AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology

Identification and Resolution of Microdiversity through Metagenomic Sequencing of Parallel Consortia

William C. Nelson,[a] Yukari Maezato,[a] Yu-Wei Wu,[b,c] Margaret F. Romine,[a] Stephen R. Lindemann[a]

Biological Sciences Division, Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington, USA[a]; Joint BioEnergy Institute, Emeryville, California, USA[b]; Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA[c]

To gain a predictive understanding of the interspecies interactions within microbial communities that govern community function, the genomic complement of every member population must be determined. Although metagenomic sequencing has enabled the *de novo* reconstruction of some microbial genomes from environmental communities, microdiversity confounds current genome reconstruction techniques. To overcome this issue, we performed short-read metagenomic sequencing on parallel consortia, defined as consortia cultivated under the same conditions from the same natural community with overlapping species composition. The differences in species abundance between the two consortia allowed reconstruction of near-complete (at an estimated >85% of gene complement) genome sequences for 17 of the 20 detected member species. Two *Halomonas* spp. indistinguishable by amplicon analysis were found to be present within the community. In addition, comparison of metagenomic reads against the consensus scaffolds revealed within-species variation for one of the *Halomonas* populations, one of the *Rhodobacteraceae* populations, and the *Rhizobiales* population. Genomic comparison of these representative instances of inter- and intraspecies microdiversity suggests differences in functional potential that may result in the expression of distinct roles in the community. In addition, isolation and complete genome sequence determination of six member species allowed an investigation into the sensitivity and specificity of genome reconstruction processes, demonstrating robustness across a wide range of sequence coverage (9× to 2,700×) within the metagenomic data set.

Microdiversity refers to the diversity of organisms that are closely related phylogenetically yet exhibit different metabolic activities and therefore occupy distinct niches. Genomic studies comparing multiple strains of the same species have revealed that while much of the genome sequence is highly conserved, significant functional variation can exist, arising from changes in gene function due to mutation, the introduction of genes by horizontal gene transfer, or changes in gene regulation due to mutation or genome rearrangement. Microbial community diversity is usually measured via sequencing of either all or a part of the 16S rRNA gene. It is well established that bacteria that have near-identical or identical 16S rRNA sequences can have significantly divergent genome complements, cell morphologies, and metabolic functions (1–4).

Recent developments in sequencing technologies and analysis have enabled cultivation-independent sequencing of intact microbial communities (metagenomics) and prediction of functional potential of individual microbial populations within a community. Early metagenomic work employed the same technology used for isolate genome sequencing, the Sanger method, which provided only a limited sequencing depth. Nonetheless, initial attempts to reconstruct member genomes in very-low-diversity communities or abundant members of complex communities were successful (5, 6). Innovations in sequence assembly (7–9), the use of differential coverage binning (10, 11), parsing mate pair linkages (12), and improved methods of evaluating sequence composition (13, 14) continue to improve our ability to reconstruct genomes from community metagenomic samples. These improvements are revolutionizing microbial ecology by enabling species-resolved genomic analysis of organisms without the need for prior cultivation, thus providing a mechanism for prediction and modeling of ecosystem function for communities that were previously intractable.

Genome reconstruction consists of two steps; first, the metagenomic sequence is assembled to create contigs (or scaffolds if clone mate information is available), and then the contigs are segregated into "bins" of sequences with similar characteristics. The characteristics used to segregate contigs include %G+C, k-mer content, estimated relative abundance, and predicted phylogeny. The resulting bins are analyzed to determine if they represent a single organism or more than one (testing the resolving power of the criteria) and to estimate how much of the total genomic complement was recovered in the reconstruction process (completeness). This is typically done by comparing the amount of sequence recovered to the genome size of sequenced near neighbors or by enumerating genes conserved across all fully sequenced bacterial or archaeal genomes.

Despite the successes in extracting reconstructed genomes from a variety of environmental metagenomes, significant challenges still exist in reconstructing entire communities as assemblages of individuals. The process of obtaining and segregating the sequence of all members of a community is difficult because many

Received 17 July 2015 **Accepted** 16 October 2015

Accepted manuscript posted online 23 October 2015

Citation Nelson WC, Maezato Y, Wu Y-W, Romine MF, Lindemann SR. 2016. Identification and resolution of microdiversity through metagenomic sequencing of parallel consortia. Appl Environ Microbiol 82:255–267. doi:10.1128/AEM.02274-15.

Editor: F. E. Löffler

Address correspondence to William C. Nelson, william.nelson@pnnl.gov.

W.C.M. and Y.M. contributed equally to this article.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.02274-15.

environments have high species richness (number of taxa present), with most members present at a low relative abundance (15). For low-abundance organisms, gathering enough sequence data to sufficiently sample their genomes for assembly may potentially be solved by simply increasing the number of sequence reads generated; however, there are concomitant increases in sequencing and computational expenses, as well as the inherent difficulty of working with large data sets. Microdiversity further hampers genome assembly and segregation, as current assembly, coverage, and composition analysis techniques are insufficient for distinguishing many closely related genomes. To date, only one community, composed of only five members, has been comprehensively sequenced and reconstructed (16).

To overcome the limitations of current approaches for identifying and resolving microdiversity in complex communities, we explored the use of parallel, enriched consortia. In this "divide-and-conquer" strategy, multiple subcultures with overlapping membership are generated from a parent community under the same culture conditions. Generating parallel consortia accomplishes (i) reduction of a community's complexity; (ii) selection for metabolically compatible organisms, whose interactions might have ecological relevance; (iii) acquisition of data sets among which direct comparisons can be made; and (iv) in some cases, expansion of rare populations. We have applied this strategy to investigate a cyanobacterial mat community that forms seasonally in the mixolimnion of an epsomitic, hypersaline lake (Hot Lake, Oroville, WA). Despite extreme variation in environmental parameters (e.g., temperature, incident light, and salinity) over the course of the year, once established, the Hot Lake mat exhibits little change in membership (17). Mat diversity ranges between ~500 and 1,000 operational taxonomic units (OTUs) throughout the seasonal cycle (17). Two unicyanobacterial consortia (UCC-A and UCC-O) were developed from the mat through physical isolation of individual filaments of two morphologically distinct cyanobacteria and serial passage (18). The heterotrophic membership of these consortia was estimated to overlap almost completely, with 14 of 15 heterotroph OTUs present in each, although at differing relative abundances.

Here we report the discovery and resolution of microdiversity using parallel consortia of tractable complexity. This work revealed 20 organisms present in the UCC-A and UCC-O consortia, uncovering two *Halomonas* species that were identical over their V4 region and therefore identified as a single OTU by amplicon analysis, and evidence of genetically distinct subpopulations of several organisms within the community. Genome analysis suggests that different mechanisms drive functional specialization in these instances of interspecies and intraspecies microdiversity. In addition, we tested the rigor of generally accepted genome reconstruction techniques by comparison against six complete isolate genomes, finding them to be quite robust. This validation of the reconstruction process lends high confidence to the accuracy of the final reconstructed genomes.

