Thiocyanate and Organic Carbon Inputs Drive Convergent Selection for Specific Autotrophic *Afipia* and *Thiobacillus* Strains Within Complex Microbiomes

Robert J. Huddy[1,2], Rohan Sachdeva[3], Fadzai Kadzinga[1,2], Rose S. Kantor[4], Susan T. L. Harrison[1,2] and Jillian F. Banfield[3,4,5,6]*

[1] Centre for Bioprocess Engineering Research, University of Cape Town, Cape Town, South Africa, [2] Future Water Institute, University of Cape Town, Cape Town, South Africa, [3] Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA, United States, [4] Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, United States, [5] Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, United States, [6] School of Earth Sciences, University of Melbourne, Melbourne, VIC, Australia

Thiocyanate (SCN^-) contamination threatens aquatic ecosystems and pollutes vital freshwater supplies. SCN^--degrading microbial consortia are commercially adapted for remediation, but the impact of organic amendments on selection within SCN^--degrading microbial communities has not been investigated. Here, we tested whether specific strains capable of degrading SCN^- could be reproducibly selected for based on SCN^- loading and the presence or absence of added organic carbon. Complex microbial communities derived from those used to treat SCN^--contaminated water were exposed to systematically increased input SCN concentrations in molasses-amended and -unamended reactors and in reactors switched to unamended conditions after establishing the active SCN^--degrading consortium. Five experiments were conducted over 790 days, and genome-resolved metagenomics was used to resolve community composition at the strain level. A single *Thiobacillus* strain proliferated in all reactors at high loadings. Despite the presence of many *Rhizobiales* strains, a single *Afipia* variant dominated the molasses-free reactor at moderately high loadings. This strain is predicted to break down SCN^- using a novel thiocyanate desulfurase, oxidize resulting reduced sulfur, degrade product cyanate to ammonia and CO_2 via cyanate hydratase, and fix CO_2 via the Calvin–Benson–Bassham cycle. Removal of molasses from input feed solutions reproducibly led to dominance of this strain. Although sustained by autotrophy, reactors without molasses did not stably degrade SCN^- at high loading rates, perhaps due to loss of biofilm-associated niche diversity. Overall, convergence in environmental conditions led to convergence in the strain composition, although reactor history also impacted the trajectory of community compositional change.

Keywords: thiocyanate, bioremediation, microbiome, metagenomics, pollution

INTRODUCTION

Thiocyanate (SCN$^-$) is an unwanted toxic by-product found in a range of wastewater streams, including those from gold mines and coal coke processing (Hung and Pavlostathis, 1999). Contamination of freshwater by industrial processes is a widespread global phenomenon (Sharma and Bhattacharya, 2017). The requirement for treatment strategies is pressing, especially in countries such as South Africa and Australia, where gold recovery is an important economic activity and freshwater supplies can be very limited. There has been some progress with remediation of SCN$^-$-containing effluents prior to discharge using mixed microbial consortia, but the systems are essentially treated as black boxes (van Buuren et al., 2011; Douglas Gould et al., 2012).

Some field reactors are inoculated with active laboratory-maintained consortia and continuously cultured with input of organic carbon, often in the form of molasses, and phosphate (van Buuren et al., 2011; Huddy et al., 2015). Molasses is routinely supplemented into commercial Activated Sludge Tailings Effluent Remediation (ASTER^{TM}, Outotec, South Africa) systems at a number of mine sites around the world. The microbial community used in this study was originally obtained from a commercial ASTER plant operated in South Africa and is maintained in the laboratory under similar conditions, including molasses supplementation, to that in the field. Previously, we characterized microbial communities derived from laboratory-maintained ASTER^{TM} systems using a shotgun sequencing-based method that enabled recovery of genomes from many consortia members (Kantor et al., 2015; Rahman et al., 2017). Different strain combinations proliferated under distinct operating conditions in these molasses-amended reactors, and the communities were very complex, with a reservoir of over 150 genotypes. Thiobacilli were the dominant community members at high SCN$^-$ loadings and were inferred to be critical to SCN$^-$ breakdown. Notably, the abundant thiobacilli were predicted to be autotrophs, bringing into question the need for added organic carbon compounds. A subsequent study showed that distinct consortia could remove SCN$^-$ from mining-contaminated solutions in the absence of added organic substrate and attributed SCN$^-$ breakdown to three thiobacilli closely related to *Thiobacillus thioparus* strain THI 111 (Watts et al., 2017). In another recent study, laboratory bioreactors were inoculated from mine tailings and the autotrophically supported consortia found to break down SCN$^-$ (Watts et al., 2019). The currently known mechanisms for SCN$^-$ degradation have been reviewed by Watts and Moreau (2016).

Here, we describe a long-term study in which five reactors were inoculated from the same reactor as used by Kantor et al. (2015) and experiments conducted to test the hypothesis that the community from the prior study could degrade SCN$^-$ without the addition of organic carbon, specifically molasses, and that autotrophic thiobacilli would dominate the reactors at high SCN$^-$ loadings. The reactors were driven to the same conditions via different experimental parameters to test whether historical contingency modulates the control of environmental conditions on community composition. We applied time-series genome-resolved metagenomics methods to investigate strain overlap (at the genotype level) among reactors that experienced a range of SCN$^-$ loadings with or without molasses. Switching reactors from heterotrophic to mixotrophic conditions consistently selected for the same SCN$^-$-degrading strain from a background of substantial strain diversity. Insights from the genome of this strain include its pathways for degradation of SCN$^-$ and breakdown products. Extensive genomic resolution of all communities enabled us to establish community convergence controlled by convergence of inputs and elucidated roles performed by community members in this laboratory model system.

MATERIALS AND METHODS

Experimental Design

Five reactors (R1–R5) were inoculated from the same mixed microbial community and were operated continuously for 790 days. Residence times of 24 h were maintained early in the experiments but were later lowered to 12 h (**Table 1**). Each reactor experienced a different combination of SCN$^-$ loading and molasses amendment, with consistent phosphate amendment over the full-time course. The experimental design is shown in **Table 1** and **Supplementary Figure 1** (also see Methods).

Reactor Setup, Inoculation, and Operation

Five 1-L water-jacketed continuous stirred tank reactors (CSTRs; Glasschem, South Africa) were inoculated with biofilm and planktonic samples harvested from a long-running SCN$^-$-degrading stock reactor at the University of Cape Town. The biofilm and planktonic samples were gently homogenized before being equally split (approximately 50 mL) across the five reactors (R1–R5) into 1,000 mL of sterile reactor media (as detailed below). These reactors, R1–R5, were stirred with a pitched-blade impeller at 270 rpm and sparged with filtered air at 900 mL/min. Four reactors (R1–R4) were initially fed 150 mg/L molasses and 0.28 mM KH$_2$PO$_4$ to provide supplemental nutrients, while the fifth reactor (R5) was only fed deionized water containing 0.28 mM KH$_2$PO$_4$. Trace element solution (1 mL; Kolmert and Johnson, 2001) was added dropwise directly to the reactors at least once every 3 weeks throughout the extended experimental run. All the reactors were further supplemented with KSCN as detailed below. The respective reactor feeds also contained increasing amounts of KOH to modulate the reactor pH as necessary, and small amounts of 5 N KOH were added dropwise directly to reactors if the observed pH was ≤6.5.

The reactors were run in batch-fed mode for 4 days, before all the reactors were switched to continuous feeding at a residence time of 48 h. The hydraulic retention time (HRT) of the reactors was sequentially lowered from 48 to 12 h, over a period of 131 days, and then maintained at 12 h, while the feed concentration of SCN$^-$ was increased stepwise over the remainder of the experimental period. The reactors were allowed to stabilize between each increase in the loading of SCN$^-$ to reach steady state. The molasses was dropped from the feed to reactors 3

TABLE 1 | Summary of experimental conditions.

