acinetobacter_lwoffii_SH145_ZP06068382
Shewanella_oneidensis_MR_1_NP715864
Vibrio_cholerae_21291_ZP04416407
Gallionella_capsiferriformans_ES_YP003846358
Dechloromonas_aromatica_RCB_YP283541
Variovorax_paradoxus_EPS_Varpa_0643_YP004152975
Acidovorax_ebreus_TPSY_YP002554675
Rhodoferax_ferrireducens_ATCCBAA_621_YP524827
Synechococcus_elongatus_PCC_6301_YP173217
Mycobacterium_abscessus_47J26_EHB99868
Megasphaera_micronuciformis_F0359_ZP07757468
Acetohalobium_arabaticum_DSM5501_YP003826771
Ammonifex_degensii_KC4_YP003239477
Thermincola_potens_JR_ADG81174
Moorella_thermoacetica_ATCC39073_YP431293
Thermoanaerobacter_ethanolicus_CCSD1_ZP05493624
Staphylococcus_aureus_04_02981_ADC36730
Bacillus_cereus_Rock1_15_ZP04237424
Bacillus_subtilis_subsp_subtilis_168_NP387988
Streptococcus_pneumoniae_670_YP003880192
Lactobacillus_acidophilus_ATCC4796_ZP04020985
Clostridium_spiroforme_DSM1552_ZP02868682
Ureaplasma_parvum_serovar_14_ATCC33697_ZP02689961
Mycoplasma_pneumoniae_FH_ADK86844
Thermobaculum_terrenum_ATCCBAA_ZP03857201
Chloroflexus_aurantiacus_J_10_fl_YP001634496
Chloroflexus_aggregans_DSM9485
Dehalococcoides_ethenogenes_195_YP181345
Dehalococcoides_sp_GT
Dehalococcoides_sp_BA
ACD22_40921_8962_21G0004
ACD24_12844_2301_18G0001
ACD25_269284_6246_7G0003
ACD25_19128_4514_8G0002
RAAC2_WWE3_53
WWE3
ACD38_61442_47416_10G0035
ACD30_3481_112666_12G0077
ACD40_10432_9945_16G0002
ACD52_283444_5857_8G0001
ACD19_97463_112753_15G0107
ACD13_11237_143636_20G0085
ACD13_11237_143636_20G0002
ACD48_397464_3299_6G0002
ACD36_3109_3202_5G0001
ACD37_227013_6591_7G0004
ACD37_465098_4363_8G0001
ACD50_194393_8884_9G0001
ACD58_44954_36057_11G0001
OP11
Cand_Saccharimonas_AGL61733
RAAC3_TM7_785
TM7
ACD28_15860_30121_12G0037
ACD51_69595_90668_11G0041
PER
SR1-OR1_2517862987
RAAC1_SR1_226
ACD80_379_11428_20G0009
SR1
ACD3_1163_15871_59G0002
ACD2_1456_17442_57G0005
ACD4_44121_2687_11G0001
ACD49_39608_179668_22G0122
BD1-5
ACD63_29683_30027_9G0002
ACD41_12491_10648_20G0002
ACD18_12617_17542_22G0008
ACD66_18318_3663_7G0001
ACD76_35740_11218_18G0003
ACD15_2121_33776_18G0016
ACD7_40_20984_80G0013
ACD1_783_44950_89G0013
ACD8_15544_30867_19G0003
AAA011_A09
ACD81_648_54561_59G0005
AAA255_P19
AAA011_A08
RAAC4_OD1_59
OD1
0.3

Figure 1.4. Phylogenetic tree of the DNA-directed RNA polymerase subunit beta protein of selected organisms including CP with partial or complete genome sequences. The tree was constructed with RAxML using 1000 bootstrap replicates from an alignment of 99 taxa with 900 aligned positions. Sequences from genomes in this study are shown in red while sequences from genomes used for comparison are shown in blue. Reference sequences and "ACD" sequences were taken from Wrighton *et al.* (2012), other CP sequences in blue are from public databases or IMG (Albertsen et al., 2013a; Campbell et al., 2013; Rinke et al., 2013).

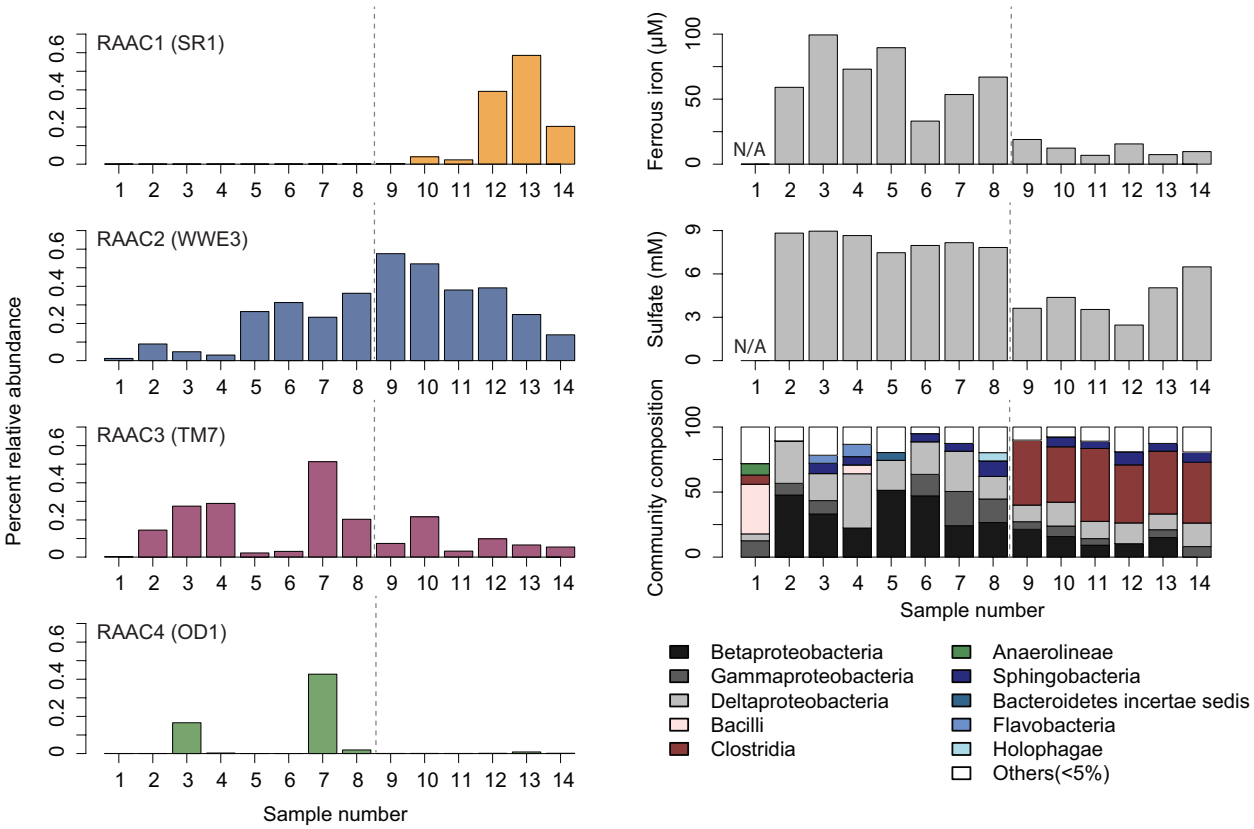

Figure 1.5. Relative abundance of the RAAC1-4 genomes varies across samples representing a range of geochemical conditions and microbial community compositions. Left: percent relative abundance (y-axis) across the 14 independent sediment columns (x-axis). All four genomes were found at less than 1% relative abundance in every sample. Right: endpoint ferrous iron and sulfate measured in column effluent and community composition (for discussion see Handley *et al.* 2015). Sample 1 represents unamended background sediment, and the dashed line roughly divides samples from columns undergoing iron reduction (2-8) versus sulfate reduction (9-14). Samples 13 and 14 were taken after peak sulfate reduction had occurred.

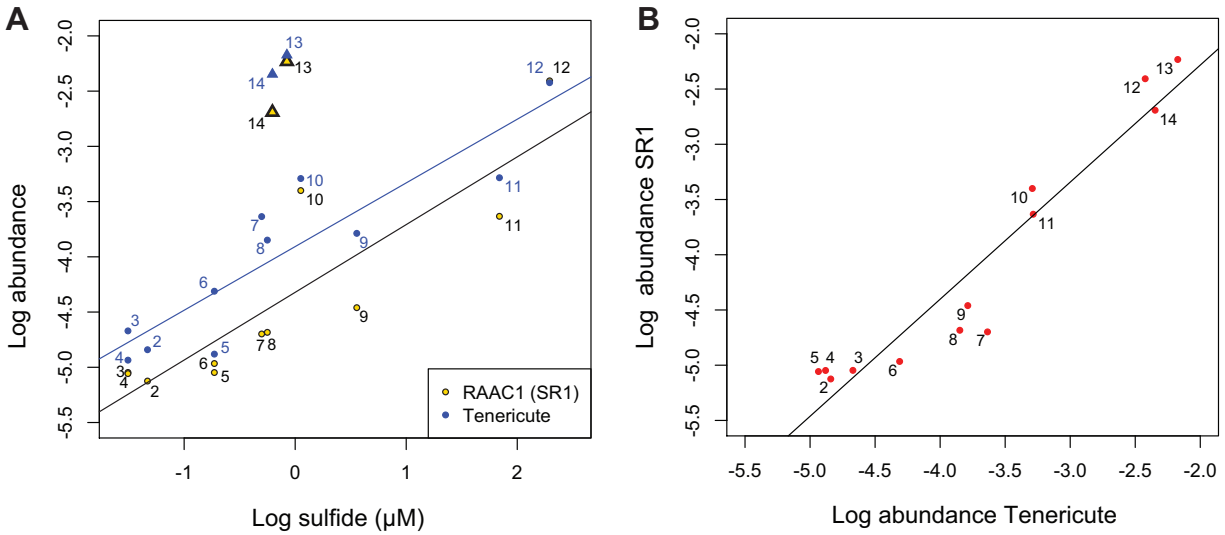

Figure 1.6. Patterns of SR1 relative abundance in relation to sulfide and relative abundance other organisms. (A) Log relative abundances of both RAAC1 (SR1) and a member of the Tenericutes correlate with log sulfide concentration (μM) in each sample until peak sulfate reduction. The two outliers, samples 13 and 14, were taken after peak sulfate reduction, when SR1 abundance was highest and were not included in the regression. (B) The log relative abundances of the SR1 and Tenericutes species are correlated when all samples are included (adjusted $R^2 = 0.9$). For all regressions $p < 0.0005$.

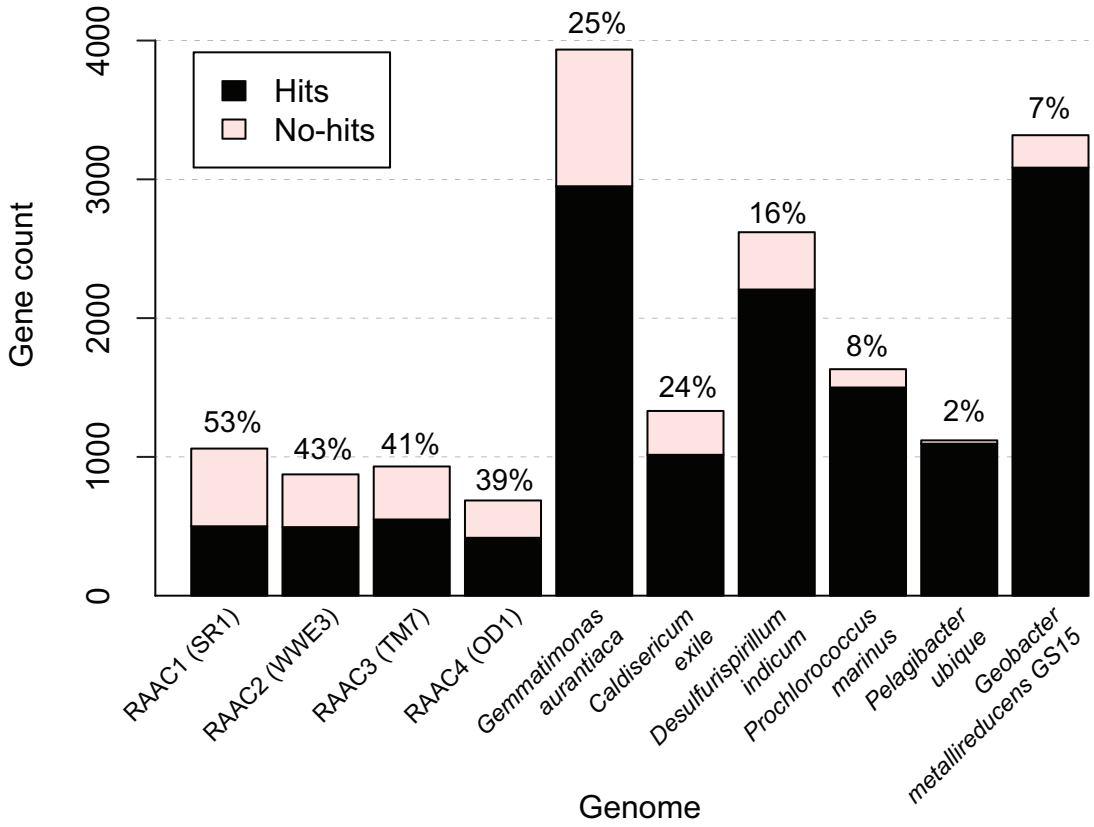

Figure 1.7. Summary of hmmsearch results against the Sfam database (29). CP genomes from this study are RAAC1-4. For comparison are 3 phyla for which only one sequenced genome

exists (*Gemmatimonas aurantiaca*, *Caldisericum exile*, and *Desulfurispirillum indicum*), two small genomes (of approximately comparable size to the CP genomes) with well-sequenced relatives are represented by *Prochlorococcus marinus* and *Pelagibacter ubique*, and a representative of a well-sequenced group with a larger genome, *Geobacter metallireducens* GS15. For each genome, the percentage of predicted proteins with no hits to the SFam database is denoted above the column). Genomes not from this study were obtained from NCBI-genomes (http://www.ncbi.nlm.nih.gov/genome/) and JGI-IMG databases (http://img.jgi.doe.gov/cgi-bin/w/main.cgi).

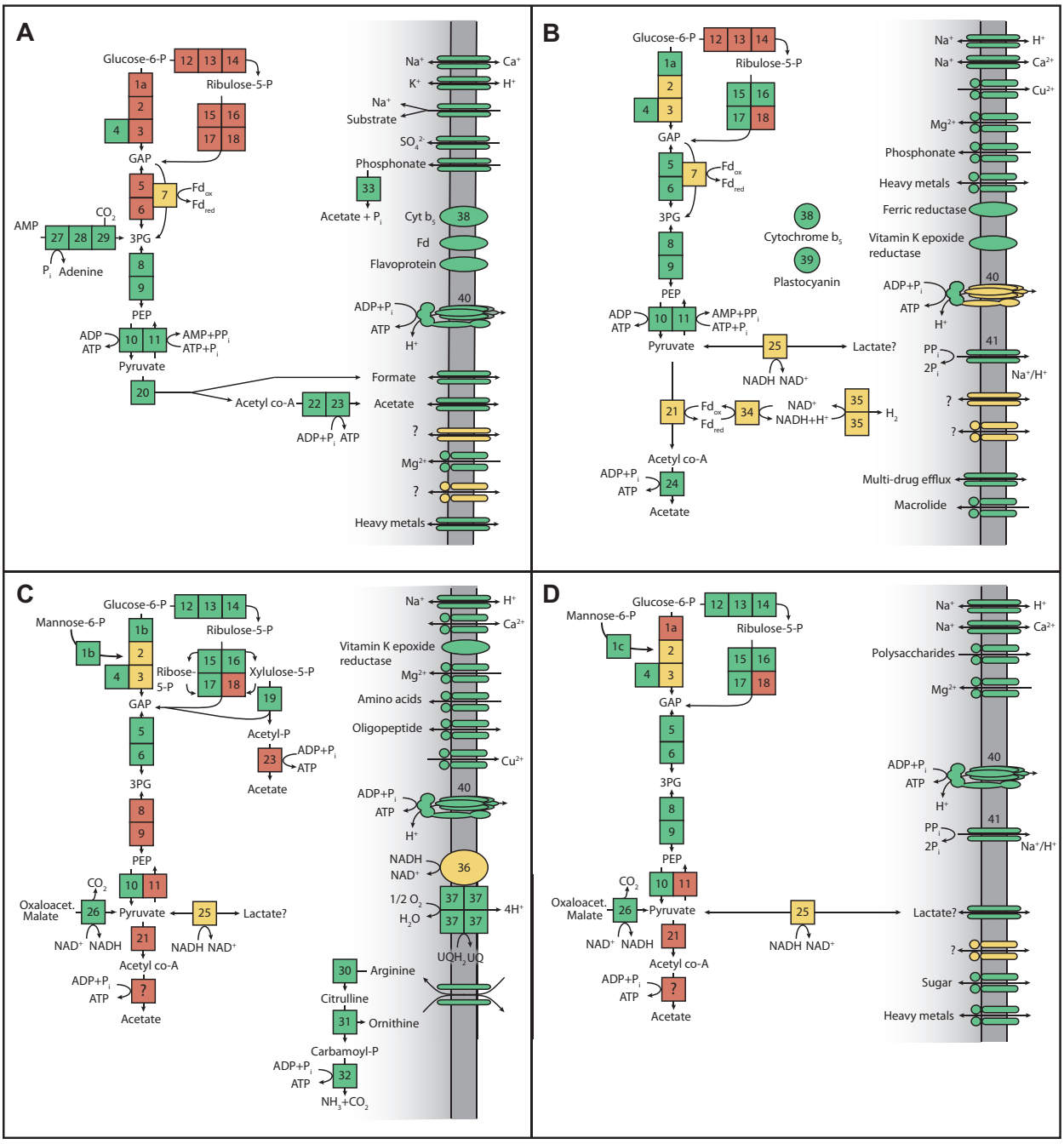

Figure 1.8. Cell diagrams depicting central carbon metabolism, proteins putatively involved in electron transfer, and transporters in the CP genomes: A) RAAC1, SR1; B) RAAC2, WWE3; C) RAAC3, TM7; and D) RAAC4, OD1. Boxes represent enzymes and are colored as follows: green, identified; yellow, homology unclear; red, not identified. Numbers correspond to enzymes listed in Table S1. Transporters: if unmarked, parallel ovals represent members of the major facilitator superfamily while circles and ovals together indicate predicted ABC transporters.

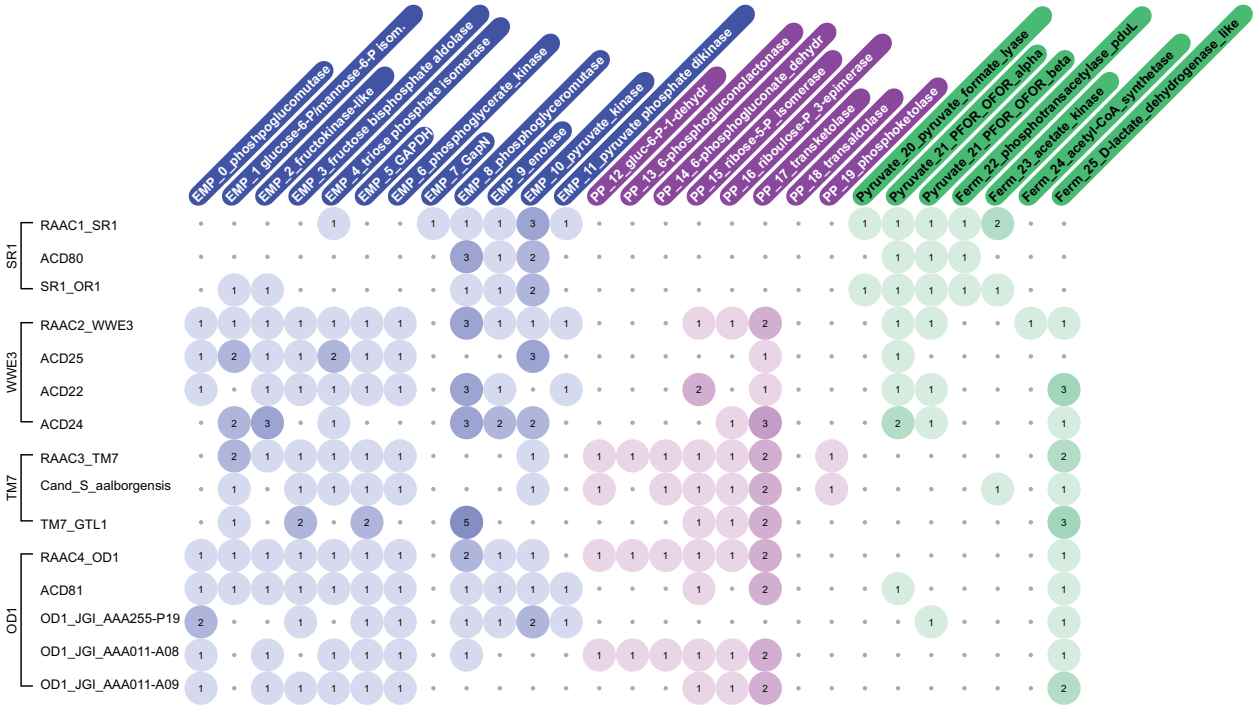

Figure 1.9. Comparison of the Embden Meyerhoff Parnas (EMP) pathway (blue), pentose phosphate pathway (purple), and fermentation pathways (green), across multiple complete and partial CP genomes and the complete RAAC1-4 genomes. We note that all SR1 genomes appear to be missing the upper EMP pathway and do not contain the pentose phosphate pathway. The TM7 genomes lack identifiable genes for enolase. The OD1 lack clear means for making acetyl-CoA from pyruvate.

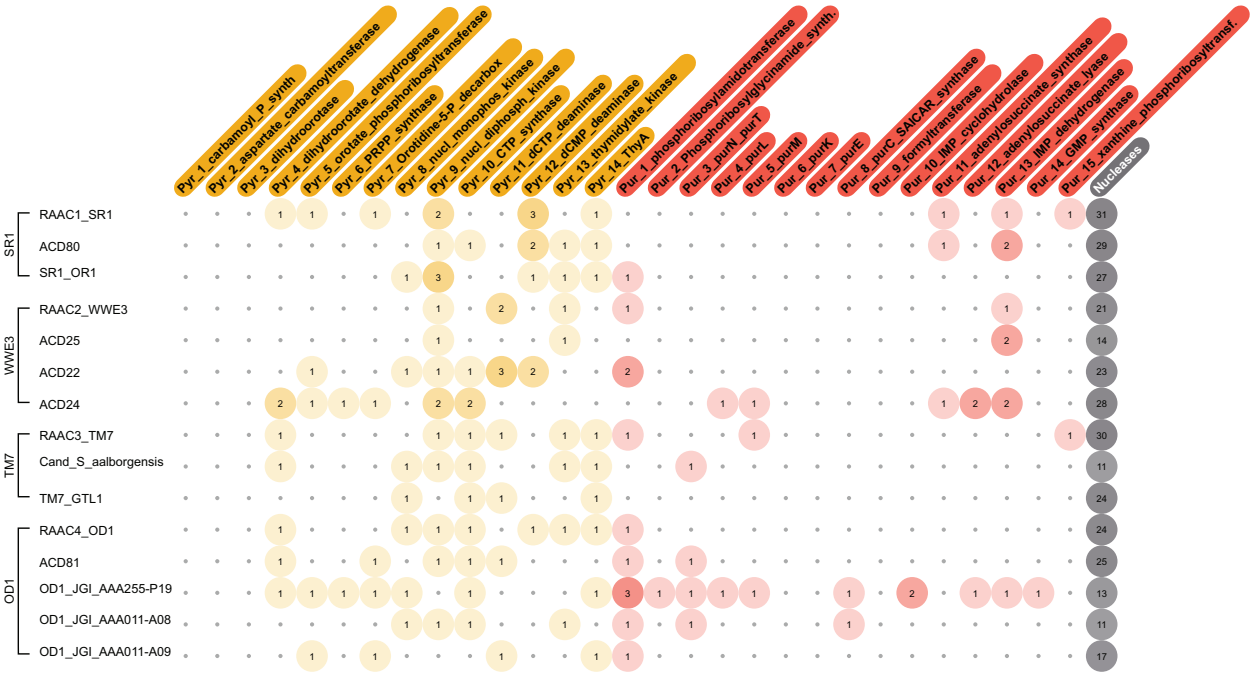

Figure 1.10. Comparison of pyrimidine (yellow) and purine (red) biosynthesis across multiple partial CP genomes and the complete RAAC1-4 genomes. Only the genome AAA255-P19 (Rinke et al., 2013)appears to have near-complete pathways for nucleotide biosynthesis. Presence or absence of this metabolism appears to vary across the CP. The large number of nucleases present in the CP genomes could provide an alternate source of nucleotides. Note that ACD22, ACD24, and ACD25 may contain multiple copies of genes due to multiple strains or species within a bin, although these genomes are incomplete.

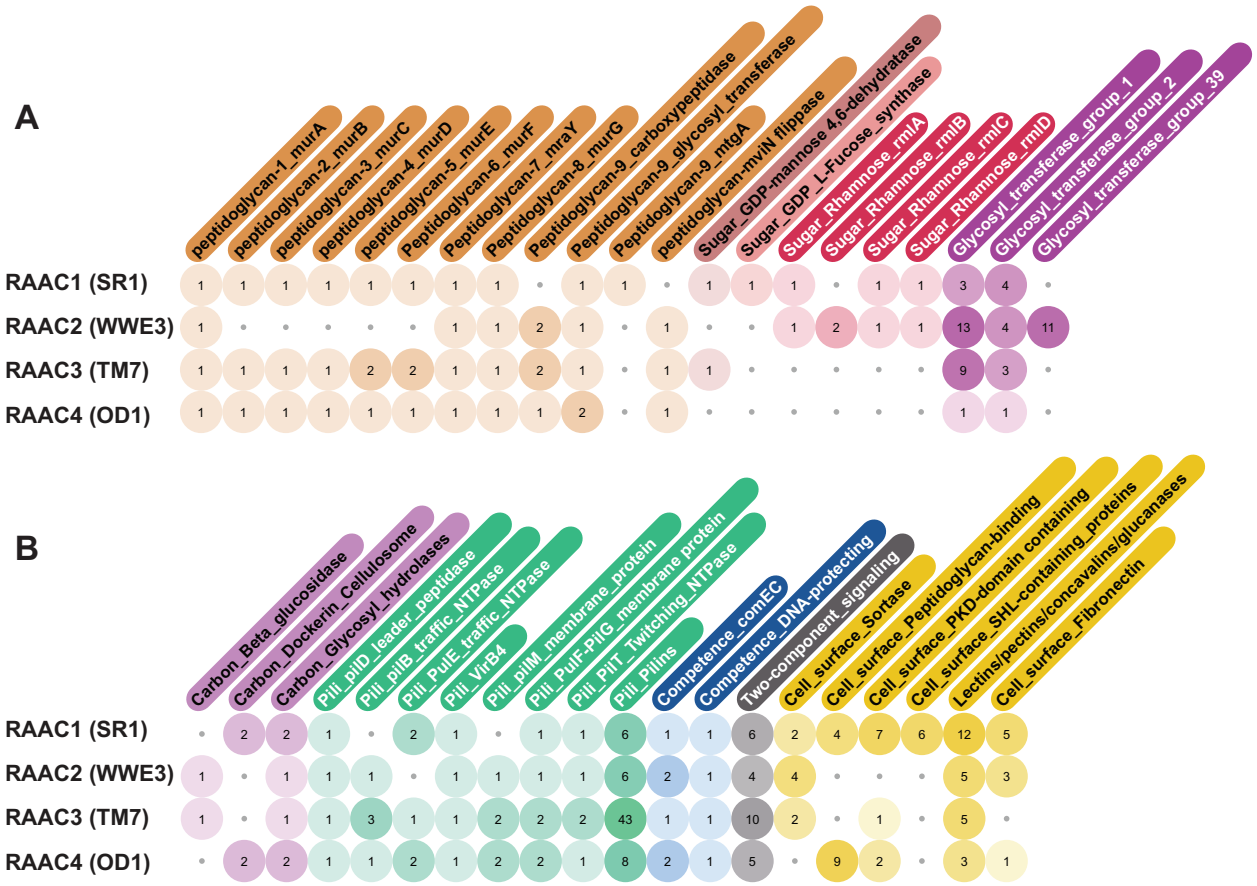

Figure 1.11. Counts by genome of proteins involved in (A) cell wall and cell surface biosynthesis and (B) cell-environment interactions. Yellow lists in (B) are not mutually exclusive as a given protein can have more than one function or domain and was counted in each appropriate category.

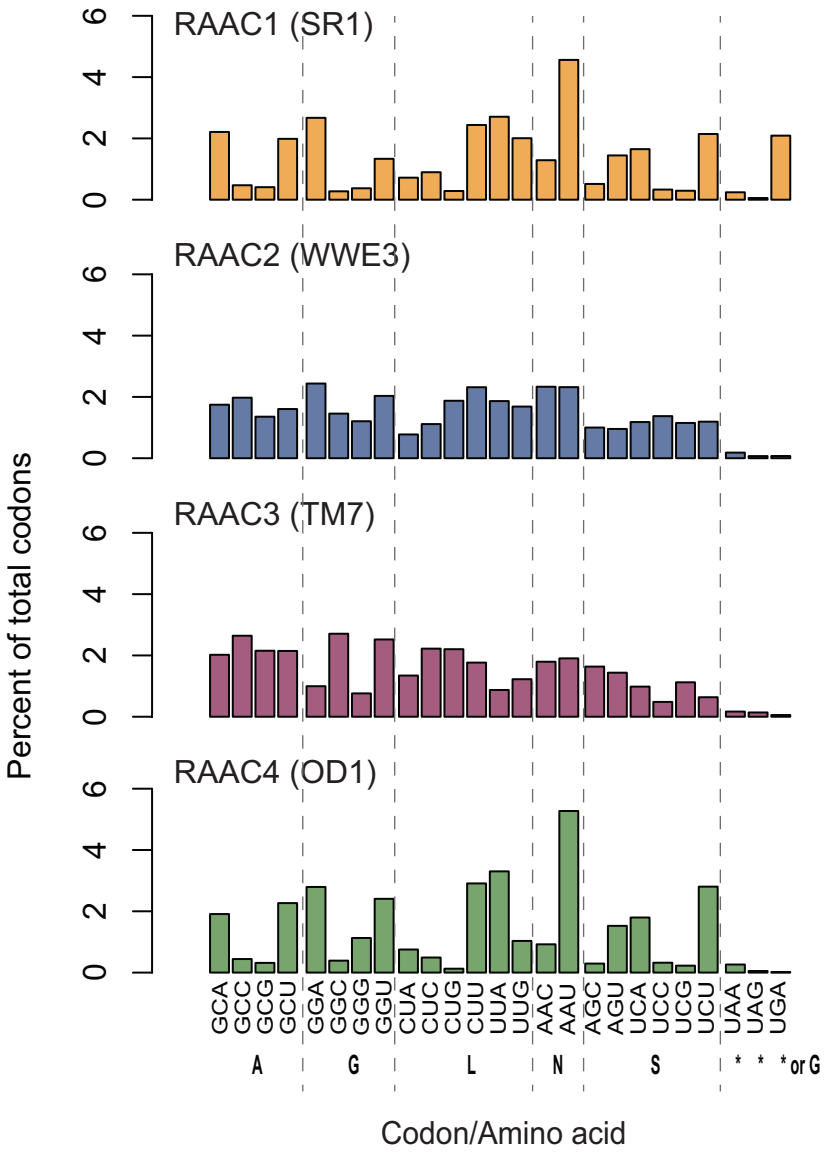

Figure 1.12. Codon usage, shown for selected amino acids reflects coding bias in SR1 and OD1 but not WWE3 and TM7 genomes. Codon was usage calculated as a percent of total codons, including termination. Dashed lines group triplets that typically encode the same amino acid. The SR1 uses UGA to encode glycine instead of termination, and hence is found much more frequently in that genome. The OD1 genome uses UGA as termination for only 34 out of 686 predicted ORFs, representing 0.02% of all codons in the genome.

Supplementary table titles

Table S1.1. Gene identifiers for enzymes numbered in Figure 1.8. Green, identified; yellow, homology and function unclear; red, not identified by blast with a diverse reference set, KAAS, or in-house annotation pipeline; n/a, not detected by in-house annotation pipeline. The relevant DNA and proteins sequences are accessible at genegrabber.berkeley.edu/aac.

Table S1.2. Complete nucleotide biosynthesis pathways were not identified in any of the four CP genomes. Gene identifiers are listed when detected. Colors are as in Table S1. All genomes were further searched using a curated reference set of Archaeal purine biosynthesis genes, collected based on the work of Brown *et al.* (2011), and no complete pathway was identified in any of the four genomes. The relevant DNA and protein sequences are accessible at genegrabber.berkeley.edu/aac.

Chapter 2

Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unraveled with genome-resolved metagenomics

Published in Environmental Microbiology, 2015

Abstract

Gold ore processing uses cyanide (CN^-), which often results in large volumes of thiocyanate-(SCN^-) contaminated wastewater requiring treatment. Microbial communities can degrade SCN^- and CN^-, but little is known about their membership and metabolic potential. Microbial-based remediation strategies will benefit from an ecological understanding of organisms involved in the breakdown of SCN^- and CN^- into sulfur, carbon and nitrogen compounds. We performed metagenomic analysis of samples from two laboratory-scale bioreactors used to study SCN^- and CN^- degradation. Community analysis revealed the dominance of *Thiobacillus* spp. whose genomes harbor a previously unreported operon for SCN^- degradation. Genome-based metabolic predictions suggest that a large portion of each bioreactor community is autotrophic, relying not on molasses in reactor feed, but using energy gained from oxidation of sulfur compounds produced during SCN^- degradation. Heterotrophs, including a bacterium from a previously uncharacterized phylum, compose a smaller portion of the reactor community. Predation by phage and eukaryotes is predicted to affect community dynamics. Genes for ammonium oxidation and denitrification were detected, indicating the potential for nitrogen removal, as required for complete remediation of wastewater. These findings suggest optimization strategies for reactor design such as improved aerobic/anaerobic partitioning and elimination of organic carbon from reactor feed.

Introduction

The use of microbes in biodegradation is widely studied as an alternative to traditional methods for remediating environmentally harmful waste. Thiocyanate (SCN^-), a cyanide derivative with lower toxicity, is produced at high concentrations in gold mining effluents and other industrial processes including coal gasification and steel processing. SCN^- also occurs naturally, both as a product of cyanide (CN^-) detoxification in many organisms (Cipollone et al., 2007a) and as a degradation product of cyanogenic glucosinolates, found in plants and insects (Jensen *et al.*, 2011). However, SCN^- at high concentrations can be harmful to human health (Erdogan, 2003) and aquatic life (Speyer and Raymond, 1988; Watson and Maly, 1987), necessitating SCN^- removal from mining wastewater. Several chemical methods have been used for SCN^- destruction, but these can be costly, inefficient, or produce other toxic chemicals (Gould *et al.*, 2012). In contrast, biological degradation is relatively inexpensive and can completely remove SCN^- and co-contaminants, such as CN^-. Known degradation pathways for SCN^- include an autotrophic pathway in which SCN^- is a source of energy, sulfur, and nitrogen. This pathway is reported to produce ammonium and carbonyl sulfide (OCS), which is subsequently converted to sulfide (Arakawa et al., 2007; Ogawa et al., 2013; Sorokin et al., 2001; 2004; Stratford et al., 1994). A heterotrophic pathway has been proposed for organisms capable of growth on organic carbon with SCN^- as the sole nitrogen source (Stratford *et al*., 1994). This pathway may produce sulfide and cyanate (OCN^-), which can be converted to ammonium via the enzyme cyanase (Sung and Fuchs, 1988). The principal end-products of microbial community SCN^- degradation are typically sulfate and ammonium or nitrate (Boucabeille *et al.*, 1994), consistent with either degradation pathway, coupled to ammonium and sulfide oxidation.

SCN^- biodegradation has been studied with a focus on reactor design and optimization for industrial applications (Boucabeille et al., 1994; Dictor et al., 1997; Hung and Pavlostathis, 1999; du Plessis et al., 2001; Stott et al., 2001), and some demonstration-scale plants have shown biodegradation to be successful (e.g. Activated Sludge Tailings Effluent Removal: ASTER™ process, Biomin Limited; Consort mine in Barberton, South Africa; Suzdal mine, Kazakhstan; and Homestake Mine, South Dakota, USA) (van Buuren *et al.*, 2011). The microbiological and molecular study of SCN^- and CN^- degradation has focused on isolates lacking sequenced genomes (Arakawa et al., 2007; Hussain et al., 2013; Katayama et al., 1998; du Plessis et al., 2001; Sorokin et al., 2012; 2007; 2004; Wood et al., 1998). Molecular fingerprinting of SCN^- degrading consortia (Felföldi et al., 2010; Huddy et al., 2015; Quan et al., 2006) provides the only knowledge at the community level to date. Hence, little is known about the metabolic potential of SCN^- degrading consortia although this is critical to understanding the degradation process.