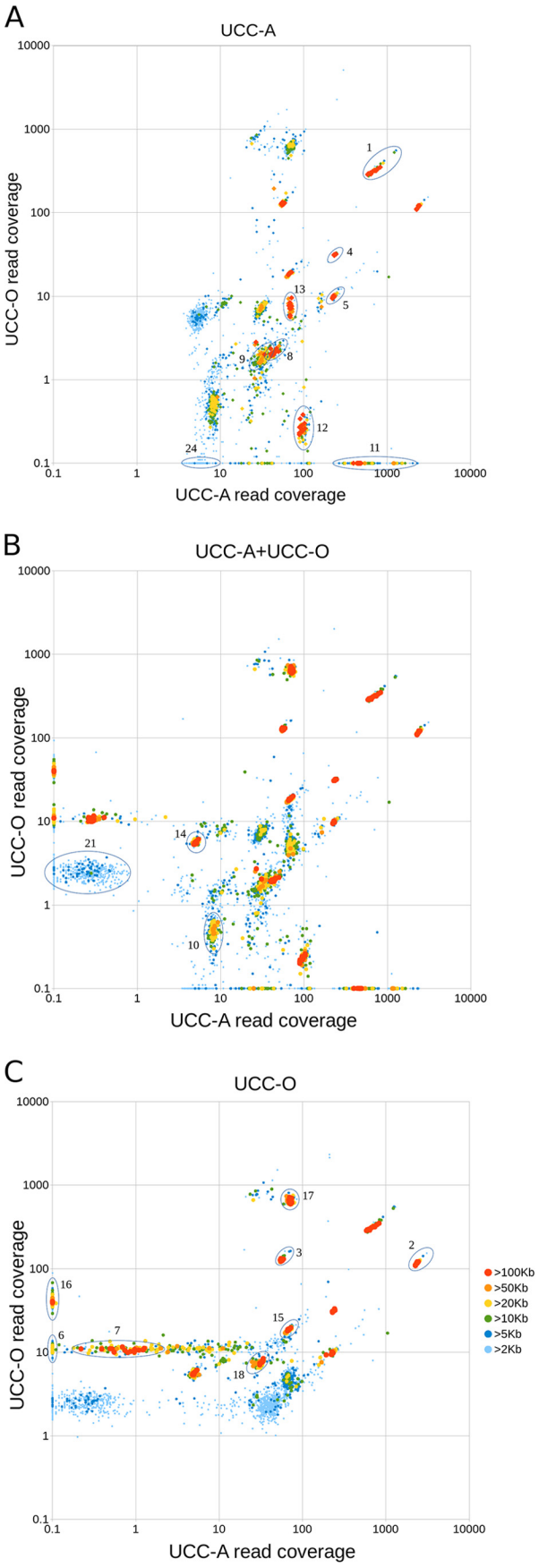
MATERIALS AND METHODS

DNA source and sequencing. Unicyanobacterial consortia UCC-A and UCC-O were generated as previously described (18). Briefly, biomass from the green lamina of a Hot Lake microbial mat sample harvested 7 July 2011 was cultivated on Hot Lake autotroph medium (HLA) plates at ~20 $\mu E/m^2/s$. Filaments of each of the two cyanobacteria growing on the plates were separated and individually enriched, forming two parallel enrichment cultures that were sequentially passaged every 28 days. To generate DNA for metagenomic sequencing, enrichment cultures of consortia UCC-O and UCC-A were harvested 3 weeks postpassage and physically homogenized, and DNA was extracted as described previously (18). Succinctly, DNA was extracted from 1 ml of homogenized culture using the MasterPure Complete DNA and RNA purification kit (Epicentre, Madison, WI) according to the manufacturer's instructions. For each sample, 0.5 lane of Illumina HiSeq 2500 paired-end (2 × 150 bp) metagenome sequencing was performed at the U.S. Department of Energy's Joint Genome Institute (JGI) under Community Sequencing Project (CSP) 701. Briefly, 500 ng of genomic DNA was sheared using the Covaris E210 instrument (Covaris) and size selected using Agencourt Ampure Beads (Beckman Coulter). The DNA fragments were treated with end repair, A-tailing, and adapter ligation using the TruSeq DNA Sample Prep kit (Illumina) and purified using Agencourt Ampure beads (Beckman Coulter). The prepared libraries were quantified using KAPA Biosystem's next-generation sequencing library quantitative PCR (qPCR) kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v3, and Illumina's cBot instrument to generate a clustered flow cell for sequencing. Sequencing of the flow cell was performed on the Illumina HiSeq2000 sequencer using a TruSeq SBS sequencing kit 200 cycles, v3, following a 2 × 150 indexed run recipe.

Strains *Porphyrobacter* sp. HL-46, *Halomonas* HL-48, *Algoriphagus marincola* strain HL-49, *Idiomarinaceae* bacterium HL-53, *Marinobacter excellens* strain HL-55, and *Marinobacter* sp. HL-58 were isolated as previously described (18), and genomic DNA was extracted from them using the MasterPure Complete DNA and RNA purification kit as described above. *Halomonas* sp. HL-93 was isolated, its genomic DNA was extracted, and its 16S rRNA gene was sequenced and analyzed in the same manner as previously described (16); sequencing was performed by Functional Biosciences (Madison, WI). Isolate genomes were sequenced and assembled by the JGI under CSP 701 using the Pacific Biosciences (PacBio) sequencing technology (19). A Pacbio SMRTbell library was constructed and sequenced on the PacBio RS platform, which generated 210,195 filtered subreads totaling 641.0 Mbp. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov. The raw reads were assembled using HGAP (version 2.3.0) (20).

Metagenome sequence assembly. Raw sequence reads were quality trimmed using Trimmomatic (21) with a quality cutoff of 20, a window of 4 nucleotides (nt), and a length cutoff of 30 nt. Only paired-end reads (~88% of the trimmed reads and ~70% of the original data set) were used in assembly. UCC-A and UCC-O reads were assembled separately and also as a coassembly, resulting in three scaffold sets. Assembly was performed using IDBA-UD (8) with a minimum contig size of 250. Resulting scaffolds with length of ≥2 kb were considered further due to the discrimination limits of the binning algorithms (22, 23).

Scaffold binning. For each assembly set, read coverage was determined for each scaffold against both the UCC-A and UCC-O trimmed read data sets. Reads were aligned against scaffolds using Bowtie 2 (24); SAMtools (25) was used to calculate per-base coverage, and a custom script was used to calculate average coverage across the length of the scaffold. For each assembly set, UCC-A read coverage was plotted against UCC-O read coverage. Manual inspection of node density on the plot was used to determine the initial coverage ranges defining bins, which were then refined using a cutoff of ±2 standard deviations from the mean. An equivalent examination of the G+C nucleotide composition was used to both verify consistency within bins and further separate bins whose read coverage overlapped. This process segregated scaffolds into consistent organism-specific bins (similar to what is described in reference 10) (Fig. 1); however, to ensure that there was no subjective bias in the approach, the assembly sets and corresponding coverage data were evaluated using MaxBin v2.0, which examines coverage and nucleotide composition and



performs expectation maximization analysis to computationally bin sequences (26). MaxBin results verified the majority of the bin relationships predicted by our manual process (see Table S1 in the supplemental material). The final scaffold set for each genome bin was chosen from the assembly (UCC-A, UCC-O, or coassembly) that resulted in the most sequence across the fewest scaffolds (Fig. 1). The accuracy of the binning process was evaluated and refined through conserved single-copy gene analysis (see below). Genome bins were evaluated for consistency using the tetra-ESOM protocol outlined in reference 22, using an evaluation window of 5 kb.

Estimation of taxonomy and completeness of genomic content. Coding regions (CDS) were predicted with Prodigal (27) using normal mode and the default translation table. An estimated taxonomy was assigned to each genome bin using AMPHORA2 (28) using the Prodigal gene predictions and the provided reference alignments. Alignments were trimmed, the default Zorro cutoff of 0.4 was used, and maximum likelihood was used for phylotyping. Some bins contained a partial or, in one case (Bin01), a complete 16S rRNA gene. These sequences were compared against the Silva database and an internal database of Hot Lake microbial mat-derived 16S rRNA clones (17, 18) to provide additional phylogenetic information.

A representative phylogenetic tree for the reconstructed genomes was constructed by first aligning RpoC protein sequences from every organism and the type strains of neighboring genera (if available) using MUSCLE (29) as implemented in and under the default parameters of MEGA5.2.2 (30). In the event that the genus type strains were not sequenced, the nearest sequenced neighbors were included. Because *rpoC*, the gene encoding the β′ subunit of RNA polymerase, is divided in cyanobacteria into *rpoC1* and *rpoC2*, only positions 667 to 2263 of the alignment, which correspond to the portion of the alignment covered by RpoC2, were used to construct the phylogenetic tree. A neighbor-joining tree was constructed in MEGA5.2.2 using a Poisson model, pairwise deletion of gaps, and uniform rates of mutation among sites and tested using 1,000 bootstrap replications. The neighbor-joining tree was compared with a corresponding maximum likelihood tree generated in MEGA5.2.2. The maximum likelihood phylogeny was generated using complete deletion of gaps, a Jones-Taylor-Thornton substitution model (31) with uniform rates of mutation among sites, and the nearest-neighbor interchange maximum likelihood heuristic method. The maximum likelihood phylogeny was tested using 500 bootstrap replications.

Completeness of genomic content was estimated using the conserved single-copy gene (CSCG) procedure previously described (32). Briefly, a library of 140 Pfam hidden Markov models (HMMs) representing 102 proteins found in a single copy in a majority of sequenced bacterial genomes was searched against the predicted coding genes (CDS) of each bin. The percentage of models matched by CDS sequences within a bin is assumed to approximate the percentage of total genome content present (see Table S2 in the supplemental material). Bins containing multiple copies of any CSCGs were reexamined for scaffolds with anomalous read coverage or G+C content. In most cases, another bin that was a better match for the coverage and/or %GC and was missing the duplicate CSCG was identified. Such scaffolds were considered misbinned, and adjustments to bin membership were made accordingly.

Identification of sequence polymorphisms. The final trimmed, screened sequence read data sets were aligned against binned scaffold sets

FIG 1 Read coverage for assembled scaffolds. Sequence reads from UCC-A and UCC-O metagenomic samples were assembled separately and together. Reads from each data set were searched against the resulting scaffolds to determine average per-base coverage values. UCC-A and UCC-O coverage values are plotted for the scaffolds resulting from the UCC-A assembly (A), the coassembly (B), and the UCC-O assembly (C). For display purposes, scaffolds with zero coverage are plotted on the axes. Ellipses approximate which scaffolds comprise each genome bin (numbered). Colors denote scaffold length: red, ≥100 kb; orange, ≥50 kb; yellow, ≥20 kb; green, ≥10 kb; blue, ≥5 kb; cyan, ≥2 kb.

TABLE 1 UCC metagenome assembly statistics

Assembly	No. of input reads (10^6)	% of reads placed	No. of scaffolds	Total length (Mb)	N_{50}
UCC-A	139	98.0	22,055	67.9	35,571
UCC-O	125	98.4	16,859	59.4	56,494
Coassembly	264	97.7	24,629	85.0	43,762

using Bowtie2 (24), as described above. Sequence polymorphism locations were identified using the BamTools piledriver script (33; see the github website for Piledriver by A. R. Quinlan). Resulting variant positions were mapped to predicted CDS and intergenic positions.

Genome bin comparison. The accuracy of the reconstructed genome bins was assessed by aligning the binned scaffolds against the complete genome sequence using nucmer (34). Scaffolds that aligned to the genome with >98% identity across >90% of the sequence were considered mapped.

Growth curves. *Halomonas* sp. strains HL-48 and HL-93 were cultured for growth phenotype experiments at 30°C in 25 ml of HLN medium (HLA medium + 5 mM NH_4Cl) (18) + 5 mM carbon source (either sucrose or glucuronate) with shaking. Overnight cultures of HL-48 and HL-93 were used as an inoculum and were diluted to optical density at 600 nm (OD_{600}) of ~0.02 in 25 ml of fresh medium to start the growth experiment. Planktonic growth was monitored at a wavelength of 600 nm using a SmartSpec Plus spectrophotometer (Bio-Rad).