SCN feed concentration (HRT)	SCN loading rate (mM/hr)	RI	R2	R3	R4	R5
50 mg/L (24 h)	0.036	M	M	M	M	W
250 mg/L (12 h)	0.36	M	M↓	M	M	W
250 mg/L (12 h)	0.36	M		W↓	W	W
500 mg/L (12 h)	0.72	M			W	W
750 mg/L (12 h)	1.08	M			W	W
750 mg/L (12 h)	1.08	M			W	W
1500 mg/L (12 h)	2.15	M			W	W
2000 mg/L (12 h)	2.87	M			W	W

The table indicates continuously supplied thiocyanate feed concentrations, hydraulic retention times (HRT), and loading rates in molasses (M) or water-based (W) reactor systems (R1–R5). The arrows in the Table indicate the points from which the experimental conditions in the reactors were held at a SCN⁻ feed concentration of 250 mg/L (0.36 mM/h).

and 4 (R3 and R4) at Day 239 for the duration of the experiment. The SCN⁻ feed concentration to R2 and R3 was maintained at 250 mg/L, from day 216, for the remainder of the experiment.

Samples of planktonic cells and biofilm attached to the walls of the bioreactor were initially sampled at a 24-h HRT and SCN⁻ feed concentration of 50 mg/L (SCN⁻ loading rate: 0.04 mM/h). The remaining samples were taken at a 12-h HRT and SCN⁻ feed concentrations of 50 to 2,000 mg/L (SCN⁻ loading rates of 0.36 to 2.87 mM/h).

Sampling

Samples of biomass from each reactor were taken for metagenomic sequencing just before increases in feed concentration. 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 approximately 300 mL of the liquid phase from each reactor through sterile Whatman paper, to remove small flocs and clumps of biofilm, before the planktonic cells were recovered onto sterile 0.2-μm filters. The initial filtration step, through Whatman paper, was performed to remove clumps of biofilm that had been dislodged during the opening of the reactors as well as flocs of biofilm found within these systems, to recover free planktonic cells. Planktonic biomass was gently washed off using the 0.22-μm filters with sterile water, harvested by centrifugation (13,000 rpm for 10 min at room temperature), and stored at −60°C until further analysis. Filtered media were returned to the reactor to maintain chemical continuity.

Analysis of Solution Chemistry

Bulk liquid was sampled daily for chemical analysis, filtered through a sterile 0.22-μm syringe filter, pH analyzed, and frozen at -20°C until further analysis. The residual SCN- concentration was measured using high-performance liquid chromatography (HPLC) as previously described (Kantor et al., 2015). Ion chromatography (IC) was performed to quantify ammonia, nitrate, and sulfate on a Dionex ICS1600 system (Thermo Scientific). Ammonia was quantified using a Dionex IonPac

CS12A column, while sulfate and nitrate were quantified using a Dionex IonPac AS16 column.

DNA Extraction and Sequencing

DNA was extracted from frozen biomass using a NucleoSpin soil genomic DNA extraction kit (Macherey-Nagel, Germany) with the inclusion of a repeated extraction step, according to the manufacturer's instructions.

Metagenomic Assembly, Binning, and Annotation

The metagenomic data processing of paired 150-bp read datasets followed methods reported previously (Kantor et al., 2017) up to the genome binning step. To generate genome bins, differential abundance information was calculated for the scaffolds of each sample using the abundance of scaffolds across all samples. The reads of each sample were mapped against the scaffolds of each sample using BBMap[1] at 98% average nucleotide identity (ANI). The series coverage information was used to create bins using ABAWACA (Brown et al., 2015), CONCOCT (Alneberg et al., 2014), MaxBin2 (Wu et al., 2016), and MetaBAT2 (Kang et al., 2019). Bins were also manually generated by leveraging information derived from matching against a dereplicated set of genomes from prior research SCN⁻-degrading communities (Kantor et al., 2015; Rahman et al., 2017) and UniProt (UniProt Consortium, 2015). The resulting taxonomic profile and other "manual binning" signals (GC content, coverage, and single-copy gene inventories) were integrated for human-guided binning. A set of optimal bins for each sample was created from the combined automated and manual bins using DASTool (Sieber et al., 2018). All of the DASTool bins were dereplicated using dRep (Olm et al., 2017) to create a near-final draft genome set. CheckM (Parks et al., 2015) was used to assess genome completeness and redundancy. The resulting dereplicated bins were further manually refined using the aforementioned "manual binning" signals. The manually refined bins were dereplicated with dRep at 99% ANI to create a final dereplicated set of bins for all subsequent analyses. Reads from each sample were mapped to the dereplicated set of genomes allowing for 1 mismatch using BBMap. Relative abundances for each genome were calculated by normalizing the coverage of each genome by the total coverage in each sample and then multiplying this fraction by the percentage of reads mapped in each sample.

The genome of *Afipia_64_1782* was selected for manual curation (Chen et al., 2020). This involved mapping of reads to the *de novo* assembled sequences, identification (based on regions of no perfect read support), and resolution of local mis-assemblies, gap filling, contig extension, and joining based on perfect overlap of a region spanned by paired reads. Curation was terminated when sequences ended in repeats that could not be uniquely spanned using short Illumina read data.

Functional predictions for all genes in the final dereplicated reference genome set were established as previously described (Kantor et al., 2017) but with additional constraints from a KEGG (Kanehisa and Goto, 2000)-based hidden Markov model

---

[1]https://sourceforge.net/projects/bbmap/

(HMM) analysis (Diamond et al., 2019). For specific genes of interest, analysis included prediction of signal peptides, domain architecture [e.g., by Interpro (Mitchell et al., 2015)]. For thiocyanate desulfurase, the best model in the Protein Data Bank (PDB) (Berman et al., 2000) was selected as the best choice by I-TASSER (Yang et al., 2015). The comparison was visualized by UCSF Chimera (Pettersen et al., 2004).