In this study, we sampled biofilm and supernatant from two long-running laboratory-scale continuous flow bioreactors, treating either SCN^- ("SCN-only") or a mixture of CN^- and SCN^- ("CN-SCN"). We used high-throughput metagenomic sequencing to reconstruct microbial draft and curated genomes from these two communities. We describe the community structure and outline potential nutrient flow paths, revealing a complex community that includes chemoautotrophs, heterotrophs, and possible predators.

Results

Reactor chemistry and observations

Prior to sampling for metagenomic analysis, the SCN-only reactor degradation performance was stable and highly efficient. Ammonium in reactor effluent accounted for between 35-82% of nitrogen input from SCN^- (**Figure 2.1a**), suggesting the possibility of nitrogen uptake or removal. Reactor biomass accumulated visibly in the form of thick biofilms (**Figure 2.1c**), beginning when SCN^- loading reached 1.72 mmol/h (100 mg/h), 173 days prior to sampling (not shown).

The CN-SCN reactor mimicked conditions of industrial reactors processing mining wastewater, including the presence of CN^- in the feed and temperature manipulations. Brief fluctuations in degradation efficiency were observed in the CN-SCN reactor following such manipulations, after which stable operation was resumed (**Figure 2.1b**). Biomass accumulated as biofilm in this reactor, similar to the SCN-only reactor, and samples were collected from both reactors for extraction of community genomic DNA (**Figure 2.1b**).

Genome recovery and community structure

High-throughput sequencing and assembly of short-read data into longer scaffolds yielded metagenomes of 97 Mbp for the SCN-only sample and 295 Mbp for the CN-SCN sample (on scaffolds > 5 Kbp; **Table 2.1**). Scaffolds were binned into genomes based on their coverage by reads in each sample and their di- and tri-nucleotide frequencies (**Figure 2.2**), generating 29 genome bins for the SCN-only reactor and 64 for the CN-SCN reactor (**Table S2.1**). These bins accounted for essentially all organisms sampled, such that all scaffolds containing > 8 key ribosomal proteins (**Figure S2.1**) and nearly all scaffolds (> 5 Kbp) containing single copy genes were binned. The overall taxonomic compositions of both reactors were similar, as the inoculum for the CN-SCN reactor was taken from the SCN-only reactor. More genomes were recovered from the CN-SCN reactor sample, perhaps owing to the greater depth of sequencing for this sample (**Table 2.1**), to sampling bias, or to community shifts due to the different conditions in each reactor.

Rank abundance analysis based on binned genomes demonstrated the dominance of *Thiobacillus* spp. in both samples (**Figure 2.3**). As the assemblies of these genomes were highly fragmented, we assembled subsamples of the reads to recover the two highest abundance *Thiobacillus* spp. genomes present in the SCN-only reactor. These, and a third lower-abundance genome from the original assembly, accounted for 27.2% of all reads in the SCN dataset. Within the CN-SCN dataset, *Thiobacillus* spp. were similarly abundant but high coverage and strain variation produced fragmented genomes that could not be satisfactorily resolved with subsampling of reads and reassembly.

In addition to *Thiobacillus* spp., six other organisms were also found at high relative abundance in both samples (**Figures 2.3** and **S2.1**). The reconstructed sequences for these organisms in the two samples shared > 98% nucleotide identity across 93% to 99% of the genomes' length. One of these six organisms was a member of the uncultivated OPB56 radiation within the Bacteroidetes-Ignavibacterium-Chlorobi superphylum (**Figures S2.1** and **S2.2**). Near-identical draft genomes for this organism were independently assembled and binned from both bioreactor metagenomic datasets. Notably, the two sequences were broken in the same regions due to repeated elements. Manual correction of local assembly errors and closure of internal gaps using reads from both datasets generated a high-quality, near-complete genome sequence. Another genome found in both samples belongs to an Actinobacteria most closely related to *Leifsonia* spp. This genome was also manually curated to near-complete status. The other four

genomes, *Pelagibacterium* sp., *Variovorax* sp., *Pseudonocardia* sp., and a member of the Xanthomonadaceae, remain as draft genomes in both datasets.

Also included in both reactors were multiple members of the genera *Rhodanobacter, Microbacterium, Pseudonocardia,* and *Pelagibacterium* as well as members of the orders Sphingomonadales, Burkholderiales, and Xanthomonadales (**Figures 2.3**, **S2.1** and **S2.2**). Near-complete genomes were recovered for several organisms found in one but not both reactors. These included members of the phyla Acidobacteria and Gemmatimonadetes that are distantly related to previously genomically sampled organisms (**Figure S2.1**). Several other genomes remained partial, and in a few instances, individual genomes could not be resolved.

In addition to bacterial genomes, we detected microbial Eukaryotes in both reactors via assembly of partial mitochondrial genomes, and, in two cases, nuclear genomes (**Table S2.1**). One of these nuclear genomes was identified as belonging to the Rhizaria while the other fell within Opisthokonta, sibling to the slime mold *Fonticula alba* (**Figure S2.3)**. Many phage, plasmid, and transposon sequences were also recovered.

Metabolic analysis and nutrient flow

In order to identify potential roles for organisms in this system and their interactions, we examined metabolisms predicted from sequences within each genome bin. Of particular interest were genes and pathways potentially involved in the breakdown of SCN^- and CN^- and their degradation products: sulfur, carbon, and nitrogen compounds (**Figure 2.3**). These findings, by pathway, are detailed below.

Thiocyanate degradation via SCNase operon in Thiobacillus spp. and Pseudonocardia spp.

We searched both datasets for the genes encoding the cobalt-coordinating enzyme thiocyanate hydrolase (SCNase), known to convert SCN^- to carbonyl sulfide (COS) and ammonium in *Thiobacillus thioparus* THI115 (Arakawa et al., 2007; Kataoka et al., 2006). These genes did not assemble well in the CN-SCN dataset, but were present on *Thiobacillus* spp. scaffolds in some subassemblies. The genes were complete in three *Thiobacillus* spp. genomes in the SCN-only reactor. Here, the SCNase genes were co-located in a conserved operon (**Figures 2.4** and **2.5a**), which also contained cyanase, the enzyme responsible for cyanate (OCN^-) metabolism to ammonium. We searched for carbonyl sulfide hydrolase (COSase), the carbonic anhydrase family enzyme responsible for degradation of COS produced from SCN^- in *T. thioparus* THI115 (Ogawa *et al*., 2013), but we did not detect this gene in either dataset. Other genes in the SCN^- operon, including three with possible roles in sulfur metabolism (**Figures 2.4** and **2.5a**), may compensate for the lack of a COSase.

A genomic region similar to the SCN^- operon was found in both of the *Pseudonocardia* spp. genomes from the CN-SCN reactor (**Figure 2.5b**). The region contains a nitrile hydratase whose metal-coordinating alpha subunit has high sequence similarity to the SCNase gamma subunit, including conserved cobalt-coordinating residues and residues that dictate substrate specificity for SCN^- rather than nitriles (Yamanaka *et al.*, 2013). Additionally, an unrelated type of SCNase, recently identified in *Afipia* sp. TH201 (Hussain *et al.*, 2013), appears to be conserved in two of the *Thiobacillus* spp. genomes in the SCN-only reactor (**Table S2.2**). This protein is found in a region of the *Thiobacillus* spp. genomes that contains genes for SoxA, SoxX, and a thioredoxin-like protein and is not associated with the main *sox* operon (see below).

Capacity for cyanide degradation and tolerance may be widespread within the community

Rhodaneses are enzymes that convert CN^- and thiosulfate to SCN^- *in vitro* (Cipollone *et al.*, 2004), and confer improved growth with CN^- *in vivo* (Cipollone et al., 2007b). Because rhodanese domains occur in a variety of proteins with functions unrelated to CN^- metabolism, we confined our search to two-domain rhodaneses *sensu stricto* (rhdA, EC: 2.8.1.1) (Cipollone et al., 2007a). A majority of genomes from both reactors were found to contain at least one of these rhodaneses (**Table S2.2**).

Other candidates for enzymatic CN^- degradation in the bioreactor system include homologs of fungal cyanide hydratase (EC: 4.2.1.66) (Cluness et al., 1993; Wang and VanEtten, 1992) and bacterial cyanide dihydratase (cynD, EC: 3.5.5.-) (Fernandez and Kunz, 2005; Jandhyala et al., 2005; 2003). These proteins are class I nitrilases (Pace and Brenner, 2001), for which the current understanding of structure-function relationship is limited (Thuku *et al.*, 2009). We searched for homologs to the bacterial and fungal proteins and identified several putative cyanide (di)hydratases in both datasets (**Table S2.2**). Additionally, genes annotated as cytochrome *bd* ubiquinol oxidases (*cydAB*) were sometimes identified adjacent to class I nitrilases in *Rhodanobacter* sp. and some Alphaproteobacteria (**Table S2.2**). These cytochromes could act as alternative terminal reductases when cytochrome C oxidases are inhibited by CN^-, although the cyanide-insensitivity of *cydAB* cytochrome oxidases is not predictable based on protein sequence alone (Borisov *et al.*, 2011). Lastly, in some CN-degrading pathways, CN^- may first be converted to a nitrile, which is suggested to be subsequently degraded by nitrile hydratases (Luque-Almagro *et al.*, 2011). Nitrile hydratases were identified in several genome bins in both datasets (**Table S2.2**).

Reduced sulfur as an energy source in the reactor system

As SCN^- is a reduced form of sulfur and sulfate is known to accumulate in the bioreactor system, we hypothesized that oxidation of some intermediate sulfur compounds (such as sulfide or thiosulfate) may serve as a key energy-generating process for some organisms in the reactors. We identified pathways for sulfur-compound oxidation in at least 8 genomes within the SCN-only dataset and 9 in the CN-SCN dataset (**Figure 2.3** and **Table S2.2**). These pathways include the Sox enzymes, which can perform complete oxidation of sulfide to sulfate, and reverse dissimilatory sulfite reductase (rDsr) enzymes, which can oxidize sulfide to sulfite. We also identified genes for APS reductase and ATP sulfurylase, which may convert sulfite to sulfate (**Table S2.2**), completing the oxidation.

The *Thiobacillus* genomes each contain partial *sox* operons (**Table S2.2**) as previously observed for *T. denitrificans* (Beller et al., 2006b), and complete operons containing rDsr were detected in all three *Thiobacillus* genomes and a *Rubrivivax sp.* genome from the SCN-only reactor. Additionally, several sulfide-quinone reductase-like genes (*sqr*), which can oxidize sulfide to sulfur, were identified within genomes in both datasets (**Table S2.2**). Although active sites and binding sites were confirmed for these predicted proteins (Cherney et al., 2010; Marcia et al., 2009), the SQR are difficult to distinguish from proteins of other functions by sequence or structural prediction (Marcia *et al.*, 2010) and therefore are not included in Figure 2.3.

Carbon flow: dominance of chemolithoautotrophy

Carbon input for the system comes from both SCN^- and molasses in the reactor feed, but due to a substantial accumulation of biomass observed as SCN^- loading was increased, we expected that some community members were capable of carrying out carbon fixation. Concordantly, we identified genes for the complete Calvin Benson Bassham (CBB) cycle in 7 genome bins in the

SCN-only reactor and 10 genome bins in the CN-SCN reactor (**Table S2.2**). Key genes for the Wood-Ljungdahl pathway (anaerobic CODH/ACS) and reverse TCA cycle (citrate lyase) were absent from all genomes. The *Thiobacillus* spp. (**Figure 2.6a**) and *Thiomonas* sp. genomes carry both type I and type II RuBisCO. The high-abundance *Thiobacillus* sp. genome in the SCN-only dataset also contains a putative type ID RuBisCO located in the same operon as the gene encoding type II RuBisCO. Only *Thiobacillus* spp., *Thiomonas* sp., and *Pseudonocardia* spp. genomes possessed genes encoding carboxysome proteins. As expected, genomes belonging to the dominant *Thiobacillus* spp. and several other relatively abundant organisms encoded both the CBB cycle and sulfur oxidation pathways, suggesting that chemolithoautotrophy is important in this system (**Figure 2.3** and **Table S2.2**).

Carbon flow: an abundant, novel heterotroph in both reactors

Each reactor sample contained many genome bins belonging to predicted heterotrophs, including five relatively abundant genomes common to both reactors (**Figure 2.3**). Metabolic reconstruction focused on the OPB56 genome, representing the first sequenced member of this phylum. Analysis suggests this bacterium is Gram-negative and capable of aerobic heterotrophic metabolism (**Figure 2.6b**). Notably, it possesses a full suite of genes for production of antibacterial microcins. These include five genes encoding microcin precursors with near-identical leader sequences, as described by Haft *et al.* (Haft et al., 2010). The OPB56 genome also encodes at least six predicted extracellular subtilisin proteases and predicted transporters for complex organics including sugars, peptides, and amino acids.

Nitrogen removal: capacity for ammonium oxidation and denitrification

The principal source of nitrogen for the bioreactor systems is SCN^-, which yields ammonium upon degradation (**Figure 2.1**). Tracing all genes responsible for ammonium oxidation and denitrification revealed that the capacity for nitrite production (at low abundance) and denitrification likely exists in both reactor communities (**Table S2.2**). Sequences for ammonium monooxygenase and hydroxylamine oxidoreductase were found on unbinned scaffolds in both datasets, corresponding to relatives of *Nitrosospira multiformis*. Genes for nitrite oxidation, anaerobic ammonium oxidation, and dissimilatory nitrate reduction to ammonium were not detected.

There were five organisms with predicted capacity for denitrification of nitrite within the SCN-only reactor and five different organisms with this potential in the CN-SCN reactor (**Figure 2.3**). Other genome bins were missing one gene in the pathway, possibly due to incomplete genome recovery or because the pathway is incomplete in these organisms. Some members of the Xanthomonadaceae appear to denitrify to N_2O, as has been observed in certain Xanthomonadaceae isolates (Chen et al., 2002; Finkmann et al., 2000). Interestingly, two other genome bins (Gemmatimonadetes and OPB56) possess only genes encoding nitrous oxide reductase, suggesting they could scavenge this intermediate to complete denitrification. A *nirK* sequence was identified in the Rhizaria (Eukaryote) genome bin within the CN-treated reactor (**Table S2.2**), suggesting this eukaryote may be able to use nitrite as an electron acceptor under anaerobic conditions, as observed in the fungus *Fusarium oxysporum* (Kim *et al*., 2009).

Of the *Thiobacillus spp.* present in the SCN-only community, two possess complete denitrification pathways (**Figures 2.3** and **2.6**). The third does not encode any denitrification genes despite containing all single-copy genes. A complete denitrification pathway, including NirS instead of NirK, was recovered from one *Thiobacillus* sp. bin in the CN-SCN reactor. The

other *Thiobacillus* bins in this dataset encoded some denitrification genes. The pathways may be incomplete in these organisms, but more likely, the remaining genes were not recovered due to poor assembly of these genomes.

Cyanate

Several genomes in each reactor encoded predicted active cyanases (Guilloton et al., 1993; Sung and Fuchs, 1988; 1992; Walsh et al., 2000) (**Table S2.2**). In most of these genomes, the *cynS* gene was located immediately downstream of genes encoding a putative cyanate transporter, as described previously for *Pseudomonas pseudoalcaligenes* (Luque Almagro et al., 2008). Possession of cyanase may allow access to nitrogen, in the form of OCN^-, for organisms without the capacity to degrade SCN^- directly, however, cyanase has been shown to be unnecessary for cyanide degradation (Luque Almagro et al., 2008).

Discussion

Community structure by energy and nutrient flow predictions

The two SCN-degrading bioreactor communities described here are enriched in several genera that have previously been detected in or isolated from SCN^- and/or CN-degrading microbial communities. These included *Thiobacillus*, *Mesorhizobium*, *Sphingomonas*, *Sphingopyxis*, *Comamonas*, *Rhodanobacter*, and *Microbacterium* (Felföldi et al., 2010; Huddy et al., 2015; du Plessis et al., 2001; Quan et al., 2006). The survival and proliferation of multiple *Thiobacillus* spp. and six other organisms under both high SCN^- and CN-SCN conditions, indicates tolerance to and/or the capacity to use SCN^-. Overall, the communities in both reactors were diverse, but some functional capacities were shared (**Figure 2.3**). This may provide resilience to fluctuations in environmental factors such as temperature, access to oxygen, pH (lowered due to sulfate production), and the presence of CN^-.

From our genome-based metabolic analysis, we can divide both communities into trophic groups (**Figure 2.7**). Of the several predicted sulfur-oxidizing chemolithoautotrophs (**Figure 2.3**), only the *Thiobacillus* spp. possess known genes for SCN^- degradation. The other chemolithoautotrophs may benefit from the production of reduced sulfur compounds from SCN^-. Several of these autotrophs also encode complete denitrification pathways (**Table S2.2**). Additional organisms with the capacity for carbon fixation include *Pseudonocardia* spp. and the ammonium oxidizing *Nitrosomonas* spp.

Heterotrophs may utilize molasses in the reactor feed, ammonium produced from SCN^- degradation, and accumulated biomass or extracellular organic compounds produced by autotrophs (**Figure 2.7**). Additionally, some heterotrophic organisms may be able to use SCN^- directly as a source of nitrogen, as has been observed for *Burkholderia phytofirmans* and *Methylobacterium thiocyanatum* (Vu et al., 2013; Wood et al., 1998). Others are predicted mixotrophs, oxidizing sulfur compounds for energy while using organic carbon. The OPB56 organism may be considered a heterotroph or a predator, based on its predicted capacity to produce bactericidal compounds.

Other predators include phage and eukaryotes, which were present in both samples. Eukaryotes may be capable of utilizing sugars in the molasses media and nitrogen products of SCN^- degradation, but they likely also consume bacterial biomass (**Figure 2.7**). In fact, light microscopy revealed the presence of unidentified grazers in the SCN-only reactor (data not shown), and a variety of eukaryotes have been observed in this system based on 18S rRNA gene

clone libraries and isolations (Huddy et al., 2015). To our knowledge the eukaryotic genomes presented here are some of the few reported from complex metagenomic data and the only ones from a high-SCN^- environment. No genome annotator assessed was able to predict genes for these genomes accurately, highlighting the need for improvement in methodology in this area and for further work to obtain transcriptomic data.

Overall, the apparent dominance of autotrophic SCN^- degraders leads us to speculate that it may be possible to eliminate molasses from the reactor feed, reducing operating costs while maintaining functionality. It may also be possible to selectively remove Eukaryotes from the system, although such alterations could affect reactor stability.

Thiocyanate degradation pathways: conservation and horizontal gene transfer

While several bacteria with SCNase activity have been reported in the literature, only two SCNase genes have been identified (Hussain et al., 2013; Katayama et al., 1998). We detected both of these in the bioreactor systems, and notably, the genes for cobalt-coordinating SCNase were found in a newly-identified operon in *Thiobacillus* spp. We found the same operon in the genomes of *Thiobacillus thioparus* DSM 505, an aerobe, and *Thioalkalivibrio thiocyanodenitrificans* ARhD 1, a Gammaproteobacteria known to be able to perform SCN-degradation under denitrifying conditions (Sorokin *et al.*, 2004) (**Figure 2.5a**). The genes in the SCN^- operon of *T. thiocyanodenitrificans* are conserved relative to the *Thiobacillus* spp. genomes, except for the substitution of the ABC-type cobalt transporter genes, *cbtAB*, in place of *cbiM*. High amino acid similarity between all genes in the *T. thiocyanodenitrificans* operon to those in *Thiobacillus* spp. and lack of SCNase in other bacteria with sequenced genomes suggests possible horizontal gene transfer, although gene order has been only partially conserved (**Figure 2.5a**). Studies examining diverse bacterial species have suggested that cyanase is co-expressed with SCNase and/or is highly expressed under SCN-degrading conditions (Bezsudnova et al., 2007; Sorokin et al., 2004; Wood et al., 1998). The presence of a cyanase gene within the SCN^- operon could account for its observed co-expression with SCNase.

Previous work suggested the existence of a heterotrophic pathway for SCN^- degradation, which produces OCN^- to be used as a nitrogen source; however, the proteins involved in this pathway are unknown. Within the bioreactor systems, only the *Thiobacillus* spp., and possibly *Pseudonocardia* spp. genome bins, possess the SCN^- operon or alternative SCNase. Thus, while a small number of species may be responsible for most SCN^- degradation here, expression data and isolation experiments are needed to provide insight into other possible pathways.

Nitrogen cycling in aerobic/anaerobic reactor phases

Given that genes for both ammonium oxidation and denitrification were detected, it may be possible to achieve near-complete removal of nitrogen compounds from SCN-contaminated wastewater if reactor operation were modified to increase the rates of these processes. The presence of thick biofilm in the reactors (**Figure 2.1c**) may already provide microaerobic or anaerobic environments suitable for denitrification, possibly accounting for some of the nitrogen released through SCN^- degradation. As performed by Kraft *et al.* (2014) mass-balances for nitrogen (and sulfur), expression data, and short-term reactor manipulations could provide insight into the dynamics of nitrogen and sulfur cycling within this complex system.

Accurate reconstruction of the genome of a novel bacterium

The genome for the novel OPB56 organism is the first near-complete sequence from this group within the Bacteroidetes-Ignavibacterium-Chlorobi radiation. Given that this organism has been maintained in the bioreactor mixed-culture and identical draft genomes were recovered independently from both reactors, we propose the name Candidatus '*Kapabacteria thiocyanatum*' for this organism. The name, "Kapa", reflects the location of cultivation, University of Cape Town. Further, we propose the name Candidatus Kapabacteria to replace OPB56 as the designation for this phylum.

Conclusion

Metagenomics has been used to understand nutrient flow in relatively low-diversity bioreactor communities degrading contaminants such as chlorinated organics (Hug *et al*., 2012) and terephthalate (Lykidis *et al*., 2010; Wu *et al*., 2012). This is the first application of genome-resolved metagenomics to characterize SCN^- and CN^- bioreactors, revealing a complex community containing novel organisms and genes. The analysis identified members with potentially important roles in the sulfur and nitrogen cycles, providing a framework for understanding a microbial community performing SCN^- degradation. Further manipulations of the reactor system will be required to elucidate the factors controlling this process and subsequent nitrogen removal.

Methods

Reactor operation and sampling

SCN-only reactor A one-liter water-jacketed reactor (Glass Chem, Stellenbosch, South Africa) was inoculated with sludge from a demonstration-scale bioreactor treating SCN-containing mining effluent (Consort Mine, Barberton, South Africa). The original source of inoculum for the demonstration-scale reactor was obtained from a mixture of sludge from an SCN-contaminated tailings pond and sludge from a domestic wastewater treatment plant. The one-liter reactor was operated in continuous flow, with a 12 hour hydraulic retention time under continuous aeration (0.8 L/min) and stirring (270 rpm). The feed SCN^- loading was increased incrementally from 0.22-1.72 mmol/h over 150 days. Microbial biomass was maintained in the reactor by means of a clarifier, which allowed settling and underflow sludge recycling. The set-up of this reactor, depicted in **Figure 2.1c**, is described further by (van Zyl et al., 2011). The reactor was sampled at day 820 for metagenomic analysis (**Figure 2.1a** and **Table 2.1**). The sample consisted of biofilm with planktonic cells in the associated fluid.

CN-SCN reactor A second one-liter reactor was inoculated with biomass from the SCN-only reactor and operated at a residence time of 14 hours degrading 0.12 mmol/h SCN^- and 0.5 mmol/h CN^-. Following steady state operation for multiple residence times, the loadings were gradually increased to 0.92 mmol/h SCN^- and 0.14 mmol/h CN^- respectively. Reactor temperature was decreased step-wise from 25 °C to 18 °C to 15 °C at which point (day 190) a reactor sample, containing biofilm and planktonic cells in associated fluid, was collected for metagenomic analysis (**Figure 2.1b** and **Table 2.1**).

HPLC analysis

Thiocyanate ion concentration was measured on a Thermo Scientific HPLC system using a UV detector (Spectasystem UV1000) at 210 nm. A reverse phase Discovery C18-HS column was

used as the stationary phase and the mobile phase consisted of 40% v/v acetonitrile in ddH_2O, containing 2 mM tetrabutyl ammonium dihydrogen phosphate. Mobile phase was pumped through the column at a rate of 0.5 mL/min. The SCN^- peak area was converted to concentration using a standard curve (1-100 mg/L SCN^-). Ammonium ion concentration was measured on a Waters 717plus system with a conductivity detector (Water model 430) and a Hamilton cation exchange column. Mobile phase (30% methanol in ddH_2O) was pumped through at a rate of 1 mL/min. A standard curve (1-100 NH_4^+ mg/L) was used to determine concentration.

CN^- analysis

Free cyanide (HCN, CN^-) in solution was measured using the Cynoprobe (Mintek, Johannesburg, South Africa). Temperature and agitation within the Cynoprobe were maintained at 20°C and 500 rpm (10 seconds) ensuring a homogeneous sample and avoiding interference of these parameters with mass transfer. The reading was converted to a cyanide concentration using a standard curve (range 0.3 – 30 mg/L CN^-) in a 2.5M NaOH solution (pH 10.0).

DNA extraction and sequencing

Genomic DNA was extracted from the samples using the High Pure PCR Template Preparation Kit (Roche Applied Sciences) with the following modifications: samples were mixed with 200 µL Tissue Lysis Buffer, vortexed, and stored at -20 °C overnight. Upon thawing, 50 µL of proteinase K and 250 µL of Binding Buffer were added and extraction proceeded via the manufacturer's protocol. Illumina library preparation and sequencing were performed at UC Berkeley Vincent J. Coates Genomics Sequencing Laboratory (Berkeley, CA), using an insert-size of 500 bp and read length of 100 bp. Sequencing results are shown in **Table 2.1**.

Read processing, assembly, and annotation

Both read sets were trimmed for quality using Sickle with default parameters (https://github.com/najoshi/sickle) and assembled independently using idba_ud with default parameters (Peng *et al.*, 2012). To improve assemblies of high-abundance organisms in the SCN-only sample, two sub-assemblies were performed with idba_ud using subsets of randomly selected reads representing either $1/5^{th}$ or $1/20^{th}$ of the full dataset. For all assemblies, open reading frames (ORFs) were predicted with Prodigal, run in metagenome mode (Hyatt et al., 2010; 2012). Annotation was performed using USEARCH (Edgar, 2010) against the Uniref100 (Suzek *et al.*, 2007), UniProt, and KEGG databases to identify the single best hit and the phylogenetic affiliation of the best hit for each ORF. Uniref100 hits were used to assign a tentative phylogenetic affiliation to each scaffold, to the lowest taxonomic level possible, based on majority rules.

Binning of the metagenomes

Bins were assigned based on the coverage, GC-content, and phylogenetic best-hit profile of each scaffold > 5 Kbp. For full assemblies, bins were confirmed using Emergent Self-Organizing Maps (ESOMs) based on di- and tri-nucleotide frequencies, and differential coverage across the two samples (Supplementary Methods) (Dick *et al.*, 2009; Sharon *et al.*, 2013).

The matrix used for ESOM training was generated using tools found at https://github.com/micronorman/bantools (A. Norman, unpublished). For each assembly, reads from both samples were mapped to the assembly using Bowtie2 (Langmead and Salzberg, 2012) with default parameters. Read-mapping information was used to calculate coverage for 10 kb

windows of each scaffold by reads from each sample. Values for di- and tri-nucleotide frequencies were also calculated for each window. Columns and rows of the resulting matrix were normalized as follows. Columns: the values for each coverage and nucleotide frequency measurement were normalized by the sum of all values for that measurement, log-transformed, standardized to follow a normal distribution, and scaled from 0 to 1; rows: the values of all measurements for each scaffold window were standardized to follow a normal distribution. ESOMs were trained on the resulting matrix for 100 epochs.

Bin information was superimposed onto the ESOMs as class files (**Figure 2.2**), and bins were checked manually for chimeric and mis-binned scaffolds (using the Databionix GUI, esomana). Chimeras were resolved using paired-read mapping with Bowtie2 (Langmead and Salzberg, 2012) and manual alignment editing in Geneious (Biomatters Ltd) to generate the correct consensus. Some genomes were manually curated using paired-read mapping to resolve assembly errors, extend scaffolds, and join scaffolds.

Genome completeness by single copy genes

Datasets were queried for 51 single copy genes (SCG) using in-house scripts. Briefly, SCG were identified in metagenomes using BLAST (Altschul et al., 1990) with a representative reference set. Ribosomal proteins were also searched by annotation, and the results of these two methods were reconciled manually by genome bin to generate information for Figure 2. The 51 genes and results from the SCG BLAST searches may be viewed using the links below:
http://genegrabber.berkeley.edu/genome_summaries/307-SCN-only_single_copy_genes
http://genegrabber.berkeley.edu/genome_summaries/176-CN-SCN_single_copy_genes

Phylogenetic analysis

Ribosomal proteins (RPs) were detected based on annotations and used to build a concatenated protein phylogenetic tree (**Figure S2.1**). We used 16 RPs that commonly co-occur in a near-contiguous block (RP L 2, 3, 4, 5, 6, 14, 15, 16, 18, 22, and 24, and RP S 3, 8, 10, 17, and 19) such that all proteins were located on one scaffold per bin. To compile a custom reference set, the RP S3 amino acid sequences from each dataset were searched against the Uniref100 database using ublast (Edgar, 2010), and RP sequences from genomes for the two best hits to each query were obtained from the NCBI genomes (http://www.ncbi.nlm.nih.gov/genome/) or JGI-IMG (http://img.jgi.doe.gov/) databases. Alignments for each of the 16 RP sets were generated using MUSCLE v3.8.31 (Edgar, 2004) and edited in Geneious to manually trim ends and automatically remove columns with greater than 95% gaps. The 16 alignments were then concatenated, resulting in an alignment containing 801 taxa with 2536 unambiguously aligned positions. Phylogenetic analysis was conducted using RAxML v8.0.26 (Stamatakis, 2014) under the LG + gamma model of evolution. Support values were generated with 100 bootstrap replicates.

Genes for 16S rRNA were detected using Rnammer (Lagesen *et al.*, 2007) and via BLAST (Altschul et al., 1990) of all scaffolds ≥ 500 bp against a condensed 16S rRNA gene reference set. For each gene identified, the best-hit non-environmental sequence from NCBI-nr was added to a custom reference dataset that included six archaeal sequences for rooting. Analysis was performed on this dataset using ssu-align (Nawrocki, 2009) to generate a nucleotide alignment to the bacterial structural model and to mask insertions. The final alignment contained 847 taxa with 1582 unambiguously aligned positions. RAxML (Stamatakis, 2014) was used to reconstruct phylogeny under the GTRCAT model of evolution (**Figure S2.2**). Support values were generated with 100 bootstrap replicates.

To identify Eukaryotic genome bins, a tBLASTn was performed (Camacho et al., 2009) with genome bins as target databases and reference sets of Eukaryotic marker proteins taken from Brown *et al.* (2013) used as queries. In order to obtain near full-length proteins, intervals of nucleotide sequences hit by each search were translated and aligned to the reference sets using MUSCLE v3.8.31 (Edgar, 2004). In total, 17 ribosomal proteins present in both Eukaryotic genome bins were chosen for analysis. These were RPL 11, 19, 20, 21, 27, 31, 32, 33, 35, 43, 44, and RPS 3, 5, 11, 16, 17, and 18. Each protein alignment was examined in Geneious, allowing exons from the same gene to be stitched together. The resulting alignments were manually trimmed and then concatenated together to generate a final alignment with 2,415 columns and 63 taxa where each taxon possessed at minimum 1,300 amino acids. ProtTest v3.0 (Darriba et al., 2011) was run (using options for -all-matricies and -all-distributions) to determine the best evolutionary model for phylogenetic reconstruction. Accordingly, RAxML (Stamatakis, 2014) was run under the LG+gamma model, and support values were generated with 100 bootstrap replicates (**Figure S2.3**).

Metabolic analysis

In order to identify metabolic pathways, gene annotations were searched by name. Additionally, sequences of biochemically and/or structurally characterized proteins of interest were used to create databases for BLAST+ (Camacho *et al.*, 2009). The SCN-only and CN-SCN datasets were used as queries with a bit-score cut off ≥60. In order to confirm conservation of active residues in the proteins identified, hits were aligned to reference sequences using MUSCLE (Edgar, 2004) and visualized in Geneious (Biomatters Ltd). The web-based tool PSORTb v3.0.2, was used to predict subcellular localization for proteins of interest (Yu *et al.*, 2010).

Access to data online

Sequence data are publicly available through the online database ggKbase at http://genegrabber.berkeley.edu/SCN-stock/organisms and http://genegrabber.berkeley.edu/CN-SCN/organisms. Read datasets are available at NCBI (http://www.ncbi.nlm.nih.gov/) under BioProject ID PRJNA279279 with BioSample identifiers SAMN03445100 (SCN-only) and SAMN03445079 (CN-SCN).

Figures and Tables

Table 2.1. Reactor conditions at time of sampling and sequencing data acquired.

	SCN stock reactor	SCN-CN reactor
Reactor volume	1 L	1 L
Hydraulic retention time	12 hours	14 hours
Thiocyanate loading (as KSCN)	1.9 mmol/h (110.0 mg/h)	0.9 mmol/h (50.6 mg/h)
Cyanide loading (as NaCN)	0 mmol/h	0.14 mmol/h (3.57 mg/h)
Feed: Thiocyanate (as KSCN)	22.7 mM (1320 mg/L)	12.2 mM (708.4 mg/L)
Cyanide (as NaCN)	0 mM	1.9 mM (50.0 mg/L)
Phosphate (as KH_2PO_4)	0.28 mM (27 mg/L)	0.28 mM (27 mg/L)
Molasses	150 mg/L	150 mg/L
Temperature	25 °C	15 °C
Reactor pH (with NaOH)	8.5	8.5
SCN degradation efficiency	93%	76%
CN degradation efficiency	NA	98%
Total sequence	5.5 Gbp	34.7 Gbp
Length of assembly in contigs ≥ 5kb	97 Mbp	295 Mbp
Contigs ≥ 5 kb	3811	13707
Genome bins	29	64

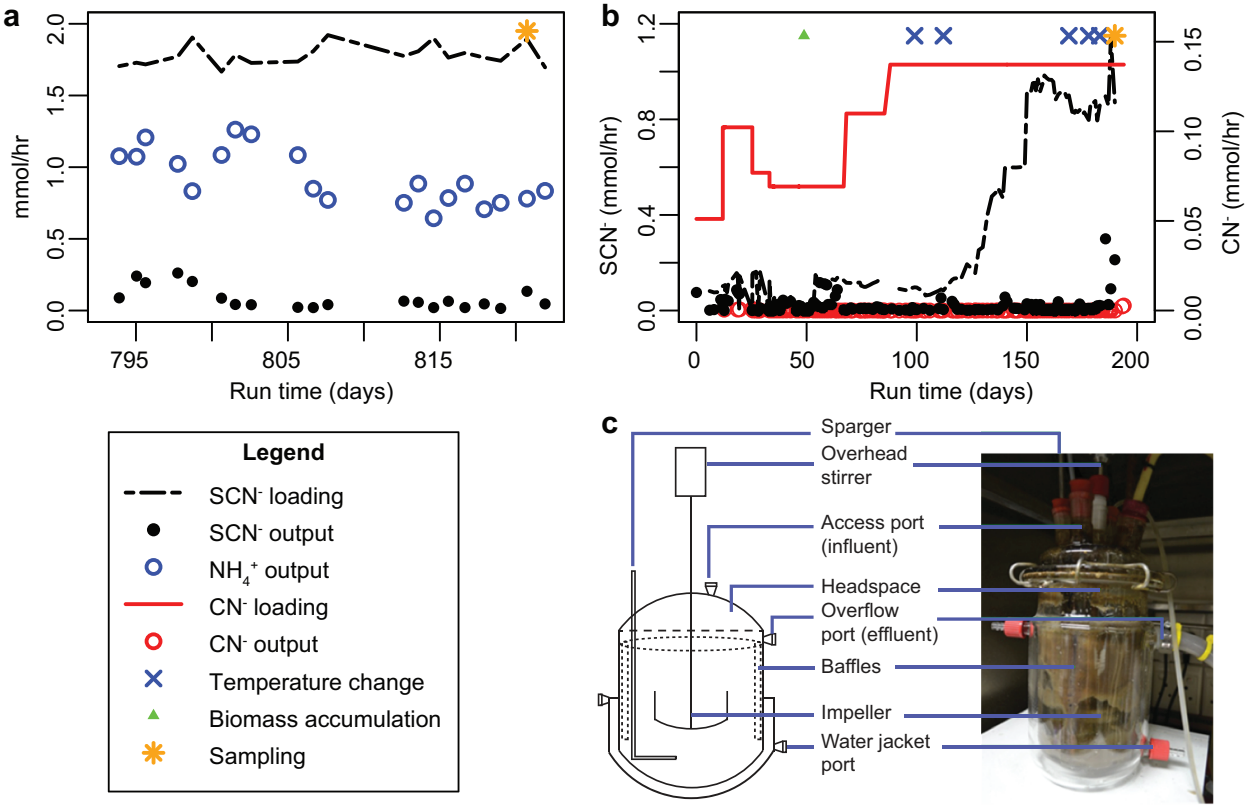

Figure 2.1. Conditions and set-up for two aerobic continuously stirred reactors treating (a) only SCN^- or (b) CN^- and SCN^-. Loadings indicate rates of compound input to reactors via media, while outputs indicate rates of compound produced in effluent. Biomass was observed and sampling for metagenomics was performed where indicated. In the CN-SCN reactor, temperature was varied, beginning at 25 °C and shifting at times marked to 40 (and back to 25), 52 (and back to 25), 20 (held), 18 (held), and 15 °C (held), respectively. Diagram and photograph of the SCN-only reactor in operation (c) shows mechanisms of stirring and aeration. Biofilm (dark color) occurred on glass reactor walls, impeller, sparger, and baffles.