Nucleotide sequence accession numbers. The sequences generated in this study are available through DDBJ/EMBL/GenBank under accession numbers NZ_JQMU00000000.1 (*Porphyrobacter* sp. HL-46), NZ_JMMC00000000.1 (*Halomonas* sp. HL-48), NZ_JAFX00000000.1 (*Algoriphagus marincola* strain HL-49), LN899469 (*Idiomarinaceae* bacterium strain HL-53), NZ_JYNR00000000.1 (GI:761631804; *Marinobacter excellens* HL-55), NZ_JMLY00000000.1 (GI:654325145; *Marinobacter* sp. HL-58), LIHN00000000 (Bin01), LIHO00000000 (Bin02), LIHP00000000 (Bin03), LJSH00000000 (Bin04), LJSG00000000 (Bin05), LJST00000000 (Bin06), LJSU00000000 (Bin07), LJSF00000000 (Bin08), LJNT00000000 (Bin09), LJXT00000000 (Bin10), LJZR00000000 (Bin11), LJSV00000000 (Bin12), LJZS00000000 (Bin13), LJZQ00000000 (Bin14), LJSW00000000 (Bin15), LJZT00000000 (Bin16), LJSX00000000 (Bin17), LJSY00000000 (Bin18), LLEX00000000 (Bin21), and LKOZ00000000 (Bin24).

RESULTS

Parallel consortia enhance scaffold segregation. Metagenomic reads derived from UCC-A and UCC-O were assembled both separately and together (Table 1). In all cases, sequences assembled extremely well, with ~98% of quality-trimmed reads being placed in scaffolds of 250 nt or more and N_{50} values of 35,571 nt, 56,494 nt, and 43,762 nt for UCC-A, UCC-O, and the coassembly, respectively. Previous amplicon sequencing revealed that the two communities have almost entirely overlapping membership, although the relative abundances of the member species vary between the two cultures (18). Because the phylotypes present in each consortium were nearly identical, we compared coverage differences between the two metagenomic data sets to segregate scaffolds into organism-specific bins (numbered 01 to 24 [Fig. 1]). Bin20, Bin22, and Bin23 comprised extrachromosomal elements that could not be associated with an organism and are not considered further here; Bin19 was determined to be a second cyanobacterial chromosome and combined with Bin11. Read coverage was calculated for the scaffolds in each data set and plotted as an estimate of relative scaffold abundance in the two cultures (Fig. 1). High-abundance (>100× read coverage) organisms assembled well without coassembly. Bin01, Bin02, Bin03, Bin04, Bin05, and Bin15 were composed of relatively few, long scaffolds in each assembly set (Fig. 1 and Table 2). Several organisms appeared to be either absent or in extremely low abundance in one UCC or the other, resulting in construction of scaffolds from only one of the data sets: Bin06, Bin07, and Bin16 were constructed from the UCC-O-only assembly, and Bin10, Bin11, and Bin12 were derived from the UCC-A-only assembly. Assembly of low-abundance organisms shared between the cultures improved modestly with coassembly. For example, Bin10 improved to 190 scaffolds covering 3.8 Mb, from 254 scaffolds covering 3.44 Mb in UCC-A (and essentially no assembly in UCC-O), while Bin21 improved to 391 scaffolds covering 1.27 Mb from 289 scaffolds covering 0.91 Mb in UCC-O, and Bin14 improved to 51 scaffolds covering 3.77 Mb from 92 scaffolds covering 3.66 Mb in UCC-O.

Identification and deconvolution of microdiversity. The presence of closely related organisms within a metagenomic data set can lead to poor assembly or misassembly (9, 35). Evidence of this sort of interference can be observed within the assembly results by examining scaffold count and length relative to read coverage (36). A distinct Bin13 was derived from UCC-A, but in the UCC-O data set the assembly was significantly worse (compare the tight cluster of long scaffolds in Fig. 1A to the diffuse cloud of shorter scaffolds in Fig. 1C). A horizontally elongated cloud of short scaffolds appears at ~10× UCC-O read coverage, and Bin13 displays reduced UCC-O read coverage (Fig. 1C) relative to the UCC-A plot (Fig. 1A), suggesting a conflated assembly of reads from more than one organism. Coassembly yielded a moderate Bin13 assembly (Fig. 1B), and Bin06 scaffolds plotted exclusively at the *y* axis. A tetranucleotide frequency-based emergent self-organizing map of the UCC-O scaffold data set shows overlapping clustering of Bin13 sequences with an additional set of scaffolds (Fig. 2) assigned to Bin06. Comparison of the scaffolds in these two sets indicates an average nucleic acid identity of ~85%, with some sequence regions displaying better than 95% identity. This level of nucleotide identity is interpreted as indicating organisms related at the genus level (37), although the value may be artificially depressed due to regions of higher identity being coassembled into Bin13 and thus absent from Bin06. Thus, establishment of the parallel consortia enabled the resolution of interspecies microdiversity through alternate community membership.

Bin18 displayed evidence of multiple related strains, which could represent distinct ecotypes of the species. Coverage analysis showed higher read coverage for Bin18 in UCC-A than in UCC-O (~30× for UCC-A versus ~7× for UCC-O). Despite this, the UCC-O scaffold set assembled into fewer, longer contigs than that of UCC-A, having 87 contigs with an average length of 42,351 and a total length of 3.7 Mb versus UCC-A's 212 contigs with an average length of 16,258 nt and a total length of 3.4 Mb. Coassembly did not improve results, yielding a scaffold set of 252 contigs with an average size of 9,716 nt and a total length of only 2.3 Mb (the decrease in total sequence length was likely due to a decrease in the number of scaffolds reaching our inclusion cutoff of 2 kb). This result is strongly suggestive of assembly interference due to the presence of intraspecies microdiversity in UCC-A. That is to say that sequence and structural differences between two or more very closely related organisms cause multiple assembly solutions (bubbles and dead ends) within the assembly graph and result in shorter contigs/scaffolds. Comparison of UCC-A reads against the

TABLE 2 Genome-resolved bins derived from UCC-A and UCC-O samples and genomes from isolated members

Taxonomic group	Bin	Identification	No. of scaffolds	Length (Mb)	Avg % G+C	CDS	Estimated % cmp[a]
Reconstructed genomes							
Alphaproteobacteria	Bin04	*Oceanicaulis*	9	2.76	62	2632	99.3
Alphaproteobacteria	Bin08	*Rhodobacteraceae*	43	3.64	66	3638	100.0
Alphaproteobacteria	Bin15	*Erythrobacteraceae*	63	3.04	68	2886	100.0
Alphaproteobacteria	Bin21	*Porphyrobacter* sp. HL-46	392	1.27	64	1597	41.0
Alphaproteobacteria	Bin17	*Rhizobiales*	47	3.80	64	3463	99.3
Alphaproteobacteria	Bin05	*Rhodobacteraceae*	24	3.04	62	2953	100.0
Alphaproteobacteria	Bin07	*Rhodobacteraceae*	34	3.42	64	3314	100.0
Alphaproteobacteria	Bin09	*Rhodobacteraceae*	206	4.11	71	4123	87.6
Alphaproteobacteria	Bin12	*Rhodobacteraceae*	52	3.73	63	3612	100.0
Alphaproteobacteria	Bin18	*Rhodobacteraceae*	87	3.68	67	3700	100.0
Alphaproteobacteria	Bin24	*Rhodobacteraceae*	145	0.41	68	592	14.6
Bacteroidetes	Bin10	*Algoriphagus marincola* HL-49	190	3.80	42	3552	97.1
Bacteroidetes	Bin01	*Bacteroidetes*	21	3.37	51	2723	98.5
Cyanobacteria	Bin11	*P. priestleyi* strain Ana	99	5.47	48	4873	99.3
		Megaplasmid	18	0.39	48		
Cyanobacteria	Bin16	*Phormidium* sp. strain OSCR	175	4.66	51	4180	99.3
Gammaproteobacteria	Bin06	*Halomonas* sp. HL-93	176	3.26	57	3504	69.3
Gammaproteobacteria	Bin13	*Halomonas* sp. HL-48	34	3.55	59	3253	99.3
Gammaproteobacteria	Bin02	*Idiomarinaceae* bacterium HL-53	15	2.69	47	2500	100.0
Gammaproteobacteria	Bin14	*Marinobacter excellens* HL-55	51	3.77	56	3515	91.2
Gammaproteobacteria	Bin03	*Marinobacter* sp. HL-58	21	4.24	57	3856	99.3
Isolate genomes							
Alphaproteobacteria		*Porphyrobacter* sp. HL-46	2	3.17	64	2956	100.0
Bacteroidetes		*Algoriphagus marincola* HL-49	1	4.18	42	3767	98.5
Gammaproteobacteria		*Halomonas* sp. HL-48	1	3.74	59	3331	99.3
Gammaproteobacteria		*Idiomarinaceae* bacterium HL-53	1	2.74	48	2528	100.0
Gammaproteobacteria		*Marinobacter excellens* HL-55	1	4.00	56	3572	100.0
Gammaproteobacteria		*Marinobacter* sp. HL-58	1	4.29	58	3873	99.3

[a] cmp, completeness.

Bin18 sequence (taken from the UCC-O assembly) revealed 17,885 single-nucleotide polymorphism (SNP) locations. Intriguingly, there are regions of Bin18 scaffolds where one set of reads aligns perfectly to the bin, and another set displays a consistent pattern of mismatches (Fig. 3A). This type of variation is very similar to that observed between the *Ferroplasma acidarmanus* fer1 isolate genome and an environmental population (6, 38) and suggests that the Bin18 scaffold set from UCC-A is a composite assembly of two ecotypes. At SNP locations, the variant base was observed in 30 to 40% of the reads, indicating that two ecotypes were present in UCC-A at an approximately 2:1 ratio. The average sequence identity between equivalent Bin18 scaffold sets from UCC-A and UCC-O was ~98%; however, each of the bins contained a subset of scaffolds with breakpoints not found in the other assembly set. This could be due to structural differences between the chromosomes of the ecotypes, in the form of either rearrangements or alternate insertion locations of mobile elements, or it could be due to assembler error.