Principal Coordinate Analysis

Heatmaps and PCA plots were generated using ClustVis (Metsalu and Vilo, 2015). Original values are $\ln(x + 1)$-transformed. No scaling is applied to rows; SVD with imputation is used to calculate the principal components. X and Y axes show principal component 1 (PC1) and principal component 2 (PC2) that explain the percentage (%) of the total variance, respectively.
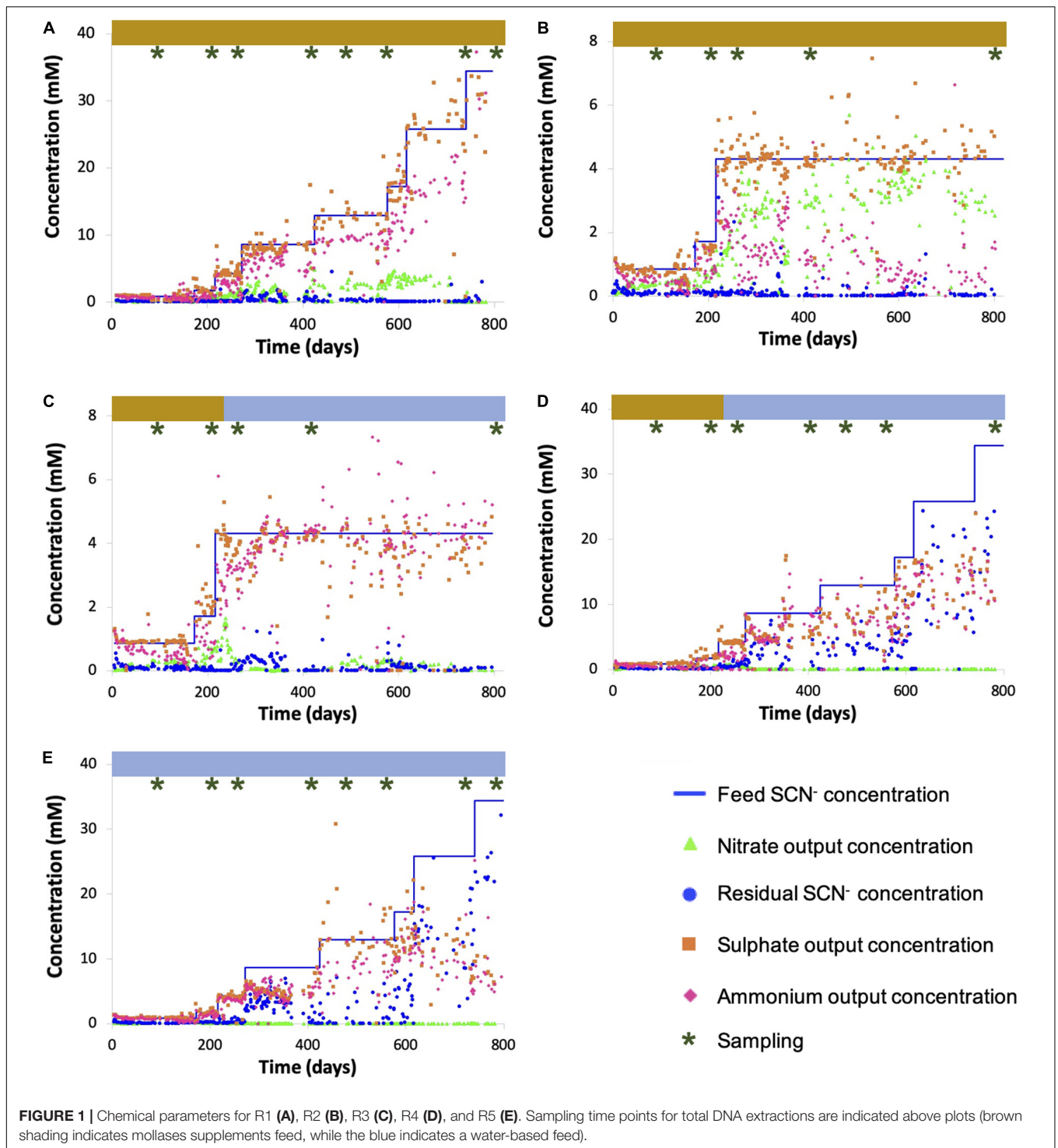
RESULTS AND DISCUSSION

Reactor Chemistry

The SCN^- degradation rate consistently matched the SCN^- loading rate over the experimental period in three of the five reactors, demonstrated by negligible residual SCN^-: R1 (which was subjected to an increasing SCN^- input concentration in the presence of molasses via the continuous feed, up to a maximum input SCN^- concentration of 2,000 ppm), R2 (which was held at an input concentration of 250 ppm SCN^- with molasses), and R3 (which was held at an input concentration of 250 ppm SCN^- before and at all time points after molasses was removed from the input solution) (**Figure 1**). The sulfate concentrations in R1, R2, and R3 closely matched those predicted based upon reaction stoichiometry, given complete oxidation of all sulfane derived from SCN^- (**Supplementary Table 1**). The concentrations of nitrogen compounds in the form of ammonia plus nitrate in R1, R2, and R3 also closely matched those predicted based on reaction stoichiometry, except in R1 after the SCN^- input concentration reached 1,500 ppm. Notably, molasses-amended consortia outperformed consortia studied previously, which lost effectiveness at lower concentrations (Kantor et al., 2017). As the experiments progressed, biofilm developed on all reactor internal surfaces. The biofilm was thick and dark brown in reactors receiving molasses and thin and creamy-white in reactors without molasses. R4 performed similarly to R1, R2, and R3 while being maintained on molasses up to an SCN^- feed concentration of 250 ppm (**Figure 1D**). After this point, molasses was removed from the input solution and the input SCN^- concentration was ramped from 250 to 2,000 ppm. The SCN^- degradation rate declined and became erratic during the ramp. The toxicity of undegraded SCN^- in reactors at higher loadings could have impacted community composition. Although the sulfate and total nitrogen compound concentrations in R4 closely matched those predicted based upon reaction stoichiometry for complete breakdown of SCN^- early in the experiment, nitrate accumulation decreased at higher SCN^- input concentrations. This finding suggests a decline in abundance or activity of ammonia-oxidizing bacteria following accumulation of residual SCN^- in solution.

In the no-molasses reactor, R5, SCN^- was degraded completely up until the input concentration reached 250 ppm, consistent with findings from R3 and R4. Subsequently, the SCN^- degradation rate did not increase in direct proportion with increased loading rate, leading to incomplete SCN^- degradation. This was particularly true at high loading rates. Further, residual SCN^- concentrations fluctuated. While the ammonia production matched that predicted given the observed extent of SCN^- degradation, little or no nitrate was produced. Again, this suggests a decline in abundance or activity of ammonia-oxidizing bacteria at elevated SCN^- concentrations in the absence of organic feed.

Genomic Description of Reactor Communities

Genome binning of metagenomic data from 55 samples taken from the five reactors (see Methods) generated 1,509 draft genomes, of which 956 were classified as near-complete (completion $\geq 90\%$ and $\leq 5\%$ contamination/redundancy) based on the predicted inventory of single-copy genes from CheckM. Given that metagenomic data were derived from samples collected from a series of related reactors at different time points, the genome dataset was dereplicated (see Methods), generating 232 genomes representative of dereplicated genome clusters, all of which correspond to bacteria. Of the dereplicated genomes, 168 were classified as near-complete. Between ~70 and ~80% of reads from each of the 55 metagenomic samples could be mapped back to the 232 genomes with high confidence (Methods and **Supplementary Table 2**). Further, based on inventories of assembled single-copy genes assigned to genomes, the abundant bacteria in all assemblies were represented by bins. Thus, we conclude that unmapped reads derived from low-abundance organisms (some of which may be strain variants of more abundant organisms). The abundances of each of the 232 bacteria across the 55 samples are provided in **Supplementary Table 3**.

We identified the abundant bacteria in R1 (ramped molasses reactor) and R5 (ramped water reactor) and calculated the abundances of these bacteria in all reactors (**Figure 2A** and **Supplementary Table 4**). We then predicted their relevant metabolic capacities (**Figure 2B**). Notably, despite a wide variety of Rhizobiales strains, one specific genotype was highly abundant or dominated reactors maintained without molasses input, especially at intermediate to high SCN^- input concentrations (**Figures 2**, **3**). This bacterium was not prominent in any molasses reactor from the current or prior studies, although it is closely related to the previously detected strain referred to as SCNPILOT_CONT_500_BF_Rhizobiales_64_17 (Kantor et al., 2017). Based on analysis of the 16S rRNA gene, this bacterium is related to organisms previously classified as either *Afipia* or *Bradyrhizobium*. We abbreviated the genome name as *Afipia*_64_1782. Given its high abundance in the community, we manually curated the genome representative of the dereplicated cluster from sample SCN18_10_11_15_R3_B, correcting many local assembly errors and filling scaffolding gaps using read sequences, extending and joining contigs (Methods).

FIGURE 1 | Chemical parameters for R1 (A), R2 (B), R3 (C), R4 (D), and R5 (E). Sampling time points for total DNA extractions are indicated above plots (brown shading indicates mollases supplements feed, while the blue indicates a water-based feed).