Figure 2.2. Emergent self-organizing maps (ESOMs) show genome bins for the SCN-only reactor (a) and the CN-SCN reactor (b). Genomes are numbered as indicated in Table S2.1. Colors correspond to individual bins, and some bins are split into more than one location on the map (denoted by the same number). Di- and trinucleotide information and differential coverage across the SCN-only and CN-SCN samples was used to generate the datapoints shown for all scaffolds > 5 Kbp (see Methods). The dashed line encloses one repeating unit of the map.

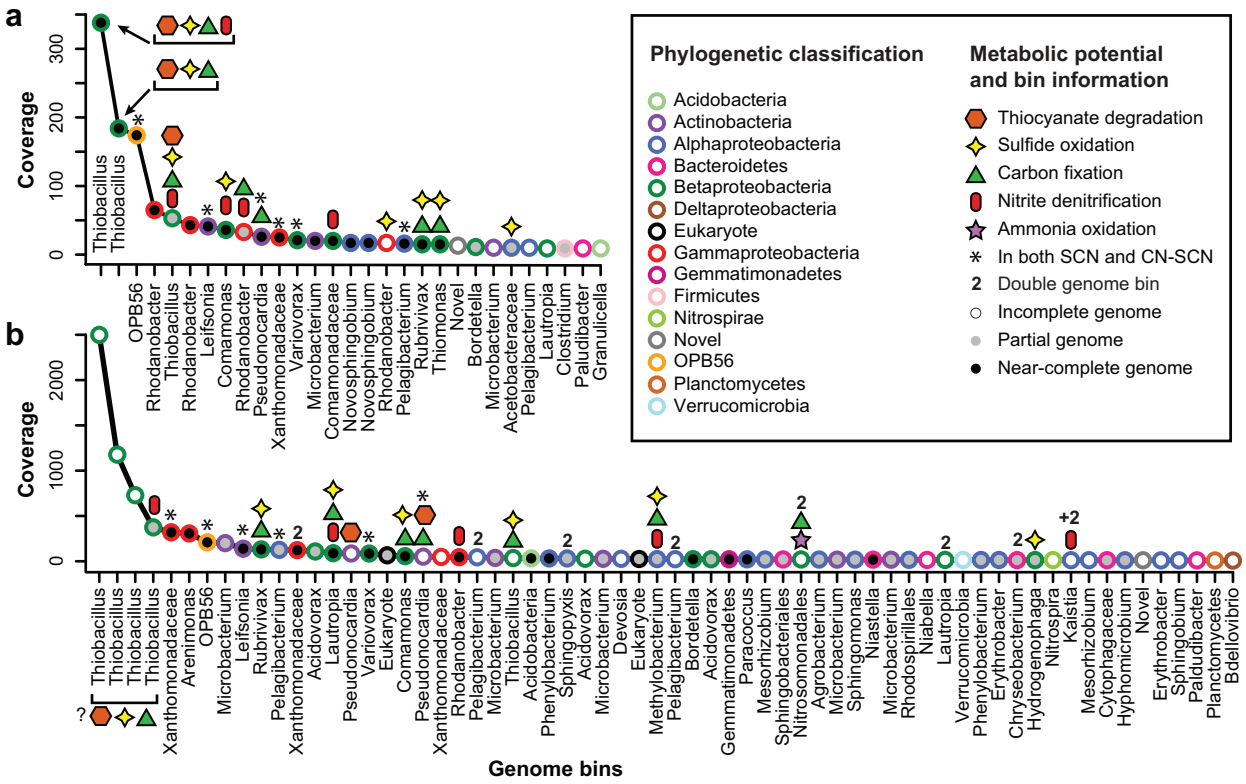

Figure 2.3. Rank abundance and metabolic potential for organisms in (a) the SCN-only or (b) the CN-SCN reactor. For each reactor, genome bins are ordered from highest to lowest coverage (the average read depth over all scaffolds in the bin). Outer circle color indicates phylogenetic affiliation, and circle fill indicates genome completeness measured by presence/absence of 51 single copy genes: white (0-25 genes), gray (26-47), and black (48-51) (see Methods).

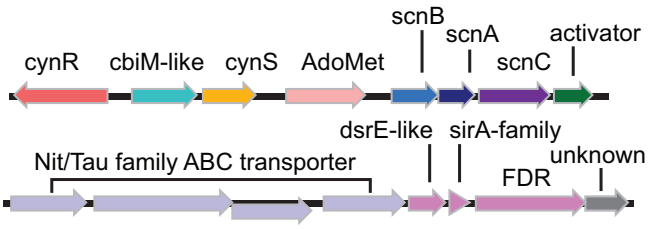

Figure 2.4. The thiocyanate operon conserved in *Thiobacillus* spp. Genes included are cyanase (*cynS*) and the cyanase transcriptional regulator (*cynR*), *cbiM* (membrane component of cobalt transporter), a methyltransferase (AdoMet), thiocyanate hydrolase subunits (*scnBAC*) and the activator (P15K), a four-subunit Nit/Tau family ABC transporter, three genes possibly related to sulfur oxidation (in pink; FDR is an FAD-dependent pyridine nucleotide-disulfide oxidoreductase family protein), and an unknown protein (conserved in all SCN operons analyzed, see Figure 2.5a).

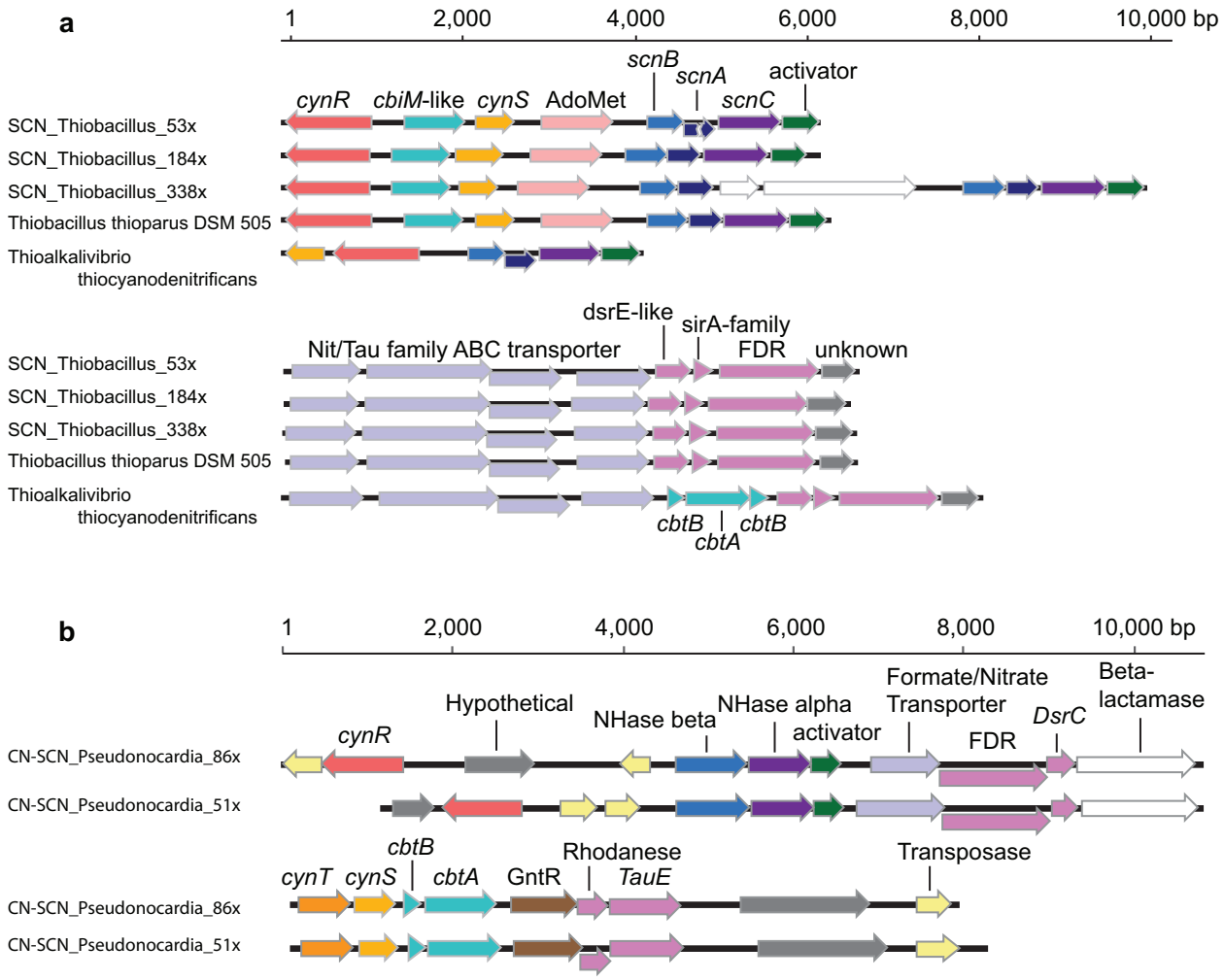

Figure 2.5. SCN operons in Thiobacillus, Thioalkalivibrio, and Pseudonocardia. The highly conserved SCN operon in Thiobacillus and Thioalkalivibrio spp. (a) suggests possible horizontal gene transfer. The SCN-only_Thiobacillus_338x operon possesses a partial duplication of SCNase including the beta and alpha subunits and the operon is interrupted by genes for acetolactate synthase and a DGC domain-containing protein (shown in white). The SCN-only Thiobacillus 53x genome contains a break in the assembly within scnA that could not be resolved with readmapping. The Pseudonocardia SCN operon (b) contains alpha and beta nitrile hydratase homologs with key residues for SCN-specificity conserved, genes for the required SCNase activator protein, a formate/nitrite transporter, FAD-dependent pyridine nucleotide disulfide oxidoreductase (FDR), beta-lactamase containing protein, carbonic anhydrase (cynT, involved in OCN- metabolism), and cyanase (cynS). Genes potentially involved in sulfur metabolism are colored pink. Both versions of this operon contain transposases (light yellow).

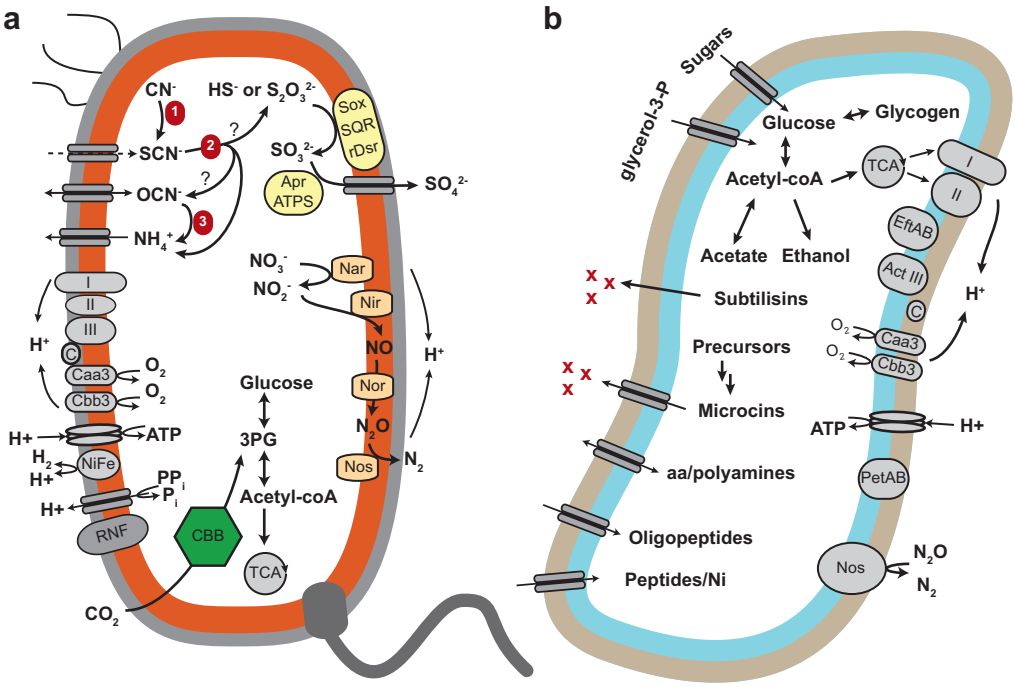

Figure 2.6. Cell diagrams depicting predicted metabolic potential for two abundant community members. The dominant, autotrophic *Thiobacillus* sp. in the SCN-only reactor (a) contains genes involved in sulfur oxidation, denitrification, carbon fixation. Key genes for SCN^- and CN^- metabolism in this genome include 1) rhodanese, 2) SCNase, and 3) cyanase. The newly described OPB56 organism (b) is a predicted heterotroph, present at high abundance in both reactors. This genome encodes proteins for scavenging and bactericidal activity. It also harbors genes for alternative electron transport proteins EftAb, alternative complex III, and PetAB as well as nitrous oxide reductase.

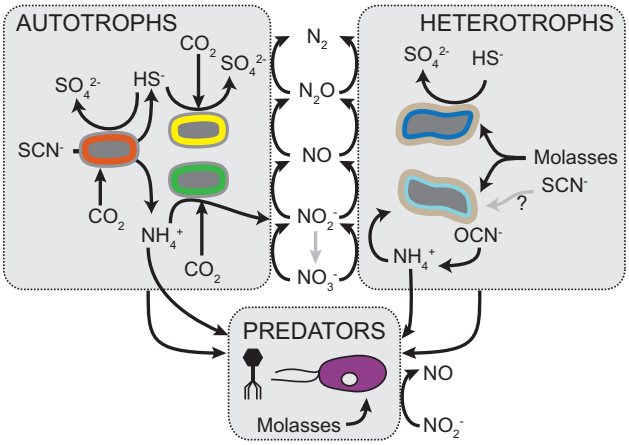

Figure 2.7. Schematic diagram of SCN^- degradation in a molasses-fed bioreactor community describing potential flow of sulfur, nitrogen and carbon through the reactor. Specific organisms are represented in different colors: *Thiobacillus* spp. (orange), other sulfur-oxidizing autotrophs (yellow), *Nitrosomonas* spp. (green), sulfur-oxidizing mixotrophs (dark blue), other heterotrophs (light blue), phage (black), and eukaryotes (purple). Gray boxes represent trophic groups as a whole. Black arrows indicate chemical transformation by predicted proteins found in the

metagenomes, while gray arrows indicate genes that were either undetected (e.g. Nxr) or unknown.

Supplementary table titles

Table S2.1. Information and statistics for genome bins generated from metagenomic sequence data assembled on scaffolds > 5 Kbp.

Table S2.2. Metabolic potential identified within each bin for SCN- and CN- degradation (orange), sulfur compound oxidation (yellow), carbon fixation (green), and nitrogen cycling (red). The presence and count of a given enzyme are indicated by numbers. If an enzyme has multiple subunits, or if a reaction can be performed by more than one enzyme, those identified may be listed by name. Genes included are thiocyanate hydrolase (scnABC; Arakawa et al. 2007), the alternative thiocyanate hydrolase identified in Afipia sp. (SCNase alt; Hussain et al. 2013), cyanase (cynS), nitrile hydratase (NHase), rhodanese sensu stricto (rhdA), cyanide dihydratase/nitrile hydratase class I (cynD), cytochrome bd ubiquinol oxidase (cydAB), sox enzymes, reverse dissimilatory sulfite reductase enzymes (dsr), APS reductase (apr), ATP sulfurylase (ATPS), sulfide:quinone oxidoreductase (sqr), Ribulose-1,6-bisphosphate carboxylase-oxygenase large subunit (cbbL), ammonium monooxygenase (amo), hydroxylamine oxidoreductase octaheme subunit (haoA), nitrate, nitrite, nitric oxide, and nitrous oxide reductases.

Chapter 3

Genome-resolved meta-omics ties microbial dynamics to process performance in biotechnology for thiocyanate degradation

Abstract

Remediation of industrial wastewater is important for preventing environmental contamination and enabling water reuse. Biological treatment for the industrial contaminant thiocyanate (SCN^-) relies upon microbial hydrolysis, but this process is sensitive to high loadings. To examine the activity and stability of a microbial community over increasing SCN^- loadings, we established and operated a continuous-flow bioreactor fed increasing loadings of SCN^-. A second reactor was fed ammonium sulfate to mimic breakdown products of SCN^-. Biomass was sampled from both reactors for metagenomics and metaproteomics, yielding a set of genomes for 144 bacteria and one rotifer that constituted the abundant community in both reactors. We analyzed the metabolic potential and temporal dynamics of these organisms across the increasing loadings. In the SCN^- reactor, *Thiobacillus* strains capable of SCN^- degradation were highly abundant, whereas the ammonium sulfate reactor contained nitrifiers and heterotrophs capable of nitrate reduction. Key organisms in the SCN^- reactor expressed proteins involved in SCN^- degradation, sulfur oxidation, carbon fixation, and nitrogen removal. Lower performance at higher loadings was linked to changes in microbial community composition. This work provides an example of how meta-omics can increase our understanding of industrial wastewater treatment and inform iterative process design and development.

Introduction

Microbial communities in biotechnology have historically been treated as a black box, but as molecular methods in microbiology have improved, our knowledge of these systems has deepened dramatically. Increasingly, 'meta-omics' methods are being used to investigate key organisms and potential weak points in biotechnology, such as nitrogen and phosphorus removal or bulking in wastewater treatment (Albertsen et al., 2013b; Sekiguchi et al., 2015; Speth et al., 2016). In particular, specialized treatment of industrial wastewater has benefited from a genome-resolved meta-omics approach, (Hug et al., 2012; Lykidis et al., 2010; Nobu et al., 2015; Taubert et al., 2012) used to identify the key species and their interactions. A better understanding of the activity and abundance of these key organisms under varying conditions is the next step in improving design and optimizing the operation of these important systems.

Thiocyanate (SCN^-) is a widespread industrial contaminant found at especially high concentrations (up to 4000 mg/L) in gold mining effluents. If not remediated, it can affect human health and aquatic organisms (Erdogan, 2003; Speyer and Raymond, 1988; Watson and Maly, 1987). Notably, SCN^- is inhibitory toward key iron- and sulfur-oxidizing microorganisms used in bio-oxidation processes at some gold mines (such as BIOX®) and therefore must be removed from wastewater before it is recycled within a mining site or discharged into the environment. SCN^- can be biodegraded by chemolithoautotrophic bacteria as a source of energy, (Hussain et al., 2013; Katayama and Kuraishi, 1978; Katayama et al., 1992; 1998; Sorokin et al., 2001) by heterotrophic organisms as a sole source of nitrogen, (Stratford et al., 1994; Wood et al., 1998) and by complex microbial consortia (Boucabeille et al., 1994). The initial products of degradation are ammonium, carbon dioxide, and reduced sulfur compounds.

A long-running thiocyanate-fed bioreactor (known as the "SCN^- stock reactor") at the University of Cape Town contains a characterized, diverse microbial community (Huddy et al., 2015; Kantor et al., 2015; van Zyl et al., 2011). Previous work on the structure and mechanism of SCN^- degradation in this community implicated several abundant *Thiobacillus* spp. in SCN^- degradation due to the presence of an SCN^- operon in the genomes of these autotrophic bacteria. Results also suggested the potential for nitrogen removal by *Thiobacillus* spp. and other community members, and the presence of heterotrophic community members (Kantor et al., 2015). Questions remained regarding community stability at different SCN^- loadings, expression of the observed metabolic potential, and the importance of inter-organism interactions, especially for nitrogen removal. The SCN^- stock reactor provided the inoculum for the bioreactors established in this study.

We used genome-resolved metagenomics and endpoint metaproteomics to track changes in the microbial community of a newly-inoculated SCN^- bioreactor operated with increasing loadings over time. To enrich for organisms that can use and remove the nitrogen produced by SCN^- degradation, a second reactor with the same inoculum was fed ammonium sulfate ($NH_4(SO_4)_{1/2}$) and molasses but no SCN^-. We describe the microbial community structure, protein expression, and replication rates in both reactors during the experiment. Our analysis linked shifts in community membership to changes in reactor function, highlighted organisms and metabolic pathways active under high-SCN^- conditions, and supported the importance of biofilm in this system.

Results and Discussion

Reactor chemistry and efficiency

In the newly-inoculated SCN-fed and $NH_4(SO_4)_{1/2}$ reactors, the loading rate was increased stepwise across 238 days. This was initially accomplished by lowering hydraulic residence time (HRT), and then from day 68, when both reactors reached 12 h HRT, the feed concentration was increased. Samples were taken for metagenomic analysis during this second phase. In the SCN^- reactor, the SCN^- removal rate consistently increased to match the increasing loading rate to a maximum of 1.07 $mmol.h^{-1}$ (**Figure 3.1A**). On further increase to 1.43 $mmol.h^{-1}$, the SCN^- removal rate decreased and efficiency dropped to near 50%. On average, the stoichiometry between the SCN^- removal rate and sulfate output was 1.05:1, near the 1:1 ratio expected based on known SCN^- degradation mechanisms coupled to complete oxidation of the sulfide released.

As loading increased, we observed a corresponding increase in the thickness of the biofilm that formed on all surfaces within the reactor. During one period early in the experiment, nitrate (NO_3^-) output reached up to 30% of nitrogen input as SCN^- (days 86-107) and fluctuated thereafter, reaching a maximum of 64% of nitrogen input. After day 200, nitrate output remained consistently low and sulfate output increased. The bioreactors had low buffering capacity and acidified as loadings were increased. Consequently, base, in the form of KOH, was added to the feed, and the pH fluctuated over the time series. The SCN^- removal rate decreased during one period of high pH that resulted from over-correction of the feed pH (**Figure 3.1A**).

In the $NH_4(SO_4)_{1/2}$ reactor, the rate of sulfate leaving the reactor rose steadily throughout the experiment, matching the sulfate loading rate and indicating that little sulfate was retained in biomass or converted to other forms (**Figure 3.1B**). The nitrate output rate increased with decreasing HRT and then increased more slowly as biofilm established and thickened. Overall, the nitrate production rate was higher than in the SCN^- reactor.

Genome recovery over the sample series

Biofilm from the two reactors was sampled at four time points (T1-T4) during the experiment, and concurrent samples of planktonic biomass were collected where possible. The inoculum for these reactors (T0), biofilm and planktonic biomass taken from the SCN^- stock reactor, was also sampled (**Figure 3.1** and **Table 3.2**). Independent metagenomic assemblies were performed for each sample, and differential coverage information across the data series resulted in clean definition of 789 bacterial genome bins (**Figure S3.1**). Two of these genomes were highly abundant in the samples in which they appeared, and subassemblies of a fraction of the reads resulted in substantially improved versions of these genomes. The bacterial genomes were de-replicated to yield a non-redundant set of 144 draft-quality genomes (**Tables 3.1, S3.1A**). Eukaryotic, mitochondrial, chloroplast, phage and plasmid genomes were also recovered and de-replicated (**Tables 3.1, S3.1B**). No Archaea were detected in this system, consistent with previous studies (Huddy et al., 2015; Kantor et al., 2015; van Zyl et al., 2011). Mapping reads from each assembly back to the non-redundant genome set demonstrated that this set accounted for between 72.5 and 93.2% of the data (**Figure 3.2**). This level of genome recovery approaches that reported for much simpler communities such as those from the infant gut (Raveh-Sadka et al., 2015). The de-replicated metagenomic dataset was used as a database for proteomic searches and accounted for 34, 32, and 15% of high-quality peptides from the SCN^- reactor biofilm, $NH_4(SO_4)_{1/2}$ reactor biofilm, and $NH_4(SO_4)_{1/2}$ reactor planktonic samples, respectively, at T4. This level of identification is comparable to that seen with the same type of analysis on infant gut metaproteomes paired to metagenomic databases (Xiong et al., 2015).

Community structure and metabolic potential

We examined community composition across the two-reactor time series. Hierarchical clustering of samples based on their community compositions grouped the samples first by the reactor, then by type of biomass and time point from which the samples were taken (**Figure 3.3**). Clustering organisms by abundance delineated several distinct groups: a small subset of organisms was present in both reactors, while other subsets were found at high-abundance in SCN^- community or the $NH_4(SO_4)_{1/2}$ community. Still other organisms were abundant primarily in the inoculum (**Figure 3.3**). In order to identify key organisms in the bioreactor communities, we characterized the metabolic potential encoded and expressed by each genome with respect to the key processes of SCN^- degradation, sulfur, ammonium, and nitrite oxidation, denitrification, and carbon fixation (**Figure 3.3**, **Table S3.1A**).

SCN^- removal and sulfur cycling

Four genomes contain one of two types of known SCN^- hydrolases (Arakawa et al., 2007; Hussain et al., 2013). These four, Thiobacillus_1, Thiobacillus_3, Thiobacillus_4 and Afipia_1, were abundant only in the SCN^- reactor (**Figure 3.4**). Proteomics data, obtained from the final time point, support activity of these organisms in SCN^- degradation, sulfur oxidation and carbon fixation (**Figure 3.5**). We also identified with proteomics nearly all proteins predicted in the recently described SCN^- operon from *Thiobacillus* spp. (Kantor et al., 2015), including an SQR-like protein at high abundance (**Table S3.2**). These observations support the involvement of all of these genes in the thiocyanate degradation pathway.

All six recovered *Thiobacillus* genomes encode the potential for autotrophic growth on sulfur compounds (**Table S3.1A**). However, relatively little is known about the process of sulfur oxidation by *Thiobacillus* spp., although its genome contains numerous sulfur-related genes from multiple pathways (Beller et al., 2006b). According to previous studies, some of these genes are constitutively expressed in *Thiobacillus denitrificans* (*sox*, *rDsr*, *apr*, *atps*) whereas others are upregulated under denitrifying conditions (SQR) (Beller et al., 2006a). We identified all sulfur oxidation genes in proteomics for several of the Thiobacilli described here (**Figure 3.5**).

In addition to these thiocyanate-degrading bacteria, twenty-two other organisms possess the Sox pathway (including at least 4 of *soxX*, *Y*, *Z*, *A*, and *B*, with or without *soxCD*; **Table S3.1A**). Of these, Burkholderiales_6, Thiobacillus_2 and Rhizobiales_3 were among the most abundant non-SCN^- degraders in the SCN^- reactor (**Figure 3.4**). Sulfur oxidation may proceed from sulfide to elemental sulfur or sulfate, likely determined by the availability of electron acceptors, as discussed below. Sulfur globules may be produced by SCN-degrading *Thiobacillus* spp., which use the reverse dissimilatory sulfite reductase (rDsr) pathway instead of soxCD (Loy et al., 2009). In turn, other sulfur oxidizers may use this elemental sulfur and any excess sulfide produced by SCN^- degradation. Since chemical data showed that SCN^- was completely converted to sulfate, and proteomic data showed that non-SCN^- degraders expressed Sox proteins (**Figure 3.5**), we suspect that sulfur species were passed from SCN-degraders to the rest of community in this type of "metabolic handoff".

Community dynamics and SCN^- removal across increased loadings

As SCN^- loadings and SCN^- degradation rate increased (**Figure 3.1**), the relative abundance of SCN-degrading *Thiobacillus* spp. also increased (**Figure 3.4**), with Thiobacillus_1 alone accounting for 38% of all reads at T2. During operation at the two final loading rates (T3 and T4), the Thiobacillus_1 population decreased in relative abundance, concordant with a decrease

in SCN^- degradation rate and reactor efficiency. Given this observation, we looked for other changes associated with loss of reactor efficiency.

First, a read mapping-based sequence variance analysis of the Thiobacillus_1 population in each sample showed that it was largely clonal throughout the time series but contained two distinct strains at the last time point. The relative proportions of the subpopulations were ~60 and 40% (**Figures 3.6A** and **3.6B)**. The genome for the second strain was not recovered, but a scaffold corresponding to the SCN^- operon was identified among the un-binned metagenomic data from this time point. We noted a few differences in the protein sequences of genes contained in this operon, which could in principle affect the efficiency of SCN^- degradation relative to the dominant strain (**Figure 3.7**).

A second change in the community was the increase in relative abundance of Burkholderiales_6, a sulfur oxidizing mixotroph that became dominant in T3 and T4 (**Figure 3.8**). No known genes for SCN^- degradation were found in the Burkholderiales_6 genome, but proteins encoded by *sox* genes from this genome were identified in proteomics (**Figure 3.5**). Hence we infer that the Burkholderiales_6 organism likely used excess reduced sulfur species produced in the reactor for growth. Thirdly, substantial algal growth was visually observed in the planktonic portion of the SCN^- reactor at the last time point, which could have affected microbial population dynamics and reactor efficiency. Lastly, the build-up of residual SCN^- in the reactor may have led to toxicity effects including lower bacterial replication rates (see below). This in turn could have reduced the SCN^- degradation rate, creating a negative feedback effect on reactor performance.

Overall, we speculate that the decline in reactor efficiency at high loading rates occurred when the capacity for SCN^- degradation was exceeded. The abundance of *Thiobacillus* cells may have been insufficient to meet the demand for SCN^- degradation owing to a maximum specific SCN^- degradation rate. Alternatively, degradation may have been inhibited or the per-cell rate of degradation may have decreased. Since metagenomic data provide relative abundance information, the apparent decrease in Thiobacillus_1 relative abundance may have been due to increases in abundance of other organisms. Further studies are needed that apply measurements of absolute, species-specific biomass and metabolic rates.

Nitrogen removal and dynamics over time

Since SCN^- degradation releases nitrogen in the form of ammonium, we looked for possible mechanisms of nitrogen cycling and removal to N_2. No anamox genes were detected in any genome and, based on identified genes and genomes, the conversion of ammonium to nitrate occurred aerobically. A single genome in the dataset, Nitrosospira_1, encoded the potential for ammonium oxidation (**Figure 3.3**). The Nitrosospira_1 organism became enriched at early time points in the $NH_4(SO_4)_{1/2}$ reactor and later in the SCN^- reactor (**Figure 3.4**) and was active at the final time point, based on proteomic data (**Figure 3.5**). Two nitrite oxidizers, Nitrobacter_1 and Nitrobacter_2, were present at low abundances in the $NH_4(SO_4)_{1/2}$ reactor (**Figure 3.4**) and were so low in abundance in the SCN^- reactor that their genomes did not assemble. However, proteins for nitrite oxidation corresponding to one of these genomes were detected in samples from both reactors (**Figure 3.5**). Low abundance but high activity has been observed previously for other nitrite oxidizing bacteria, and some have hypothesized that high nitrite oxidation activity may be a requirement for growth, given the low energy yield of this metabolism (Baker et al., 2013). Despite the low relative abundance of both ammonium and nitrite oxidizers, nitrate was detected in the effluent of both reactors during the initial ramping phases (**Figure 3.1**).

Searching for mechanisms of nitrate removal, we identified 92 genomes that contained at least one gene involved in denitrification (including *nar*, *nap*, *nirS*, *nirK*, *norB*/*norZ*, and *nosZ*; **Table S3.1A**). Fourteen of these genomes encoded the capacity for complete nitrate reduction to N_2, while 49 had only one gene in the pathway. Complete denitrifiers included three predicted autotrophs implicated in SCN^- removal (**Figure 3.4**). SCN^- hydrolysis and concomitant sulfide oxidation coupled to denitrification may be possible in these organisms, as was described for *Thioalkalivibrio* (Sorokin et al., 2004). The sulfide oxidizing Burkholderiales_6, which became abundant in T3 and T4 in the SCN^- reactor, also contributed to denitrification (**Figure 3.4** and **3.3**). All denitrification genes except *norB* were identified in proteomics from the SCN^- reactor biofilm at T4 (**Figure 3.5**). The limited detection of NorB may be an extraction bias artifact due to numerous transmembrane domains in these proteins (Hino et al., 2010; Matsumoto et al., 2012).

In the $NH_4(SO_4)_{1/2}$ reactor, three of most abundant bacteria, Rhodanobacter_1, Xanthomonadales_1, and Novosphingobium_1 may have roles in denitrification (**Figure 3.4** and **Table S3.1A**). The potential for dissimilatory nitrate reduction to ammonia (via *nrfA)* was detected in genomes from several members of the Bacteroidetes and one *Aeromonas* sp. but most of these were abundant only in the inoculum (**Table S3.1A**) and NrfA was not detected in the proteomic data.

Changes in bacterial replication rates across the time series

We used a recently established approach to investigate bacterial replication rates from metagenomics (Korem et al., 2015) with a new implementation that reports rates as index of replication (iRep) values (Brown et al., 2016). In the SCN^- reactor biofilm, iRep values increased between T0 and T1, suggesting replication proceeded more quickly in newly-forming biofilm than in inoculum biofilm taken from the long-running SCN^- stock reactor (**Figure 3.9** and **Table S3.3**). Over the remainder of the experiment (T2-T4), iRep values decreased for most organisms, especially toward the end of the experiment. This may have been due to toxicity of residual SCN^- in the reactor media or to spatial limitations for growth within the thickening biofilm.

In contrast, bacterial growth rates in the $NH_4(SO_4)_{1/2}$ reactor biofilm were initially low, but increased from T1 to T3. This is suggestive of a period of adaptation, as organisms adjusted to the new conditions relative to the SCN^- stock reactor (**Figure 3.9**).

Biofilm and planktonic communities

The planktonic and biofilm portions of the reactor were sampled separately in order to understand whether these represented independent communities. At T1, the planktonic fraction of each reactor was very dilute, yielding inadequate amounts of DNA for sequencing, and at T4, the planktonic portion of the SCN^- reactor was overgrown with algae. At T2, metagenomes for planktonic samples from both reactors were highly enriched in a rotifer genome, which accounted for 45 and 25 % of the sequence data in the $NH_4(SO_4)_{1/2}$ and SCN^- reactors, respectively. With microscopy, rotifers were observed grazing on biofilm (**Figure 3.10**). In the SCN^- reactor, planktonic samples taken at T2 and T3 were similar to corresponding biofilm samples with the notable difference that the bacterium TM7_2, a putative symbiont belonging to the Saccharibacteria (Bor et al., 2015; He et al., 2015), was highly enriched in planktonic samples (**Figure 3.4** and **Figure 3.8**).

Other eukaryotes and symbionts were observed in both reactors, and many of these organisms were at higher relative abundance in the planktonic samples compared to biofilm

(**Table S3.1B**). Recurring mitochondrial sequences recovered in the metagenomes corresponded to relatives of *Vermamoeba vermiformis*, *Acanthamoeba* spp., *Naegleria fowleri*, and *Chlorella* sp., identified based on the phylogenetic profile of their proteins and searches against NCBI-nr (**Table S3.1B**). Eleven bacterial genomes in the dataset derived from predicted symbionts, as indicated by their phylogenetic affiliations and/or reduced genome sizes (**Figure 3.3** and **Table S3.1A**). These included complete genomes for two Saccharibacteria (formerly TM7). One of these, TM7_2 (noted above), was the only putative symbiont found at high abundances in both reactors (**Figure 3.4**).

Overall, the majority of the biomass in the reactor was likely in the form of biofilm, and sloughing may have contributed to the planktonic community. While SCN^- degradation itself does not rely on biofilm (van Zyl et al., 2015), the formation of biofilm effectively uncoupled the HRT from bacterial growth rates, preventing wash-out as the HRT was decreased. This may have allowed *Thiobacillus* spp., and nitrifiers to reach higher population sizes than would otherwise have been possible, thereby converting higher loadings of SCN^- to nitrate.