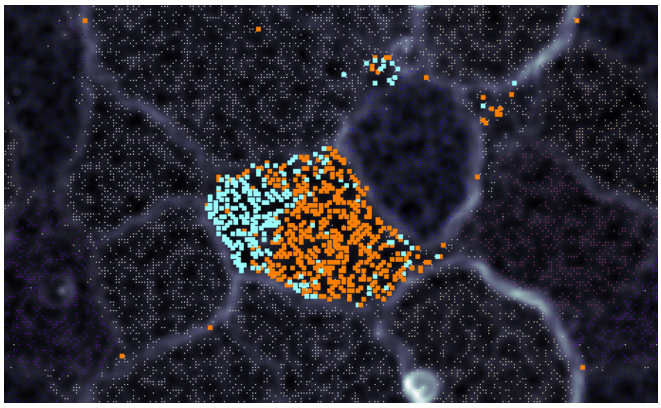


FIG 2 Region of an emergent self-organizing map based on tetranucleotide content showing Bin06/Bin13 scaffolds from UCC-O assembly. Analysis was performed on 5-kb windows. Cyan, bin06 scaffold windows; orange, bin13 scaffold windows. The grouping and overlap of the two sets indicate similar sequence composition, which complicates both assembly and binning.

We examined the genome bins for other examples of intraspecies microdiversity by aligning the UCC-A and UCC-O reads against the scaffolds and examining sequence variance. Similar to the Bin18 *Rhodobacteraceae* species, the Bin17 *Rhizobiales* species contains sites that display an elevated frequency of SNPs within the UCC-A community (Fig. 3B), while in UCC-O, *Halomonas* sp. HL-48 (Bin13) shows a pattern of elevated SNP frequency (Fig. 3C). One possibility is that these SNPs represent closely related sequence regions between Bin13 and the other *Halomonas* isolate, Bin06; however, Bin06 does not show a similar pattern, and thus we conclude that this is evidence of microdiversity within Bin13. Reexamination of the assembly plots (Fig. 1) shows that all three of these genome bins display similar behavior: good assembly from the metagenomic data set in which they do not show micro-

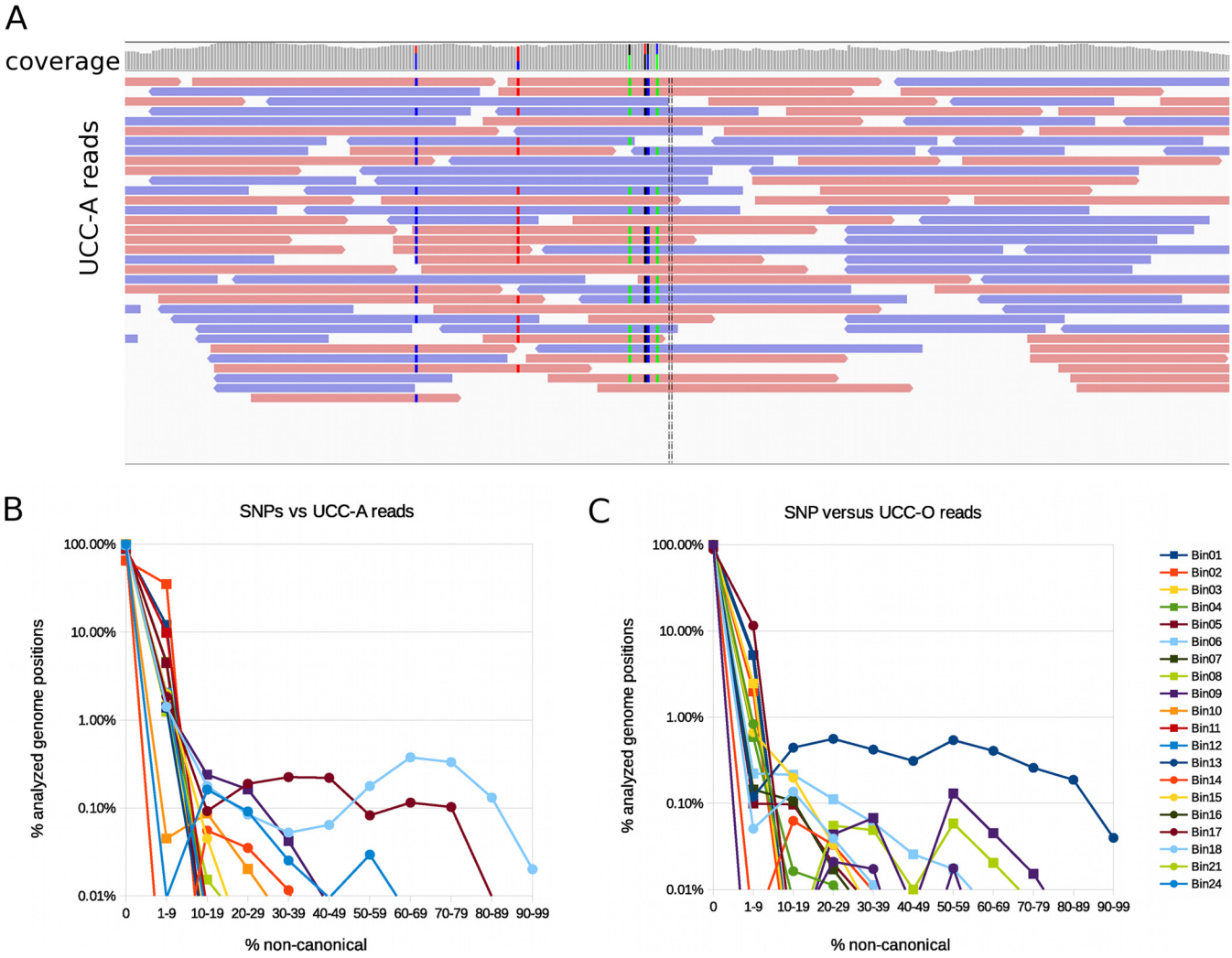


FIG 3 Sequence polymorphisms suggest the presence of microdiversity. (A) The UCC-A metagenomic read data set was searched against the Bin18 scaffolds reconstructed from the UCC-O assembly, and the alignments were visualized using IGV (50). Consistent patterns of polymorphisms (blue, red, green, and black bars) are apparent in a subset of the reads, indicating the presence of two Bin18-like organisms in UCC-A. (B and C) Scaffolds from genome bins were searched against the UCC-A (B) and UCC-O (C) read sets, and positions displaying variation (SNP or indel) were tallied. Bin17 and Bin18 show evidence of microdiversity in UCC-A, while Bin13 shows evidence of microdiversity in UCC-O. Bins for which a majority of genome positions could not be evaluated due to low coverage were not plotted.

diversity and poor assembly (regardless of high sequence coverage) in the metagenomic set in which there appear to be multiple variant strains. Curiously, Bin14 also displays assembly anomaly, assembling unusually well for low-coverage data from UCC-O, and assembling poorly from nearly equivalently low coverage data from UCC-A. In spite of this, the variance analysis does not indicate that microdiversity is present in UCC-A, suggesting that the modest difference in coverage between the two data sets resulted in significantly improved assembly. Therefore, not all situations in which an organism assembles well from one consortium and poorly from another are indicative of microdiversity.

Resolution of microdiversity reveals true community composition. Binning distinguished 24 distinct sequence sets. Several methods were employed to quality check and refine bin membership. The predicted protein complement of each bin was searched against a select Pfam HMM database of conserved single-copy genes (32), and the number of models with members and the number of members per model were tallied for each bin. The percentage of models with members estimates the completeness of genomic information within a bin, and the number of members per model measures specificity: each bin should have one protein and one protein only that is a member of each family; more than one protein per family would indicate that genomic information from another species is present in the bin. Bins were refined using CSCG data (see Materials and Methods), eliminating scaffolds containing duplicate genes, and adding scaffolds with complementary genes (that also met compositional and coverage criteria). We also assessed the CSCGs as phylogenetic markers, ensuring taxonomic consistency within each bin. From this analysis, 15 of the 20 genome bins were estimated to contain >95% of their respective genome complement (Table 2).

Previous amplicon-based analysis (18) indicated that the UCC communities each contained a single cyanobacterium and associated cohorts of 14 (UCC-A) and 15 (UCC-O) heterotroph OTUs.