The essentially complete 3.549-Mbp genome has all expected single-copy genes present in single copy and is in just 8 fragments (most fragments are terminated in unresolvable repeats).

Based on the prevalence of abundant bacteria in the five reactors, the molasses-free reactor communities cluster together and away from communities from reactors with molasses in the input solution (**Figure 3**). *Thiobacillus_3-i* and *Thiobacillus_3-iii*,

which are closely related to strains abundant previously in molasses-amended reactors (Kantor et al., 2017), were almost undetectable in initial samples, but they proliferated in later samples, especially those from reactors not receiving molasses. *Thiobacillus_63_590*, shown as *Thiobacillus*-i in **Figure 3** and **Figure 2A**, was barely detectable at the first time point but it generally increases in abundance with increasing SCN⁻ input

FIGURE 2 | (A) read mapping-based relative abundance of (high red, low yellow to green; see **Supplementary Table 4**) of the 26 most abundant organisms (rows) in R1 (molasses SCN ramp reactor, shown in shades of brown at the bottom of the table) and/or R5 (SCN ramp reactor without molasses, i.e., "water" reactor, shown in shades of blue at the bottom of the table) across all reactors. Genomes are colored by taxonomic grouping. For the full names of bacteria see **(B)** Black boxes group together samples from the same time point (biofilm and p = planktonic and f = floc sample). **(B)** Overview of functionalities encoded in the 26 genotypes representing the most abundant organisms in R1 and/or R5.

concentration and is the dominant organism in R1 and R2 at later time points.

Predicted Metabolism and Capacity for SCN⁻ Degradation

Given the proliferation of *Afipia*_64_1782 in reactors without molasses that achieved SCN⁻ degradation, we searched the genome for genes capable of degrading SCN⁻ and its breakdown products. We identified a 489-amino acid protein (SCN18_10_11_15_R3_B_scaffold_20_388) that was initially annotated as thiocyanate hydrolase, but the three subunits expected (*scnABC* of the carbonyl sulfide pathway)

were not present. Further analyses (Methods) indicated that the gene encodes thiocyanate desulfurase of the cyanate pathway (Tikhonova et al., 2020). This identification was strongly supported by selection of threading templates using I-TASSER (Yang et al., 2015) as well as by SWISS-MODEL (Waterhouse et al., 2018) (33.7% identity, top match to PDB 5F75, with four Cu ligand sites identified although the GMQE score is below expectation relative to the value expected for experimental structures of similar size) and Phyre2 (Kelley et al., 2015) (template fold library ID c5f30B, thiocyanate desulfurase from *Thioalkalivibrio paradoxus*) (**Figure 4**). The sequence is closely related to that of mesophilic alphaproteobacteria strain THI201, which was
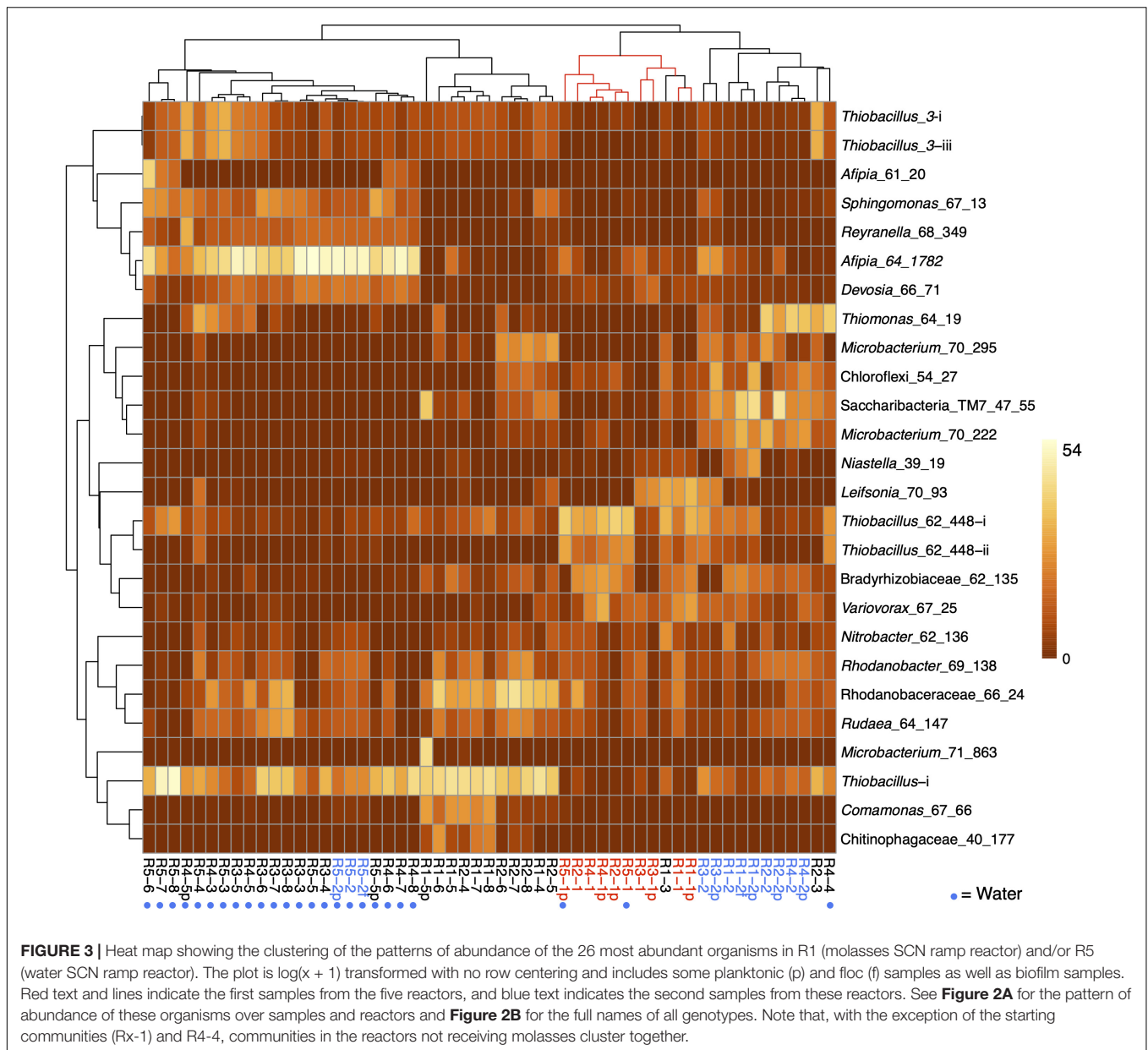
FIGURE 3 | Heat map showing the clustering of the patterns of abundance of the 26 most abundant organisms in R1 (molasses SCN ramp reactor) and/or R5 (water SCN ramp reactor). The plot is log(x + 1) transformed with no row centering and includes some planktonic (p) and floc (f) samples as well as biofilm samples. Red text and lines indicate the first samples from the five reactors, and blue text indicates the second samples from these reactors. See **Figure 2A** for the pattern of abundance of these organisms over samples and reactors and **Figure 2B** for the full names of all genotypes. Note that, with the exception of the starting communities (Rx-1) and R4-4, communities in the reactors not receiving molasses cluster together.

shown to degrade thiocyanate and is also well modeled by the PDB 5F75 structure. The genome also encodes a cyanate hydratase (SCN18_10_11_15_R3_B_scaffold_20_38), which could convert cyanate (NCO⁻) produced by thiocyanate desulfurase to carbamate that could spontaneously decompose down to ammonia and CO_2 or may be used in pyrimidine metabolism. The *Afipia*_64_1782 also has sulfide dehydrogenase, sulfide:quinone oxidoreductase, and genes of the SOX pathway to oxidize sulfur compounds released by SCN⁻ degradation (as well as genes for sulfur assimilation). Although *Afipia*_64_1782 has the genes necessary to oxidize ammonia (produced by thiocyanate desulfurase and cyanate hydratase), the genome encodes a nitrate/nitrite transporter and is capable of converting nitrate to nitrite and form nitric oxide via NirK, as well as nitric

oxide reductase and nitrous oxide reductase (thus, it is predicted to be capable of complete denitrification). Given that essentially all nitrogen predicted to be produced from SCN⁻ degradation is accounted for by ammonia in reactors without molasses (**Figure 1**), this bacterium may not be reducing nitrate or nitrite *in situ*.