Long-term community stability and phage susceptibility

We compared the 114 bacterial genomes in this study to 86 genomes reconstructed in a prior study of the SCN^- stock reactor and a daughter reactor fed cyanide and SCN^- (CN-SCN reactor) conducted two years earlier (Kantor et al., 2015). Thirty-one genomes were matched, overlapping by at least a total of 1 Mbp at 98% nucleotide identity (**Table S3.1A**). These included close relatives of the three SCN-degrading *Thiobacillus* spp. enriched in the SCN-treated reactor studied here. Given the importance of these three populations, we looked for evidence of recent phage infections, based on changes to CRISPR loci over time. Thiobacillus_1 has no CRISPR locus, perhaps making it more susceptible to acquisition of mobile elements and to phage attack. The CRISPR locus in the Thiobacillus_3 genome was identical in all versions of this genome recovered from the current study, but was not recovered in the previous study. The recovered Thiobacillus_4 genomes belonged to two distinct CRISPR sub-types that differed from one another in 12 spacers at the 3' end of the array: sequences from biofilm and planktonic inoculum samples comprised one version, while sequences from later in the time series (and those recovered previously in the SCN^- stock reactor and the CN-SCN reactor) comprised the second version. Importantly, no spacers from Thiobacillus_3 and Thiobacillus_4 targeted any sequence in the metagenomes (or previous metagenomes from the SCN^- stock or CN-SCN reactors), suggesting little recent phage interaction.

System overview

SCN-degrading chemolithoautotrophs (*Thiobacillus* and *Afipia*) can oxidize the SCN-sulfur as their sole energy source under both aerobic and anaerobic conditions (**Figure 3.11A**). Sulfide oxidation may stop at elemental sulfur when parts of the reactor become anaerobic (Moraes et al., 2012) (**Figure 3.11B**), and proteomic data suggest that several *Thiobacillus* spp., as well as Afipia_1, and Bukrholderiales_6 coupled this sulfur oxidation to denitrification (**Figure 3.5**). In fact, these organisms are inferred to be the key denitrifiers in the system. Other sulfur oxidizing autotrophs and mixotrophs may use reduced sulfur compounds produced during SCN^- degradation, including elemental sulfur. The combination of sulfide oxidation activities by SCN-degraders and non-degraders may explain the observed complete conversion of sulfur from SCN^- to sulfate (**Figure 3.4B)**. The breakdown of SCN^- produces ammonium that is converted to nitrate by autotrophic ammonium and nitrite oxidizers. Overall, heterotrophs in the system

contributed to sulfur oxidation and denitrification, and likely also to biofilm formation and biofilm integrity (perhaps via filamentous morphology, see **Figure 3.10**). Heterotrophs may also break down SCN^- as a source of nitrogen (via an unknown pathway). However, this SCN^- degradation may be inhibited if there is abundant nitrogen available as ammonium, as observed in some alkaliphiles (Sorokin et al., 2001). Lastly, eukaryotes such as rotifers and amoeba are predators and thus contribute to carbon turnover in the system.

Engineering of SCN^- degradation by a microbial community

We conclude that the consortia can completely hydrolyze SCN^- and oxidize sulfide under a range of SCN^- loadings, but that reduced performance can result at higher loadings where residual SCN^- accumulates. Smaller increases in concentration to reach higher loadings may lead to sustained reactor performance by allowing microbial cell numbers and associated volumetric degradation rates to keep pace with input SCN^-. Having an understanding of the community structure, which organisms perform which functions, and how organisms respond to changes in reactor conditions allows us to provide suggestions for further reactor design. Specifically, *Thiobacillus* spp. depend on sulfide generated by SCN^- degradation for energy. The SCN^- degradation rate may be faster under aerobic conditions, where the energy yield for sulfide oxidation is highest. Thus, for maximum conversion of SCN^- to ammonium and sulfate, it is desirable to promote a system dominated by autotrophs and aerobic growth. Conversely, to promote nitrogen removal in the presence of high SCN^- loading and degradation rates, anaerobic denitrification is required. The bioreactor communities can form dense biofilms, enabling retention of biomass and shorter HRTs, and providing microenvironments for nitrification and denitrification. At the industrial scale, attached growth could reduce the reactor footprint by increasing volumetric degradation rates (product of specific degradation rate and biomass concentration) and removing the requirement for a separator.

Biofilm formation may be encouraged by increasing surface area to provide sufficient access to O_2 for biofilm-associated aerobes as well as anaerobic zones for denitrification (e.g., introducing solid carriers). This is common in many wastewater treatment processes and could be implemented with the SCN-degrading community studied here. Alternatively, where nitrogen removal is a high priority, a two-stage reactor system would allow SCN^- degradation, sulfur, ammonium, and nitrite oxidation occur aerobically, with denitrification occurring in the second stage in a manner similar to that established for wastewater treatment.

Our work highlights the applicability of bioinformatics tools to gain a mechanistic understanding of contaminant degradation by a microbial community, to assess community stability, and ultimately, to inform engineering design choices. Others have called for broader use of metagenomics to advance biotechnology, including in wastewater treatment (Ju and Zhang, 2015; Roume et al., 2015), and this study represents a step toward the use of such techniques in the field. The level of resolution achieved using metagenomics combined with metaproteomics allowed us to access not only phylogenetic classifications and diversity of community members, but also which members express key proteins involved in the degradation process. The dataset and analysis provide valuable information that can be used to generate primers or probes for on-site measurements.

Methods

Reactor set-up, inoculation, and operation

Two continuous stirred tank reactors were inoculated with homogenized biofilm and planktonic samples from the long-running SCN^- stock reactor at the University of Cape Town. Reactors were stirred with a pitched-blade impeller at 270 rpm and sparged with filtered air at 900 mL/min. One reactor was fed KSCN while the other was fed $NH_4(SO_4)_{1/2}$ at equivalent nitrogen loadings in order to mimic the end-products of thiocyanate degradation. Both reactors were also fed molasses (150 mg/L) and KH_2PO_4 (0.28 mM) to provide supplemental nutrients. Feed contained increasing amounts of KOH to modulate reactor pH as necessary (**Figure 3.1**) and small amounts of 5 N KOH were added directly to reactors if observed pH was ≤ 6.5. Bicarbonate (4 g/L) was added to the feed to buffer the system from day 112 to day 136.

The reactors were run in batch-fed mode until SCN^- degradation was stably observed in the SCN^- reactor, at which time both reactors were switched to continuous feeding at a residence time of 42 hours (day 5). Subsequently, the hydraulic retention time (HRT) of both reactors was lowered from 42 hours to 12 hours (days 5-68). Upon reaching 12 hours HRT, this was maintained for the remainder of the experiment and the feed concentration of SCN^- or equivalent $NH_4(SO_4)_{1/2}$ was increased stepwise. The reactor was allowed to stabilize between each step to reach steady state.

Sampling

Samples of biomass from each reactor were taken for metagenomic sequencing just before increases in feed concentration (**Figure 3.1** and **Table 3.2**). Approximately 0.5 g (wet-weight) of biofilm was scraped from the wall of each reactor with sterile spatula and stored at -60 °C. Paired samples of planktonic biomass were collected by filtering 300 mL of the liquid phase from each reactor onto a sterile 0.22 μm filter. Biomass was gently washed off the filter with sterile water, pelleted, and stored at -60 °C until further analysis. Filtered media was returned to the reactor to maintain chemical continuity.

Chemical analysis

Bulk liquid was sampled daily for chemical analysis, filtered through a 0.22 μm filter, pH analyzed, and frozen at -20 °C until further analysis. SCN^- was measured using High Performance Liquid Chromatography as described previously (Kantor et al., 2015). Ion chromatography for anions was performed using a Dionex ICS – 1600 Ion Chromatography system fitted with Dionex IonPac AS16 column and a conductivity detector. The system was run isocratically using a 22 mM sodium hydroxide eluent at a flow rate of 1 ml/min at ambient temperature. The pH was measured with a Cyberscan 2500 micro pH meter.

DNA extraction and sequencing

DNA was extracted using a NucleoSpin® soil genomic DNA extraction kit (Machery-Nagel, Germany) with the inclusion of a repeated extraction step, according to the manufacturer's instructions. Paired-end Illumina TruSeq libraries with either tight insert fragment sizes of 800 bp or regular insert sizes of 500 bp, depending on the sample, were prepared at the Joint Genome Institute (Walnut Creek, CA) (**Table 3.2**). Libraries were sequenced on an Illumina HiSeq-2500 in rapid run mode to yield 250 bp paired-end reads.

Metagenomic assembly, binning, and annotation

Reads from each sample were trimmed based on quality scores using sickle (https://github.com/najoshi/sickle) and then assembled independently with idba_ud (Peng et al.,

2012). Binning of the assembled scaffolds was performed using ggKbase tools (ggkbase.berkeley.edu) based on scaffold taxonomy, percent GC, and sequencing coverage. Within each assembly, bins were refined and added using differential abundance data visualized in emergent self-organizing maps (ESOMs) as in Sharon *et al.* (2013). Data from two additional samples from a reactor with the same inoculum were used to assist in binning only, and are not described further here. In order to generate each ".lrn" file for ESOM training, reads from each of the 17 samples were mapped to one assembly using Bowtie 2 (Langmead and Salzberg, 2012) with default parameters. Mappings were parsed using prepare_esom_files.pl (https://github.com/CK7/esom/blob/master/prepare_esom_files.pl) (Sharon et al., 2015) to gather coverage information for 10 kb windows of each scaffold, with a minimum scaffold size of 5 kb. The resulting matrix contained 17 columns of normalized coverage information (one for each sample used) and an additional column representing the log of the sum of all normalized coverages for a given 10 kb window. Each ESOM was trained and visualized using databionic ESOM tools (http://databionic-esom.sourceforge.net/index.html) (Ultsch and Mörchen, 2005). In two samples (planktonic inoculum and SCN^- reactor T2 biofilm), subassemblies using $1/60^{th}$ or $1/50^{th}$ of the reads, respectively, were performed to improve assembly of the most abundant organism as previously described by Hug *et al*. (Hug et al., 2015).

Many bins were redundant given the recurrence of organisms across the time series experiment. Nucmer (Kurtz et al., 2004) was used to align sequences and identify sets of bacterial genomes sharing ≥ 98% nucleotide identity across > 50% of the sequence. The best bin from each set of replicates was chosen for inclusion in a de-replicated dataset based on genome completeness and length. Genome bins that had no replicates were also included in the de-replicated set, except for two known contaminant genomes ('Candidatus Altiarchaeum hamiconexum' and an Epsilonproteobacterium) from another sequencing run on the same lane, which were excluded from further analysis. Bins were excluded from the final de-replicated dataset if they contained < 36 of 51 single copy genes (SCG) or > 8 multi-copy SCG. One recurring eukaryotic genome bin, one chloroplast, several mitochondria, phages, eukaryotic viruses, and plasmid bins were included in the de-replicated dataset. De-replicated bacterial genomes were curated using ra2.py, an automated curation method that makes use of coverage and paired-end read information to find and reassemble or mask regions with mis-assemblies (https://github.com/christophertbrown/fix_assembly_errors/releases/tag/2.00) (Brown et al., 2015). Curation used the reads of the sample from which the genome originated.

Annotation of genome bins used reciprocal ublast (Edgar, 2010) searches against KEGG (Kanehisa et al., 2016) and UniRef100, (Suzek et al., 2015) as well as single-direction searches against UniProt (The UniProt Consortium, 2015). Functional genes were identified by annotations and using hmmsearch (HMMER 3.1b2; http://hmmer.org/) with Hidden Markov Models (HMMs) from TIGRFAM (v15.0), PFAM (Finn et al., 2016), and with custom HMMs (accessible at https://github.com/banfieldlab) (Anantharaman et al., 2016b).

Community composition

Bowtie2 (Langmead and Salzberg, 2012) was used with default settings to map reads from each sample to the de-replicated dataset (**Figure 3.2**). The resulting mapping files were filtered using mapped.py (https://github.com/christophertbrown/mapped) to remove reads that mapped with > 3 mismatches. Coverage for each genome in each sample was calculated and values ≤ 1x were converted to zero. Genome coverage values were then normalized by dividing by the number of reads for each sample and then multiplying by the number of reads in the largest sample.

Normalized coverage was used as a proxy for the relative abundances of organisms across samples (**Figure 3.3**).

Variant analysis

Reads from each sample were mapped to the de-replicated set of sequences using Bowtie2 with default settings. The mapping files were subsetted with mapped.py (https://github.com/christophertbrown/mapped), filtering for only those scaffolds belonging to a genome of interest. The resulting files were converted to bam format, sorted, and indexed using samtools (Li et al., 2009). Bam files were then passed individually to Freebayes (Garrison and Marth, 2012) with the parameters --min-alternate-fraction 0.02 and --pooled-continuous. The counts of reference alleles and alternate alleles reported in the vcf file were used to plot strain abundances. Depth across individual regions of the genome was calculated using samtools, and read mapping was visualized with Geneious (Biomatters Ltd).

Replication rate calculations

Indices of replication (iRep values) were calculated for each genome in each sample using iRep.py (Brown et al., 2016) (https://github.com/christophertbrown/iRep). The results were filtered to collect iRep values where genomes had > 10x coverage, the R^2 value for fitting the iRep line to the ranked coverage curve was > 0.90, 98% of all 5 kb windows along the genome were used, and genomes were assembled into no more than 175 fragments per 1 Mbp of sequence. To avoid non-specific mapping, only reads with ≤ 3 mismatches in each mate-pair were used. We excluded values from analysis when a given genome was known to derive from multiple strains in the same sample.

Protein extraction and proteomic data analysis

Proteins were extracted as previously described (Chourey et al., 2010). An aliquot consisting of ~1 mg of protein was subjected to trichloroacetic acid precipitation and subsequent digestion with trypsin. Proteolytic peptides were analyzed via an online nano 2D LC–MS/MS system interfaced with hybrid LTQ-Orbitrap-Velos MS (ThermoFisher Scientific). A 25 µg aliquot of peptides was loaded onto a biphasic column consisting of reverse phase followed by strong cation exchange and analysed by eleven step MudPIT (Multidimentional protein identification technology) as described previously (Lochner et al., 2011). The instruments were operated in a data-dependent mode. MS1 was performed in Orbitrap and data dependent MS/MS was performed in LTQ (top twenty), 1 microscan for both full and MS/MS scans; normalized collision energy 35% and dynamic exclusion time of 30 seconds. MS2 mass spectra were analyzed using the following software protocol: Thermo RAW files were converted to mzML peaklists by ProteoWizard msConvert (Chambers et al., 2012) and database searches used Myrimatch (Tabb et al., 2007), with the dereplicated set of genomes as the database. Configuration parameters were as follows: fully tryptic peptides with any number of missed cleavages, an average precursor mass tolerance of 1.5 *m*/*z*, a mono precursor mass tolerance of 10 ppm, a fragment mass tolerance of 0.5 *m*/*z*, a static cysteine modification (+57.0214 Da), an N-terminal dynamic carbamylation modification (+43.0058 Da), and a dynamic oxidation modification (+15.9949 Da). Peptide identifications were filtered with IDPicker v3.1 (Ma et al., 2009) to < 1% peptide false discovery rate (at the peptide level: maximum *Q* value <2%, minimum one spectra per peptide, and minimum one spectra per match; at the protein level:

minimum two distinct peptides, minimum one additional peptide, and minimum two spectra per protein). ScanRanker (Ma et al., 2011) was used to assess spectral quality.

Analysis of proteins involved in key metabolic pathways considered spectral counts for unique peptides and total spectral counts for each protein from two technical replicates.

Data availability

Raw read data is accessible at NCBI under accession number SRP056932 (http://www.ncbi.nlm.nih.gov/sra/SRP056932). Genome bins and sequences for scaffolds, genes, and proteins may be viewed and downloaded at http://ggkbase.berkeley.edu/scnpilot-dereplicated/organisms. Proteomics data is available at https://massive.ucsd.edu/ProteoSAFe/datasets.jsp (MassIVE ID MSV000080104).

Table 3.1. Counts and completeness estimated by single copy gene (SCG) inventories for de-replicated bins resulting from 15 metagenome assemblies.

Bacterial genomes (144)	**Count**
Genomes ≥ 96% complete (49/51 SCG)	111
Genomes in ≤ 30 scaffolds and 49/51 SCG	37
Genomes with ≥ 90% of sequence in scaffolds > 10 kb	117
Non-bacterial bins (90)	**Count**
Plasmids and mobile elements	45
Phage	25
Eukaryotes	1
Mitochondria	15
Chloroplasts	1
Viruses	3

Table 3.2. Sequencing depth (untrimmed reads) and insert size used for each of the 15 samples.

Time point (d)	**Feed concentration**	**Thiocyanate reactor**		**Ammonium sulfate reactor**	
		Biofilm	**Planktonic**	**Biofilm**	**Planktonic**
T0: 0 d	Inoculum	7.49 Gbp*	17.69 Gbp	same inoculum	same inoculum
T1: 152 d	0.43 mM/h (300 ppm SCN feed)	5.79 Gbp*	-	14.60 Gbp*	-
T2: 172 d	0.72 mM/h (500 ppm SCN feed)	12.20 Gbp*	16.13 Gbp	15.82 Gbp*	16.30 Gbp
T3: 199 d	1.08 mM/h (750 ppm SCN feed)	15.81 Gbp*	17.05 Gbp	17.12 Gbp*	12.01 Gbp
T4: 217 d	1.43 mM/h (1000 ppm SCN feed)	38.54 Gbp*+	-	19.26 Gbp*+	28.94 Gbp+

*indicates 800 bp insert size, otherwise a 500 bp insert size was used
"+indicates corresponding proteomics"

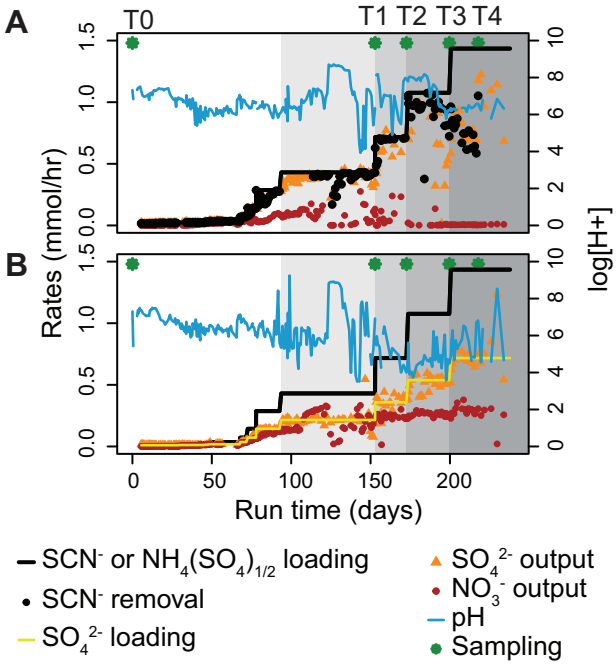

Figure 3.1. Chemical parameters of operation for the SCN^- reactor (**A**), and $NH_4(SO_4)_{1/2}$ reactor (**B**). Sampling time points are indicated above plots. Gray shading indicates the different loading regimes.

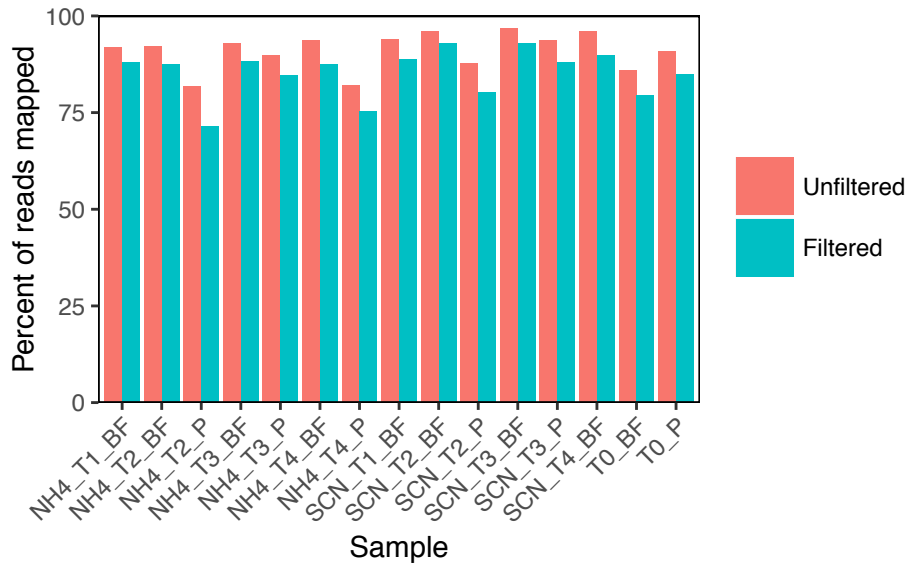

Figure 3.2. Reads accounted for by the dereplicated dataset of bacterial, eukaryotic, phage, and plasmid genomes. Paired reads were mapped against the de-replicated dataset using bowtie2, and the percentage of mapped reads was calculated before (red) and after filtering to remove reads with > 3 mismatches per 250 bp read (blue). Filtering also required mapping of both mates in a pair.

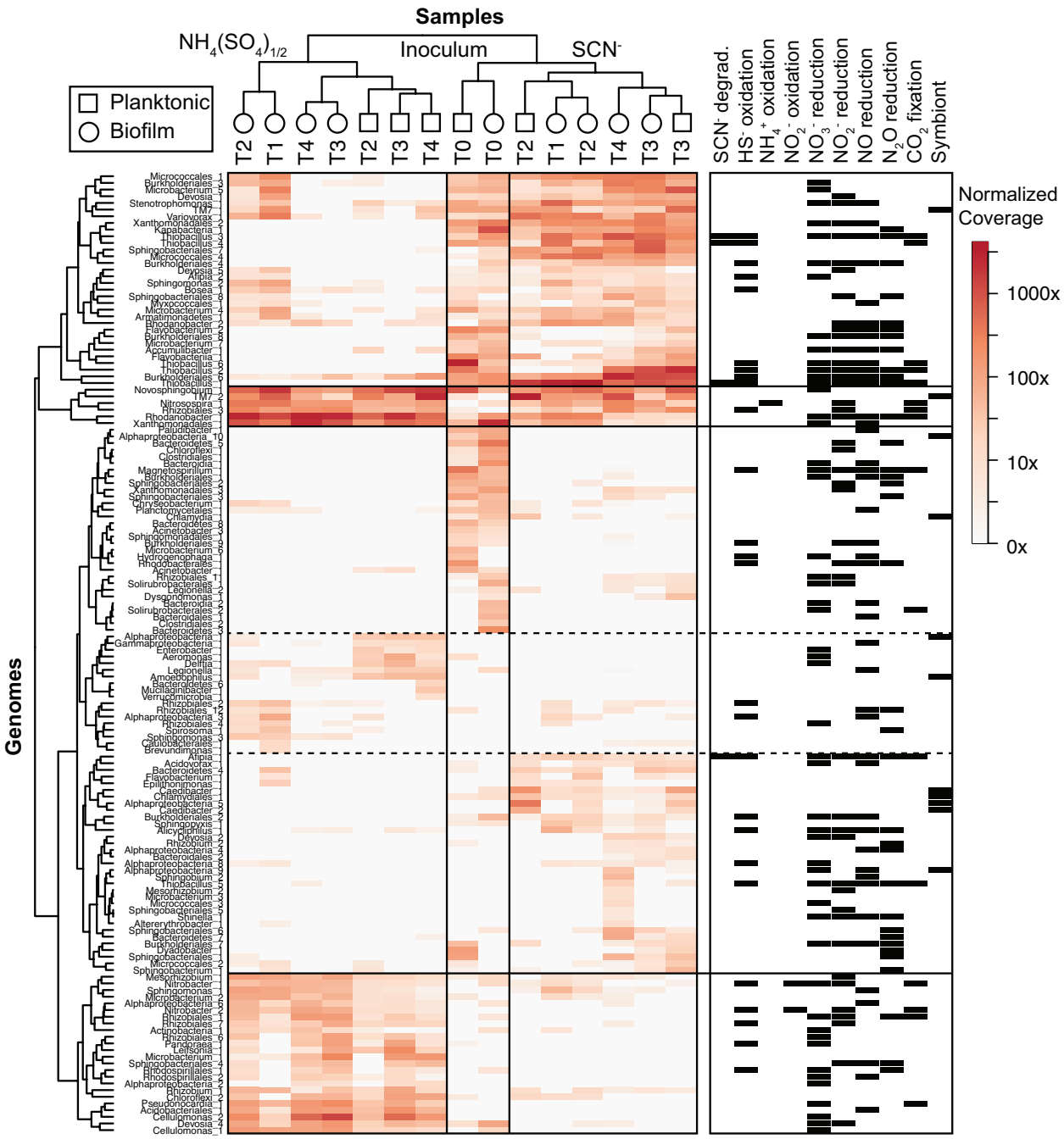

Figure 3.3. Relative abundances of the 144 recovered bacterial genomes across 15 samples and relevant metabolic pathways encoded by each genome. Color corresponds to log-scaled normalized coverage data for each bacterial genome in each sample. The presence of genes indicating metabolic potential is shown in black for each genome, and symbionts (based on taxonomy) are also noted here (see Table S2A). Hierarchical clustering of samples used Bray-Curtis dissimilarity and figure was constructed using HeatPlus in Rstudio (https://github.com/alexploner/Heatplus).

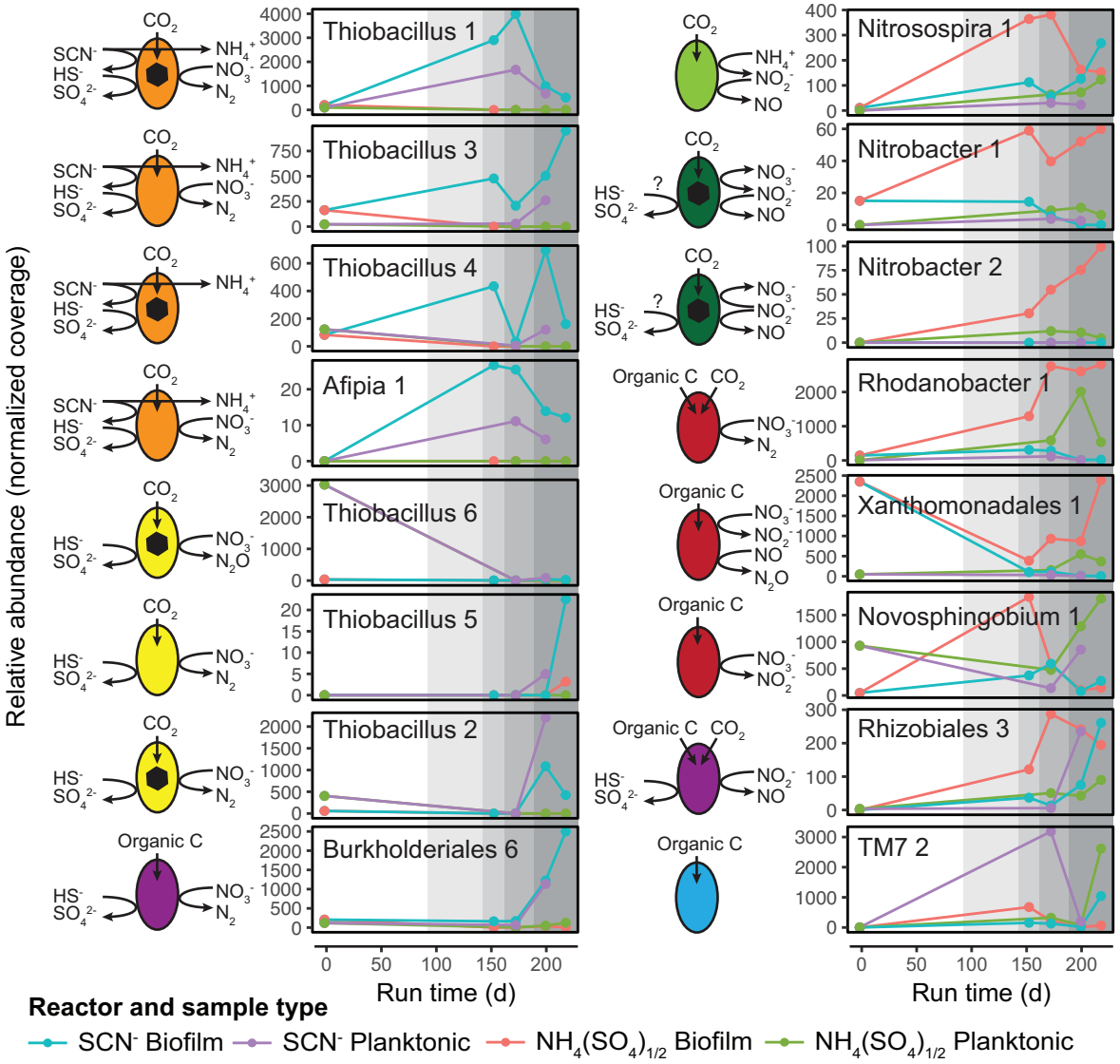

Figure 3.4. Metabolic potential and relative abundances of key organisms of interest over time in biofilm and planktonic samples from both reactors. Gray shading corresponds to increasing loading rates of SCN^- or $NH_4(SO_4)_{1/2}$ as in Figure 3.1. SCN^- degraders (orange) and several key sulfur oxidizers (yellow and purple) were found at higher relative abundance in the SCN^- reactor while ammonium and nitrite oxidizers (green) were at higher relative abundances in the $NH_4(SO_4)_{1/2}$ reactor. Several highly abundant heterotrophs (red and blue) and one possible sulfur oxidizing mixotroph (purple) were present in both reactors. Note different y-axis scales. Hexagons indicate carboxysomes where annotations support this prediction.

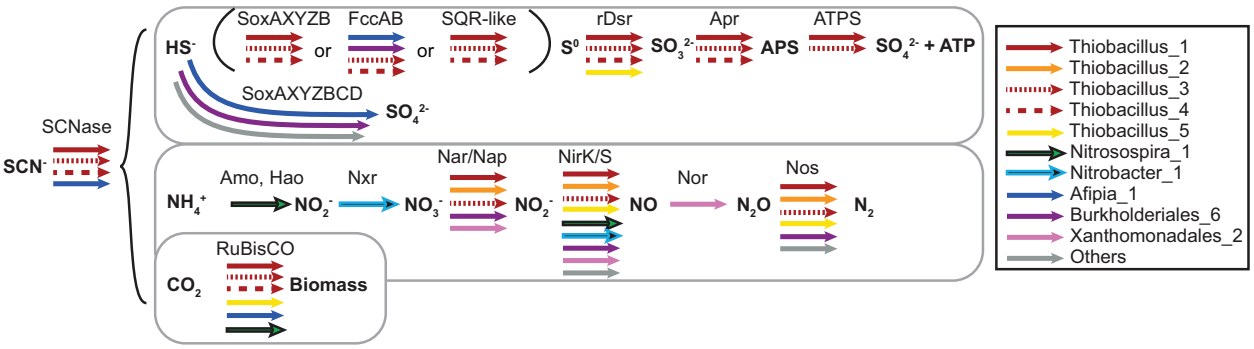

Figure 3.5. Metaproteomics at end point (T4) in SCN^- reactor shows expression of genes involved in SCN^- degradation and byproduct breakdown in key organisms. Each arrow indicates that the average of unique spectral counts across two technical replicates was ≥ 2 for at least one subunit or component of the enzyme complex involved in the chemical transformation.

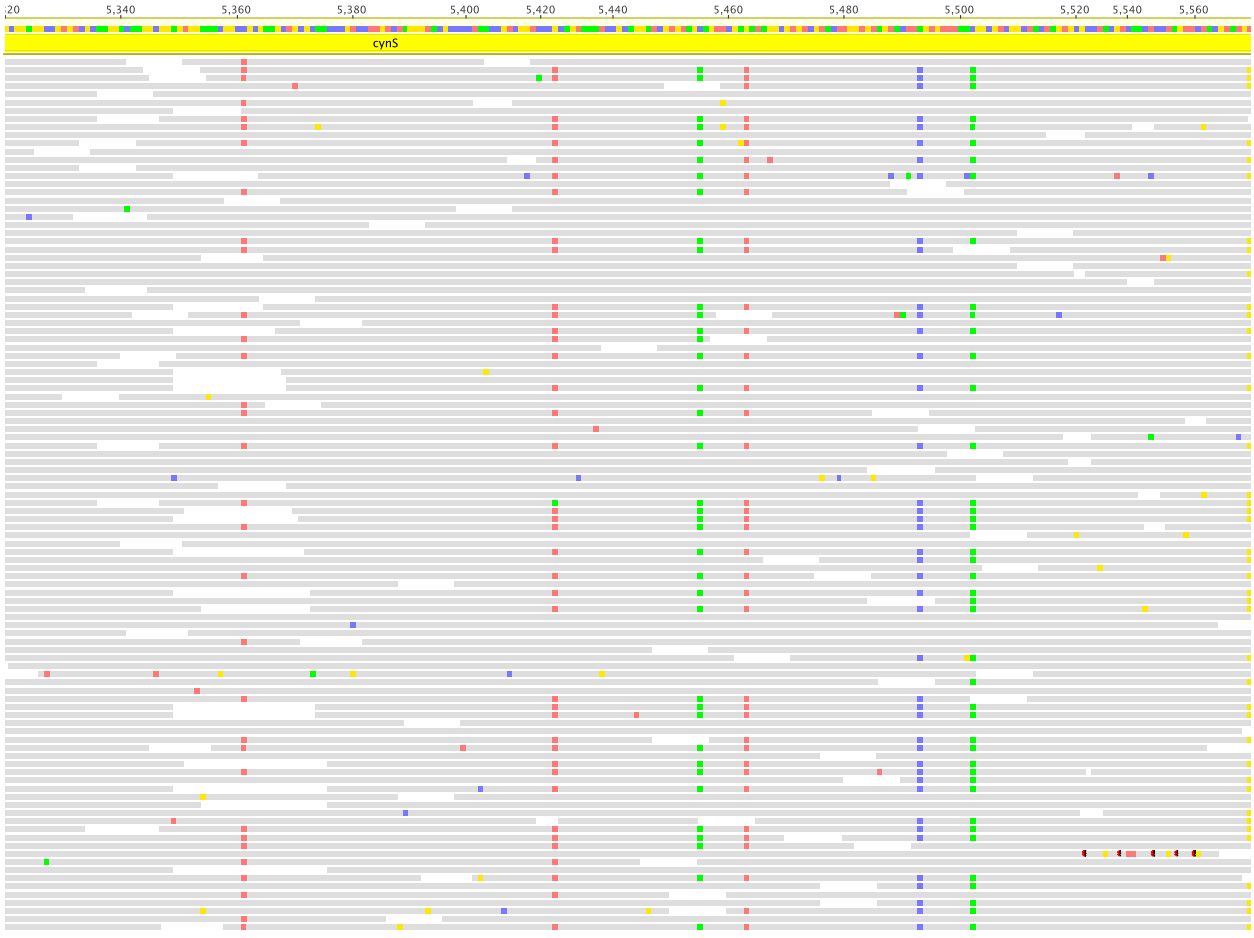

Figure 3.6A. Readmapping to the SCN^- operon of Thiobacillus_1 shows two strain variants in the T4 biofilm sample from the SCN^- reactor. Color denotes bases matching consensus (gray) and variant bases (red=A, green=T, blue=C, yellow=G). Figure shows zoom-in on region containing cyanase (*cynS*).