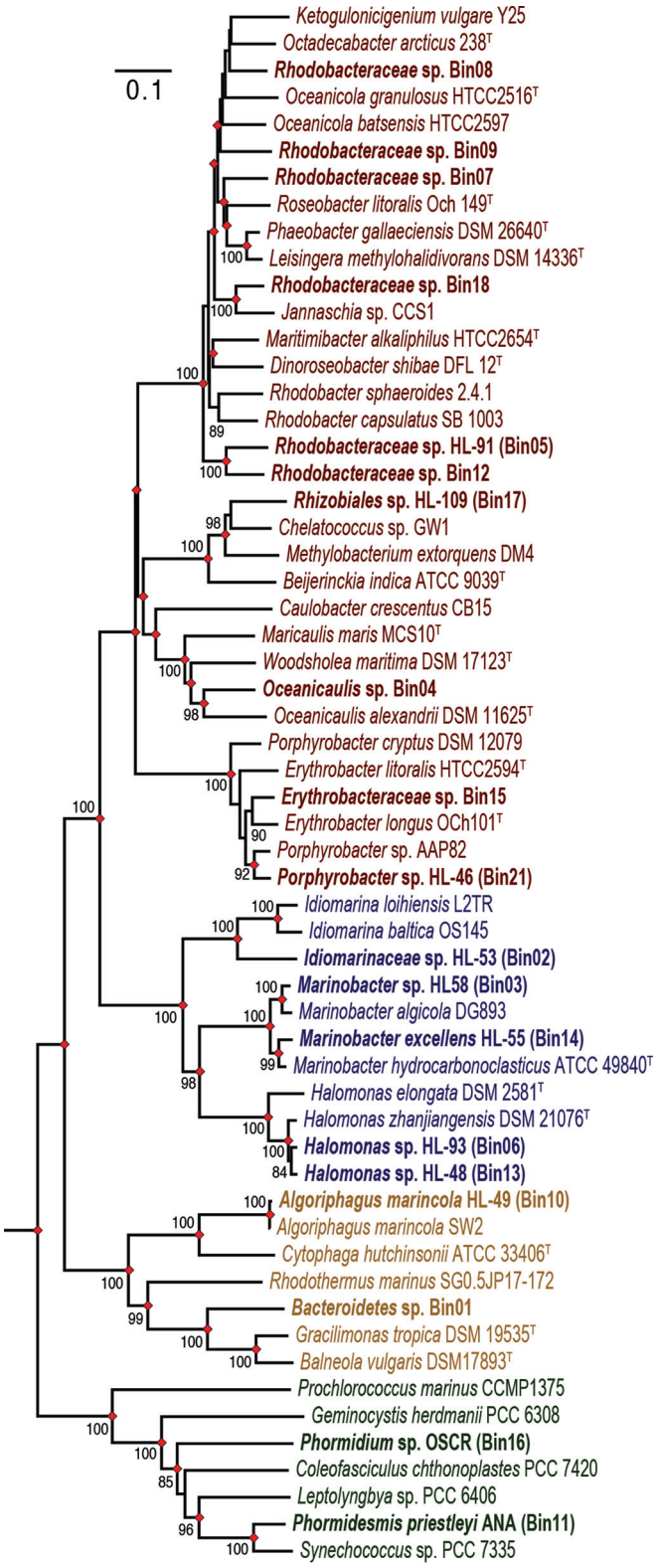


FIG 4 Neighbor-joining tree of *rpoC* demonstrating phylogenetic position of reconstructed genome sequences relative to model organisms. In cyanobacteria, the *rpoC* gene (encoding the β′ subunit of RNA polymerase) is divided into two genes, *rpoC1* and *rpoC2*; positions 667 to 2263 of the alignments were used, corresponding to the range of *rpoC2* in the cyanobacteria. Names are color coded by taxonomy: members of *Alphaproteobacteria* are in red, *Gammaproteobacteria* in blue, *Bacteroidetes* in yellow, and *Cyanobacteria* in green; members of the consortia are in bold. Where possible, the type strains of genera were included as references and are denoted by a superscript T; in cases where the type strains were not sequenced, the nearest sequenced neighbor was included. Bootstrap values exceeding 80% are reported next to the nodes to which they refer. Nodes of the neighbor-joining tree that were duplicated in the maximum likelihood reconstruction are denoted with a red diamond; as expected, most nodes that were not duplicated displayed low bootstrap values in either the neighbor-joining or maximum likelihood trees, or both.

Taxonomic analysis of the bins based on a CSCG (DNA-directed RNA polymerase subunit β′ [*rpoC*]) (Fig. 4) identified the 24 bins as the two key cyanobacteria (one with a megaplasmid), two members of *Bacteroidetes*, 5 members of *Gammaproteobacteria*, 12 members of *Alphaproteobacteria*, and three putative megaplasmids without a known host (Table 2). The greater resolution provided by the metagenomic data distinguished two halomonads (Bin06 and Bin13) that clustered into one OTU in the amplicon analysis (see below). Also, two additional *Rhodobacteraceae* species were revealed; each either had clustered into other OTUs classified as *Rhodobacteraceae* in the amplicon analysis or did not meet the abundance criteria previously employed. Only one bin could confidently be taxonomically assigned at the species level based upon AMPHORA analysis alone (Bin10 to *Algoriphagus marincola*), underscoring the genomic novelty present within these communities.

Previously, a halomonad was isolated from UCC-O as *Halomonas* sp. HL-48 (18), and here it has been confirmed as identical to Bin13 by comparing the *rpoC* sequence of the isolate with the genome bin (data not shown). The genomic evidence of two *Halomonas* spp., where amplicon analysis had identified only one, prompted isolation of the Bin06 halomonad (isolated as *Halomonas* sp. HL-93). The full-length 16S sequence is >98.8% identical to that of HL-48 and exactly matches a clone sequence derived from UCC-O (OCL_P1G6) that was previously thought to be an error-containing representation of the HL-48 sequence (see Fig. S1 in the supplemental material). The two 16S sequences are identical across the V4 region queried by amplicon analysis in our previous analysis (18); thus, the most abundant read from that OTU (representing 89% of the total reads) exactly represents both species.

Predicted functional impact of microdiversity. Gause's competitive exclusion principle (39) states that species competing for resources within an ecological niche cannot coexist at stable population abundances. Thus, it was of interest to see if the observed microdiversity in the *Halomonas* and *Rhodobacteraceae* species translated into significant differences in predicted function for these organisms that suggested they may occupy distinct niches. Bin06 (3,423 predicted CDS) contains 759 genes not found in the complete *Halomonas* sp. HL-48 (Bin13) genome (3,331 CDS). The reverse comparison is not informative, since the genome content of Bin06 is incomplete, and thus any genes absent in comparison with HL-48 could exist in the genome but be located within a sequence gap. Of these genes unique to Bin06, 362 were assigned to a specific function and functional role (Table 3; see also Table S3 in the supplemental material). In particular, an operon predicted to encode metabolism of glucuronate was identified. Growth assays using HL-48 and HL-93 (the cognate isolate for Bin06) have demonstrated that HL-93 can grow using glucuronate as a sole carbon source while HL-48 cannot (Fig. 5).

The presence of both strains of the Bin18 *Rhodobacteraceae*

TABLE 3 Functional role categorization of genes found in Bin06 but not *Halomonas* sp. HL-48

Functional role category[a]	No. of genes found to be Bin06 specific
Amino acid metabolism	15
Carbohydrate metabolism	13
Cell motility and adherence	9
Cell structure, growth, and death	18
Defense and invasion systems	13
Energy metabolism	2
Fatty acid and lipid metabolism	3
Intracellular trafficking, assembly, and processing	19
Metabolism of other amino acids and amines	1
Mobile and extrachromosomal element functions	28
Nucleic acid metabolism	4
Prosthetic groups, cofactors, and carriers	4
Regulatory functions	71
Signal transduction	12
Translation	5
Transport and binding proteins	131
Xenobiotics biodegradation and metabolism	27
Unknown function/no function assigned	384

[a] Role categories based on KEGG assignment.

TABLE 4 Functional role categorization of *Rhodobacteraceae* Bin18 genes containing SNPs

Functional role category[a]	No. of genes		
	Silent	Disrupted	Altered
Amino acid metabolism	57	2	122
Carbohydrate metabolism	57	3	109
Cell motility and adherence	15	1	21
Cell structure, growth, and death	14	3	15
Defense and invasion systems	25	0	35
Energy metabolism	70	4	72
Fatty acid and lipid metabolism	25	3	43
Glycan biosynthesis and metabolism	14	1	66
Intracellular trafficking, assembly, and processing	48	8	90
Mobile and extrachromosomal element functions	5	2	14
Natural products biosynthesis	0	0	1
Nucleic acid metabolism	52	6	112
Prosthetic groups, cofactors, and carriers	36	1	63
Regulatory functions	76	3	76
Signal transduction	34	0	26
Transcription	11	0	5
Translation	52	4	98
Transport and binding proteins	156	14	250
Xenobiotics biodegradation and metabolism	6	0	19
No assigned function	139	48	411
Total	892	103	1,648

[a] Role categories based on KEGG assignment.

species in UCC-A and only a single strain in UCC-O suggests that these organisms exhibit distinct functions. The strong sequence conservation between the two supports the assumption that gene content should also be highly conserved; however, gene content analysis similar to the above is not possible due to the incomplete status of both genomes. Another possible mechanism driving functional distinction between the strains may be alternate regulation of gene expression. Mapping SNPs identified within the UCC-A read data set (see above) against the annotated Bin18 sequence revealed that ~9% of the SNPs map evenly across intergenic regions. This figure approximates the 9.2% of genome sequence that is intergenic, indicating that there is generally not a strong bias of SNPs to putatively regulatory intergenic regions. The rest of the SNPs map to 2,630 of the 3,700 predicted coding regions and 1 of the 36 identified tRNA genes. Of these, 892 genes

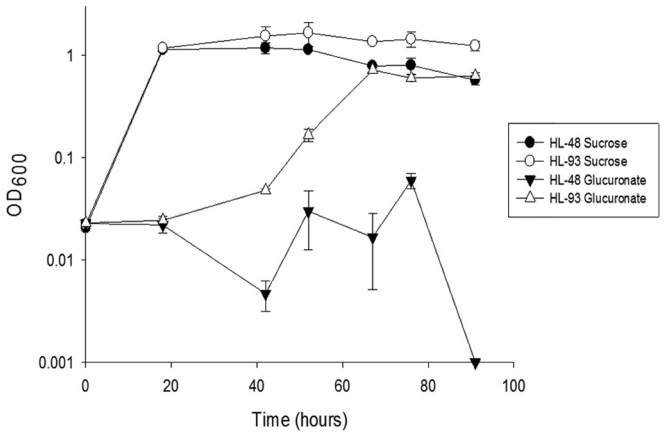


FIG 5 Growth of *Halomonas* sp. strains HL-48 and HL-93 on sucrose and glucuronate. Cultures of *Halomonas* sp. strains HL-48 and HL-93 in HLN medium were supplemented with 5 mM sucrose or glucuronate and grown for 90 h.

contained only silent mutations, even though some carried up to 31 SNPs. However, in 89 genes, SNPs or indels were predicted to prematurely truncate translation by introducing stop codons or frame shifts. The effect of this potential loss of function is difficult to predict. The plurality of affected genes are transporters with unspecified substrates or have no annotated function (Table 4; see also Table S4 in the supplemental material). In addition, SNPs altered the protein sequences of 1,648 genes with annotated functions in amino acid, nucleic acid, and carbohydrate metabolism and transport. Thus, it appears more likely that functional diversity in the Bin18 strains might arise from alteration or loss due to sequence polymorphism rather than due to changes in regulation.