The abundant bacteria in R5 are presumably autotrophs, given the absence of added organic carbon. *Afipia*_64_1782 encodes two genes annotated as ribulose-1,5-bisphosphate carboxylase (RuBisCO) (confirmed to be the small and large subunits of Form ID, based on phylogenetic analysis of the large subunit) and other genes of the Calvin–Benson–Bassham (CBB) cycle for CO_2 fixation. *Afipia*_64_1782 is predicted to be capable of aerobic growth. Its genome encodes a
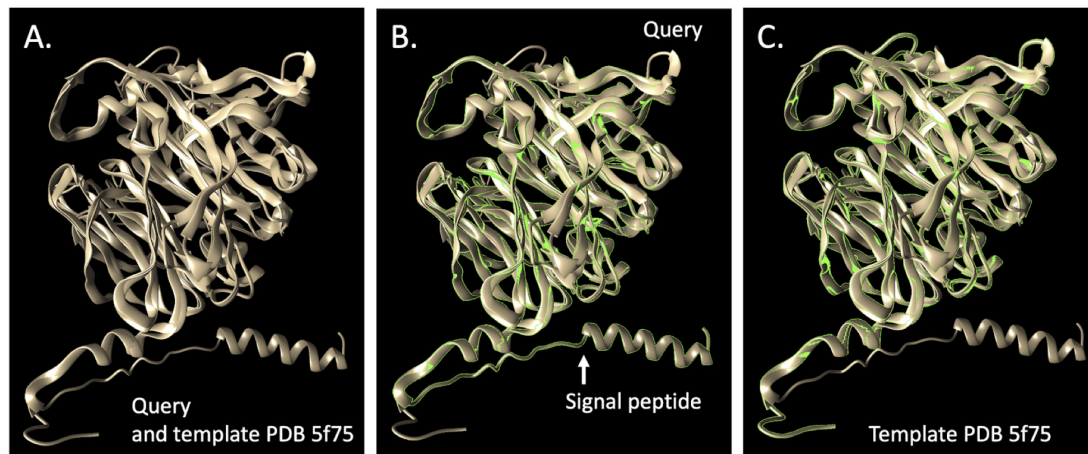
FIGURE 4 | Comparison of the putative thiocyanate desulfurase (EC 1.8.2.7) from Afipia_64_1782 and the most closely related protein, Model 5F75 (Tsallagov et al., 2016), in the PDB. **(A)** shows the query superimposed upon the template, **(B)** shows the query and **(C)** the template. The TM-score of the structural alignment between the query and 5f75 was 0.901, compared to the second choice score, PDB 2ECE (Yamada et al., 2008), a hypothetical selenium binding protein from Sulfolobus tokodaii of 0.608.

complete glycolysis/gluconeogenesis pathway, TCA cycle, and glyoxylate cycle, as well as genes of the pentose phosphate pathway, including those required to synthesize precursors for nucleic acid synthesis and for purine and pyrimidine biosynthesis. The genome also encodes genes for production of trehalose, starch/glycogen, and amino sugars including UDP-N-acetylmuramate precursor for peptidoglycan production and for peptidoglycan production. Also, identified are genes for synthesis of a Gram-negative cell envelope, acetate production via acetyl-P, and interconversion of pyruvate and lactate. Interestingly, although the genome encodes genes for the biosynthesis of folate, NAD, pantothenate, CoA, thiamine, riboflavin, and probably biotin, this bacterium appears to be unable to produce cobalamin.

As expected based on prior work (Kantor et al., 2015; Rahman et al., 2017; Watts et al., 2019), various thiobacilli were prominent in the reactors. However, _Thiobacillus_63_590_, which was highly abundant especially in the molasses reactors and in the molasses-free reactor at the highest SCN$^-$ input concentrations (_Thiobacillus_-i in **Figure 2A**), is distinct from strains/species abundant in prior studies. Similar to previously described abundant thiobacilli in SCN$^-$-degrading reactors, this bacterium is an autotroph with both Forms I and II RuBisCO. Its genome encodes the three-subunit thiocyanate hydrolase, the necessary transporters, cyanate hydratase, and carbonic anhydrase. Carbonic anhydrase can potentially be used to form bicarbonate, which is necessary for the conversion of cyanate and bicarbonate into carbamate by cyanate hydratase. It also has genes that may break down organic sulfur compounds, genes of the SOX, and the reverse DSR pathways for inorganic sulfur compound oxidation, for uptake of nitrate and nitrite, complete denitrification, and recovery of ammonia from nitriles. This bacterium has genes for biosynthesis of nucleic acids, a complete pathway for synthesis of peptidoglycan, isoprenoid precursors via the MEP pathway and the ability to store carbon compounds as starch. Like _Afipia_64_1782, it cannot synthesize cobalamin _de_

novo but it is predicted to synthesize other cofactors, including thiamine, biotin, and riboflavin.

Overall, the reactors contain numerous other bacteria that likely confer community-relevant functions. Several other bacteria abundant in the molasses-free reactor that proliferated at moderate to high SCN$^-$ input concentration also have Form I and/or Form II RuBisCOs and are predicted to fix CO_2 via the CBB cycle (**Figure 2B**). Three relatively abundant bacteria belonging to the Rhodanobacteraceae (pink names in **Figure 2B**), _Sphingomonas_ and Chloroflexi are predicted to have genomes that encode numerous enzymes involved in degradation of complex carbon compounds. Many genomes of coexisting _Mesorhizobium_ and _Bosea_, _Reyranella_, various other Alphaproteobacteria (especially Rhizobiales), as well as _Pseudonocardia_, organisms that are present at relatively low abundance in the communities, appear to have genes required to produce cobalamin. Only two bacteria, Rhizobiales_68_13 and Verrucomicrobia_61_8_61_17, are predicted to have the capacity to fix N_2. However, many bacteria are predicted to have various combinations of genes for nitrate, nitrite, nitric oxide, and nitrous oxide reduction. Similarly, many have SOX genes for sulfide/sulfur compound oxidation, but the thiobacilli appear to be the main group using the reverse dissimilatory sulfite reductase pathway. Only a few bacteria, including some Actinobacteria (various _Microbacterium_ species) and Candidatus Saccharibacteria (TM7), are predicted anaerobes. Similar abundance patterns of the TM7 and _Microbacterium ginsengisoli_ in R2 and R3 may indicate a Saccharibacteria (TM7)–actinobacterial host association similar to that reported previously for oral Saccharibacteria (He et al., 2015). The clustering in **Figure 3** supports this inference, although the pattern is not apparent in all analyses (**Supplementary Figure 2**).