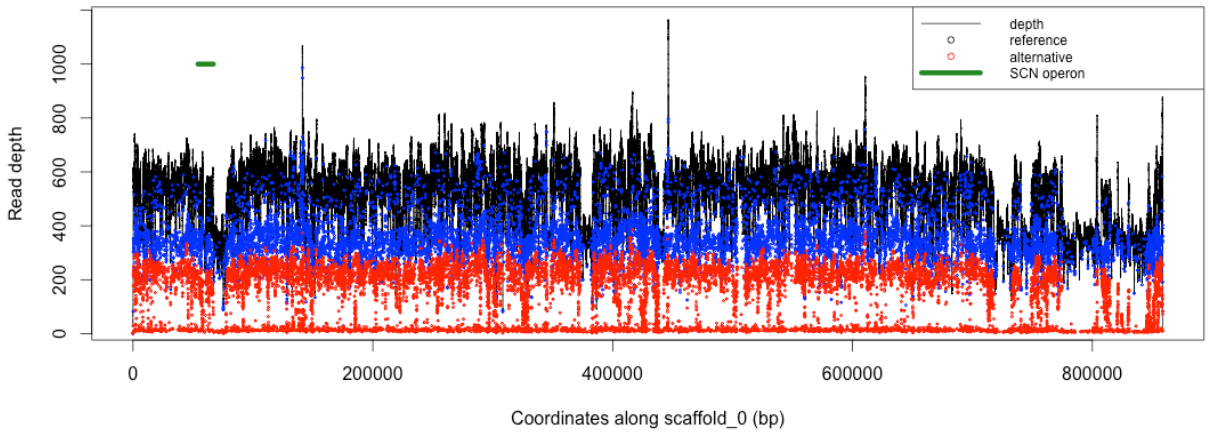

Figure 3.6B. Read depth in biofilm sample T4 from the SCN^- reactor across the longest contig (858717 bp) of Thiobacillus_1, a genome derived from the T2 biofilm sample. Total depth at each coordinate (black line); reads supporting the Thiobacillus_1 genome (blue); reads which map to the Thiobacillus_1 genome but contain an alternative sequence at a given site (red). Reference and alternative read counts are shown only where variants were detected. Drops in total depth occur where reads from the secondary strain were too divergent to map. The SCN operon (54,000-67,000 bp) is indicated (green line) and is immediately upstream from a region that appears to have been horizontally transferred (based on taxonomy of best-hits to open reading frames) into the primary strain but not the secondary strain.

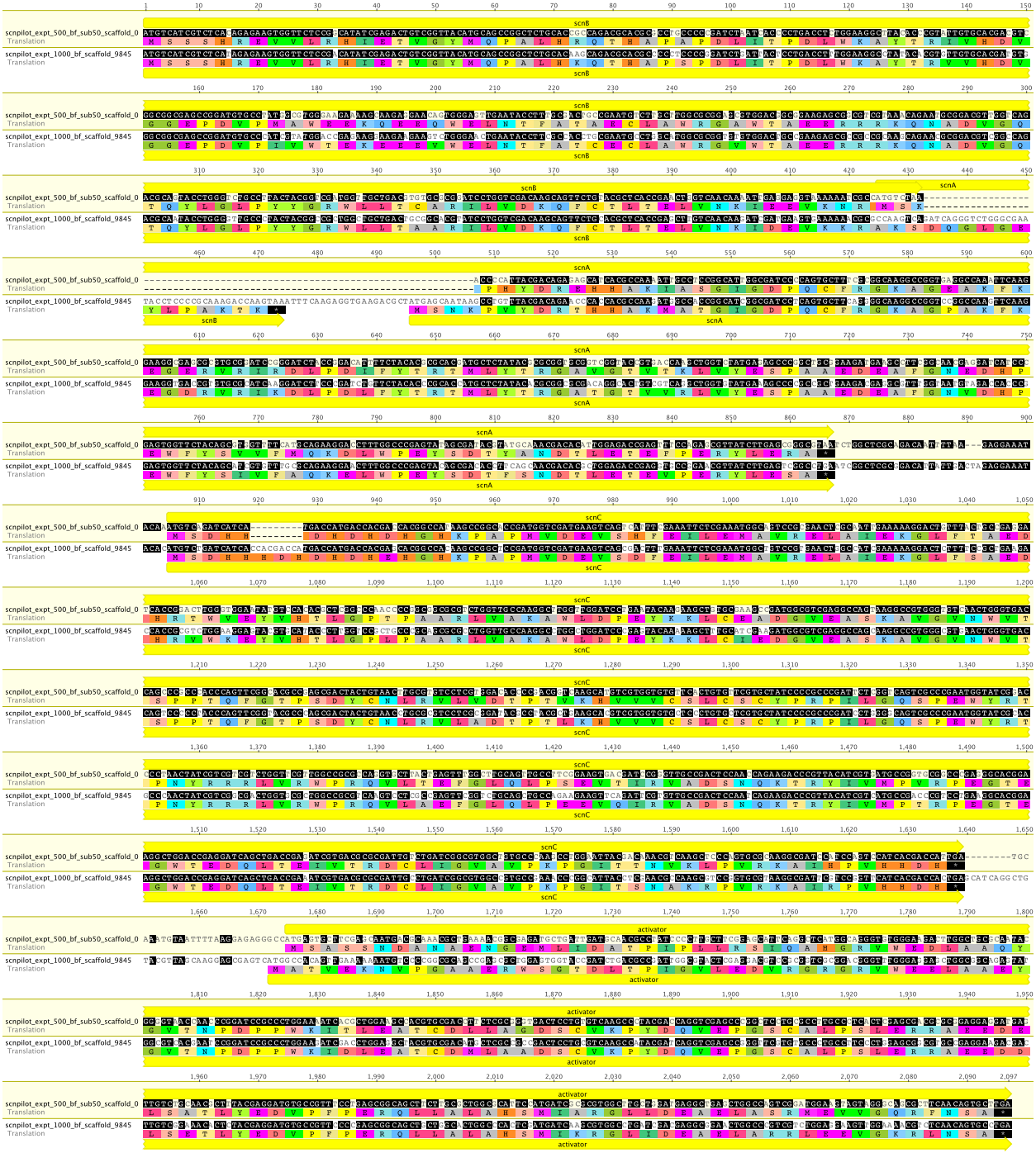

Figure 3.7. Nucleotide alignment of the SCN^- hydrolases from two strains of Thiobacillus_1 with translations for each open reading frame shown below. Thiobacillus_1 sequence as recovered from T2 biofilm SCN^- reactor (top track) and an alternative sequence recovered from an unbinned scaffold in the T4 biofilm SCN^- reactor (bottom track). The alignment was made using the map-to-reference function with manual correction in Geneious v.7.0.6 (Biomatters Limited).

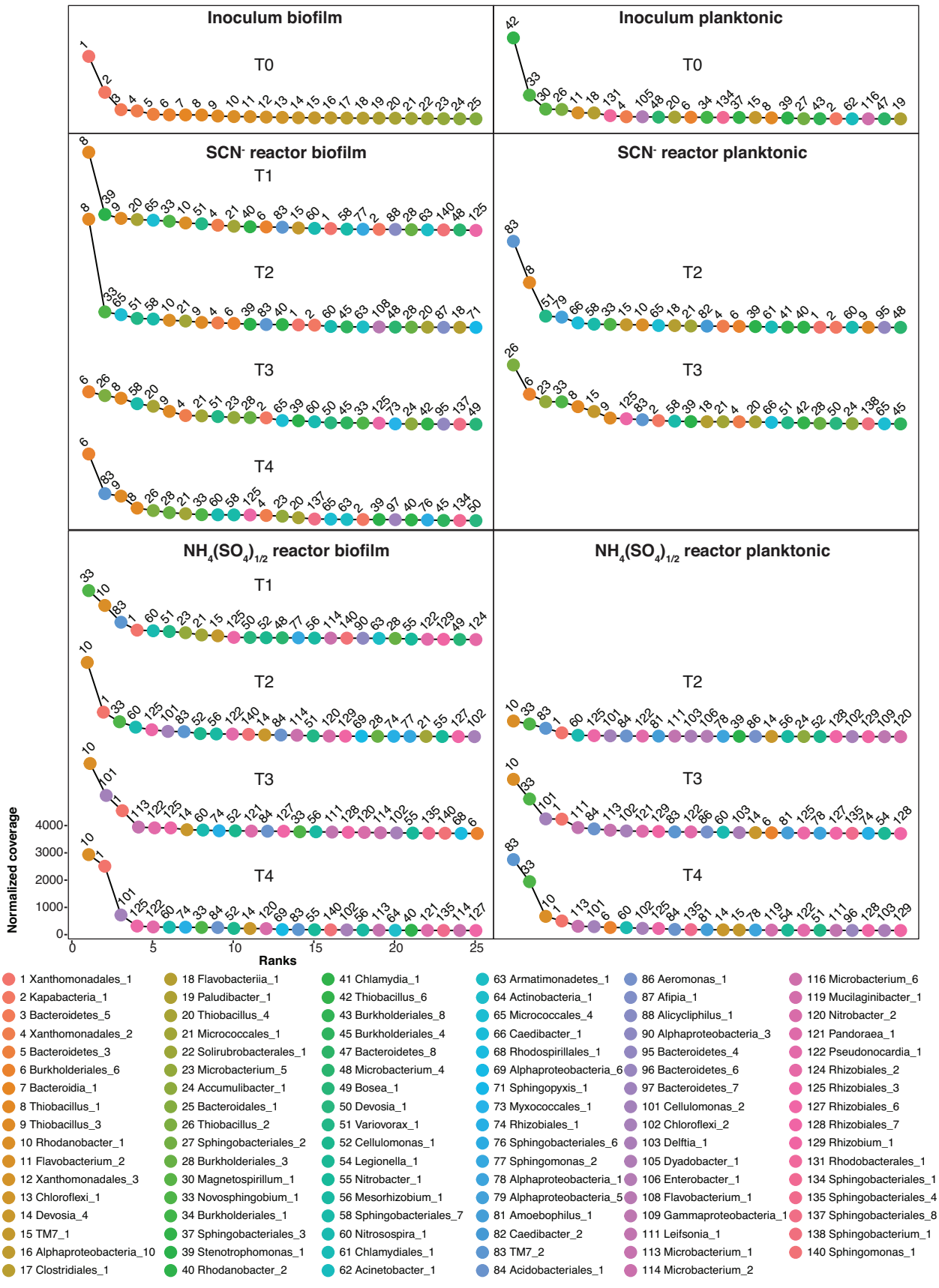
Inoculum biofilm
T0
Inoculum planktonic
T0
SCN- reactor biofilm
T1
T2
T3
T4
SCN- reactor planktonic
T2
T3
NH4(SO4)1/2 reactor biofilm
T1
T2
T3
T4
NH4(SO4)1/2 reactor planktonic
T2
T3
T4
Normalized coverage
0
1000
2000
3000
4000
0
5
10
15
20
25
Ranks
1 Xanthomonadales_1
2 Kapabacteria_1
3 Bacteroidetes_5
4 Xanthomonadales_2
5 Bacteroidetes_3
6 Burkholderiales_6
7 Bacteroidia_1
8 Thiobacillus_1
9 Thiobacillus_3
10 Rhodanobacter_1
11 Flavobacterium_2
12 Xanthomonadales_3
13 Chloroflexi_1
14 Devosia_4
15 TM7_1
16 Alphaproteobacteria_10
17 Clostridiales_1
18 Flavobacteriia_1
19 Paludibacter_1
20 Thiobacillus_4
21 Micrococcales_1
22 Solirubrobacterales_1
23 Microbacterium_5
24 Accumulibacter_1
25 Bacteroidales_1
26 Thiobacillus_2
27 Sphingobacteriales_2
28 Burkholderiales_3
30 Magnetospirillum_1
33 Novosphingobium_1
34 Burkholderiales_1
37 Sphingobacteriales_3
39 Stenotrophomonas_1
40 Rhodanobacter_2
41 Chlamydia_1
42 Thiobacillus_6
43 Burkholderiales_8
45 Burkholderiales_4
47 Bacteroidetes_8
48 Microbacterium_4
49 Bosea_1
50 Devosia_1
51 Variovorax_1
52 Cellulomonas_1
54 Legionella_1
55 Nitrobacter_1
56 Mesorhizobium_1
58 Sphingobacteriales_7
60 Nitrosospira_1
61 Chlamydiales_1
62 Acinetobacter_1
63 Armatimonadetes_1
64 Actinobacteria_1
65 Micrococcales_4
66 Caedibacter_1
68 Rhodospirillales_1
69 Alphaproteobacteria_6
71 Sphingopyxis_1
73 Myxococcales_1
74 Rhizobiales_1
76 Sphingobacteriales_6
77 Sphingomonas_2
78 Alphaproteobacteria_1
79 Alphaproteobacteria_5
81 Amoebophilus_1
82 Caedibacter_2
83 TM7_2
84 Acidobacteriales_1
86 Aeromonas_1
87 Afipia_1
88 Alicycliphilus_1
90 Alphaproteobacteria_3
95 Bacteroidetes_4
96 Bacteroidetes_6
97 Bacteroidetes_7
101 Cellulomonas_2
102 Chloroflexi_2
103 Delftia_1
105 Dyadobacter_1
106 Enterobacter_1
108 Flavobacterium_1
109 Gammaproteobacteria_1
111 Leifsonia_1
113 Microbacterium_1
114 Microbacterium_2
116 Microbacterium_6
119 Mucilaginibacter_1
120 Nitrobacter_2
121 Pandoraea_1
122 Pseudonocardia_1
124 Rhizobiales_2
125 Rhizobiales_3
127 Rhizobiales_6
128 Rhizobiales_7
129 Rhizobium_1
131 Rhodobacterales_1
134 Sphingobacteriales_1
135 Sphingobacteriales_4
137 Sphingobacteriales_8
138 Sphingobacterium_1
140 Sphingomonas_1

Figure 3.8. Rank abundance curves for biofilm samples from the inoculum (T0), and four timepoints during loading increases in the SCN^- and $NH_4(SO_4)_{1/2}$ reactors. The 25 bacterial genomes with the highest normalized coverage in each sample are shown and organisms are numbered by rank order in the biofilm inoculum sample (T0).

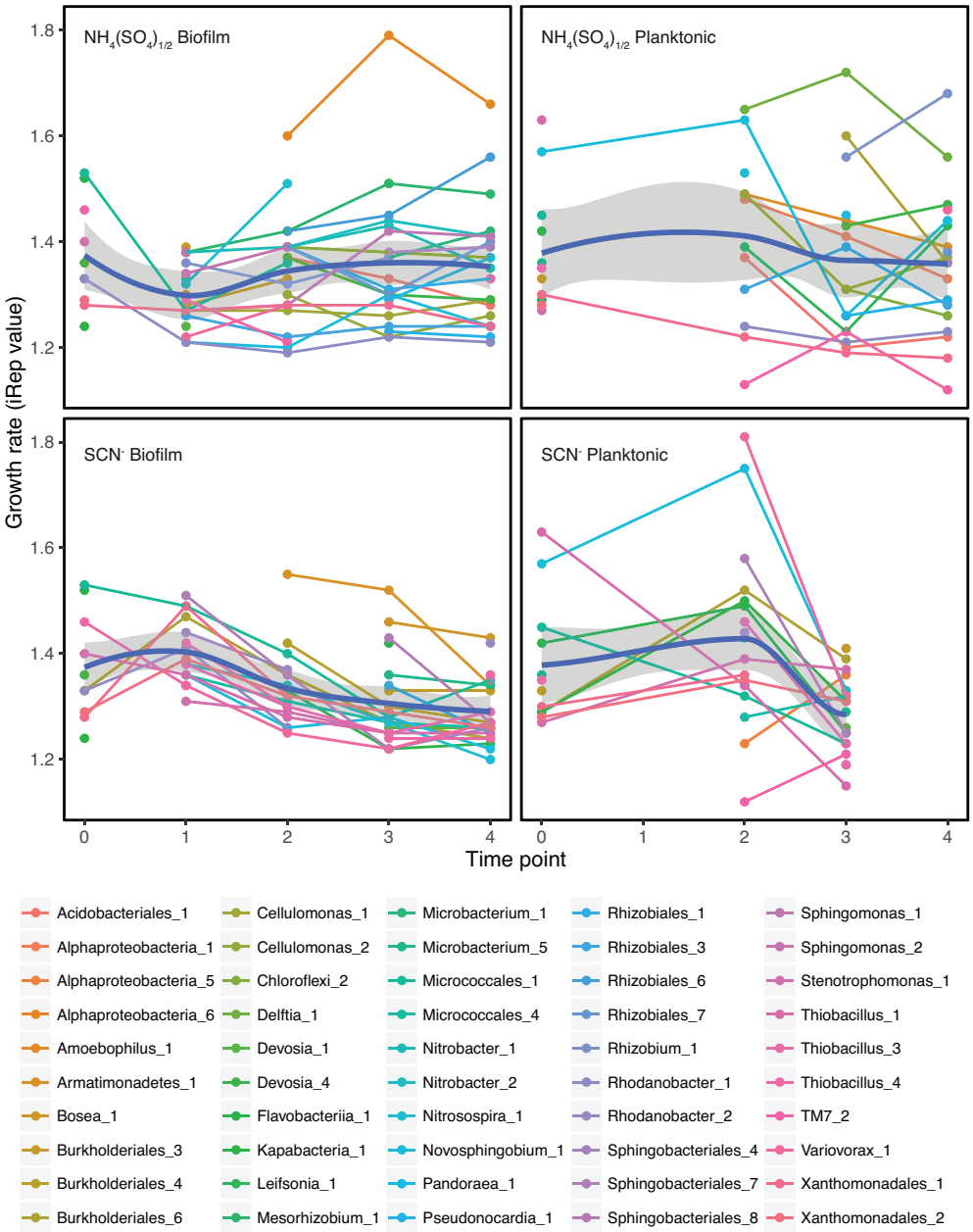

Figure 3.9. Replication rate information for organisms in biofilm samples from each reactor. The index of replication (iRep) value was calculated for all organisms in all samples and was included if it passed coverage and quality filters (see Methods) and the organism had a value for two or more samples in either reactor. Dark blue line represents a Loess curve fitted to the data (see Table S4 for all iRep values).

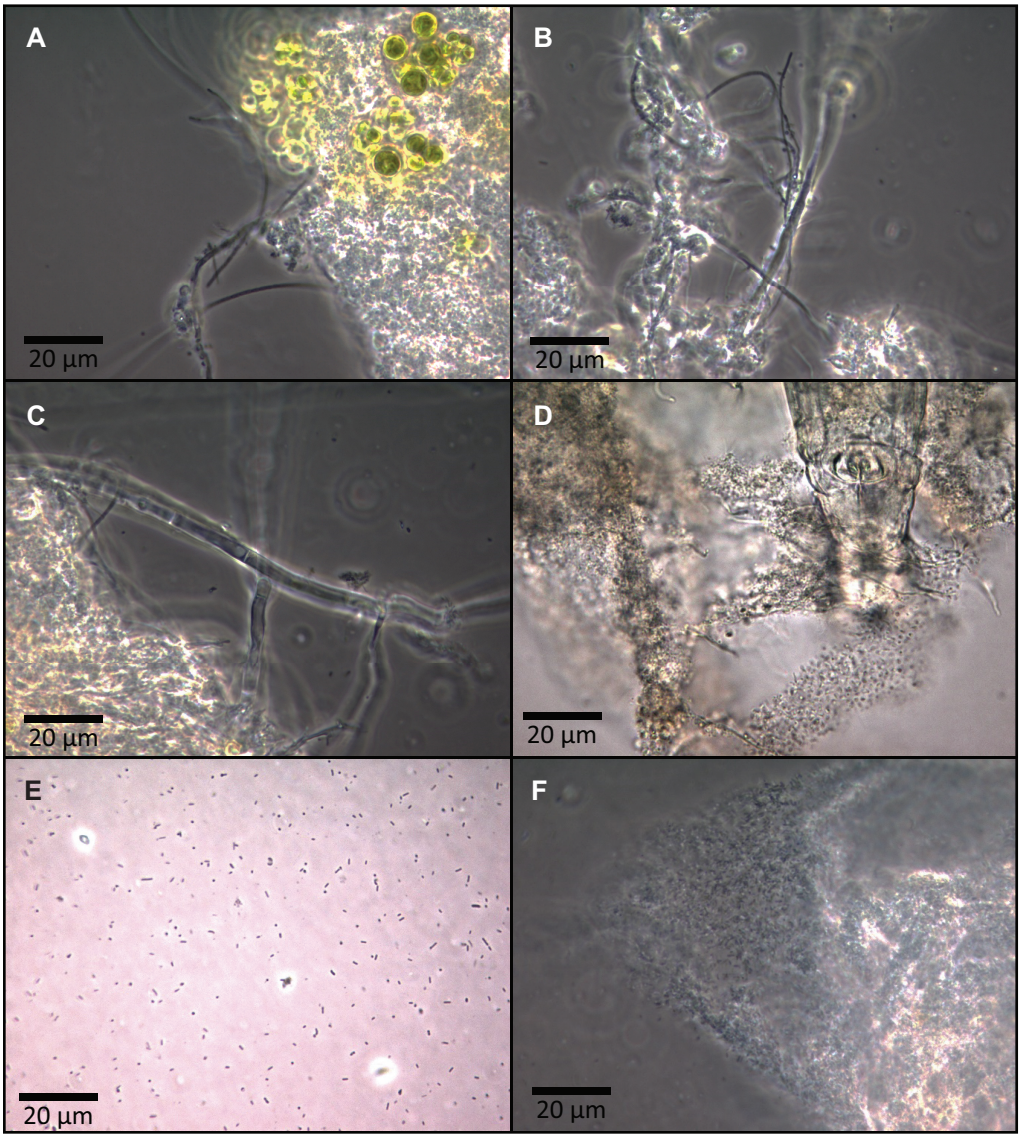

Figure 3.10. Light microscopy images at 1000x magnification of samples from the SCN^- reactor. Algae, filamentous organisms and edge of biofilm, (day 68) (A); Filamentous organisms, 6/11/2013 (day 68) (B); branched filamentous organisms, (day 68) (C); Rotifer (head only) grazing on biofilm, (day 147) (D), planktonic bacilli, (day 216) (E); bacteria in thick biofilm, (day 216) (F).

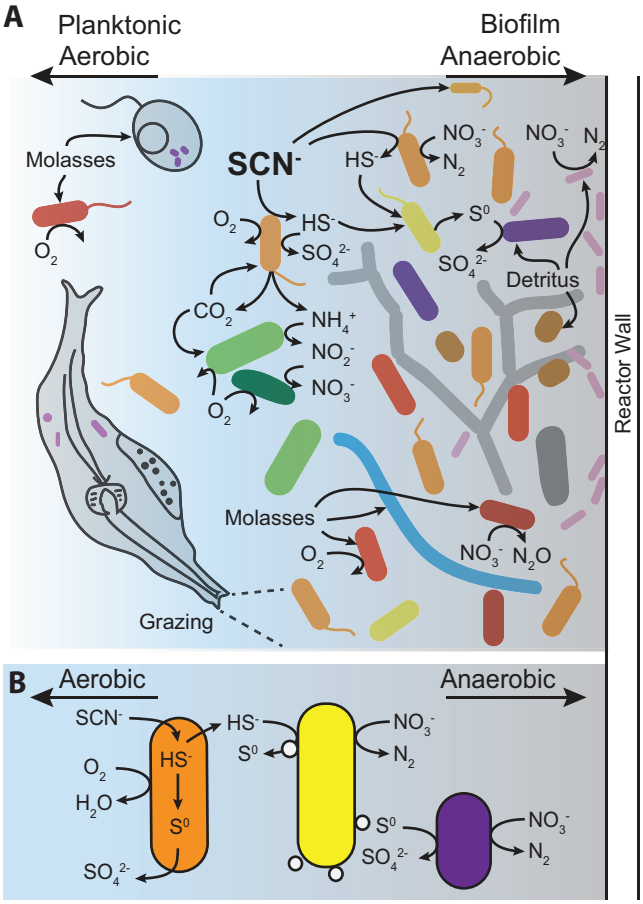

Figure 3.11. SCN^- removal and sulfur, nitrogen, and carbon cycling in the reactor system depicted based on metagenomic analysis. Organism colors match those used in Figure 3.3.

Supplementary table titles

Table S3.1A. Information on non-redundant set of recovered bacterial genomes. Overlap with previously detected genomes is based on number of basepairs accounted for within alignments ≥ 500 bp in length with 98% nucleotide identity, although some previously detected genomes were only partially complete. Genomes can be found at: http://ggkbase.berkeley.edu/scnpilot-dereplicated/organisms.

Table S3.1B. Information on non-redundant set of recovered eukaryotic nuclear, mitochondrial, and chloroplast genomes as well as phage, plasmids, other mobile elements, and predicted eukaryotic viruses. Genomes can be found at: http://ggkbase.berkeley.edu/scnpilot-dereplicated/organisms.

Table S3.2. Proteomics data from two technical replicates for key proteins involved in thiocyanate, sulfur, nitrogen, and carbon metabolism.

Table S3.3. All indices of replication (iReps) for de-replicated set of bacterial genomes.

Chapter 4

A genome-based analysis of in situ sulfur remediation across a mining landscape

Abstract

Landscape-scale environmental impacts of mining are a serious and costly problem. Waste rock and tailings generate wastewater containing intermediate sulfur compounds that must be oxidized prior to release or reuse of the water. While successful management practices exist, the microbial communities responsible for sulfur removal are understudied. To examine these communities, we performed genome-resolved metagenomic analysis across a mining-impacted system in Ontario, Canada. A dedicated oxidation reservoir and wastewater inputs to the reservoir were sampled to investigate how microbial processes shape environmental sulfur chemistry. Sites varied in terms of pH, solids, and retention times, but each contained high concentrations of colloidal sulfur and other sulfur species. All sites were enriched in sulfur oxidizing bacteria, with metabolisms variably linked to aerobic growth, denitrification, phototrophy, methylotrophy and CO_2 fixation. Waste rock run-off and external mine drainage were dominated by acidophilic iron oxidizing bacteria and heterotrophic sulfur oxidizers but also included iron and sulfate reducers. Neutral tailings dewatering run-off was enriched in methylotrophic bacteria and contained sulfur oxidizers at lower abundances. In contrast, communities sampled mid-reservoir near a tailings input pipe were dominated by autotrophic sulfur-oxidizing bacteria. Within the reservoir there was a shift in community composition from late summer to early winter, however, the capacities for sulfur oxidation and denitrification were preserved. Overall, genome-informed insight into microbial processes that modulate geochemical inputs to the reservoir could provide a foundation for design of environmental manipulations to improve site management.

Introduction

Mining activities contribute significantly to many nations' economies but with these activities come significant amounts of potentially harmful wastes. These wastes can generate extremely low pH acid mine drainage (AMD) resulting from the exposure of sulfidic rock and can release heavy metals into local water. Two main types of mine wastes are tailings, finely ground rock that has undergone extraction, and waste rock, coarser material removed to access the ore. Both wastes may produce AMD when exposed to oxygen and water (Johnson, 2014). Therefore, treatment of mine wastes aims to completely oxidize sulfur compounds from wastewater and run-off and to neutralize solutions before this water is released to the surrounding environment.

Microorganisms play an important role in sulfur cycling in natural systems and at mine waste sites and their metabolisms are diverse. These include well-known AMD organisms belonging to the genera *Acidithiobacillus*, *Acidiphilium*, *Sulfobacillus*, and several thermophilic archaea (Dopson and Johnson, 2012). Other sulfur oxidizers are adapted to neutral, alkaline, or salty environments and can utilize sulfur compounds including hydrogen sulfide, thiosulfate, tetrathionate, and elemental sulfur as a source of electrons (Ghosh and Dam, 2009). Sulfur compound oxidation can be coupled to aerobic metabolism or to denitrification. In addition to oxidation, some organisms perform sulfur disproportionation and others use sulfate as an electron acceptor under anaerobic conditions.

Our knowledge of the roles of microorganisms in accelerating AMD (Baker and Banfield, 2003; Denef et al., 2010) and also in bioremediation has dramatically increased in recent decades, with the aid of both culture-dependent and -independent approaches. This understanding has led to the development of various active (bioreactors and *in situ* amendments) and passive (wetlands) treatments to remediate mining waste (Johnson, 2014), as well as to bioleaching technologies for improved metals extraction. Garris *et al*. (Garris et al., 2016) suggest that genomic and metagenomic tools are important for monitoring these microbial-based treatment systems. Additionally, metagenomic and geochemical investigations of successful *in situ* sulfur remediation could provide new information to shape management strategies.

In this study, we investigated the microbiology and geochemistry of several sites in the Glencore mining district, located near Sudbury, Ontario, Canada (**Figure 4.1**). In this system, a large reservoir accumulates water from waste rock sites, a tailings dewatering site, nearby mines, and tailings inputs directly from a mill. Recently, extensive geochemical characterization of the reservoir and associated waste streams was carried out, which highlighted the need for investigation of the microbial community to provide perspective on the biological processes occurring at these sites. We used a genome-resolved metagenomic approach to profile the phylogenetic and metabolic diversity present in waste inputs and in the reservoir itself. Population genomic data from different sites were compared to identify organisms shared between sites and to assess the likelihood of inter-site dispersal versus geochemical selection. Samples were collected from the same sites in late summer and early winter to examine seasonal shifts in microbial communities and associated geochemistry.

Results

Waste rock and tailings were comprised of the sulfidic minerals pyrrhotite and pyrite, and wastewaters sampled contained high levels of sulfur species. Geochemical characterization of the oxidation reservoir and input waters revealed that colloidal sulfur (S^0) was between 2.6 and 16.4

times higher in abundance than all combined aqueous sulfur compounds. The pH of the waste inputs ranged from 3.47 in waste rock run-off to 7.01 in tailings dewatering run-off, while the reservoir pH was between 6 and 7 (**Table 4.1)**. Metagenomic assemblies from all input waters were fragmented and accounted for a relatively low proportion of the reads (47 – 66 %). In contrast, assemblies from the reservoir accounted for much more of the data (83 - 95% of reads), suggesting adequate reconstruction of the sampled community (**Table 4.2**). Binning of these assemblies recovered 290 partial and near-complete bacterial genome bins from 12 environmental samples and 3 enrichments. These bins were clustered at 95% genome-wide average nucleotide identity (ANI) into 170 distinct clades. The metabolic capacities of each clade were inventoried to generate a community profile of each site (**Table S4.1, Figure 4.2**), and phylogenetic affiliations of all bins containing marker genes were determined (**Figure 4.3**).

Waste Rock 1

Moderately acidic, green water was sampled from a pool collecting waste rock and tailings deposit run-off (**Table 4.1, Figure 4.1**). The pool was partially depleted of oxygen, measuring 61% saturation in September and 82% in November (**Table 4.1**). The metagenomes in September and November were dominated by a genome most closely related to the iron-oxidizing *Ferrovum* sp. (Ullrich et al., 2016) and multiple fragmented *Acidiphilium* spp. genomes. Genes for the Calvin Benson Bassham cycle for carbon fixation were identified in the *Ferrovum* sp. and one *Acidiphilium* sp. The *Acidiphilium* genomes also encoded chlorophyll biosynthesis, the sox pathway for thiosulfate, sulfide and sulfur oxidation, and pyrroloquinoline quinone (PQQ) methanol dehydrogenases for methylotrophy (**Table S4.1**). A genome corresponding to the common AMD iron- and sulfur-oxidizing genus *Acidithiobacillus* was recovered from the September metagenome, but was at low relative abundance (**Table S4.1**). Photosynthetic eukaryotes were present in both seasons, and identified based on 16S rRNA gene sequences corresponding to chloroplasts of Chlorophyta and Ochrophyta. Another chloroplast sequence contained an intron and was closest to *Dunaliella tertiolecta* (**Table S4.2**).

Several anaerobic metabolisms were inferred in the waste rock 1 site, with the potential to recycle oxidized sulfur and iron and reduced carbon in the system. The genome of a novel Acidobacteria abundant in both seasons (cluster 123) contained a *dsrABD* operon for dissimilatory sulfite reduction and a nickel-iron type I uptake hydrogenase. A novel Deltaproteobacteria genome recovered from the November sample encoded a 30-heme cytochrome c protein, which could be involved in iron reduction, similar to multiheme cytochromes in *Geobacter sulfurreducens* (Methe et al., 2003). Lastly, partial genomes were recovered for members of the Parcubacteria superphylum and one of these (cluster 162) was at high coverage in the September metagenome. The metabolic profile of these genomes suggests potential use of organic carbon and a symbiotic lifestyle, consistent with previously characterized members of the superphylum (Brown et al., 2015; Kantor et al., 2013).

Waste Rock 2

Clear, moderately acidic water was sampled in September from a second waste rock pool receiving the same run-off as the first. The two sites were separated by an embankment, and oxygen saturation in the second pool was higher, at 99.5%. The recovered genomes with the highest coverage corresponded to *Acidocella* sp. and *Acidiphilium* sp., and encoded the sox pathway for thiosulfate oxidation as well as chlorophyll biosynthesis pathways. A third abundant genome corresponding to *Halothiobacillus* sp. also carried the sox pathway and a sulfur

oxygenase reductase (SOR) for sulfur disproportionation to hydrogen sulfide, sulfite, and thiosulfate (Veith et al., 2012). Other community members identified by their genomes included relatives of Rhodospirillales, *Thiomonas*, Parcubacteria, and additional species or strains of *Acidocella* and *Halothiobacillus*. No iron-oxidizing *Ferrovum* spp. were detected, but *Halothiobacillus* genomes encoded the blue-copper protein rusticyanin, putatively involved in iron oxidation. The taxonomic identification of this gene corresponded to *Acidocella aminolytica* and the surrounding gene neighborhood included transposases suggestive of horizontal gene transfer. Nitrate reductase genes identified in genomes corresponding to *Acidocella*, *Thiomonas*, and Rhizobiales indicate the potential to couple sulfur oxidation to nitrate reduction. Photosynthetic eukaryotes were detected based on analysis of assembled 18S rRNA genes and included an *Ochromonas* sp. and two dinoflagellates belonging to the genus *Peridinium* (**Table S4.2**).

External mine discharge ("piped water")

Water was sampled from a pipe conveying moderately acidic mine wastewater discharge directly to the oxidation reservoir (**Table 4.1**). This effluent had extremely high elemental sulfur (82 mM) and very low sulfate concentrations (0.19 mM). Overall, the community in the external drainage was similar to that in Waste Rock 1, and dominated by *Ferrovum* and *Acidiphilium* species. In addition to *Acidiphilium*, two other genomes at lower coverages, corresponding to members of the Thiotrichales and Rhodobacterales, encoded sulfur oxidation pathways. The Rhodobacterales genome also harbored genes for complete denitrification and chlorophyll biosynthesis. Eukaryotes and plastids detected by small subunit rRNA genes included a relative of *Ochromonas* sp. (**Table S4.2**) as well as two fungi, one ciliate (classified as Euplotes), and one mitochondrion of unknown taxonomy.

Tailings

We collected water entering the oxidation reservoir from nickel and copper mine tailings undergoing dewatering. The water passes through the dewatering site and then enters a small storage area before flowing into the oxidation reservoir (**Figure 4.1**). Oxygen saturation in this water was near 100% in both September and November and it contained 19-24 mM S^0, the lowest of all input waters sampled (**Table 4.1**). Unlike waste rock sites, the water was of neutral pH and had higher dissolved organic carbon (DOC) and suspended solids. Nitrate in the water dropped from 0.9 mM to 0.07 mM between September and November (**Table 4.1**). The microbial community in both seasons was dominated by a single *Hyphomicrobium* sp., with read coverage for this genome accounting for 28.8% and 9.7% of the summed coverage of all recovered genomes in September and November respectively.

The *Hyphomicrobium* sp. genome harbored the pathway for methylotrophy via a PQQ methanol dehydrogenase, consistent with other characterized members of this genus (Martineau et al., 2013; 2014). Genes encoding biosynthesis of bacteriochlorophyll and a flavocytochrome c sulfide dehydrogenase were also present (**Table S4.1**), possibly allowing energy generation via aerobic anoxygenic photoheterotrophy coupled to sulfide oxidation. Other low-coverage genomes were predicted to correspond to heterotrophs based on lack of carbon fixation pathways. Some of these genomes encoded rhodopsins or bacteriochlorophyll, possibly enabling photoheterotrophic metabolism (**Table S4.1**). Apart from the *Hyphomicrobium* genome, two other genomes, a *Brevundimonas*, and a member of the Sphingomonadales, were recovered in metagenomes from in both seasons. Genomes corresponding to sulfur oxidizers (possessing the

sox pathway and a variety of other sulfur-oxidation genes) were present at lower relative abundances, and some populations, including Hydrogenophilaceae, *Bosea* sp., and Thiotrichales, persisted from September to November. Nitrogen fixation genes were identified in three Rhizobiales genomes and one Hydrogenophilaceae. Lastly, small-subunit rRNA genes were identified corresponding to chloroplasts, mitochondria, and nuclear genomes of Ochrophyta.

Reservoir

Geochemical analysis of the reservoir across multiple seasons showed that oxidation of aqueous sulfur species was most complete in summer and least complete under ice-on conditions in winter (L. Warren, pers. comm.). Consistent with this analysis, a comparison of samples from September and November showed that sulfate accounted for 24 - 27 % of total aqueous sulfur in late summer but only 4.5 % in early winter (**Table 4.1**). In September, the lake was stratified and dissolved oxygen and nitrate concentrations decreased substantially with depth (**Table 4.1, Figure 4.2**). In November, turnover had occurred, and there was no difference in dissolved oxygen between depths, while nitrate was much higher at 2 m than at 21 m. High concentrations of suspended solids were measured at this time, whereas total aqueous sulfur and elemental sulfur concentrations were lower than in September. As in all waste inputs sampled, elemental sulfur far exceeded aqueous sulfur concentrations in all reservoir samples, with an extremely high value of 117 mM S^0 at the shore in November (**Table 4.1**).