Evaluation of reconstructed genomes. To validate the robustness of the genome reconstruction method and CSCG analysis, confirm that our isolates were identical to the dominant members of the consortia, and identify portions of the genome that did not assemble well enough to be included in genome bins, we compared the reconstructed genome bins to the complete genome sequences of six member species isolated from the consortia: *Idiomarinaceae* bacterium strain HL-53, *Marinobacter* sp. HL-58, *Algoriphagus marincola* strain HL-49, *Halomonas* sp. HL-48, *Marinobacter excellens* strain HL-55, and *Porphyrobacter* sp. HL-46. These organisms correspond to Bin02, Bin03, Bin10, Bin13, Bin14, and Bin21, respectively. Importantly, these organisms were present at a range of abundances within the UCCs, allowing us to examine reconstruction accuracy as a function of sequence coverage. Only in the case of *Porphyrobacter* sp. HL-46 did the addition of the isolate sequence significantly improve the completeness of our genomic information for the organism; Bin21 had only ~4×

read coverage between the two UCCs, resulting in moderate genome reconstruction (see above). The complete genomes were searched against their cognate bins (Fig. 6) and the total scaffold sets from the assembly from which the bin had been derived (Fig. 1), and the accuracy of binning procedures was evaluated using principles of signal detection theory. We measured the number of correct detections (CD; scaffolds correctly assigned to a genome bin), the number of false positives (FP; scaffolds incorrectly assigned to a genome bin), and the number of missed detections (MD; scaffolds not assigned to a genome bin that should have been). Specificity is measured by the ratio CD/*N*, where *N* is the total number of scaffolds in the bin. Sensitivity is measured by the ratio CD/(CD + MD). Three bins (Bin02, Bin10, and Bin03) were fully precise, containing all scaffolds that belonged and only those scaffolds (Table 5). Another two bins (Bin13 and Bin14) contained only scaffolds that belonged but were missing a few, totaling 76 kb and 110 kb, respectively. Even the low-coverage Bin21 had 388 of 391 (99%) scaffolds correctly assigned, with the three false positives totaling 7,578 bp, and was missing only 22 scaffolds, totaling 109 kb. Thus, the binning process outlined here was highly accurate, even for low-abundance organisms. Errors in the process appear more likely to occur for shorter sequences, suggesting that binning accuracy will be lower when assembly yields mainly short (<5-kb) scaffolds. The specificity and sensitivity of the reconstruction process demonstrated by this analysis lend confidence that the other bins are of equal quality and thus good representatives of the genomic content present in these community members.

We also used isolate genomes to evaluate the accuracy of the CSCG-estimated completeness analysis. First, it should be noted that running the analysis on complete genomes can result in a prediction of less-than-fully complete genomic information. For example, the *Marinobacter* sp. HL-58 complete genome scored as only 99.3% complete. This was due to the lack of a match to the IF3_N model, although HL-58 does have a protein identified as translation initiation factor IF-3 and that hits the IF3_C model. Thus, in this case, a lack of sensitivity in the model resulted in an underestimation of completeness. This type of problem is expected to be amplified in phylogenetic branches that are poorly represented in the sequence databases and thus not present in the seed alignments from which the CSCG HMMs are built. The accuracy of CSCG-based completeness analysis was evaluated by comparing completeness estimates to the percentage of total genome sequence contained within a bin. The agreement between the two measures was quite good (Table 5); however, it should be noted that some of the genes in this analysis, in particular the ribosomal protein genes, are frequently found together in long operons, and thus misbinning of a relatively short segment of genomic sequence could have an exaggerated effect on completeness estimations.

DISCUSSION

We have reconstructed near-complete genomic information from short-read metagenomic data for 17 of 20 species present in the UCC-A and UCC-O communities. The use of parallel consortia was critical to distinguishing closely related organisms at both the species and subspecies taxonomic levels. This strategy was able to resolve microdiversity between organisms classified within a single OTU in our previous amplicon analysis. For example, a second *Halomonas* species was discovered in UCC-O that was either absent or at very low abundance in the UCC-A metagenomic sample. Reconstruction also uncovered two previously undiscovered *Rhodobacteraceae* species, resulting in seven genomes of *Rhodobacteraceae* species, in contrast to amplicon assessments that suggested only five *Rhodobacteraceae* species in the consortia.

These results demonstrate the limitations of 16S rRNA-based amplicon surveys in evaluating the composition of even simple communities or in estimating community function. The implications of this problem are illustrated by reexamining previously reported community successional dynamics in these consortia based upon 16S rRNA abundances (18). The genome reconstructions now suggest that our quantitative PCR assay targeted against the 16S rRNA of *Halomonas* sp. HL-48 likely represented the combined abundance of the two halomonads. Therefore, OTU clustering may have masked any difference in the dynamics of the two organisms as the consortial biofilms assembled, which is significant in that the halomonads differ over ~25 to 30% of their gene content and likely have distinct metabolic roles. Based upon the moderate abundance of the closely related halomonads and the difficulty of separating them by compositional analysis, it is likely that attempts to reconstruct their genomes directly from the mat instead of through parallel consortia would have lumped them together into "metaorganisms." Such conflations hamper our ability to understand the relationship between member functional capacity and dynamics in community structure.

Assembly dynamics and SNP analysis suggested the presence of multiple variants of the Bin18 *Rhodobacteraceae* species, one being present in UCC-O and two being present in UCC-A. The level of sequence similarity between the two organisms (estimated to be ~98% nucleic acid identity across ~93% of the gene content) made it impossible to disentangle the two genomes in the UCC-A assembly; however, the maintenance of both variants over continued serial passage suggests that they occupy distinct realized niches in UCC-A. Studies of coresident, closely related species have demonstrated various optimal growth conditions (e.g., nutrient concentration, light intensity or wavelength, pH, temperature, etc.) or differences in susceptibility to predation by viruses that circumvent direct competitive exclusion (40–42). It should be noted, however, that the species examined in these studies were resolvable by 16S rRNA analysis and thus were not as closely related as the Bin18 variants. Another possibility is that one of the two subpopulations could be a "social cheater," benefiting from an activity that the other population expresses and gaining a fitness advantage by not incurring the cost of expressing that activity (43, 44). Examining the abundance and expression dynamics of the variant *Rhodobacteraceae* Bin18 strains in both consortia will permit the resolution of functional differences between these two populations and provide insight into why both populations persist in UCC-A. Although their functional potential appears to be similar because of similar gene content, SNP analysis indicates a large number of genes with potentially altered function and, in some cases, loss of expression. Thus, these modest changes in genome sequence could have significant impact on function and may define distinct niches for these organisms. It is also possible that activities specific to one or the other of the subpopulations are encoded by genomic islands or other variable regions that were not captured in our genome reconstructions. It has been hypothesized that the biogeochemical function of microbial communities can be predicted by 16S rRNA amplicon sequencing (45), but our observations and those of others (e.g., references 4, 46, 47, and