As expected based on the chemical data that show complete oxidation of sulfane derived from thiocyanate degradation to sulfate, many of the more abundant non-SCN$^-$-degrading
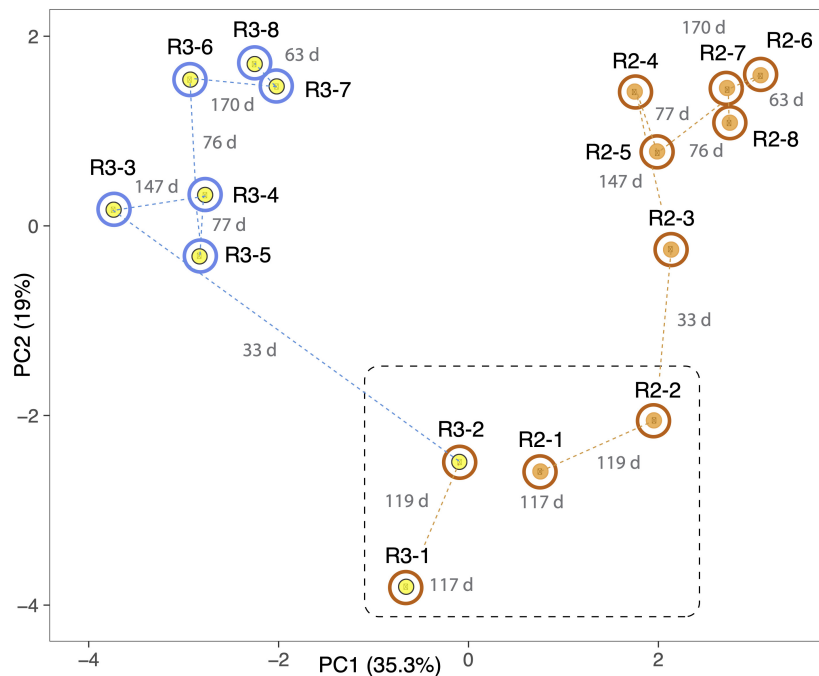
FIGURE 5 | Principal component analysis based on the relative abundances of 232 organisms (accounting for the most abundant 70–80% of the community) in R2 (orange dots) and R3 (yellow dots). Both reactors were ramped to 250 ppm SCN and held at this concentration for the duration of the experiment. Outer circles distinguish reactors receiving molasses (brown) from those receiving water (blue) at the time of sampling. The dashed black line separates samples taken prior to the switch of R3 to water. Dashed lines indicate sequential samples, and numbers in gray indicate the number of days (d) between sampling times. Note the strong and rapid deviation of community after the switch to water in R3. Samples 6, 7, and 8 in both reactors group away from 3, 4, and 5, suggesting that the community shift was incomplete at the middle time points.

bacteria in all reactors have genes for sulfur compound oxidation (**Figure 2B**). However, organisms potentially responsible for conversion of ammonia to nitrate that were detected primarily in molasses-amended reactors are less apparent. In fact, *amoA,B,C* genes involved in ammonia oxidation were not identified in any of the abundant bacteria (**Figure 2B**). Similarly, nitrifying microbes were not identified in the study system of Watts et al. (2019), despite the potential value of ammonium as an energy source. In the study of Kantor et al. (2017), it was inferred that *Nitrosospira* were responsible for ammonia oxidation in molasses-amended reactors, based on unbinned sequences. Here, we identified three draft genomes of relatively low relative abundance *Nitrosospira* strains with genes annotated as *amoA,B,C*. One genome for a strain related to *Nitrosospira multiformis* also encodes hydroxylamine oxidoreductase (HaoAB), as occurs in *Nitrosomonas europaea* (Chain et al., 2003). Interestingly, *Comamonas* and Burkholderiales genomes also have HaoAB. All HaoA-predicted proteins have the anticipated 8 CxxCH motifs. Genes for ammonia oxidation were not identified in the *Comamonas* and Burkholderiales genomes. *Comamonas* were only relatively abundant at later time points in R1 and R2, but this may be unrelated to the presence of HaoAB.

The strains of *Nitrosospira* implicated in ammonia oxidation switch in relative abundance over time. The strain represented by SCN18_31_3_15_R1_B_Nitrosospira_56_42 is most prominent

at the start of the experiments as well as at low to intermediate SCN$^-$ input concentrations, regardless of whether molasses was added. However, *N. multiformis*, which is at very low abundance in all reactors at the first time point, increases in abundance in molasses-amended reactors at longer times, especially at higher SCN$^-$ loadings. The proliferation of this strain only in molasses-amended reactors may explain the very low production of nitrate in reactors without added molasses. Overall, the findings suggest that the activity of a variety of *Nitrospira* strains, some apparently at quite low abundance, is responsible for nitrification when it occurred in the reactors. Some ammonia released by SCN$^-$ degradation is presumably assimilated for biomass production, and, as noted above, a fraction of SCN$^-$-derived N partitioned into cyanate may also be assimilated.

Community Shifts With Reactor Conditions and Time

We used principal component analysis (PCA) to investigate whole community compositions over time and as a function of reactor conditions. At the first time point, when R1–R4 had received molasses and R5 had been maintained over the startup period of 117 days without molasses at a low SCN$^-$ feed concentration of 50 ppm and a high residence time of 48–24 h, the communities generally clustered together. However, the planktonic fraction of R5 was displaced toward that of the R5 consortia sampled at time point 2 after ramping the
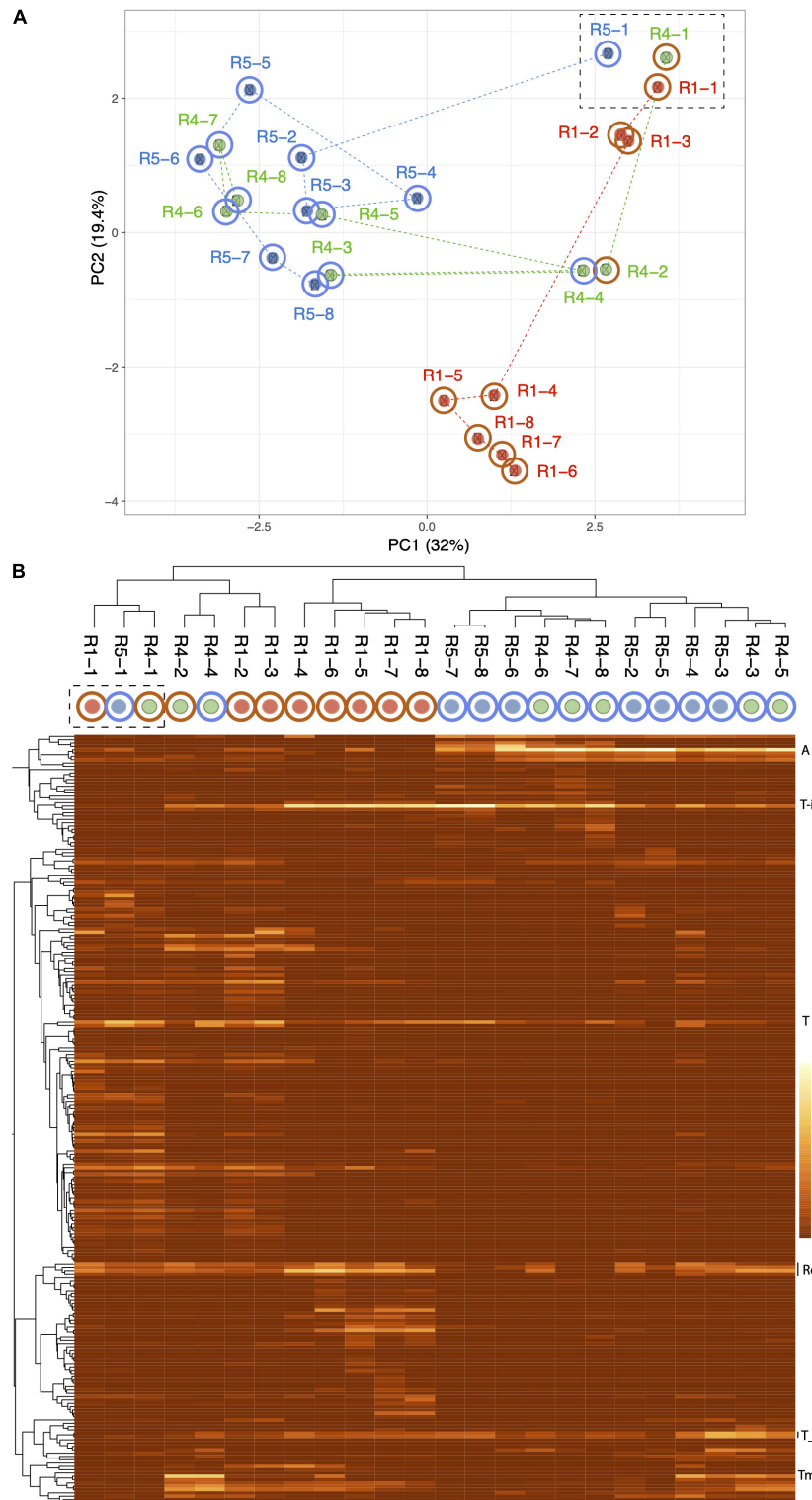
FIGURE 6 | Principal component analysis based on the relative abundances of 232 organisms (accounting for the most abundant 70–80% of the community) in reactors 1 (red dots), 4 (green dots), and 5 (blue dots). The dashed box indicates the first samples from each reactor. Note the dramatic divergence of R5 after ramping from 50 to 250 ppm SCN^- in the absence of molasses, and the convergence of R4 with R5 after termination of molasses input. **(B)** Heat map showing clustering of samples based on the abundance of each of the 232 organisms. The proliferation of *Afipia*_64_1782 **(A)** separates most molasses-free (blue outer circle) from molasses-amended (brown outer circle) communities. Other abundant organisms include *Thiobacillus*-i (T), two strains related to *Thiobacillus*_3 (T_3), and several Rhodanobacteraceae (Ro). *Thiomonas* (Tm) that is unusually abundant in R4-4 and may contribute to its unexpected clustering with R4-2.