In September, metagenomes from open water at both depths contained an *Acidovorax* sp. genome (cluster 31) at high coverage, which encoded a complete set of genes for denitrification. Also at high coverages were numerous genomes corresponding to sulfur-oxidizing autotrophs belonging to the *Sulfurovum*, Hydrogenophilales, *Polynucleobacter*, Thiotrichales, *Polaromonas*, and *Sulfuricurvum*. Several of these genomes encoded *nar* or *nap* genes, for nitrate reduction coupled to sulfur-compound oxidation (**Tables S4.3** and **S4.5**). Genomes related to *Methylotenera* and Methylophilaceae were predicted to be methylotrophs based on taxonomy and the presence of PQQ methanol dehydrogenase genes. The shore community was dominated by predicted heterotrophs with sulfur oxidizers at lower abundances.

In November, after the reservoir turned over, all three depths including open water and surface/shore shared almost all taxa at similar relative abundances (**Table 4.3**). Most abundant was a *Polaromonas* sp. (cluster 3), whose genome encoded the capacity for denitrification to nitrous oxide. A type I RuBisCO was identified in this genome cluster but sox genes for sulfur oxidation were not found. Given that all recovered genomes corresponding to this cluster were extremely fragmented (> 500 scaffolds and ≤ 82 % complete), it is possible these genes were not well-assembled or remained unbinned. Two sox-containing genome clusters, *Polynucleobacter* (cluster 35) and *Sideroxydans* (cluster 61), were highly abundant in the open water. Genomes corresponding to the *Polynucleobacter* were highly fragmented but contained flavocytochrome c dehydrogenase and in some genomes, both type I and type II RuBisCO. The *Sideroxydans* possessed a suite of sulfur oxidation genes including *sox*, flavocytochrome c dehydrogenase, reverse dissimilatory sulfite reductase (*rDsr*), adenylylsulfate reductase (*apr*), and ATP sulfurylase, to oxidize compounds ranging from sulfide and thiosulfate to elemental sulfur and sulfite. The *Sideroxydans* genomes also possess the *nar* genes for coupling sulfur oxidation to anaerobic respiration of nitrate. Other genomes from November that contained the sox pathway included a lower abundance *Polaromonas,* Rhodospirillales, Rhizobiales, *Sulfurovum*, Thiotrichales, Burkholderiales, and *Halothiobacillus*.

Seasonal shifts were clearly observed in all guilds in the reservoir (**Table 4.3**). With the exception of a low-abundance *Polaromonas* (cluster 4), nitrate reducing sulfur oxidizers shifted completely from Hydrogenophilales and *Sulfurovum* to *Sideroxydans*, Rhodospirillales, and Rhizobiales (**Table 4.3**). Nitrogen fixation, a key metabolic function in the reservoir ecosystem was only identified in one population per season in the open water and this was different between seasons. In September, the *nif* genes were identified in a Hydrogenophilales genome cluster (55) and in November, *nif* genes were in a Rhizobiales cluster (67). Both of these genome clusters were also categorized as sulfur oxidizers. Predicted methylotrophs (Methylophilaceae) belonging to genome cluster 18 were present in both seasons while those in cluster 34 were present in September only (**Table 4.3**). Lastly, eukaryotes corresponding to the group Ochrophyta were detected by ribosomal RNA genes in both seasons. The 18S rRNA genes from September were classifiable only to the family level, Chrysophyceae, while in November, they were confidently assigned to the genus *Paraphysomonas* (**Table S4.2**).

Enrichments

The September shore sample from the oxidation reservoir served as inoculum for three thiosulfate-fed enrichments. The dominant organisms included *Halothiobacillus* sp. and *Thiomonas* sp., whose genomes encoded the sox pathway for thiosulfate and sulfide oxidation. Within the parent sample these were not the most abundant populations or the most abundant sulfur oxidizers (**Table S4.1**). Lower abundance genomes from non-sulfur oxidizers included a Xanthomonadales (cluster 72) found in all three enrichments, a *Microbacterium* sp. (cluster 82) present in the two neutral pH enrichments, and a *Stenotrophomonas* sp. and a *Cellulomonas* sp. found only in enrichment 2.

Community variation across sites

Given that genomes from metagenomics represent composites of the most abundant strain(s) within a population, we can use these genomes as proxies to compare populations from different samples to one another. Based on genome clusters generated at 95% nucleotide identity, which is near the threshold for microbial species (Richter and Rossello-Mora, 2009), related populations were present at many sites across landscape (**Figure 4.4**). Genomes from different sites tended to have lower ANI than genomes recovered from within the same sites at different seasons or depths (**Figure 4.4A** and **Table S4.3**). Overall, the most cross-site genome matches were between the tailings and reservoir, which shared 12 clusters (**Figure 4.4B**). Some diverse clades were more cosmopolitan, including the order Thiotrichales (cluster 51). Genomes in this cluster encoded sulfur oxidation and were present in 8 samples from three different sites in both seasons (**Figure 4.4C**), but by ANI, none were identical to each other.

Two genome clusters contained members with very high pairwise ANI (≥ 99.7 %) across sites (**Figure 4.4A**). The Hydrogenophilales cluster 53 contained one genome from the tailings and two the reservoir. All three genomes shared at least 2619 orthologous genes (out of a total of 3079 open reading frames) and were largely contiguous. Differences consisted of mobile elements such as prophage and several heavy-metal related proteins. The other cluster of genomes with high identity across the reservoir and tailings was a *Novosphingobium* sp. (cluster 113). These genomes shared 2842 orthologs across a total of 3927 open reading frames. This *Novosphingobium* sp. was one of the most abundant organisms in the reservoir shore sample in September (**Figure 4.2** and **Table 4.5**) although it was not highly abundant in the tailings.

Phage and CRISPR-Cas loci

In addition to bacterial genomes, we identified phage genomes and found that many similar phages (> 95% ANI) occur at multiple sites (**Figure 4.5**). Sharing of phages was most prominent among the reservoir samples, although in some cases the same phages were present across multiple site types. To trace bacteria-phage interactions, we analyzed the CRISPR loci involved in bacterial immunity. Genomes from 44 clusters encode one or more CRISPR-Cas loci. The number of matches between spacers from all arrays (binned and unbinned) and putative targets varied by site, ranging from 3 in the tailings to 69 in the waste rock (**Table S4.4**). The large number of spacers with targets identified in the samples from the acidic sites was due the presence of multiple, large arrays in the genomes of *Alishewanella* (cluster 2) and *Acidocella* (cluster 64) and many of the spacers with identified phage targets were located near the new end of the locus, suggesting recent incorporation.

The CRISPR loci were also used to analyze the relatedness of bacterial populations from different samples. The Hydrogenophilales strains (cluster 53) recovered from the September reservoir and tailings samples (described above) have the same CRISPR repeat sequence, and all of the spacers in the two reservoir genomes are identical. Although the spacer locus for the tailings strain was truncated, mapping reads from the tailings sample to the reservoir assembly demonstrated that the populations from these two sites share many old-end spacers, but nine new-end spacers in the reservoir strain are absent from the tailings strain (**Figure 4.6)**. Furthermore, the new-end spacers that were recovered from the tailings strain are absent in the reservoir populations and one of these spacers exactly matches to a phage in the reservoir. Likewise, two new-end spacers in the reservoir strain have exact matches to phages present in the tailings.

In the November reservoir samples, identical CRISPR loci from three versions of a Rhizobiales genome (cluster 67) account for the majority of all spacers matching to phage, including both new end and old end spacers. Interestingly, the new-end spacers match only phage present in the reservoir whereas old end spacers match to phage detected in the tailings (where related Rhizobiales are present at low abundance).

Discussion

Mining waste microbial communities have diverse physiologies

Genomes recovered from the acidic waste sites indicated the presence of iron- and sulfur-oxidizing species as well as iron and sulfate reducers, and expanded the metabolic range known for some phyla. Overall, the dominance of a *Ferrovum* sp. in multiple samples is consistent with observations of other AMD systems at pH > 2.3 (Kuang et al., 2013) and *Acidiphilium* species are known to co-occur with *Ferrovum* species. Prior genome analyses of co-cultures have suggested that *Acidiphilium* utilizes the carbon fixed by *Ferrovum* for heterotrophic growth and may also be capable of photoheterotrophy (Ullrich et al., 2016; 2015). *Acidiphilium* and *Acidocella* are also known to reduce iron and oxidize sulfur, and some *Acidiphilium* are capable of fixing carbon (Coupland and Johnson, 2008; Dopson and Johnson, 2012; Küsel et al., 1999).

In addition to these relatively well-characterized populations, we recovered the genome for a novel Acidobacteria population from the waste rock run-off that appears to encode sulfate reduction. Its nearest relative with a sequenced genome (95% identity across the 16S rRNA gene), *Thermoanaerobaculum aquaticum*, does not share this metabolic potential (Losey et al., 2013; Stamps et al., 2014), nor do any other known Acidobacteria. Lastly, the observation of

genomes from the superphylum Parcubacteria (OD1) in several acidic samples extends the known geochemical range of this group to low pH (Brown et al., 2015; Kantor et al., 2013).

In the tailings dewatering run-off, the microbial metabolism appeared to be driven by aerobic, heterotrophic consumption of organic matter augmented by phototrophy. The organic carbon to support this metabolism may come from natural sources, entering the run-off site from upstream in the watershed. While the capacity for complete sulfur oxidation to sulfate was present in some populations, these were at lower abundances than in the other waste sites. Concordantly, the tailings dewatering run-off was at neutral pH, confirming that AMD processes were not occurring.

At all waste sites, flagellated protists related to the mixotrophic grazer *Ochromonas* were detected. Depending on their rates of predation and photosynthesis, these flagellates can act to remove carbon from or add carbon to the wastewater, and may also be an important part of the phosphorus cycle, as observed for an *Ochromonas* sp. in an acidic mining lake (Schmidtke et al., 2006). Ribosomal RNA genes but not genomes were recovered for these eukaryotes. Lack of marker genes made an exact taxonomic identification difficult, but further investigation is warranted given their potential contribution to acidic waste systems.

Microbial diversity and functional capacities in a tailings-impacted freshwater reservoir

In undisturbed freshwater ecosystems, photosynthesis by algae and photosynthetic bacteria provides dissolved organic matter (DOM) as the main source of available chemical energy. To our knowledge, no algal blooms occur in the oxidation reservoir, but high concentrations of sulfur compounds derived from waste inputs provide ample chemical energy to the system. These inputs appear to have selected for autotrophic sulfur oxidizing bacteria that are typically not abundant in natural freshwater lakes. In support of this interpretation, many common freshwater bacteria were detected including *Polynucleobacter*, *Polaromonas*, and Methylophilales (Newton et al., 2011), but there was strong enrichment for populations capable of sulfur oxidation such as Hydrogenophilales*, Thiotrichales, Sideroxydans, Sulfurovum,* and *Sulfuricurvum*. Some of these populations, including Hydrogenophilales and *Sideroxydans* sp. may also contribute to iron oxidation, as is observed for related organisms (Beller *et al.*, 2013; Emerson, 2013). It appears that these organisms are providing the *in situ* sulfur oxidation service required for successful management of mine wastes at this site, and that they are influenced by physical conditions such as shifting temperature and turnover, as would be expected in undisturbed freshwater lakes.

Sulfur oxidation was more extensive in September than in November, likely because of higher microbial activity under warmer conditions, but perhaps also due to changes in waste streams entering the reservoir. Community shifts reflect the change in temperature, including the high abundance of the cold-tolerant *Polaromonas* spp. in November. Geochemistry and microbial communities of all depths look similar in November, likely owing to the mixing effect of lake turnover. Notably, despite the community shift, important functions including sulfur oxidation, nitrogen fixation, methylotrophy, and heterotrophy were conserved.

A cross-landscape view of microbial communities in a mining-impacted environment

Studies of mining waste microbiology often focus on a specific type of waste or a single site, however, large-scale mining contains many unique environments, and water flow across landscapes can allow these environments to interact with one another. Comparing organisms at different sites using whole-genome ANI revealed that most species-level overlaps were actually

distinct at the strain level. This suggests strains were independently enriched at each site rather than distributed across sites by fluid flow. It is worth noting that because these genomes are composite representations of populations, some low-abundance members of the populations maybe identical across sites, but cannot be detected with this method. There were two exceptions to this pattern of independent enrichment across sites: Hydrogenophilales cluster 53 and Novosphingobium cluster 113 had high ANI between genomes from different sites. Shared CRISPR loci and a high degree of synteny between genomes in cluster 53 suggested that the tailings and reservoir bacteria and two associated phage may have been recently dispersed from a common source.

Sulfur dynamics and implications for management strategies

To date, characterizations of mining-impacted sites have focused on AMD (Chen et al., 2015; Kuang et al., 2013; Tyson et al., 2004) , radioactive waste, and heavy metals contamination (Kang et al., 2013), sediment and tailings dams or piles (Wakelin et al., 2012; Wielinga et al., 1999). Sulfur contamination has been addressed by means of bioreactor studies (Liljeqvist et al., 2011), but not studied in freshwater systems. Geochemical investigations of this mining landscape, including the freshwater reservoir, revealed unexpectedly high levels of elemental sulfur, and low concentrations of sulfur oxidation intermediates (including thiosalts). The enrichment for sulfur oxidizers in the reservoir and some maintenance of a natural ecosystem have combined to create conditions for sulfur oxidation, generally achieving the desired remediation at the Glencore site. Given our observations, various amendments may improve remediation, with a focus on elemental sulfur. The use of surfactants could help solubilize elemental sulfur, increasing its accessibility to microorganisms. Additionally, nitrogen may be limiting in this system (**Table 4.1**), and nitrogen fixation is limited to very few organisms present in the reservoir (**Table 4.3**). Amendment with nitrogen or enrichment for nitrogen fixers could increase microbial activity and improve rates of sulfur oxidation. Nitrogen added as nitrate could have the additional benefit of providing an alternative electron acceptor for sulfur oxidation under anoxic conditions, which occur in the reservoir in both summer and winter.

Methods

Sites and sampling

Samples were collected within an active nickel and copper mining district located near Onaping, Ontario, Canada. Water was sampled from a natural, 34-meter deep freshwater lake that is used as an oxidation reservoir. The reservoir receives ~50% of its water from higher in the watershed and additional inputs from mining sources including waste rock run-off, tailings dewatering run-off, external discharge piped from other mines, and tailings (in the form of total tailings and high-pyrrhotite tailings) piped from a mill directly into the middle of the reservoir (**Figure 4.1**). Three of these inputs were sampled: (1) green- and clear-colored pools containing run-off from a waste rock site, (2) outfall from a pipe transporting acidic tailings effluent from a mill to the reservoir, and (3) effluent from a tailings deposit dewatering site (**Figure 4.1**). Sites were sampled in September (late summer) and/or November (early winter) of 2014. Samples from open water in the reservoir were collected at 2 m and 21 m depths. The latter depth was chosen because it represents the lower hypolimnion, and the sampling location was near the input pipes conveying mill tailings. Samples from the surface at the shore of the reservoir were also collected, where the total depth was 3 m. At each sampling site, water was pumped through a 0.2

μm filter until the collected biomass clogged the filter. This process was repeated for 1-3 filters, representing approximately 1 - 11 L of water per sample in total.

Geochemistry

Physicochemical profiles of the water column in the oxidation reservoir and point samples of the inputs were measured *in situ* with a YSI 600 XLM (YSI Incorporated, Yellow Springs, Ohio, USA). Sample water for total dissolved S analysis was filtered (Pall Corp., 0.45μm Supor® membrane) directly into 50 mL acid-clean polypropylene centrifuge tubes that were pre-spiked with HNO_3 (0.2% v/v, Optima grade, Fisher Chemical). Samples were stored at 4°C until shipment to the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia for analysis by ICP-AES. Water samples for $S_2O_3^{2-}$ and SO_3^{2-} were derivatized immediately using bromobimane (Sigma Aldrich, B4380), and S^0 subsequently extracted in chloroform mixed with equal parts derivatized sample based on the method by Rethmeier et al., 1997 (Rethmeier et al., 1997). Samples were stored at -20°C until analysis. Samples for dissolved organic carbon (DOC) measurements were collected into carbon-clean amber glass 120 mL bottles that were pre-rinsed with sample. Samples were frozen at -20°C until filtration (0.45 μm) for DOC characterization. Analyses on the filtrate were conducted using a Shimadzu TOC-L Total Organic Carbon Analyzer using the 680°C combustion catalytic oxidation method and non-dispersive infrared detection. DOC was determined as the difference between total dissolved carbon and total dissolved inorganic carbon. Determination of $\Sigma H_2S_{(aq)}$, SO_4^{2-}, NO_3^-, NO_2^-, NH_4^+, suspended solids, and color was made by spectrophotometric analysis using a HACH DR2800 (HACH Company, Loveland, CO, USA). Ferrous and Ferric iron samples were preserved by the addition of HCl (2% v/v Optima-Grade) and analyzed colorimetrically by the modified ferrozine method in Viollier *et al.* (Viollier et al., 2000).

Enrichments

The reservoir shore sample from September served as inoculum for three thiosulfate-fed enrichment cultures, (metagenome samples 13, 14, and 15), which were fed 5 g $Na_2S_2O_3$ and treated as follows: Enrichment 13 was fed standard neutrophilic sulfur oxidizing media (NSOM) (Kelly and Wood, 1998) and the pH began at 7 and was allowed to drop until it reached pH 5. At this time, the solution was neutralized to pH 7 and the pH cycling repeated 15 times (107 days). Enrichment 14 was fed standard acidophilic sulfur oxidizing media (ASOM) (Staley et al., 1989) starting at pH 7 and decreasing to pH 5 before neutralization and repeated cycling (14 cycles, 118 days). Enrichment 15 was fed ASOM but pH began at 5 and was allowed to drop to 3 before being raised to 5 again (15 cycles, 107 days). Cultures were grown at 28 ˚C and sampled for metagenomic sequencing after all pH cycles were completed.

DNA extraction and sequencing

Whole community DNA was extracted from filters using a MO BIO Laboratories PowerSoil® DNA Isolation kit (Carlsbad, CA, USA). Each filter was divided into 4-6 separate extractions (to improve yield) and then extractions were condensed (2-6) onto one filter, and pooled. Final DNA extracts were dried and resuspended in 25μL of water. Library construction and sequencing were performed at the Farncombe Metagenomics Facility at McMaster University. All available DNA (up to 1 μg) from each sample was fragmented using the Covaris S220 Ultrasonicator. Parameters for 500 bp shearing with 50 μl input were: 175W PIP, 5% duty factor, 200 cpb, 35 seconds. Dual-indexed shotgun libraries were prepared with the NEBNext Ultra DNA Library

Prep Kit for Illumina (New England Biolabs Inc.). The libraries were quantitated by qPCR, pooled in equimolar amounts and sequenced using the Illumina HiSeq 1500 platform (Rapid v2 chemistry with onboard cluster generation, 151 bp paired-end reads). Raw data was processed with HCS v2.2.58 (RTA v1.18.64). File conversion and demultiplexing were performed with CASAVA v1.8.2 allowing 1 mismatch in the indexes.

Reads processing and assembly

Raw reads were trimmed with Sickle (https://github.com/najoshi/sickle) using default parameters to remove low-quality sequence, and in some cases bbmap (https://sourceforge.net/projects/bbmap/) was used remove adapter sequence. Assembly was performed with idba_ud (Peng et al., 2012). In order to determine coverage of each scaffold, reads were mapped to the full assembly using Bowtie 2 with default parameters.

Metagenome annotation and binning

Prodigal (Hyatt et al., 2010; 2012) was used to predict open reading frames on all scaffolds > 1 kbp in length. Predicted proteins were searched with USEARCH (Edgar, 2010) against KEGG (Kanehisa et al., 2016) and Uniref (Suzek et al., 2007) to identify best reciprocal hits and against Uniprot (The UniProt Consortium, 2015) to identify best forward-searching hits. The tRNAs were predicted using tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/)(Lowe and Eddy, 1997), and small subunit rRNA gene sequences were identified using SSU-ALIGN (Nawrocki, 2009). All rRNA genes identified by this method were classified using SILVA (Pruesse et al., 2012) and, where unclassified, BLAST against the NCBI-nr database (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Coverages were calculated by mapping all reads against the full assembly using Bowtie 2 (Langmead and Salzberg, 2012) with default settings. Binning was performed using the percent GC, coverage, and phylogenetic profile of each scaffold (ggkbase.berkeley.edu) and fine-tuned using emergent self-organizing maps (ESOMs) based on series data. In order to construct ESOMs for each sample (scaffolds > 5kbp), reads from every sample were mapped to the sample of interest with Bowtie 2 (Langmead and Salzberg, 2012). Abundance information was collected from mappings and was used to generate a ".lrn" file using prepare_esom_files.pl (https://github.com/CK7/esom/blob/master/prepare_esom_files.pl) with settings "--log_transformed_coverage -w 10000 -m 5000 -k 0". This file was used for training the ESOM using databionic ESOM tools (http://databionic-esom.sourceforge.net/index.html). For most ESOMs, abundance information was included only from samples containing sufficient overlap with the sample of interest.

Bin completeness and taxonomy

Bins were evaluated for completeness with CheckM v1.0.7 using the lineage workflow (https://github.com/Ecogenomics/CheckM/wiki) (Parks et al., 2015). Bins were assigned to the lowest possible taxonomy within ggkbase using the phylogenetic profile of genes on each scaffold. One genome corresponding to a member of the fungal order Hypocreales was detected in multiple samples and found in datasets from other concurrent studies at different sites. This likely contaminant was excluded from further analyses.

Genome clustering and comparison

Genomes were clustered at 95% average nucleotide identity (ANI) using Mash (Ondov et al., 2016) to pairwise-compare all genomes. To get a more precise average nucleotide identity (ANI) within clusters, the pyANI package average_nucleotide_identity.py (https://github.com/widdowquinn/pyani) was run on each cluster with the NUCmer method (ANIm). Pairs of genomes were reported as matched if they aligned across at least 75% of the smaller genome and ANI was reported for each. Selected genomes were compared using USEARCH (Edgar, 2010) to identify reciprocal best-hit proteins.

Metabolic predictions

In addition to text searches of annotations, Hidden Markov Model (HMM) searches (HMMER 3.1b2, http://hmmer.org/) were used to identify some key proteins of interest in nitrogen, sulfur, and carbon cycling. A combination of TIGRFAM (with noise cut-off) and custom models (with customized trusted cut-off) (Anantharaman et al., 2016b) (https://github.com/banfieldlab/metabolic-hmms) were used. To confirm *narG* and *napA* hits from HMM searches, an alignment with DMSO reductase family reference sequences from Castelle *et al.* (2013) was constructed with MUSCLE (drive5.com) and used to build a tree with FastTree v1.0 (Price et al., 2009) in Geneious (Biomatters Ltd.). Pyrroloquinoline quinone methanol dehydrogenases were confirmed in the same manner, using reference sequences from Butterfield *et al.* (2016). Metabolic pathways related to genes in operons (*soxAXYZBCD*, *narGHJI*, *rDsrAB*, chlorophyll biosynthesis) were also confirmed by checking for the presence of multiple components within the operon in addition to the active subunit(s).

Phylogenetic reconstruction

USEARCH with a custom database was used to identify scaffolds > 1 kbp that contained at least 6 of 16 ribosomal proteins (S3, S8, S10, S17, S19, L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, and L24). These scaffolds were curated with ra2.py to fix misassemblies, using reads from their sample of origin (https://github.com/christophertbrown/fix_assembly_errors/releases/tag/2.00). Two mismatches were allowed (-m 2) in mapping steps, and regions that could not be fixed were masked (--add-Ns). Genes were re-predicted with Prodigal (Hyatt et al., 2010; 2012) and the 16 ribosomal proteins were identified again via USEARCH against a custom database. Individual alignments for each of the 16 proteins were created with MUSCLE (Edgar, 2004), trimmed to remove N- and C- termini and stripped of all columns with ≥ 99% gaps in Geneious (Biomatters Ltd.). Alignments were concatenated to generate a final alignment with 2818 columns and scaffolds containing ≤ 1409 amino acids in the alignment were removed. Phylogenetic reconstruction was performed using FastTree with default settings (Price et al., 2009) and visualized in FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Phage and CRISPR identification

Phage scaffolds were identified by VirSorter, using as input all scaffolds > 1 kbp, with the "virome db" (Roux et al., 2015). Scaffolds > 5 kbp from all assemblies that were confidently identified as phage (categories 1 and 2) were clustered at 95% ANI using Mash, and pairwise ANI values were calculated within clusters using pyANI and requiring at least 75% coverage to be considered in the analysis. CRISPR loci were identified using a local version of CRISPRFinder (Grissa et al., 2007). CRISPR spacers were extracted and used as queries to search all assemblies (scaffolds > 1 kbp) with BLASTn-short (Camacho et al., 2009), allowing 1

mismatch and excluding all hits to scaffolds containing identified CRISPR arrays. All spacers were clustered with USEARCH (uclust) at 100% identity to count unique spacers.

Data accessibility

All metagenomic datasets are available online at http://ggkbase.berkeley.edu/project_groups/glencore.

Figures and Tables

Table 4.1. Physical and geochemical characteristics for each site and sample.

		Physical characteristics						Carbon	Oxygen		Iron (µM)	
Name	Sample	Depth	pH	Temp (C)	Conductivity (uS.cm/s)	Redox (mv)	Susp. Solids (ppm)	DOC (ml/L)	O_2 (ppm)	O_2 % sat	Fe_III	Fe_II
Waste_rock1_Sept	21	N/A	3.47	15.27	N/A	424.5	14	6.24	6	61.3	0.089	0.006
Waste_rock1_Nov	37	N/A	4.05	4.20	2303	397.5	1	5.23	10.65	82.3	0.099	0.011
Waste_rock2_Sept	20	N/A	3.64	15.44	N/A	349.1	2	2.35	9.87	99.5	0.071	0.008
Piped_water_Nov	34	N/A	3.56	6.29	1824	410.9	9	3.29	N/A	N/A	0.026	0.014
Tailings_Sept	12	N/A	7.01	13.17	2673	119.3	30	10.56	10.13	97.6	BD	0.004
Tailings_Nov	32	N/A	6.66	7.28	N/A	139.5	80.33	17.41	12.1	100.4	0.006	0.010
Reservoir_parent_Sept	28	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Reservoir_2m_Sept	16	2	6.92	16.37	2244	151	2	3.1	2.75	27.8	BD	0.002
Reservoir_21m_Sept	17	21	7.66	15.89	2262	-81	2	3.26	0.48	4.9	BD	0.002
Reservoir_2m_Nov	24	2	6.63	8.17	2232	140.9	43	3.12	6.56	N/A	BD	0.001
Reservoir_21m_Nov	39	21	7.23	8.13	2231	137.8	34.25	5.53	6.06	N/A	BD	0.001
Reservoir_shore_Nov	35	0	6.10	7.75	2191	227	2	8.23	6.72	56.8	BD	0.002

	Sulfur (mM - S)								Nitrogen (µM)			
Name	S (Total Aqueous)	SO_4^{2-}	H_2S	$S_2O_3^{2-}$	SO_3^{2-}	other SOI	S^0	% SO_4^{2-} of total (aq)	Total N	NO_3^-	NO_2^-	NH_3^+
Waste_rock1_Sept	10.865	3.896	0.000	2.462	0.000	4.507	33.400	35.9	433	143	0	290
Waste_rock1_Nov	8.071	0.305	0.000	0.671	0.015	7.080	74.900	3.8	1121	857	0	264
Waste_rock2_Sept	9.659	2.407	0.000	2.279	0.000	4.973	25.300	24.9	752	500	0	252
Piped_water_Nov	7.391	0.187	0.000	0.606	0.009	6.589	82.000	2.5	505	286	0	219
Tailings_Sept	7.852	1.860	0.001	1.847	0.000	4.145	24.200	23.7	1252	857	0	395
Tailings_Nov	5.317	0.328	0.002	0.000	0.000	4.986	19.400	6.2	450	71	22	357
Reservoir_parent_Sept	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Reservoir_2m_Sept	10.307	2.552	0.000	2.051	0.000	5.705	39.600	24.8	364	71	2	290
Reservoir_21m_Sept	9.392	2.576	0.000	2.361	0.000	4.455	29.100	27.4	294	2	2	290
Reservoir_2m_Nov	7.120	0.319	0.000	0.020	0.000	6.777	22.800	4.5	1196	928	1	267
Reservoir_21m_Nov	7.120	0.314	0.000	0.029	0.000	6.773	19.200	4.4	204	3	1	200
Reservoir_shore_Nov	7.116	0.252	0.000	0.616	0.018	6.230	116.700	3.5	1020	785	2	233

Table 4.2. Sequencing, assembly, and binning statistics for each metagenome.

		Sequencing		Assembly			Genome bins	
Name	Sample	DNA (ng/mL sample)	Read count (1E6 reads)	Total sequence (1E6 bp)	N50	Reads mapped (%)	Total bins	> 70% complete with < 10% contamination
Reservoir_parent_Sept	28	-	17.87	146.27	7680	83.72	19	18
Reservoir_2m_Sept	16	0.028	26.59	120.41	16985	89.27	16	10
Reservoir_21m_Sept	17	0.058	26.95	162.30	1837	89.22	27	16
Reservoir_shore_Nov	35	0.643	31.42	113.20	3869	92.88	28	15
Reservoir_2m_Nov	24	0.053	20.03	87.01	7911	95.19	19	13
Reservoir_21m_Nov	39	0.730	24.60	105.51	3271	94.44	22	13
Enrichment_1	13	NA	13.98	23.93	10241	98.11	6	4
Enrichment_2	14	NA	18.30	29.66	9285	98.13	8	6
Enrichment_3	15	NA	18.30	25.49	4246	98.23	4	3
Tailings_Sept	12	2.412	13.34	154.30	3685	66.11	33	13
Tailings_Nov	32	0.485	29.86	318.65	2307	57.03	47	27
Waste_rock2_Sept	20	0.094	21.79	140.48	3311	63.74	19	6
Waste_rock1_Sept	21	0.121	18.35	181.19	1026	47.01	22	11
Waste_rock1_Nov	37	0.091	16.56	170.76	1158	53.32	9	5
Piped_water_Nov	34	0.141	24.82	141.72	1238	59.11	13	9

Table 4.3. Metabolism and abundance information for genomes from the reservoir across time and depth. The most complete representative from each cluster was chosen and only clusters where the representative genome is ≥ 65% complete are shown. The percent relative abundance

at each depth is given for each member of the cluster (from each sample) as the percent contribution of genome coverage to total bacterial coverage in that sample.

					Sept			Nov		
Representative	**Phylogeny**	**Cluster**	**Completeness**	**Metabolism**	**0m**	**2m**	**21m**	**0m**	**2m**	**21m**
04302015_28_RIFOXYD1_Hydrogenophilaceae_61_11	Hydrogenophilales	25	93.12	sulfur_ox*	2.1	16.0	8.2			
04302015_16_Hydrogenophilaceae_64_40	Hydrogenophilales	55	99.49	sulfur_ox*+		5.7	7.0			
04302015_39_Polarimons-related_63_27	Polaromonas	4	99.84	sulfur_ox*		4.3	3.5	3.5	3.2	3.2
04302015_17_Sulfuricurvum-related_40_25	Sulfuricurvum-like	60	99.19	sulfur_ox*		3.0	2.7			
04302015_17_Thiobacillus_64_11	Hydrogenophilales	136	92.9	sulfur_ox*			1.1			
04302015_17_Thiobacillus_61_9	Hydrogenophilales	168	86.4	sulfur_ox*			0.9			
04302015_39_Gallionellales_52_133	Sideroxydans	61	99.52	sulfur_ox*				11.6	18.2	17.1
04302015_39_Rhodospirillales_66_50	Rhodospirillales	91	97.97	sulfur_ox*				8.5	4.9	6.4
04302015_39_Rhizobiales_66_18	Rhizobiales	67	89.92	sulfur_ox*+				3.1	1.8	2.3
04302015_28_Betaproteobacteria_61_22	Betaproteobacteria	145	95.73	sulfur_ox*+	4.4					
04302015_39_Polynucleobacteria-related_46_119	Polynucleobacteria	35	69.92	sulfur_ox		9.8	6.4	18.9	24.1	17.4
04302015_24_Sulfurovum_sp_AR-rel_42_9	Sulfurovum	46	98.16	sulfur_ox	1.2	7.4	10.1	2.0	1.4	1.7
04302015_17_Thiobacillus_63_34	Hydrogenophilales	53	98.58	sulfur_ox		6.6	3.6			
04302015_35_Thiotrichales_rel_46_9	Thiotrichales	51	98.17	sulfur_ox		3.1	5.1	0.9	0.7	0.7
04302015_39_Burkholderiales_55_53	Burkholderiales	22	99.14	sulfur_ox				4.8	7.6	6.8
04302015_24_Halothiobacillus_neapolitanus_54_40	Halothiobacillus	33	99.14	sulfur_ox	1.0			6.2	5.9	5.8
04302015_35_Acidocella_58_6	Acidocella	64	81.72	sulfur_ox				0.6		
04302015_28_Gammaproteobacteria_57_27	Chromatiales-like	132	96.55	sulfur_ox+	5.5					
04302015_28_Methylophilaceae_bacterium_11_44_11	Methylophilaceae	34	93.68	methylotroph	2.3	1.5	0.9			
04302015_24_Methylotenera_45_7	Methylophilales	18	96.79	methylotroph		1.3	0.8	0.6	1.0	0.9
04302015_39_Acidovorax_64_12	Acidovorax	31	98.99	heterotroph*	2.9	23.0	30.2	1.6	1.2	1.7
04302015_28_Polaromonas_63_22	Polaromonas	3	81.98	heterotroph?*	4.4			24.3	21.2	26.1
04302015_17_Pedobacter_ruber-related_39_43	Sphingobacteriales	58	99.03	heterotroph		7.1	4.6	0.5	0.6	0.6
04302015_16_Alphaproteobacteria_39_46	Alphaproteobacteria	1	96.77	heterotroph		6.6	5.6			
04302015_17_Novosphingobium_62_19	Novosphingobium	65	94.35	heterotroph		1.4	2.1	0.5		0.5
04302015_28_Novosphingobium_62_57	Novosphingobium	113	99.66	heterotroph	12.3		0.9			
04302015_35_Rhizobiales_68_8	Rhizobiales	85	65.3	heterotroph			0.7	0.8		
04302015_28_Alphaproteobacteria_55_17	Sphingomonadales	124	97.41	heterotroph	3.3		0.6			
04302015_39_Sphingomonadales_57_19	Sphingomonadales	71	97.55	heterotroph	8.7			2.2	2.0	2.4
04302015_24_RIFOXYD2_Sphingobacteria_36_13	Sphingobacteria	6	96.92	heterotroph				1.5	1.9	1.8
04302015_39_Polynucleobacteria-related_46_11_partial	Polynucleobacteria	10	66.38	heterotroph				1.3	1.6	1.6
04302015_24_RIFOXYD2_Sphingobacteria-related_39_8	Sphingobacteria	48	91.13	heterotroph				0.8	1.2	1.0
04302015_28_Sphingomonadales_64_96	Sphingomonadales	41	98.7	heterotroph	19.5					
04302015_28_RIFOXYD2_Bacteroidetes_36_52	Sphingobacteria	81	99.01	heterotroph	10.6					
04302015_28_Bdellovibrio_41_41	Bdellovibrio	143	96.43	heterotroph	8.2					
04302015_28_Sphingomonas-related_62_20	Sphingomonas	133	98.89	heterotroph	4.1					
04302015_28_Sphingomonadales-related_66_16	Sphingomonas	102	83.3	heterotroph	3.2					
04302015_28_Sphingomonas-like_63_12	Sphingomonas	49	98.06	heterotroph	2.4					
04302015_28_Sphingomonas_62_11	Sphingomonas	15	97.61	heterotroph	2.2					
04302015_28_Burkholderiales_67_8	Burkholderiales	17	69.26	heterotroph	1.6					

*Nitrate reduction
+Nitrogen fixation

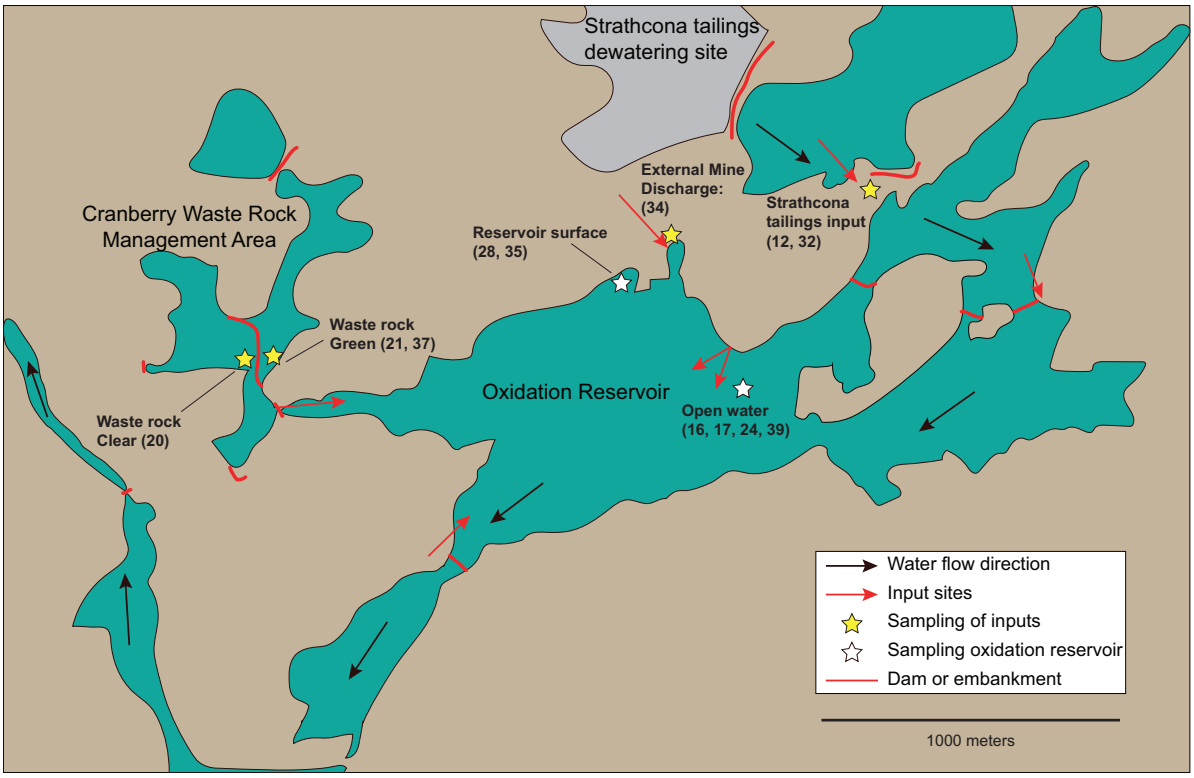

Figure 4.1. Detailed map of the study site in Onaping, Ontario, Canada. Sampling sites are indicated by white stars. Flow paths for tailings sources to enter the reservoir are shown with red arrows, while flow of water within the reservoir is indicated by black arrows. Red lines highlight dammed areas.