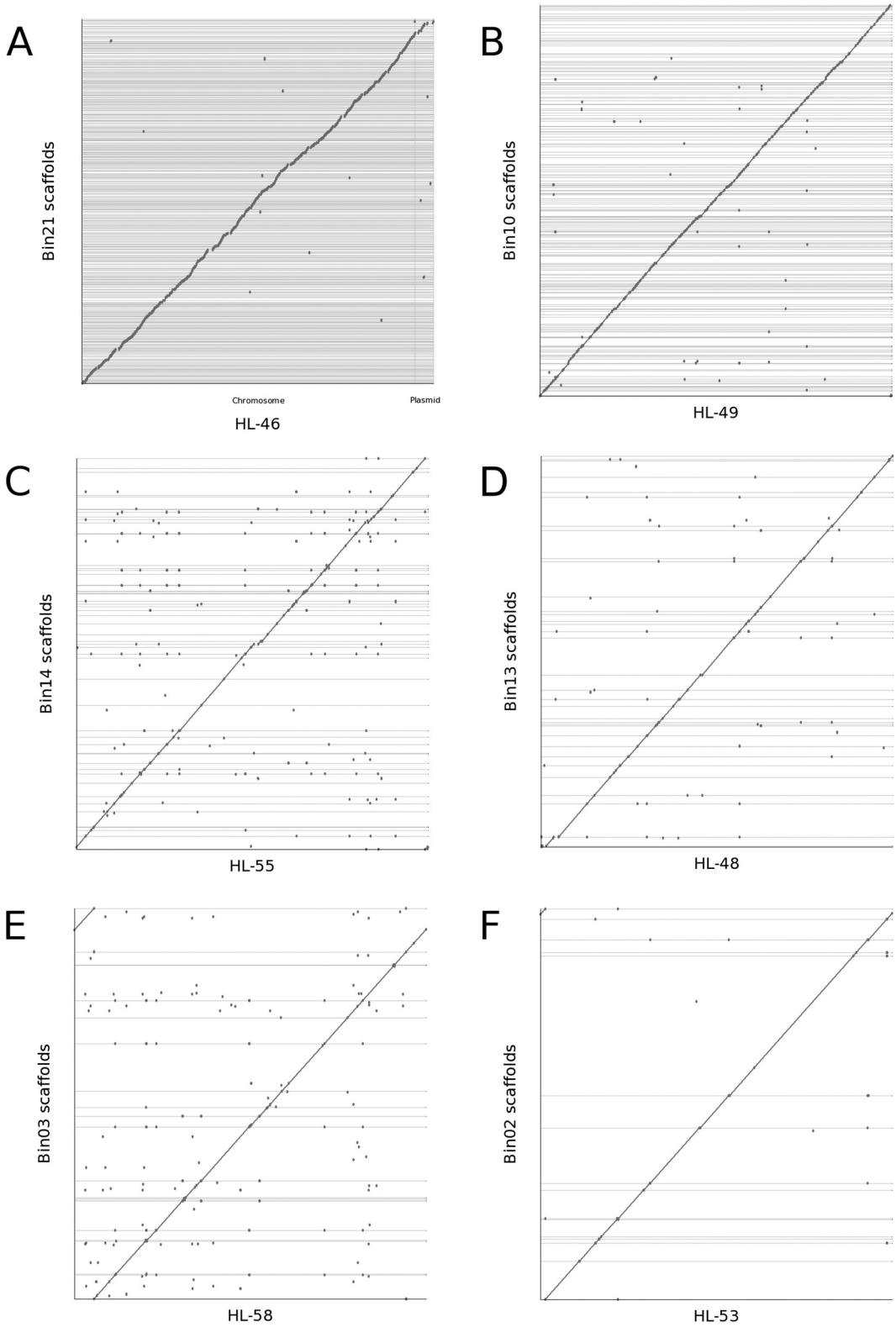


FIG 6 Comparison of binned scaffolds versus cognate complete genome sequences. Binned scaffold sequences were searched against complete genome sequences from cultured isolates using nucmer, and alignment regions were plotted using mummerplot. (A) Bin21 (3.8×) versus *Porphyrobacter* sp. HL-46; (B) Bin10 (10×) versus *Algoriphagus marincola* strain HL-49; (C) Bin14 (11×) versus *Marinobacter excellens* strain HL-55; (D) Bin13 (30×) versus *Halomonas* sp. HL-48; (E) Bin03 (120×) versus *Marinobacter* sp. HL-58; (F) Bin02 (2,700×) versus *Idiomarinaceae* sp. HL-53. The strong diagonals demonstrate consistent coverage across the length of the genome and an absence of sequence that does not map to the genome. Horizontal and vertical dotted lines represent scaffold boundaries.

TABLE 5 Comparison of genome-resolved bins to complete genome sequences

Isolate bin (read coverage)	% of isolate genome binned from metagenome[a]	Bin estimated completeness (%)	No. of mapped scaffolds (specificity)	No. of missed scaffolds (length; sensitivity)
Porphyrobacter sp. HL-46 Bin21 (3×)	40	41	388[b]/391 (0.99)	22 (109 kb; 0.95)
Halomonas sp. HL-48 Bin13 (70×)	95	99.3	34/34 (1.00)	7 (110 kb; 0.83)
Algoriphagus marincola HL-49 Bin10 (10×)	91	97.1	190/190 (1.00)	0 (0 kb; 1.00)
Idiomarinaceae strain HL-53 Bin02 (2,380×)	98	100	15/15 (1.00)	0 (0 kb; 1.00)
Marinobacter excellens HL-55 Bin14 (17×)	95	91.3	55/55 (1.00)	4 (76 kb; 0.93)
Marinobacter sp. HL-58 Bin03 (130×)	99	99.3	21/21 (1.00)	0 (0 kb; 1.00)

[a] Ratio of the total length of binned scaffolds to the total length of the isolate genome sequence.
[b] Ten of the 388 Bin21 scaffolds only partially aligned against HL-46 (>50% scaffold length), but with ≥99% identity.

51) suggest that metabolic diversity hidden within OTU designations makes such predictions unreliable.

The diversity and microdiversity present in natural communities complicate all aspects of genome reconstruction and make evaluation of the accuracy of the processes and results difficult. Studies performed in very simple communities have benefited from restricted diversity; however, even in the low-diversity Iron Mountain mat communities, microdiversity within the *Ferroplasma* type II populations was detected (6). One question that arises is how microdiversity affects assembly. In this study, we used a deBruijn graph assembler (IDBA_UD) (8). SNPs or indels within the read data set cause paths in the graph to split and merge (referred to as "bubbles"). Most deBruijn graph assemblers perform "bubble merging" by default to achieve a consensus assembly, wherein the shortest path with the greatest read support is chosen and the other is disregarded. This process will mask SNP level microdiversity. Many assemblers also perform a "read error correction." In IDBA_UD, "erroneous" reads (i.e., those that are not identical to the consensus sequence) that align against the consensus with >95% identity and 3 or fewer mismatches are candidates for correction. If every position in the consensus contig is confirmed by ≥80% of the supporting reads, the read is corrected to the confirmed sequence and considered support for the contig. Many of the SNP positions observed in Bin18 would not meet these correction criteria because the minority allele frequency averages 30%, and thus the contigs would not be considered "confirmed."

One method used for evaluating results from genome reconstruction is examining the content of conserved single-copy genes. Various forms of this analysis have been used since the earliest genome reconstructions, and it has proven to be fairly robust. Reported CSCG lists, however, including the one used in this study, usually include a large number of ribosomal proteins, many of which are typically linked in long operons. This can lead to skewed completeness estimates with the presence or absence of a relatively small amount of sequence. For example, *Marinobacter excellens* strain HL-55 has a region of 13,625 bases that contains 16 of the genes in the CSCG list. Thus, if only that region was omitted in a genome reconstruction, the CSCG analysis would indicate only 88% completeness, whereas the reconstruction would in fact be 99.7% complete. Another problem is that gene variability can cause low scores versus the models, falsely lowering completeness estimations. In addition, our concept of which genes are universally conserved shifts as more genome sequences are collected, especially from poorly characterized phyla. CSCG analysis has also been used phylogenetically to evaluate consistency in taxonomic assignment across a binned scaffold set. This type of analysis, however, does not have the resolution to resolve misassignment between closely related species. Also, novel phyla that have recently been targeted for reconstruction (11, 48) frequently do not yet have a good set of reference sequences, making it difficult to determine if such consistency is present.

Because we had access to organisms isolated from our experimental cultures, we were able to obtain complete genome sequences and assess the quality of our reconstructed genomes by direct comparison. It is important to note that these sequences were not available at the time of binning, so both were entirely *de novo*. Still, for all the bins tested, save Bin21, binning specificity was perfect. Because of low relative abundance in the sample, *Porphyrobacter* sp. HL-46 was poorly represented in the metagenomic data set and therefore yielded shorter scaffolds with lower coverage. Shorter sequences are more susceptible to read coverage and compositional skew, and this is likely the reason for the reduced specificity. We have also demonstrated that our process was unable to detect only 3% or less of sequence present within the set of scaffolds under consideration (i.e., >2 kb). These missed detections may be due to either constrained gene sequences skewing compositional analysis or sampling bias skewing read coverage. Overall, these results demonstrate that binning by differential coverage and nucleotide composition, supplemented by evaluation of CSCG content, results in highly accurate genome bins across organisms ranging over 2 orders of magnitude in coverage levels (2,300×, *Idiomarinaceae* bacterium HL-53; 9×, *A. marincola* strain HL-49).

To date, UCC-A and UCC-O are the most complex communities for which a comprehensive species-resolved genomic data set exists. Until now, genome reconstruction techniques have been applied almost exclusively toward investigation of particular species within a community, rather than attempts to describe a community *in toto*. While this can be critical to understanding the function and role of the specific organism in its environment, in order to understand the principles governing community formation and dynamics, the specific interactions between different members, and the impact of community function upon the environment, a complete foundational knowledge of the genomic potential of each member organism is necessary. Such information is critical to the interpretation of other omics technologies that are now being applied to environmental samples, such as transcriptomics and proteomics. This allows the identification of which species is performing which function within the community, enabling investigation into how energy and nutrients enter a system and flow between community members. Understanding the contribution of microdiversity to community functional responses to

shifts in environmental conditions requires a genome-resolved understanding of function.

Organismal diversity remains the major hurdle to comprehensive reconstruction of complex natural communities. Species richness in most natural communities is on the order of hundreds to thousands of organisms (49). Microdiversity, the presence of closely related strains that share high average nucleotide identity yet have distinct physiology, can confound both assemblers and binning protocols and thus make a species-resolved understanding of community function a near impossibility. Here we have evaluated a strategy to investigate the complex interactions that drive community formation and dynamics by cultivating and sequencing parallel consortia. This approach has the benefit of simplifying overall community complexity to a tractable level and also, in some cases, selecting for a single variant of multiple closely related species that would otherwise impede assembly and binning. This helps bring us closer to fulfilling the promise of metagenomics—the ability to gain a species-resolved ecological understanding of communities directly from environmental samples—and serves as a useful intermediate between environmental metagenomics and single-cell amplified genomics.

ACKNOWLEDGMENTS

We thank Jessica Cole for laboratory assistance and Jim Fredrickson for critical review of the manuscript during preparation. We also acknowledge the U.S. Bureau of Land Management, Wenatchee Field Office, for their assistance in authorizing the research and providing access to the Hot Lake Research Natural Area.

FUNDING INFORMATION

U.S. Department of Energy (DOE) provided funding to William C. Nelson under grant number 56812.