input solution from 50 ppm to 250 ppm and reducing the residence time to 12 h (**Supplementary Figure 3**). This dramatic shift occurred within an additional 119 days. Differences in community composition in R1–R4 biofilm samples at time point 2 (despite experiencing the same conditions) may be due to stochastic processes during community development or may reflect within-reactor heterogeneity. However, within-reactor heterogeneity is unlikely to account for differences in the planktonic communities, as the reactors were well mixed.

At both the first and second time points, the biofilm, planktonic samples, and samples of flocculated biomass ("flocs") from each reactor are typically more similar to each other than to communities in other reactors (and planktonic and floc samples are essentially indistinguishable). PCA plots for individual reactors indicate some separation of biofilm and planktonic samples (**Supplementary Figure 4**). Of the relatively abundant bacteria in R1, strong partitioning into the planktonic relative to biofilm fraction occurred only for *Microbacterium_71_863* and Saccharibacteria at time point 5. This Saccharibacteria is also enriched in the planktonic fraction in R2 (second time point). Interestingly, *Afipia*_64_1782, which dominates molasses-free reactors and is at relatively low abundance in R1, is partitioned into the biofilm fraction of R1. R4 at time point 5 shows strong enrichment of a different *Microbacterium* (*Microbacterium_71_138*) in the biofilm compared to planktonic fraction, consistent with the prediction that it is an anaerobe (**Figure 2B**). The only strong fractionation in the molasses-free reactor, R5, occurred at time point 5 and involved partitioning of three abundant thiobacilli and a moderately abundant WPS-2 [candidate phylum Euryarchaeota (Ji et al., 2017)] into the planktonic fraction.

Independent PCA analyses of R1, R4, and R5 (**Supplementary Figure 4**) show step-like community compositional change as reactor SCN⁻ input concentrations were ramped. In contrast, the later timepoints in both R2 and R3 (timepoints 4 to 8 at constant input SCN⁻ concentration of 250 ppm) cluster together. Both of these observations would be expected if input SCN⁻ concentration is the primary compositional control within each reactor. Given these results, we compared the community compositions of R2 and R3 (**Figure 5**) and of R1, R4, and R5 (**Figure 6**). The PCA plot for R2 and R3 illustrates dramatic divergence attributable to termination of molasses addition. Notably, this separation occurred within 33 days of removal of molasses from the input solution for R3 (**Figure 5**). In detail, communities at timepoints 3, 4, and 5 separate from those at timepoints 6, 7, and 8 in both R2 and R3, a result that suggests ongoing evolution in community composition independent of SCN⁻ input concentration. The PCA plot for R1, R4, and R5 (**Figure 6A**) also clearly shows differences in community composition in molasses-free (R5) compared to molasses-amended reactors after reactors were ramped from SCN⁻ input concentrations of 50 to 250 ppm. The shift in R5 is generally tracked by communities in R4 after molasses was removed from input solution at time point 3. However, unexpectedly, R4-4 clusters with R4-2, possibly due to the unusually high abundance of *Thiomonas* in R4-4 (Tm, **Figure 6B**). The detected *Thiomonas* (Thiomonas_64_19

and Thiomonas_64_14) have genes encoding SoxABCXYZ, suggesting that their proliferation was linked to the oxidation of sulfur compounds from degraded SCN. Overall, the clustering of communities at moderate to high input SCN⁻ concentrations is driven by the proliferation of *Afipia*_64_1782 strain in the water reactors and one specific *Thiobacillus* strain in both molasses-free and molasses reactors (**Figure 6B**).

CONCLUSION

By combining a substantial set of long-term reactor-based experiments with resolution of the strain composition achieved via genome-resolved metagenomics, we demonstrate consistent (across reactors) and largely reproducible (across experiments) selection for specific bacteria capable of autotrophic growth and thiocyanate degradation from a background of considerable strain diversity. Nonetheless, there is evidence for ongoing evolution in community composition in reactors maintained under constant conditions and some divergence in reactors experiencing the same conditions may be evidence for historical contingency effects. The reactors clearly diverge based on the presence or absence of added organic substrate in the form of molasses. Interestingly, despite the association of key genes for SCN⁻ degradation with autotrophs in both reactors and the prominence of a variety of autotrophs in molasses-free reactors, SCN⁻ degradation is destabilized in the absence of molasses, especially at high-input SCN⁻ concentrations. We attribute this to the reduced ability of the molasses-free reactor community to manage the disbalance between sulfur oxidation and assimilated nitrogen following thiocyanate degradation, leading to excessive ammonia accumulation, as well as the importance of heterotrophs that provide community-essential functions, including production of cobalamin, and creation of biofilm and its associated niche diversity.

DATA AVAILABILITY STATEMENT

Metagenomic reads and assembled genomes are available at National Center for Biotechnology Information (NCBI) under project accession: PRJNA629336.

AUTHOR CONTRIBUTIONS

The reactor experiments were designed by RH, SH, and RK, with input from JB, and conducted at CeBER by RH and FK. FK and RH performed the chemical analyses with input from SH. RH extracted the DNA, which was sequenced at the UC Berkeley QB3 facility. RS conducted the metagenomic assembly, performed the functional annotations, and performed the bioinformatics analysis for automated binning using previous genomes from the study system. JB performed manual genome binning, bin curation, and genome curation. JB, RS, SH, and RH analyzed the

data. JB and RH wrote the manuscript, with input from RS and RK. All authors commented on and approved the final version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.643368/full#supplementary-material

Supplementary Figure 1 | Diagram summarizing the operation and sampling of the five SCN⁻ reactors over the course of the experimental period. **(A)** shows the experimental SCN⁻ feed loading regime over time, while **(B)** shows the five SCN⁻ degrading reactors in operation. The vertical dotted lines and asterisks show the points at which biomass was sampled for subsequent total genomic DNA extractions and sequencing.