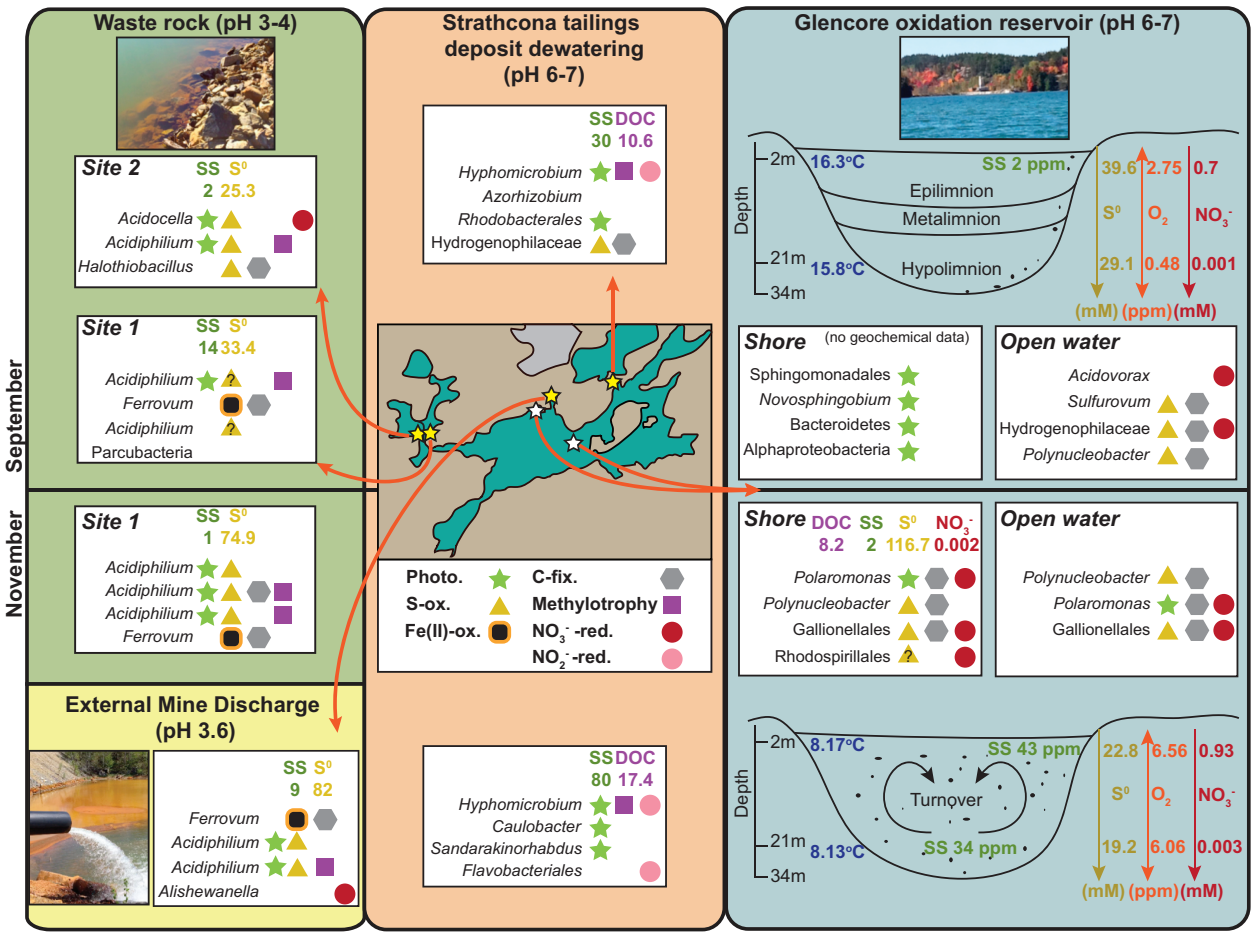

Figure 4.2. Relevant geochemistry, dominant microorganisms and metabolisms shown by site and season. Map in center panel shows locations of sampling and a key for metabolic capacities. Geochemistry for the open water samples at 2 and 21 m is shown alongside a diagram of the reservoir.

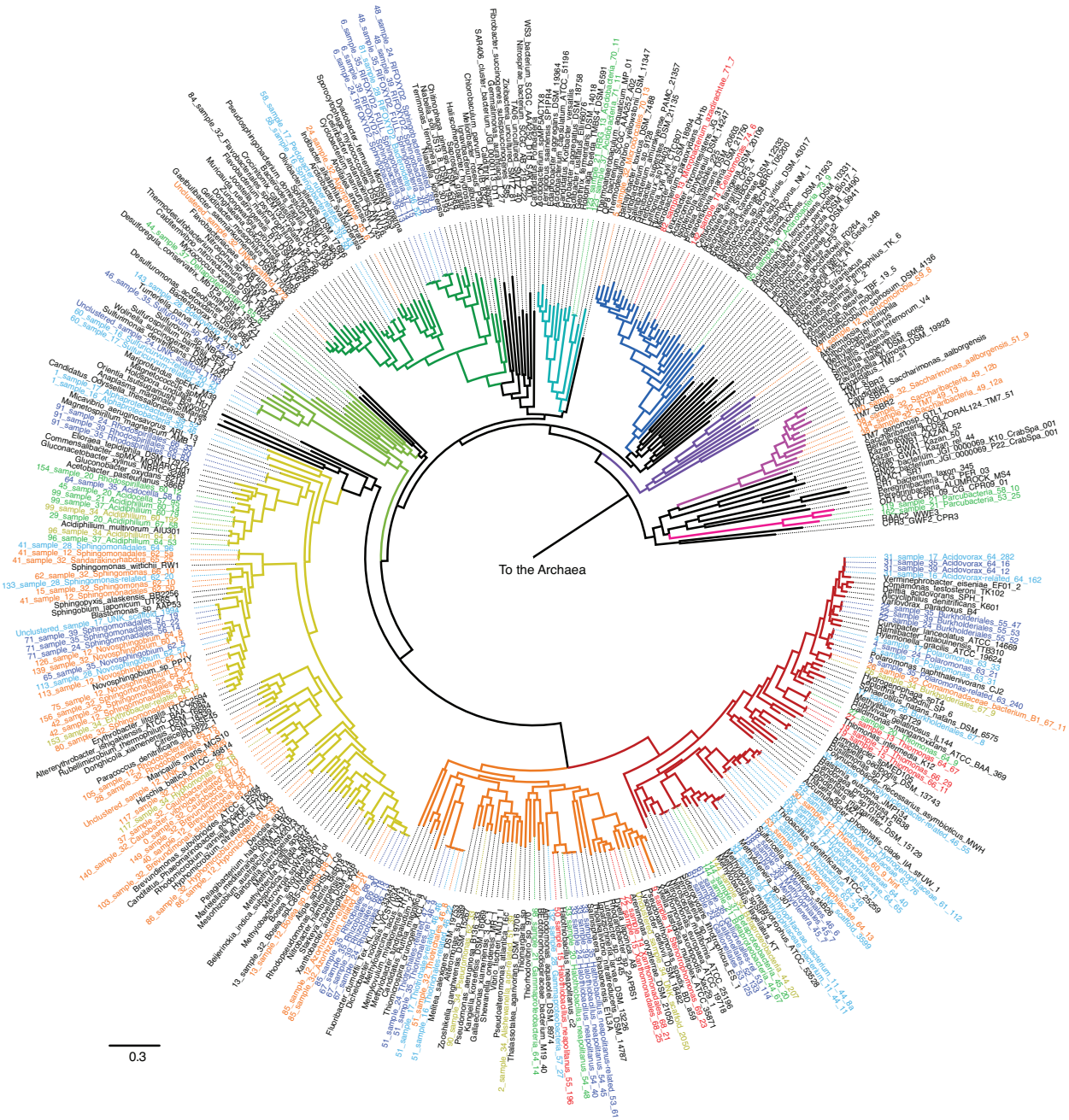

Figure 4.3. Phylogenetic tree based on 16 concatenated ribosomal proteins. Branches representing phylum-level groups containing genomes from this study are colored. Genome names are colored by sample: black, reference; orange, tailings; yellow, piped external mine drainage; green, waste rock; light blue, oxidation reservoir in September; dark blue, oxidation reservoir in November; and red, enrichments. Genome names begin with the cluster number as assigned at 95% ANI. Scale bar represents 0.3 amino acid substitutions.

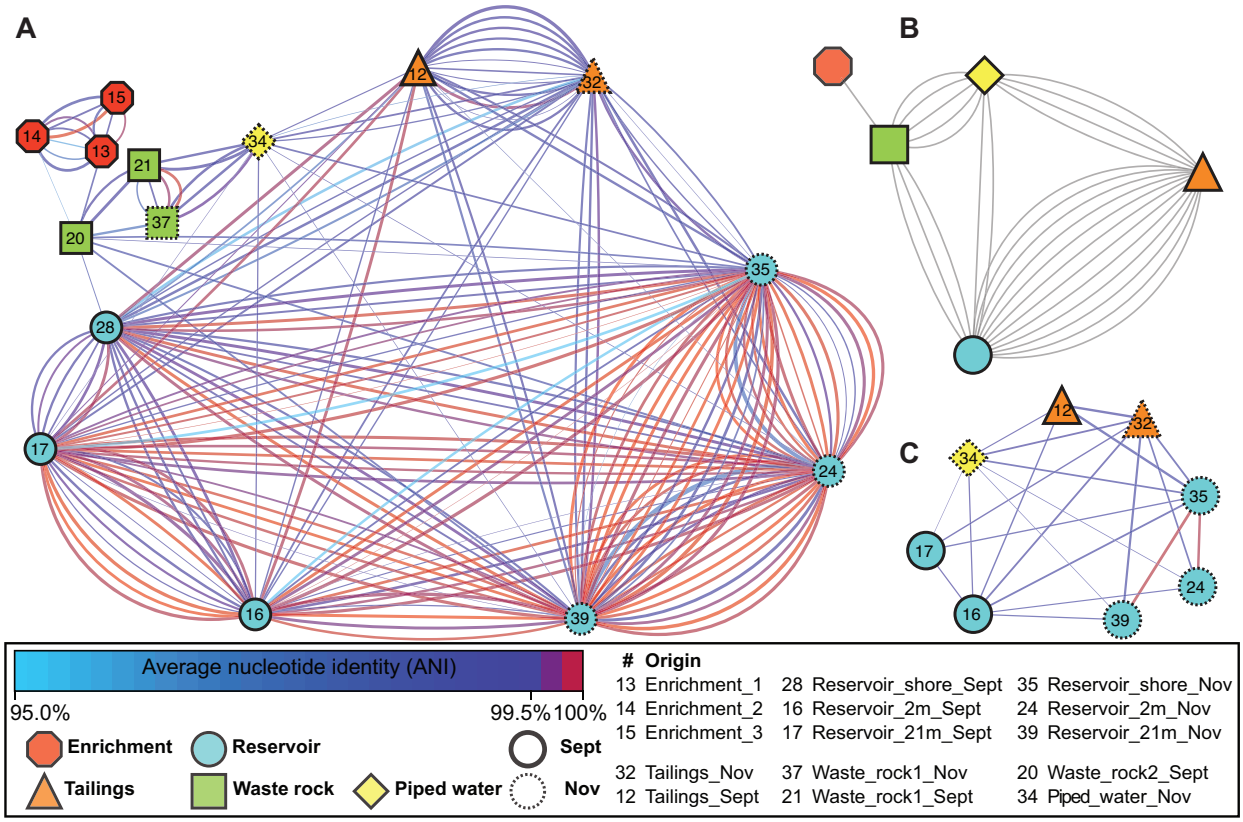

Figure 4.4. Network graph of samples connected by matching recovered genomes with high average nucleotide identity (ANI). Pairwise genome comparisons (199) with alignment coverage > 75% (across at least one genome) and ANI > 95% across aligned regions (A). Samples collapsed into sites and genomes collapsed into clusters (B). A single cluster, *Thiotrichales* spp., showing ANI between 8 different recovered genomes, each from a different sample (C). Line thickness indicates % of genome covered by the alignment used to calculate ANI for the larger genome in each pair (see Table S4.3).

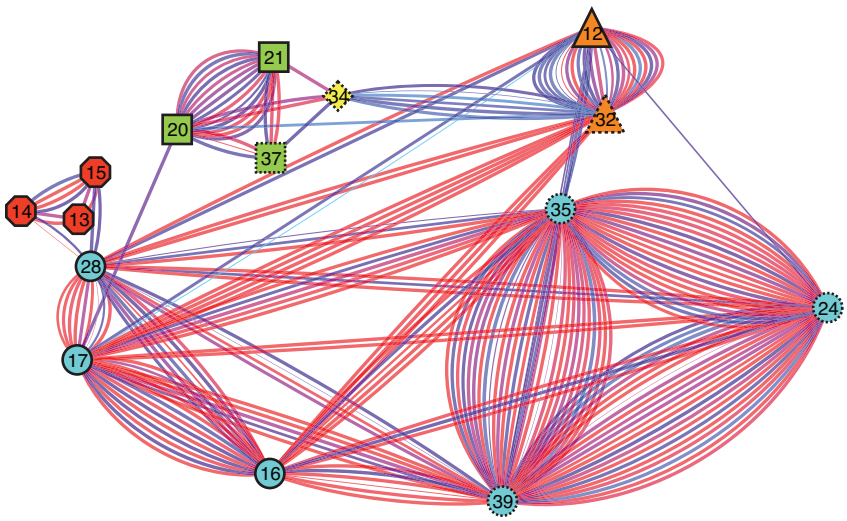

Figure 4.5. Average nucleotide identity of predicted phage across sites. Pairwise alignment of all scaffolds identified as belonging to phage genomes, using the same thresholds as for bacterial genomes (Figure 4.4).

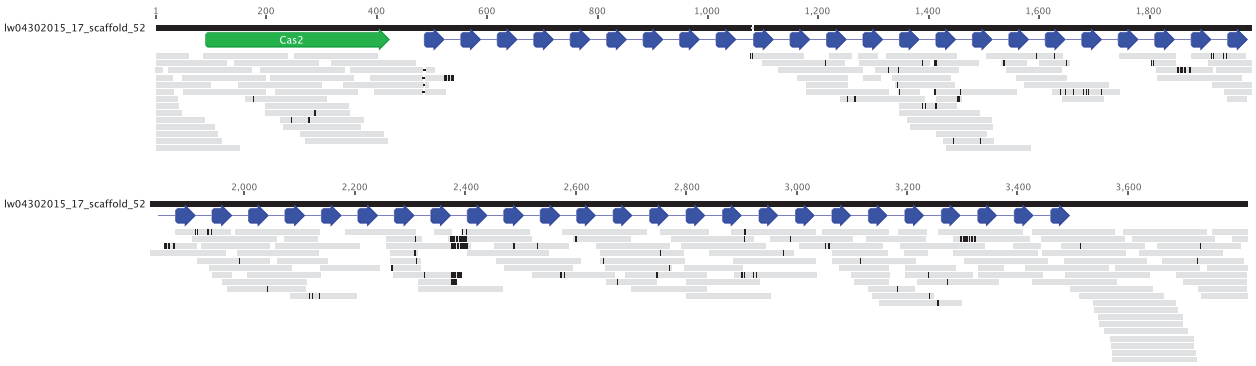

Figure 4.6. Reads from the September tailings sample mapped to the CRISPR locus of the cluster 53 Hydrogenophilales genome from the September reservoir 21 m depth sample. Green arrow indicates Cas2 open reading frame and blue arrows show CRISPR repeats with spacers in between these blue arrows.

Supplementary table titles

Table S4.1. Information about the abundance (coverage), completeness, and metabolism of each bin. Confirmed presence of a gene/pathway of interest is indicated by "y" and pathways that are partially complete are indicated by "p". Methods used for to identify and confirm each gene/pathway are indicated in the header.

Table S4.2. Taxonomic classification of small subunit rRNA genes belonging to eukaryotes (18S) or eukaryotic organelles (16S).

Table S4.3. Average nucleotide identity determined using nucmer/MUMmer (ANIm) for pairs of genomes within the same cluster.

Table S4.4. CRISPR spacers identified in genome bins and corresponding targets identified by BLAST.

Concluding remarks and future work

The studies presented here include metagenomes from three mining-impacted systems and contribute to several active areas of research: (1) elucidating roles of candidate phyla in diverse environments, (2) improving biotechnology design for mine waste treatment, and (3) understanding *in situ* processes for bioremediation of mine waste. In each case, metagenomic analysis has suggested how microorganisms contribute to biogeochemical cycles in their environments and interact with one another. These analyses led to new hypotheses in the form of qualitative models for both individual metabolisms and community-based processes. Future work will test these hypotheses and, in some cases, collect more data to allow generation of quantitative models. Subsequently, these models may be used in design of mine waste remediation systems.

Analyses of four genomes from organisms belonging to four different phyla within the candidate phyla radiation (CPR) indicated that key biosynthetic pathways were most likely absent in these organisms. Additionally, their small genome sizes were consistent with the hypothesis that these organisms are symbionts (chapter 1). Since this work was performed, deep sequencing at several sites has vastly expanded the number of complete and near complete CPR genomes, and the trends of small genome size and limited metabolisms appear to hold (Anantharaman et al., 2016a; Brown et al., 2015; Probst et al., 2016). At least two members of the CPR have now been characterized as symbionts and linked to hosts: an OD1 was reported to be an intracellular symbiont of the protist *Paramecium bursaria* (Gong et al., 2014), while a Saccharibacteria (TM7) was described as an epibiont of the bacterium *Actinomyces odontolyticus* (He et al., 2015).

Further work is needed to identify hosts of CPR organisms in metagenomic data and may be possible by using time-series abundance information to derive co-correlations. This was unsuccessful in the SCN^- bioreactor study in which a member of the CPR lineage Saccharibacteria (TM7) was highly abundant with no obvious host. Indeed, some members of the CPR may not have a specific host but instead may rely on multiple species within a community. Another area of interest is improving the annotation of proteins of unknown function, or so-called hypothetical proteins. CPR genomes contain an unusually high portion of hypothetical proteins, which can hamper investigations of metabolic potential. Hypothetical proteins are prevalent in part because metagenomics relies on pre-existing linkages between protein sequences and protein functions in the literature, but numerous protein families exist for which there has been no experimental characterization. In the future, a combination of approaches will address this knowledge gap including canonical isolate-based techniques, high-throughput isolate-based assays such as BarSeq (Wetmore et al., 2015), and high-throughput functional metagenomics.

Metagenomic studies of a thiocyanate (SCN^-) degrading microbial community used in a successful industrial process pointed to several possible design and operational changes to this process. First, the reactors were originally designed to promote an activated sludge process but biofilm spontaneously formed on all surfaces and likely promoted SCN^- degradation. Second, culture-based methods had suggested that SCN^- degradation relied on heterotrophic metabolism and the ASTER™ process is fed low concentrations of molasses. However, the metagenomic and proteomic studies (chapters 2 and 3) pointed to sulfur oxidizing autotrophs, *Thiobacillus* spp., as major contributors to this biodegradation. Third, the process is highly aerated at the

industrial scale to provide oxygen and mixing. While mixing is important, anaerobic SCN^- degradation may be possible since some *Thiobacillus* spp. genomes encoded both SCN^- hydrolase and complete denitrification pathways, and this metabolism has been observed in alkaliphilic sulfur-oxidizing bacteria (Bezsudnova et al., 2007). Overall, the bioreactor-based metagenomic and proteomic studies suggest possible process modifications including a biofilm-based reactor design, removal of organic carbon from the feed, and lower levels of aeration. These could improve the efficiency and decrease the costs of the process, although they could introduce instability.

Future work on system design could address the question of anaerobic thiocyanate hydrolysis by *Thiobacillus* and other community members. Studies could also test whether certain components of molasses such as vitamins and trace minerals are essential to SCN^- hydrolysis. Identifying which of these are important could lead to more strategic augmentation of the feed. Most importantly, studies will need to sample the industrial-scale, on-site bioreactors to ascertain the degree of similarity between their microbial communities and those present in the laboratory-scale reactors. While the industrial- and laboratory-scale reactors shared a common inoculum, the communities may have diverged. Further work in laboratory-scale bioreactors is still important as it will allow linkage of metagenomic data to several other key pieces of information needed for quantitative modeling: absolute abundances of individual strains within a community and rates of reactions performed by these strains. These quantitative studies will ideally include more replicates to allow for significance testing, reflecting recent discussions on the use of metagenomics in biotechnology (Johnson et al., 2015; Ju and Zhang, 2015).

From a basic biology perspective, much remains unknown about the biochemical pathways underpinning thiocyanate degradation. The genes involved in the "OCN^- pathway" remain uncharacterized (**Figure i.3**). Heterotrophic SCN^- degradation, presumed to use this pathway, has been observed (Stafford and Callely, 1969 and unpublished data), but the genes involved have not been identified. This means that metagenomic datasets cannot yet be searched to identify these genes and the contribution of heterotrophs to thiocyanate degradation in the bioreactor communities cannot be quantified. In the "OCS pathway", it was previously observed that SCN^- degradation by SCN^- hydrolase produces carbonyl sulfide gas (OCS), which is converted to sulfide by carbonyl sulfide hydrolase (Ogawa et al., 2013). However, the gene for this enzyme was not found in any genome from the bioreactors studied here, and it is unclear what reactions occur immediately following SCN^- hydrolysis (**Figure i.3**). Genetic and biochemical studies of other genes in the SCN^- operon (chapter 2) could help address this knowledge gap.

The approach of natural *in situ* attenuation of sulfidic mine wastes is fairly common, in conjunction with chemical treatments such as neutralization. However, there is limited information on contaminated freshwater microbial communities, and most studies in this area have focused on heavy metals (Kang et al., 2013; Reis et al., 2016). Furthermore, while there is awareness of thiosalts, there is no characterization of how they alter native microbial communities. Given the extent of the thiosalts problem and the associated costs to the mining industry and the environment, there is a great need for more work on microbial remediation. From a biological perspective, future work on mine tailings in freshwater could examine genetic diversity of natural populations and could investigate whether specific groups are associated with faster or slower thiosalts removal rates. From an engineering viewpoint, a key weakness of natural attenuation is that the rate of thiosalts oxidation decreases with colder temperatures.

Bioreactor systems are a promising approach to this problem, and genome-guided biostimulation may also improve remediation in natural systems.

In parallel with system redesign, monitoring mine waste remediation in industrial bioreactors and mine waste sites is important. Already, chemical readouts are used to check that wastewater is treated before it is released, but monitoring the microbial communities and correlating this with performance could give a clearer picture of the remediation process (Garris et al., 2016). Microbiologists in Denmark have created a database of genomes from wastewater treatment plants (McIlroy et al., 2015) to help track organisms and understand system performance. In the future, a similar database will be developed for mining-related organisms. An obvious place to start is with well-known acid mine drainage-associated organisms present across specific pH ranges. The database can be extended to include organisms involved in thiocyanate-degradation, thiosalts oxidation and other important functions such as heavy metals precipitation. This database could contain genomes, gene expression data, probe sequences and all other associated geochemical and kinetic data, and would allow for improved modeling and monitoring of diverse mining systems. Finally, the database could increase collaboration on mine waste remediation around the globe, accelerating progress in this critical area.

Literature cited

Abramson, J., Riistama, S., Larsson, G., Jasaitis, A., Svensson-Ek, M., Laakkonen, L., Puustinen, A., Iwata, S., and Wikström, M. (2000). The structure of the ubiquinol oxidase from Escherichia coli and its ubiquinone binding site. Nat. Struct. Biol. *7*, 910–917.

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

Albertsen, M., Saunders, A.M., Nielsen, K.L., and Nielsen, P.H. (2013b). Metagenomes obtained by "deep sequencing" – what do they tell about the enhanced biological phosphorus removal communities? Water Sci. Technol. *68*, 1959.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology *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.

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. Nature Communications *7*, 13219.

Anderson, R.T., Vrionis, H.A., Ortiz-Bernad, I., Resch, C.T., Long, P.E., Dayvault, R., Karp, K., Marutzky, S., Metzler, D.R., Peacock, A., et al. (2003). Stimulating the in situ activity of Geobacter species to remove uranium from the groundwater of a uranium-contaminated aquifer. Applied and Environmental Microbiology *69*, 5884–5891.

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

Aukema, K.G., Kron, E.M., Herdendorf, T.J., and Forest, K.T. (2005). Functional dissection of a conserved motif within the pilus retraction protein PilT. Journal of Bacteriology *187*, 611–618.

Baker, B.J., and Banfield, J.F. (2003). Microbial communities in acid mine drainage. FEMS Microbiol Ecol *44*, 139–152.

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. Proc. Natl. Acad. Sci. U.S.A. *107*, 8806–8811.

Baker, B.J., Sheik, C.S., Taylor, C.A., Jain, S., Bhasi, A., Cavalcoli, J.D., and Dick, G.J. (2013).

Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. The ISME Journal *7*, 1962–1973.

Beller, H.R., Letain, T.E., Chakicherla, A., Kane, S.R., Legler, T.C., and Coleman, M.A. (2006a). Whole-genome transcriptional analysis of chemolithoautotrophic thiosulfate oxidation by *Thiobacillus denitrificans* under aerobic versus denitrifying conditions. Journal of Bacteriology *188*, 7005–7015.

Beller, H.R., Chain, P.S.G., Letain, T.E., Chakicherla, A., Larimer, F.W., Richardson, P.M., Coleman, M.A., Wood, A.P., and Kelly, D.P. (2006b). The genome sequence of the obligately chemolithoautotrophic, facultatively anaerobic bacterium *Thiobacillus denitrificans*. Journal of Bacteriology *188*, 1473–1488.

Beller, H.R., Zhou, P., Legler, T.C., Chakicherla, A., Kane, S., Letain, T.E., and A O'Day, P. (2013). Genome-enabled studies of anaerobic, nitrate-dependent iron oxidation in the chemolithoautotrophic bacterium *Thiobacillus denitrificans*. Front. Microbio. *4*, 249.

Bezsudnova, E.Y., Sorokin, D.Y., Tikhonova, T.V., and Popov, V.O. (2007). Thiocyanate hydrolase, the primary enzyme initiating thiocyanate degradation in the novel obligately chemolithoautotrophic halophilic sulfur-oxidizing bacterium *Thiohalophilus thiocyanoxidans*. Biochim. Biophys. Acta *1774*, 1563–1570.

Bor, B., Poweleit, N., Bois, J.S., Cen, L., Bedree, J.K., Zhou, Z.H., Gunsalus, R.P., Lux, R., McLean, J.S., He, X., et al. (2015). Phenotypic and physiological characterization of the epibiotic interaction between TM7x and its basibiont *Actinomyces*. Microb Ecol *71*, 243–255.

Borisov, V.B., Gennis, R.B., Hemp, J., and Verkhovsky, M.I. (2011). The cytochrome bd respiratory oxygen reductases. BBA - Bioenergetics *1807*, 1398–1413.

Borrel, G., Lehours, A.-C., Bardot, C., Bailly, X., and Fonty, G. (2010). Members of candidate divisions OP11, OD1 and SR1 are widespread along the water column of the meromictic Lake Pavin (France). Arch Microbiol *192*, 559–567.

Boucabeille, C., Bories, A., Ollivier, P., and Michel, G. (1994). Microbial degradation of metal complexed cyanides and thiocyanate from mining wastewaters. Environ. Pollut. *84*, 59–67.

Brown, A.M., Hoopes, S.L., White, R.H., and Sarisky, C.A. (2011). Purine biosynthesis in archaea: variations on a theme. Biology Direct *6*, 63.

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. Nature Biotechnology.

Brown, M.W., Sharpe, S.C., Silberman, J.D., Heiss, A.A., Lang, B.F., Simpson, A.G.B., and Roger, A.J. (2013). Phylogenomics demonstrates that breviate flagellates are related to

opisthokonts and apusomonads. Proceedings of the Royal Society B: Biological Sciences *280*, 20131755–20131755.

Butterfield, C.N., Li, Z., Andeer, P.F., Spaulding, S., Thomas, B.C., Singh, A., Hettich, R.L., Suttle, K.B., Probst, A.J., Tringe, S.G., et al. (2016). Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. PeerJ *4*, e2687.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

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

Castelle, C.J., Hug, L.A., Wrighton, K.C., Thomas, B.C., Williams, K.H., Wu, D., Tringe, S.G., Singer, S.W., Eisen, J.A., and Banfield, J.F. (2013). Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. Nature Communications *4*, 1–10.

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nature Biotechnology *30*, 918–920.

Chen, I., and Dubnau, D. (2004). DNA uptake during bacterial transformation. Nature Reviews Microbiology *2*, 241–249.

Chen, L.-X., Hu, M., Huang, L.-N., Hua, Z.-S., Kuang, J.-L., Li, S.-J., and Shu, W.-S. (2015). Comparative metagenomic and metatranscriptomic analyses of microbial communities in acid mine drainage. The ISME Journal *9*, 1579–1592.

Chen, M.-Y., Tsay, S.-S., Chen, K.-Y., Shi, Y.-C., Lin, Y.-T., and Lin, G.-H. (2002). *Pseudoxanthomonas taiwanensis* sp. nov., a novel thermophilic, N_2O-producing species isolated from hot springs. International Journal of Systematic and Evolutionary Microbiology *52*, 2155–2161.

Cherney, M.M., Zhang, Y., Solomonson, M., Weiner, J.H., and James, M.N.G. (2010). Crystal structure of sulfide:quinone oxidoreductase from *Acidithiobacillus ferrooxidans*: insights into sulfidotrophic respiration and detoxification. Journal of Molecular Biology *398*, 292–305.

Chourey, K., Jansson, J., VerBerkmoes, N., Shah, M., Chavarria, K.L., Tom, L.M., Brodie, E.L., and Hettich, R.L. (2010). Direct Cellular Lysis/Protein Extraction Protocol for Soil Metaproteomics. J. Proteome Res. *9*, 6615–6622.

Cipollone, R., Ascenzi, P., and Visca, P. (2007a). Common themes and variations in the rhodanese superfamily. IUBMB Life *59*, 51–59.

Cipollone, R., Bigotti, M.G., Frangipani, E., Ascenzi, P., and Visca, P. (2004). Characterization of a rhodanese from the cyanogenic bacterium *Pseudomonas aeruginosa*. Biochemical and Biophysical Research Communications *325*, 85–90.

Cipollone, R., Frangipani, E., Tiburzi, F., Imperi, F., Ascenzi, P., and Visca, P. (2007b). Involvement of *Pseudomonas aeruginosa* rhodanese in protection from cyanide toxicity. Applied and Environmental Microbiology *73*, 390–398.

Cluness, M.J., Turner, P.D., Clements, E., Brown, D.T., and O'Reilly, C. (1993). Purification and properties of cyanide hydratase from *Fusarium lateritium* and analysis of the corresponding chy1 gene. J. Gen. Microbiol. *139*, 1807–1815.

Comfort, D., and Clubb, R.T. (2004). A comparative genome analysis identifies distinct sorting pathways in gram-positive bacteria. Infection and Immunity *72*, 2710–2722.

Coupland, K., and Johnson, D.B. (2008). Evidence that the potential for dissimilatory ferric iron reduction is widespread among acidophilic heterotrophic bacteria. FEMS Microbiology Letters *279*, 30–35.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. J Gerontol *27*, 1164–1165.

Davis, J.P., Youssef, N.H., and Elshahed, M.S. (2009). Assessment of the diversity, abundance, and ecological distribution of members of candidate division SR1 reveals a high level of phylogenetic diversity but limited morphotypic diversity. Applied and Environmental Microbiology *75*, 4139–4148.

Denef, V.J., Mueller, R.S., and Banfield, J.F. (2010). AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. The ISME Journal *4*, 599–610.

Dewhirst, F.E., Chen, T., Izard, J., Paster, B.J., Tanner, A.C., Yu, W.-H., Lakshmanan, A., and Wade, W.G. (2010). The human oral microbiome. Journal of Bacteriology *192*, 5002–5017.

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.

Dictor, M.C., Battaglia-Brunet, F., Morin, D., Bories, A., and Clarens, M. (1997). Biological treatment of gold ore cyanidation wastewater in fixed bed reactors. Environ. Pollut. *97*, 287–294.

Dinis, J.M., Barton, D.E., Ghadiri, J., Surendar, D., Reddy, K., Velasquez, F., Chaffee, C.L., Lee, M.-C.W., Gavrilova, H., and Ozuna, H. (2011). In search of an uncultured human-associated TM7 bacterium in the environment. PLoS ONE *6*, e21280.

Dopson, M., and Johnson, D.B. (2012). Biodiversity, metabolism and applications of acidophilic sulfur-metabolizing microorganisms. Environ Microbiol *14*, 2620–2631.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I.M., Barbe, V., Duprat, S., Galperin, M.Y., Koonin, E.V., Le Gall, F., et al. (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc. Natl. Acad. Sci. U.S.A. *100*, 10020–10025.

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

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

Emerson, D. (2013). Comparative genomics of freshwater Fe-oxidizing bacteria: implications for physiology, ecology, and systematics. 1–17.

Erdogan, M.F. (2003). Thiocyanate overload and thyroid disease. Biofactors *19*, 107–111.

Fauvart, M., Braeken, K., Daniels, R., Vos, K., Ndayizeye, M., Noben, J.-P., Robben, J., Vanderleyden, J., and Michiels, J. (2007). Identification of a novel glyoxylate reductase supports phylogeny-based enzymatic substrate specificity prediction. Biochim. Biophys. Acta *1774*, 1092–1098.

Felföldi, T., Székely, A.J., Gorál, R., Barkács, K., Scheirich, G., András, J., Rácz, A., and Márialigeti, K. (2010). Polyphasic bacterial community analysis of an aerobic activated sludge removing phenols and thiocyanate from coke plant effluent. Bioresource Technology *101*, 3406–3414.

Feng, Y., Li, W., Li, J., Wang, J., Ge, J., Xu, D., Liu, Y., Wu, K., Zeng, Q., Wu, J.-W., et al. (2012). Structural insight into the type-II mitochondrial NADH dehydrogenases. Nature *491*, 478–482.

Fernandez, R.F., and Kunz, D.A. (2005). Bacterial cyanide oxygenase is a suite of enzymes catalyzing the scavenging and adventitious utilization of cyanide as a nitrogenous growth substrate. Journal of Bacteriology *187*, 6396–6402.

Finkmann, W., Altendorf, K., Stackebrandt, E., and Lipski, A. (2000). Characterization of N_2O-producing Xanthomonas-like isolates from biofilters as *Stenotrophomonas nitritireducens* sp. nov., *Luteimonas mephitis* gen. nov., sp. nov. and *Pseudoxanthomonas broegbernensis* gen. nov., sp. nov. International Journal of Systematic and Evolutionary Microbiology *50 Pt 1*, 273–282.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research *44*, D279–D285.

Garris, H.W., Baldwin, S.A., Van Hamme, J.D., Gardner, W.C., and Fraser, L.H. (2016). Genomics to assist mine reclamation: a review. Restoration Ecology *24*, 165–173.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv Preprint arXiv:1207.3907 [Q-Bio.GN] 1–9.

Ghosh, W., and Dam, B. (2009). Biochemistry and molecular biology of lithotrophic sulfur oxidation by taxonomically and ecologically diverse bacteria and archaea. FEMS Microbiol. Rev. *33*, 999–1043.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., et al. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. Science *309*, 1242–1245.

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.

Gould, W.D., King, M., Mohapatra, B.R., Cameron, R.A., Kapoor, A., and Koren, D.W. (2012). A critical review on destruction of thiocyanate in mining effluents. Minerals Engineering *34*, 38–47.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Research *35*, W52–W57.

Grote, J., Thrash, J.C., Huggett, M.J., Landry, Z.C., Carini, P., Giovannoni, S.J., and Rappé, M.S. (2012). Streamlining and core genome conservation among highly divergent members of the SAR11 clade. mBio *3*.

Guermazi, S., Daegelen, P., Dauga, C., Rivière, D., Bouchez, T., Godon, J.J., Gyapay, G., Sghir, A., Pelletier, E., Weissenbach, J., et al. (2008). Discovery and characterization of a new bacterial candidate division by an anaerobic sludge digester metagenomic approach. Environ Microbiol *10*, 2111–2123.

Guilloton, M.B., Lamblin, A.F., Kozliak, E.I., Gerami-Nejad, M., Tu, C., Silverman, D., Anderson, P.M., and Fuchs, J.A. (1993). A physiological role for cyanate-induced carbonic-anhydrase in *Escherichia coli*. Journal of Bacteriology *175*, 1443–1451.

Haft, D.H., Basu, M.K., and Mitchell, D.A. (2010). Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family. BMC Biol. *8*.

Handley, K.M., Wrighton, K.C., Miller, C.S., Wilkins, M.J., Kantor, R.S., Thomas, B.C., Williams, K.H., Gilbert, J.A., Long, P.E., and Banfield, J.F. (2015). Disturbed subsurface microbial communities follow equivalent trajectories despite different structural starting points. Environ Microbiol *17*, 622–636.

Handley, K.M., Wrighton, K.C., Piceno, Y.M., Andersen, G.L., Desantis, T.Z., Williams, K.H., Wilkins, M.J., N'Guessan, A.L., Peacock, A., Bargar, J., et al. (2012). High-density PhyloChip profiling of stimulated aquifer microbial communities reveals a complex response to acetate amendment. FEMS Microbiol Ecol.

Harris, J.K., Kelley, S.T., and Pace, N.R. (2004). New Perspective on Uncultured Bacterial Phylogenetic Division OP11. Applied and Environmental Microbiology *70*, 845–849.

Harris, J.K., Caporaso, J.G., Walker, J.J., Spear, J.R., Gold, N.J., Robertson, C.E., Hugenholtz, P., Goodrich, J., McDonald, D., Knights, D., et al. (2013). Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat. The ISME Journal *7*, 50–60.

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. Proc. Natl. Acad. Sci. U.S.A. *112*, 244–249.

Hino, T., Matsumoto, Y., Nagano, S., Sugimoto, H., Fukumori, Y., Murata, T., Iwata, S., and Shiro, Y. (2010). Structural basis of biological N_2O generation by bacterial nitric oxide reductase. Science *330*, 1666–1670.

Huddy, R.J., van zyl, A.W., van Hille, R.P., and Harrison, S.T.L. (2015). Characterisation of the complex microbial community associated with the ASTER^{TM} thiocyanate biodegradation system. Minerals Engineering *76*, 65–71.

Huddy, R.J., van zyl, A.W., van Hille, R.P., and Harrison, S.T.L. Characterisation of the complex microbial community associated with the ASTER™ thiocyanate biodegradation system (Minerals Engineering).

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. Nature Microbiology *1*, 16048.

Hug, L.A., Beiko, R.G., Rowe, A.R., Richardson, R.E., and Edwards, E.A. (2012). Comparative metagenomics of three Dehalococcoides-containing enrichment cultures: the role of the non-dechlorinating community. BMC Genomics *13*, 327.

Hug, L.A., Thomas, B.C., Sharon, I., Brown, C.T., Sharma, R., Hettich, R.L., Wilkins, M.J., Williams, K.H., Singh, A., and Banfield, J.F. (2015). Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. Environ Microbiol.

Hugenholtz, P., Goebel, B.M., and Pace, N.R. (1998a). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. Journal of Bacteriology *180*, 4765–4774.

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

Hugenholtz, P., Tyson, G.W., Webb, R.I., Wagner, A.M., and Blackall, L.L. (2001). Investigation of candidate division TM7, a recently recognized major lineage of the domain Bacteria with no known pure-culture representatives. Applied and Environmental Microbiology *67*, 411–419.

Hung, C.-H., and Pavlostathis, S.G. (1999). Kinetics and modeling of autotrophic thiocyanate biodegradation. Biotechnol Bioeng *62*, 1–11.

Hussain, A., Ogawa, T., Saito, M., Sekine, T., Nameki, M., Matsushita, Y., Hayashi, T., and Katayama, Y. (2013). Cloning and expression of a gene encoding a novel thermostable thiocyanate-degrading enzyme from a mesophilic alphaproteobacteria strain THI201. Microbiology *159*, 2294–2302.

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.

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

Iglesias, N., Romero, R., and Mazuelos, A. (2016). Treatment of tetrathionate effluents by continuous oxidation in a flooded packed-bed bioreactor. International Journal of Mineral Processing.

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.

Jandhyala, D.M., Willson, R.C., Sewell, B.T., and Benedik, M.J. (2005). Comparison of cyanide-degrading nitrilases. Appl Microbiol Biotechnol *68*, 327–335.

Jandhyala, D., Berman, M., Meyers, P.R., Sewell, B.T., Willson, R.C., and Benedik, M.J. (2003). CynD, the cyanide dihydratase from *Bacillus pumilus*: gene cloning and structural studies. Applied and Environmental Microbiology *69*, 4794–4805.

Jensen, N.B., Zagrobelny, M., Hjernø, K., Olsen, C.E., Houghton-Larsen, J., Borch, J., Møller, B.L., and Bak, S. (2011). Convergent evolution in biosynthesis of cyanogenic defence compounds in plants and insects. Nature Communications *2*, 273.

Jeong, Y.-S., and Chung, J.S. (2006). Biodegradation of thiocyanate in biofilm reactor using fluidized-carriers. Process Biochemistry *41*, 701–707.

Jing, H., Takagi, J., Liu, J.-H., Lindgren, S., Zhang, R.-G., Joachimiak, A., Wang, J.-H., and Springer, T.A. (2002). Archaeal surface layer proteins contain beta propeller, PKD, and beta helix domains and are related to metazoan cell surface proteins. Structure/Folding and Design *10*, 1453–1464.

Johnson, D. (2014). Recent Developments in Microbiological Approaches for Securing Mine Wastes and for Recovering Metals from Mine Waters. Minerals *4*, 279–292.

Johnson, D.R., Helbling, D.E., Men, Y., and Fenner, K. (2015). Can meta-omics help to establish causality between contaminant biotransformations and genes or gene products? Environmental Science: Water Research & Technology *1*, 272–278.

Ju, F., and Zhang, T. (2015). Experimental design and bioinformatics analysis for the application of metagenomics in environmental sciences and biotechnology. Environ. Sci. Technol. *49*, 12628–12640.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research *44*, D457–D462.

Kang, H.J., and Baker, E.N. (2012). Structure and assembly of Gram-positive bacterial pili: unique covalent polymers. Curr. Opin. Struct. Biol. *22*, 200–207.

Kang, S., Van Nostrand, J.D., Gough, H.L., He, Z., Hazen, T.C., Stahl, D.A., and Zhou, J. (2013). Functional gene array-based analysis of microbial communities in heavy metals-contaminated lake sediments. FEMS Microbiol Ecol *86*, 200–214.

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*, e00708–13–e00708–13.

Kantor, R.S., van zyl, A.W., van Hille, R.P., Thomas, B.C., Harrison, S.T.L., and Banfield, J.F. (2015). Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unravelled with genome-resolved metagenomics. Environ Microbiol *17*, 4929–4941.

Kataoka, S., Arakawa, T., Hori, S., Katayama, Y., Hara, Y., Matsushita, Y., Nakayama, H., Yohda, M., Nyunoya, H., Dohmae, N., et al. (2006). Functional expression of thiocyanate hydrolase is promoted by its activator protein, P15K. FEBS Lett. *580*, 4667–4672.

Katayama, Y., and Kuraishi, H. (1978). Characteristics of *Thiobacillus thioparus* and its thiocyanate assimilation. Can. J. Microbiol. *24*, 804–810.

Katayama, Y., Matsushita, Y., Kaneko, M., Kondo, M., Mizuno, T., and Nyunoya, H. (1998). Cloning of genes coding for the three subunits of thiocyanate hydrolase of *Thiobacillus thioparus* THI 115 and their evolutionary relationships to nitrile hydratase. Journal of Bacteriology *180*, 2583–2589.

Katayama, Y., Narahara, Y., Inoue, Y., Amano, F., Kanagawa, T., and Kuraishi, H. (1992). A thiocyanate hydrolase of Thiobacillus thioparus. A novel enzyme catalyzing the formation of carbonyl sulfide from thiocyanate. J. Biol. Chem. *267*, 9170–9175.

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.

Kelly, D.P., and Wood, A.P. (1998). Microbes of the Sulfur Cycle. In Techniques in Microbial Ecology, R.S. Burlage, ed. (Oxford Univ Press on Demand), pp. 37–38.

Kim, S.-W., Fushinobu, S., Zhou, S., Wakagi, T., and Shoun, H. (2009). Eukaryotic *nirK* genes encoding copper-containing nitrite reductase: originating from the protomitochondrion? Applied and Environmental Microbiology *75*, 2652–2658.

Knight, R.D., Freeland, S.J., and Landweber, L.F. (2001). Rewiring the keyboard: evolvability of the genetic code. Nat. Rev. Genet. *2*, 49–58.

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.

Kraft, B., Tegetmeyer, H.E., Sharma, R., Klotz, M.G., Ferdelman, T.G., Hettich, R.L., Geelhoed, J.S., and Strous, M. (2014). The environmental controls that govern the end product of bacterial nitrate respiration. Science *345*, 676–679.

Kuang, J.-L., Huang, L.-N., Chen, L.-X., Hua, Z.-S., Li, S.-J., Hu, M., Li, J.-T., and Shu, W.-S. (2013). Contemporary environmental variation determines microbial diversity patterns in acid mine drainage. The ISME Journal *7*, 1038–1050.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol *5*, R12.

Kuyucak, N., and Yaschyshyn, D. (2007). Managing thiosalts in mill effluents: studies conducted at the Kidd metallurgical site. pp. 1–11.

Küsel, K., Dorsch, T., Acker, G., and Stackebrandt, E. (1999). Microbial reduction of Fe(III) in acidic sediments: isolation of Acidiphilium cryptum JF-5 capable of coupling the reduction of Fe(III) to the oxidation of glucose. Applied and Environmental Microbiology *65*, 3633–3640.

Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research *35*, 3100–3108.

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

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lasken, R.S. (2012). Genomic sequencing of uncultured microorganisms from single cells. Nature Reviews Microbiology *10*, 631–640.

Lee, J., Sperandio, V., Frantz, D.E., Longgood, J., Camilli, A., Phillips, M.A., and Michael, A.J. (2009). An alternative polyamine biosynthetic pathway is widespread in bacteria and essential for biofilm formation in Vibrio cholerae. J. Biol. Chem. *284*, 9899–9907.

Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, J.S., Schlegel, M.L., Tucker, T.A., Schrenzel, M.D., Knight, R., et al. (2008). Evolution of mammals and their gut microbes. Science *320*, 1647–1651.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liang, Y., Van Nostrand, J.D., N'guessan, L.A., Peacock, A.D., Deng, Y., Long, P.E., Resch, C.T., Wu, L., He, Z., Li, G., et al. (2012). Microbial functional gene diversity with a shift of subsurface redox conditions during In situ uranium reduction. Applied and Environmental Microbiology *78*, 2966–2972.

Liljeqvist, M., Sundkvist, J.-E., Saleh, A., and Dopson, M. (2011). Low temperature removal of inorganic sulfur compounds from mining process waters. Biotechnol Bioeng *108*, 1251–1259.

Lochner, A., Giannone, R.J., Keller, M., Antranikian, G., Graham, D.E., and Hettich, R.L. (2011). Label-free quantitative proteomics for the extremely thermophilic bacterium *Caldicellulosiruptor obsidiansis* reveal distinct abundance patterns upon growth on cellobiose, crystalline cellulose, and switchgrass. J. Proteome Res. *10*, 5302–5314.

Losey, N.A., Stevenson, B.S., Busse, H.-J., Sinninghe Damsté, J.S., Rijpstra, W.I.C., Rudd, S., and Lawson, P.A. (2013). *Thermoanaerobaculum aquaticum* gen. nov., sp. nov., the first cultivated member of Acidobacteria subdivision 23, isolated from a hot spring. International Journal of Systematic and Evolutionary Microbiology *63*, 4149–4157.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research *25*, 955–964.

Loy, A., Duller, S., Baranyi, C., Mussmann, M., Ott, J., Sharon, I., Béjà, O., Le Paslier, D., Dahl, C., and Wagner, M. (2009). Reverse dissimilatory sulfite reductase as phylogenetic marker for a subgroup of sulfur-oxidizing prokaryotes. Environ Microbiol *11*, 289–299.

Luque Almagro, V.M., Huertas, M.-J., Sáez, L.P., Luque-Romero, M.M., Moreno Vivián, C., Castillo, F., Roldán, M.D., and Blasco, R. (2008). Characterization of the *Pseudomonas pseudoalcaligenes* CECT5344 Cyanase, an enzyme that is not essential for cyanide assimilation. Applied and Environmental Microbiology *74*, 6280–6288.

Luque-Almagro, V.M., Merchan, F., Blasco, R., Igeno, M.I., Martínez-Luque, M., Moreno-Vivián, C., Castillo, F., and Roldan, M.D. (2011). Cyanide degradation by *Pseudomonas pseudoalcaligenes* CECT5344 involves a malate : quinone oxidoreductase and an associated cyanide-insensitive electron transfer chain. Microbiology *157*, 739–746.

Lykidis, A., Chen, C.-L., Tringe, S.G., McHardy, A.C., Copeland, A., Kyrpides, N.C., Hugenholtz, P., Macarie, H., Olmos, A., Monroy, O., et al. (2010). Multiple syntrophic interactions in a terephthalate-degrading methanogenic consortium. The ISME Journal *5*, 122–130.

Ma, Z.-Q., Chambers, M.C., Ham, A.-J.L., Cheek, K.L., Whitwell, C.W., Aerni, H.-R., Schilling, B., Miller, A.W., Caprioli, R.M., and Tabb, D.L. (2011). ScanRanker: Quality assessment of tandem mass spectra via sequence tagging. J. Proteome Res. *10*, 2896–2904.

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.

Madsen, E.L. (2011). Microorganisms and their roles in fundamental biogeochemical cycles. Current Opinion in Biotechnology *22*, 456–464.

Marcia, M., Ermler, U., Peng, G., and Michel, H. (2009). The structure of *Aquifex aeolicus*

sulfide:quinone oxidoreductase, a basis to understand sulfide detoxification and respiration. Proc. Natl. Acad. Sci. U.S.A. *106*, 9625–9630.

Marcia, M., Ermler, U., Peng, G., and Michel, H. (2010). A new structure-based classification of sulfide:quinone oxidoreductases. Proteins *78*, 1073–1083.

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. Proc. Natl. Acad. Sci. U.S.A. *104*, 11889–11894.

Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., et al. (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Research *40*, D115–D122.

Martineau, C., Villeneuve, C., Mauffrey, F., and Villemur, R. (2013). *Hyphomicrobium nitrativorans* sp. nov., isolated from the biofilm of a methanol-fed denitrification system treating seawater at the Montreal Biodome. International Journal of Systematic and Evolutionary Microbiology *63*, 3777–3781.

Martineau, C., Villeneuve, C., Mauffrey, F., and Villemur, R. (2014). Complete genome sequence of *Hyphomicrobium nitrativorans* strain NL23, a denitrifying bacterium isolated from biofilm of a methanol-fed denitrification system treating seawater at the Montreal Biodome. Genome Announc *2*.

Matsumoto, Y., Tosha, T., Pisliakov, A.V., Hino, T., Sugimoto, H., Nagano, S., Sugita, Y., and Shiro, Y. (2012). Crystal structure of quinol-dependent nitric oxide reductase from *Geobacillus stearothermophilus*. Nat Struct Mol Biol *19*, 238–245.

McCutcheon, J.P., and Moran, N.A. (2012). Extreme genome reduction in symbiotic bacteria. Nature Reviews Microbiology *10*, 13–26.

McCutcheon, J.P., McDonald, B.R., and Moran, N.A. (2009). Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc. Natl. Acad. Sci. U.S.A. *106*, 15394–15399.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., Desantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME Journal *6*, 610–618.

McIlroy, S.J., Saunders, A.M., Albertsen, M., Nierychlo, M., McIlroy, B., Hansen, A.A., Karst, S.M., Nielsen, J.L., and Nielsen, P.H. (2015). MiDAS: the field guide to the microbes of activated sludge. Database *2015*, bav062.

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. Proc. Natl. Acad. Sci. U.S.A. *110*, E2390–E2399.

Methe, B.A., Nelson, K.E., Eisen, J.A., Paulsen, I.T., Nelson, W., Heidelberg, J.F., Wu, D., Wu, M., Ward, N., Beanan, M.J., et al. (2003). Genome of *Geobacter sulfurreducens*: metal reduction in subsurface environments. Science *302*, 1967–1969.

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

Miranda, M., and Sauer, A. (2010). Mine the Gap: Connecting Water Risks and Disclosure in the Mining Sector. World Resources Institute 1–30.

Miranda, M., Burris, P., Bingcang, J.F., Shearman, P., Briones, J.O., La Vina, A., and Menard, S. (2003). Mining and critical ecosystems: mapping the risks. World Resources Institute 1–72.

Miranda-Trevino, J.C., Pappoe, M., Hawboldt, K., and Bottaro, C. (2013). The importance of thiosalts speciation: review of analytical methods, kinetics, and treatment. Critical Reviews in Environmental Science and Technology *43*, 2013–2070.

Moraes, B.S., Souza, T., and Foresti, E. (2012). Effect of sulfide concentration on autotrophic denitrification from nitrate and nitrite in vertical fixed-bed reactors. Process Biochemistry 1395–1401.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research *35*, W182–W185.

Müller, V., and Grüber, G. (2003). ATP synthases: structure, function and evolution of unique energy converters. Cell. Mol. Life Sci. *60*, 474–494.

Nawrocki, E.P. (2009). Structural RNA homology search and alignment using covariance models. Ph.D. Thesis, Washington University in Saint Louis, School of Medicine.

Newton, R.J., Jones, S.E., Eiler, A., McMahon, K.D., and Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. Microbiol. Mol. Biol. Rev. *75*, 14–49.

Nobu, M.K., Narihiro, T., Rinke, C., Kamagata, Y., Tringe, S.G., Woyke, T., and Liu, W.-T. (2015). Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. The ISME Journal *9*, 1710–1722.

Ogawa, T., Noguchi, K., Saito, M., Nagahata, Y., Kato, H., Ohtaki, A., Nakayama, H., Dohmae, N., Matsushita, Y., Odaka, M., et al. (2013). Carbonyl sulfide hydrolase from *Thiobacillus thioparus* strain THI115 is one of the β-carbonic anhydrase family enzymes. Journal of the American Chemical Society *135*, 3818–3825.

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 1–14.

Pace, H.C., and Brenner, C. (2001). The nitrilase superfamily: classification, structure and function. Genome Biol *2*, REVIEWS0001.

Pace, N.R. (2009). Mapping the Tree of Life: Progress and Prospects. Microbiology and Molecular Biology Reviews *73*, 565–576.

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 Research *25*, 1043–1055.

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.

Perner, M., Seifert, R., Weber, S., Koschinsky, A., Schmidt, K., Strauss, H., Peters, M., Haase, K., and Imhoff, J.F. (2007). Microbial CO_2 fixation and sulfur cycling associated with low-temperature emissions at the Lilliput hydrothermal field, southern Mid-Atlantic Ridge (9 degrees S). Environ Microbiol *9*, 1186–1201.

Perona, J.J., and Hadd, A. (2012). Structural diversity and protein engineering of the aminoacyl-tRNA synthetases. Biochemistry *51*, 8705–8729.

Peura, S., Eiler, A., Bertilsson, S., Nykänen, H., Tiirola, M., and Jones, R.I. (2012). Distinct and diverse anaerobic bacterial communities in boreal lakes dominated by candidate division OD1. The ISME Journal *6*, 1640–1652.

Pierce, E., Xie, G., Barabote, R.D., Saunders, E., Han, C.S., Detter, J.C., Richardson, P., Brettin, T.S., Das, A., Ljungdahl, L.G., et al. (2008). The complete genome sequence of *Moorella thermoacetica* (f. *Clostridium thermoaceticum*). Environ Microbiol *10*, 2550–2573.

du Plessis, C.A., Barnard, P., Muhlbauer, R.M., and Naldrett, K. (2001). Empirical model for the autotrophic biodegradation of thiocyanate in an activated sludge reactor. Lett. Appl. Microbiol. *32*, 103–107.

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. Applied and Environmental Microbiology *73*, 3205–3214.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular Biology and Evolution *26*, 1641–1650.

Probst, A.J., Castelle, C.J., Singh, A., Brown, C.T., Anantharaman, K., Sharon, I., Hug, L.A., Burstein, D., Emerson, J.B., Thomas, B.C., et al. (2016). Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO_2 concentrations. Environ Microbiol.

Proft, T., and Baker, E.N. (2009). Pili in Gram-negative and Gram-positive bacteria - structure,

assembly and their role in disease. Cell. Mol. Life Sci. *66*, 613–635.

Pruesse, E., Peplies, J., and Glöckner, F.O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics *28*, 1823–1829.

Quan, Z.X., Rhee, S.K., Bae, J.W., Baek, J.H., Park, Y.H., and Lee, S.T. (2006). Bacterial community structure in activated sludge reactors treating free or metal-complexed cyanides. Journal of Microbiology and Biotechnology *16*, 232–239.

Rappe, M.S., and Giovannoni, S.J. (2003). The uncultured microbial majority. Annu. Rev. Microbiol. *57*, 369–394.

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*.

Reis, M.P., Dias, M.F., Costa, P.S., Ávila, M.P., Leite, L.R., de Araújo, F.M.G., Salim, A.C.M., Bucciarelli-Rodriguez, M., Oliveira, G., Chartone-Souza, E., et al. (2016). Metagenomic signatures of a tropical mining-impacted stream reveal complex microbial and metabolic networks. Chemosphere *161*, 266–273.

Rethmeier, J., Rabenstein, A., Langer, M., and Fischer, U. (1997). Detection of traces of oxidized and reduced sulfur compounds in small samples by combination of different high-performance liquid chromatography methods. Journal of Chromatography A *760*, 295–302.

Richter, M., and Rossello-Mora, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U.S.A. *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.

Roume, H., Heintz-Buschart, A., Muller, E.E.L., May, P., Satagopam, V.P., Laczny, C.C., Narayanasamy, S., Lebrun, L.A., Hoopmann, M.R., Schupp, J.M., et al. (2015). Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. Npj Biofilms and Microbiomes *1*, 15007.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: mining viral signal from microbial genomic data. PeerJ *3*, e985.

Sato, T., Atomi, H., and Imanaka, T. (2007). Archaeal type III RuBisCOs function in a pathway for AMP metabolism. Science *315*, 1003–1006.

Schmidtke, A., Bell, E.M., and Weithoff, G. (2006). Potential grazing impact of the mixotrophic flagellate *Ochromonas* sp. (Chrysophyceae) on bacteria in an extremely acidic lake. Journal of Plankton Research *28*, 991–1001.

Scow, K.M., and Hicks, K.A. (2005). Natural attenuation and enhanced bioremediation of

organic contaminants in groundwater. Current Opinion in Biotechnology *16*, 246–253.

Sekiguchi, Y., Ohashi, A., Parks, D.H., Yamauchi, T., Tyson, G.W., and Hugenholtz, P. (2015). First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking. PeerJ *3*, e740.

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 Research *25*, 534–543.

Sharon, I., Morowitz, M.J., Thomas, B.C., Costello, E.K., Relman, D.A., and Banfield, J.F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Research *23*, 111–120.

Sharpton, T., Jospin, G., Wu, D., Langille, M., Pollard, K., and Eisen, J. (2012). Sifting through genomes with iterative-sequence clustering produces a large, phylogenetically diverse protein-family resource. BMC Bioinformatics *13*, 264.

Sorokin, D.Y., Tourova, T.P., Lysenko, A.M., and Kuenen, J.G. (2001). Microbial thiocyanate utilization under highly alkaline conditions. Applied and Environmental Microbiology *67*, 528–538.

Sorokin, D.Y., Muntyan, M.S., Panteleeva, A.N., and Muyzer, G. (2012). *Thioalkalivibrio sulfidiphilus* sp. nov., a haloalkaliphilic, sulfur-oxidizing gammaproteobacterium from alkaline habitats. International Journal of Systematic and Evolutionary Microbiology *62*, 1884–1889.

Sorokin, D.Y., Tourova, T.P., Bezsoudnova, E.Y., Pol, A., and Muyzer, G. (2007). Denitrification in a binary culture and thiocyanate metabolism in *Thiohalophilus thiocyanoxidans* gen. nov. sp. nov. - a moderately halophilic chemolithoautotrophic sulfur-oxidizing Gammaproteobacterium from hypersaline lakes. Arch Microbiol *187*, 441–450.

Sorokin, D.Y., Tourova, T.P., Antipov, A.N., Muyzer, G., and Kuenen, J.G. (2004). Anaerobic growth of the haloalkaliphilic denitrifying sulfur-oxidizing bacterium *Thialkalivibrio thiocyanodenitrificans* sp. nov. with thiocyanate. Microbiology (Reading, Engl.) *150*, 2435–2442.

Speth, D.R., In 't Zandt, M.H., Guerrero-Cruz, S., Dutilh, B.E., and Jetten, M.S.M. (2016). Genome-based microbial ecology of anammox granules in a full-scale wastewater treatment system. Nature Communications *7*, 11172.

Speyer, M.R., and Raymond, P. (1988). The acute toxicity of thiocyanate and cyanate to rainbow trout as modified by water temperature and pH. Environmental Toxicology and Chemistry *7*, 565–571.

Stafford, D.A., and Callely, A.G. (1969). The utilization of thiocyanate by a heterotrophic bacterium. J. Gen. Microbiol. *55*, 285–289.

Staley, J.T., Bryant, M.P., Pfennig, N., and Holt, J.G. (1989). Acidithiobacillus. In Bergey's

Manual of Systematic Bacteriology, W. Williams, ed. (Baltimore), pp. 1842–1858.

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

Stamps, B.W., Losey, N.A., Lawson, P.A., and Stevenson, B.S. (2014). Genome Sequence of *Thermoanaerobaculum aquaticum* MP-01T, the First Cultivated Member of Acidobacteria Subdivision 23, Isolated from a Hot Spring. Genome Announc *2*.

Stott, M.B., Franzmann, P.D., Zappia, L.R., Watling, H.R., Quan, L.P., Clark, B.J., Houchin, M.R., Miller, P.C., and Williams, T.L. (2001). Thiocyanate removal from saline CIP process water by a rotating biological contactor, with reuse of the water for bioleaching. Hydrometallurgy *62*, 93–105.

Stratford, J., Dias, A.E., and Knowles, C.J. (1994). The utilization of thiocyanate as a nitrogen source by a heterotrophic bacterium: the degradative pathway involves formation of ammonia and tetrathionate. Microbiology (Reading, Engl.) *140*, 2657–2662.

Sung, Y.C., and Fuchs, J.A. (1988). Characterization of the cyn operon in *Escherichia coli* K12. Journal of Biological Chemistry.

Sung, Y.C., and Fuchs, J.A. (1992). The *Escherichia coli* K-12 cyn operon is positively regulated by a member of the lysR family. Journal of Bacteriology *174*, 3645–3650.

Sutcliffe, I.C. (2010). A phylum level perspective on bacterial cell envelope architecture. Trends Microbiol. *18*, 464–470.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., The UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926–932.

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.

Taubert, M., Vogt, C., Wubet, T., Kleinsteuber, S., Tarkka, M.T., Harms, H., Buscot, F., Richnow, H.-H., Bergen, von, M., and Seifert, J. (2012). Protein-SIP enables time-resolved analysis of the carbon flux in a sulfate-reducing, benzene-degrading microbial consortium. The ISME Journal *6*, 2291–2301.

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

Thompson, A.W., Foster, R.A., Krupke, A., Carter, B.J., Musat, N., Vaulot, D., Kuypers, M.M.M., and Zehr, J.P. (2012). Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. Science *337*, 1546–1550.

Thuku, R.N., Brady, D., Benedik, M.J., and Sewell, B.T. (2009). Microbial nitrilases: versatile, spiral forming, industrial enzymes. J. Appl. Microbiol. *106*, 703–727.

Tripp, H.J., Bench, S.R., Turk, K.A., Foster, R.A., Desany, B.A., Niazi, F., Affourtit, J.P., and Zehr, J.P. (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. Nature *464*, 90–94.

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.

Ullrich, S.R., Poehlein, A., Tischler, J.S., González, C., Ossandon, F.J., Daniel, R., Holmes, D.S., Schlömann, M., and Mühling, M. (2016). Genome analysis of the biotechnologically relevant acidophilic iron oxidising strain JA12 indicates phylogenetic and metabolic diversity within the novel genus " *Ferrovum* ." PLoS ONE *11*, e0146832.

Ullrich, S.R., Poehlein, A., Voget, S., Hoppert, M., Daniel, R., Leimbach, A., Tischler, J.S., Schlömann, M., and Mühling, M. (2015). Permanent draft genome sequence of *Acidiphilium* sp. JA12-A1. Stand Genomic Sci *10*, 56.

Ultsch, A., and Mörchen, F. (2005). ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. Data Bionics Research Group, University of Marburg, Germany.

US Dept of Energy (2015). Fact Sheet: UMTRCA Title I, Rifle, Colorado, Processing Sites and Disposal Site.

van Buuren, C., Makhotla, N., and Olivier, J.W. (2011). The ASTER process: technology development through to piloting, demostration, and comercialization. Proceedings of the ALTA 2011 Nickel-Cobalt-Copper, Uranium and Gold Conference.

van Hille, R.P., Dawson, E., Edward, C., and Harrison, S.T.L. (2015). Effect of thiocyanate on BIOX (R) organisms: Inhibition and adaptation. Minerals Engineering *75*, 110–115.

van Zyl, A.W., Harrison, S.T.L., and van Hille, R.P. (2011). Biodegradation of thiocyanate by a mixed microbial population. Proceedings of the International Mine Water Association Conference, 2011 119–124.

van Zyl, A.W., Huddy, R., Harrison, S.T.L., and van Hille, R.P. (2015). Evaluation of the ASTER^{TM} process in the presence of suspended solids. Minerals Engineering *76*, 72–80.

Veith, A., Botelho, H.M., Kindinger, F., Gomes, C.M., and Kletzin, A. (2012). The sulfur oxygenase reductase from the mesophilic bacterium *Halothiobacillus neapolitanus* is a highly active thermozyme. Journal of Bacteriology *194*, 677–685.

Viollier, E., Inglett, P.W., Hunter, K., Roychoudhury, A.N., and Van Cappellen, P. (2000). The ferrozine method revisited: Fe(II)/Fe(III) determination in natural waters. Applied Geochemistry

15, 785–790.

Vu, H.P., Mu, A., and Moreau, J.W. (2013). Biodegradation of thiocyanate by a novel strain of *Burkholderia phytofirmans* from soil contaminated by gold mine tailings. Lett. Appl. Microbiol. *57*, 368–372.

Wakelin, S.A., Anand, R.R., Reith, F., Gregg, A.L., Noble, R.R.P., Goldfarb, K.C., Andersen, G.L., Desantis, T.Z., Piceno, Y.M., and Brodie, E.L. (2012). Bacterial communities associated with a mineral weathering profile at a sulphidic mine tailings dump in arid Western Australia. FEMS Microbiol Ecol *79*, 298–311.

Walsh, M.A., Otwinowski, Z., Perrakis, A., Anderson, P.M., and Joachimiak, A. (2000). Structure of cyanase reveals that a novel dimeric and decameric arrangement of subunits is required for formation of the enzyme active site. Structure/Folding and Design *8*, 505–514.

Wang, P., and VanEtten, H.D. (1992). Cloning and properties of a cyanide hydratase gene from the phytopathogenic fungus *Gloeocercospora sorghi*. Biochemical and Biophysical Research Communications *187*, 1048–1054.

Warren, L.A., Colenbrander Nelson, T.E., Bennett, D., Marshall, S., and Apte, S. (In prep.). Sulfur geochemistry of mining wastewaters.

Wasserlauf, M., and Dutrizac, J.E. (1984). CANMET's project on the chemistry, generation and treatment of thiosalts in milling effluents. Canadian Metallurgical Quarterly.

Watson, S.J., and Maly, E.J. (1987). Thiocyanate toxicity to *Daphnia magna*: modified by pH and temperature. Aquatic Toxicology *10*, 1–8.

Wetmore, K.M., Price, M.N., Waters, R.J., Lamson, J.S., He, J., Hoover, C.A., Blow, M.J., Bristow, J., Butland, G., Arkin, A.P., et al. (2015). Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. mBio *6*, e00306–e00315.

White, D. (2000). The Physiology and Biochemistry of Prokaryotes (Oxford: Oxford Univ Press).

Whitlock, J.L. (1990). Biological detoxification of precious metal processing wastewaters. Geomicrobiology Journal *8*, 241–249.

Wielinga, B., Lucy, J., Moore, J., Seastone, O., and Gannon, J. (1999). Microbiological and geochemical characterization of fluvially deposited sulfidic mine tailings. Applied and Environmental Microbiology *65*, 1548–1555.

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., and Kerkhof, L.J. (2011). Acetate availability and its influence on sustainable bioremediation of uranium-contaminated groundwater. Geomicrobiology Journal *28*, 519–539.

Wood, A.P., Kelly, D.P., McDonald, I.R., Jordan, S.L., Morgan, T.D., Khan, S., Murrell, J.C., and Borodina, E. (1998). A novel pink-pigmented facultative methylotroph, *Methylobacterium*

thiocyanatum sp. nov., capable of growth on thiocyanate or cyanate as sole nitrogen sources. Arch Microbiol *169*, 148–158.

Woyke, T., Teeling, H., Ivanova, N.N., Huntemann, M., Richter, M., Gloeckner, F.O., Boffelli, D., Anderson, I.J., Barry, K.W., Shapiro, H.J., et al. (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. Nature *443*, 950–955.

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. The ISME Journal *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, J.-H., Wu, F.-Y., Chuang, H.-P., Chen, W.-Y., Huang, H.-J., Chen, S.-H., and Liu, W.-T. (2012). Community and proteomic analysis of methanogenic consortia degrading terephthalate. Applied and Environmental Microbiology *79*, 105–112.

Xiong, W., Giannone, R.J., Morowitz, M.J., Banfield, J.F., and Hettich, R.L. (2015). Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut. J. Proteome Res. *14*, 133–141.

Yamanaka, Y., Arakawa, T., Watanabe, T., Namima, S., Sato, M., Hori, S., Ohtaki, A., Noguchi, K., Katayama, Y., Yohda, M., et al. (2013). Two arginine residues in the substrate pocket predominantly control the substrate selectivity of thiocyanate hydrolase. Journal of Bioscience and Bioengineering *116*, 22–27.

Youssef, N.H., Blainey, P.C., Quake, S.R., and Elshahed, M.S. (2011). Partial genome assembly for a candidate division OP11 single cell from an anoxic spring (Zodletone Spring, Oklahoma). Applied and Environmental Microbiology *77*, 7804–7814.

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J., et al. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics *26*, 1608–1615.

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