Supplementary Figure 2 | Heatmap based on normalized relative abundance patterns of bacteria in R2 and R3 indicating co-occurrence of Saccharibacteria (TM7) and Microbacterium bins.

Supplementary Figure 3 | Principal coordinate analysis of community composition at the first two time points in all reactors. The rectangular box shows that communities in all reactors were similar at the first time point, despite reactor 5 receiving no molasses for 117 days. The rounded box indicates communities in all reactors receiving molasses and 250 ppm SCN⁻ (R1, R2, R3, and R4, the second time point). Despite experiencing the same treatments, R1, R2, R3, and R4 are somewhat distinct from each other. Reactor 5 (molasses-free reactor) after the second time point has a very different community composition compared to the other reactors. Included are samples from biofilm, the planktonic fraction (p) and, for reactors 1 and 5, floc samples (f). In most cases, the planktonic fraction is somewhat different from the biofilm (and floc samples).

Supplementary Figure 4 | Principal coordinate analysis of community composition for each of the five reactors. Boxes indicate biofilm, planktonic (p) and floc (f) samples from the same time point.

Supplementary Figure 5 | Version of **Figure 6B** showing organism names.

Supplementary Table 1 | Solution chemistry from the five reactors.

Supplementary Table 2 | Percentage of reads from each sample that map to the 232 genomes.

Supplementary Table 3 | Abundances of organisms represented by 232 genomes across the samples from the five reactors.

Supplementary Table 4 | The relative abundances of bacteria that were abundant in R1 (molasses) and R5 (no molasses) that were amended with thiocyanate and ramped to 2000 mg/L in the five reactors and over time.

REFERENCES

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146. doi: 10.1038/nmeth.3103

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., et al. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523, 208–211. doi: 10.1038/nature14486

Chain, P., Lamerdin, J., Larimer, F., Regala, W., Lao, V., Land, M., et al. (2003). Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea. J. Bacteriol.* 185, 2759–2773. doi: 10.1128/jb.185.9.2759-2773.2003

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., and Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. *Genome Res.* 30, 315–333. doi: 10.1101/gr.258640.119

Diamond, S., Andeer, P. F., Li, Z., Crits-Christoph, A., Burstein, D., Anantharaman, K., et al. (2019). Mediterranean grassland soil C–N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms. *Nat. Microbiol.* 4, 1356–1367. doi: 10.1038/s41564-019-0449-y

Douglas Gould, W., 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. *Miner. Eng.* 34, 38–47. doi: 10.1016/j.mineng.2012.04.009

He, X., McLean, J. S., Edlund, A., Yooseph, S., Hall, A. P., Liu, S.-Y., 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. doi: 10.1073/pnas.1419038112

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. *Miner. Eng.* 76, 65–71. doi: 10.1016/j.mineng.2014.12.011

Hung, C. H., and Pavlostathis, S. G. (1999). Kinetics and modeling of autotrophic thiocyanate biodegradation. *Biotechnol. Bioeng.* 62, 1–11. doi: 10.1002/(sici)1097-0290(19990105)62:1<1::aid-bit1>3.0.co;2-q

Ji, M., Greening, C., Vanwonterghem, I., Carere, C. R., Bay, S. K., Steen, J. A., et al. (2017). Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature* 552, 400–403.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., et al. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7:e7359.

Kantor, R. S., Huddy, R. J., Iyer, R., Thomas, B. C., Brown, C. T., Anantharaman, K., et al. (2017). Genome-resolved meta-omics ties microbial dynamics to process performance in biotechnology for thiocyanate degradation. *Environ. Sci. Technol.* 51, 2944–2953. doi: 10.1021/acs.est.6b04477

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. doi: 10.1111/1462-2920.12936

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858. doi: 10.1038/nprot.2015.053

Kolmert, Å, and Johnson, D. B. (2001). Remediation of acidic waste waters using immobilised, acidophilic sulfate-reducing bacteria. *J. Chem. Technol. Biotechnol.* 76, 836–843. doi: 10.1002/jctb.453

Metsalu, T., and Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic Acids Res.* 43, W566–W570.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., et al. (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* 43, D213–D221.

Olm, M. R., Brown, C. T., Brooks, B., and Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* 11, 2864–2868. doi: 10.1038/ismej.2017.126

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Rahman, S. F., Kantor, R. S., Huddy, R., Thomas, B. C., van Zyl, A. W., Harrison, S. T. L., et al. (2017). Genome-resolved metagenomics of a bioremediation system for degradation of thiocyanate in mine water containing suspended solid tailings. *Microbiologyopen* 6:e00446. doi: 10.1002/mbo3.446

Sharma, S., and Bhattacharya, A. (2017). Drinking water contamination and treatment techniques. *Appl. Water Sci.* 7, 1043–1067. doi: 10.1007/s13201-016-0455-7

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., et al. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* 3, 836–843. doi: 10.1038/s41564-018-0171-1

Tikhonova, T. V., Sorokin, D. Y., Hagen, W. R., Khrenova, M. G., Muyzer, G., Rakitina, T. V., et al. (2020). Trinuclear copper biocatalytic center forms an active site of thiocyanate dehydrogenase. *Proc. Natl. Acad. Sci. U.S.A.* 117, 5280–5290. doi: 10.1073/pnas.1922133117

Tsallagov, S. I., Polyakov, K. M., Tikhonova, T. V., Trofimov, A. A., Shabalin, I. G., Popov, A. N., et al. (2016). *Thiocyanate Dehydrogenase From Thioalkalivibrio Paradoxus.* Protein Data Bank. doi: 10.2210/pdb5f75/pdb

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

van Buuren, C., Makhotla, N., and Olivier, J. W. (2011). "The ASTER process: technology development through to piloting, demonstration and commercialization," in *Paper Presented at the Proceedings of the ALTA 2011 Nickel–Cobalt–Copper, Uranium and Gold Conference*, Perth,WA, 23–28.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303.

Watts, M. P., and Moreau, J. W. (2016). New insights into the genetic and metabolic diversity of thiocyanate-degrading microbial consortia. *Appl. Microbiol. Biotechnol.* 100, 1101–1108. doi: 10.1007/s00253-015-7161-5

Watts, M. P., Spurr, L. P., Gan, H. M., and Moreau, J. W. (2017). Characterization of an autotrophic bioreactor microbial consortium degrading thiocyanate. *Appl. Microbiol. Biotechnol.* 101, 5889–5901. doi: 10.1007/s00253-017-8313-6

Watts, M. P., Spurr, L. P., Lê Cao, K.-A., Wick, R., Banfield, J. F., and Moreau, J. W. (2019). Genome-resolved metagenomics of an autotrophic thiocyanate-remediating microbial bioreactor consortium. *Water Res.* 158, 106–117.

Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607. doi: 10.1093/bioinformatics/btv638

Yamada, M., Yoshida, H., Kuramitsu, S., and Kamitori, S. (2008). *X-ray Structure of Hypothetical Selenium-Binding Protein from Sulfolobus Tokodaii, ST0059.* Protein Data Bank. doi: 10.2210/pdb2ece/pdb

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER suite: protein structure and function prediction. *Nat. Methods* 12, 7–8. doi: 10.1038/nmeth.3213