This work was supported by the U.S. Department of Energy (DOE) Genome Sciences Program (GSP), Office of Biological and Environmental Research (OBER), and is a contribution of the Pacific Northwest National Laboratory (PNNL) Foundational Scientific Focus Area. Sequencing was done at the DOE Joint Genome Institute under contract no. DE-AC02-05CH11231 and Community Science Project 701. This work was part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy.

REFERENCES

1. **Moore LR, Rocap G, Chisholm SW.** 1998. Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. Nature **393:**464–467. http://dx.doi.org/10.1038/30965.
2. **Jaspers E, Overmann J.** 2004. Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. Appl Environ Microbiol **70:**4831–4839. http://dx.doi.org/10.1128/AEM.70.8.4831-4839.2004.
3. **Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL, Carmel G, Tummino PJ, Caruso A, Uria-Nickelsen M, Mills DM, Ives C, Gibson R, Merberg D, Mills SD, Jiang Q, Taylor DE, Vovis GF, Trust TJ.** 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. Nature **397:**176–180. http://dx.doi.org/10.1038/16495.
4. **Haverkamp TH, Schouten D, Doeleman M, Wollenzien U, Huisman J, Stal LJ.** 2009. Colorful microdiversity of Synechococcus strains (picocyanobacteria) isolated from the Baltic Sea. ISME J **3:**397–408. http://dx.doi.org/10.1038/ismej.2008.118.
5. **Pope PB, Smith W, Denman SE, Tringe SG, Barry K, Hugenholtz P, McSweeney CS, McHardy AC, Morrison M.** 2011. Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. Science **333:**646–648. http://dx.doi.org/10.1126/science.1205760.
6. **Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428:**37–43. http://dx.doi.org/10.1038/nature02340.
7. **Namiki T, Hachiya T, Tanaka H, Sakakibara Y.** 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res **40:**e155. http://dx.doi.org/10.1093/nar/gks678.
8. **Peng Y, Leung HC, Yiu SM, Chin FY.** 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28:**1420–1428. http://dx.doi.org/10.1093/bioinformatics/bts174.
9. **Wu YW, Rho M, Doak TG, Ye YZ.** 2012. Stitching gene fragments with a network matching algorithm improves gene assembly for metagenomics. Bioinformatics **28:**i363–i369. http://dx.doi.org/10.1093/bioinformatics/bts388.
10. **Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH.** 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol **31:**533–538. http://dx.doi.org/10.1038/nbt.2579.
11. **Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF.** 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science **337:**1661–1665. http://dx.doi.org/10.1126/science.1224041.
12. **Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV.** 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science **335:**587–590. http://dx.doi.org/10.1126/science.1212665.
13. **Patil KR, Roune L, McHardy AC.** 2012. The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. PLoS One **7**(6)**:**e38581. http://dx.doi.org/10.1371/journal.pone.0038581.
14. **Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW.** 2014. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ **2:**e603. http://dx.doi.org/10.7717/peerj.603.
15. **Pedros-Alio C.** 2012. The rare bacterial biosphere. Annu Rev Marine Sci **4:**449–466. http://dx.doi.org/10.1146/annurev-marine-120710-100948.
16. **Hanke A, Hamann E, Sharma R, Geelhoed JS, Hargesheimer T, Kraft B, Meyer V, Lenk S, Osmers H, Wu R, Makinwa K, Hettich RL, Banfield JF, Tegetmeyer HE, Strous M.** 2014. Recoding of the stop codon UGA to glycine by a BD1-5/SN-2 bacterium and niche partitioning between Alpha- and Gammaproteobacteria in a tidal sediment microbial community naturally selected in a laboratory chemostat. Front Microbiol **5:**231. http://dx.doi.org/10.3389/fmicb.2014.00231.
17. **Lindemann SR, Moran JJ, Stegen JC, Renslow RS, Hutchison JR, Cole JK, Dohnalkova AC, Tremblay J, Singh K, Malfatti SA, Chen F, Tringe SG, Beyenal H, Fredrickson JK.** 2013. The epsomitic phototrophic microbial mat of Hot Lake, Washington: community structural responses to seasonal cycling. Front Microbiol **4:**323. http://dx.doi.org/10.3389/fmicb.2013.00323.
18. **Cole JK, Hutchison JR, Renslow RS, Kim YM, Chrisler WB, Engelmann HE, Dohnalkova AC, Hu D, Metz TO, Fredrickson JK, Lindemann SR.** 2014. Phototrophic biofilm assembly in microbial-mat-derived unicyanobacterial consortia: model systems for the study of autotroph-heterotroph interactions. Front Microbiol **5:**109. http://dx.doi.org/10.3389/fmicb.2014.00109.
19. **Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, et al.** 2009. Real-time DNA sequencing from single polymerase molecules. Science **323:**133–138. http://dx.doi.org/10.1126/science.1162986.
20. **Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J.** 2013. Nonhybrid, finished microbial genome assemblies from long-read

SMRT sequencing data. Nat Methods **10:**563–569. http://dx.doi.org/10.1038/nmeth.2474.
21. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:**2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.
22. **Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF.** 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biol **10:**R85. http://dx.doi.org/10.1186/gb-2009-10-8-r85.
23. **McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I.** 2007. Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods **4:**63–72. http://dx.doi.org/10.1038/nmeth976.
24. **Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods **9:**357–359. http://dx.doi.org/10.1038/nmeth.1923.
25. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.
26. **Wu YW, Tang YH, Tringe SG, Simmons BA, Singer SW.** 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome **2:**26. http://dx.doi.org/10.1186/2049-2618-2-26.
27. **Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11:**119. http://dx.doi.org/10.1186/1471-2105-11-119.
28. **Wu M, Scott AJ.** 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics **28:**1033–1034. http://dx.doi.org/10.1093/bioinformatics/bts079.
29. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
30. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol **28:**2731–2739. http://dx.doi.org/10.1093/molbev/msr121.
31. **Jones DT, Taylor WR, Thornton JM.** 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci **8:**275–282.
32. **Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T.** 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature **499:**431–437. http://dx.doi.org/10.1038/nature12352.
33. **Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT.** 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics **27:**1691–1692. http://dx.doi.org/10.1093/bioinformatics/btr174.
34. **Delcher AL, Salzberg SL, Phillippy AM.** 2003. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics **Chapter 10:**Unit 10.3. http://dx.doi.org/10.1002/0471250953.bi1003s00.
35. **Pop M.** 2009. Genome assembly reborn: recent computational challenges. Brief Bioinform **10:**354–366. http://dx.doi.org/10.1093/bib/bbp026.
36. **Charuvaka A, Rangwala H.** 2011. Evaluation of short read metagenomic assembly. BMC Genomics **12**(Suppl 2)**:**S8.
37. **Kim M, Oh HS, Park SC, Chun J.** 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol **64:**346–351. http://dx.doi.org/10.1099/ijs.0.059774-0.
38. **Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF.** 2007. Genome dynamics in a natural archaeal population. Proc Natl Acad Sci U S A **104:**1883–1888. http://dx.doi.org/10.1073/pnas.0604851104.
39. **Gause GF.** 1934. The struggle for existence. Williams & Wilkins, Baltimore, MD.
40. **Acinas SG, Haverkamp TH, Huisman J, Stal LJ.** 2009. Phenotypic and genetic diversification of Pseudanabaena spp. (cyanobacteria). ISME J **3:**31–46. http://dx.doi.org/10.1038/ismej.2008.78.
41. **Gruber-Dorninger C, Pester M, Kitzinger K, Savio DF, Loy A, Rattei T, Wagner M, Daims H.** 2015. Functionally relevant diversity of closely related Nitrospira in activated sludge. ISME J **9:**643–655. http://dx.doi.org/10.1038/ismej.2014.156.
42. **Verbeke TJ, Dumonceaux TJ, Wushke S, Cicek N, Levin DB, Sparling R.** 2011. Isolates of Thermoanaerobacter thermohydrosulfuricus from decaying wood compost display genetic and phenotypic microdiversity. FEMS Microbiol Ecol **78:**473–487. http://dx.doi.org/10.1111/j.1574-6941.2011.01181.x.
43. **Hibbing ME, Fuqua C, Parsek MR, Peterson SB.** 2010. Bacterial competition: surviving and thriving in the microbial jungle. Nat Rev Microbiol **8:**15–25. http://dx.doi.org/10.1038/nrmicro2259.
44. **Sandoz KM, Mitzimberg SM, Schuster M.** 2007. Social cheating in Pseudomonas aeruginosa quorum sensing. Proc Natl Acad Sci U S A **104:**15876–15881. http://dx.doi.org/10.1073/pnas.0705653104.
45. **Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C.** 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol **31:**814–821. http://dx.doi.org/10.1038/nbt.2676.
46. **Denef VJ, Kalnejais LH, Mueller RS, Wilmes P, Baker BJ, Thomas BC, VerBerkmoes NC, Hettich RL, Banfield JF.** 2010. Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. Proc Natl Acad Sci U S A **107:**2383–2390. http://dx.doi.org/10.1073/pnas.0907041107.
47. **Wilmes P, Simmons SL, Denef VJ, Banfield JF.** 2009. The dynamic genetic repertoire of microbial communities. FEMS Microbiol Rev **33:**109–132. http://dx.doi.org/10.1111/j.1574-6976.2008.00144.x.
48. **Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, Thomas BC, Banfield JF.** 2013. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. mBio **4:**e00708-13. http://dx.doi.org/10.1128/mBio.00708-13.
49. **Torsvik V, Ovreas L, Thingstad TF.** 2002. Prokaryotic diversity—magnitude, dynamics, and controlling factors. Science **296:**1064–1066. http://dx.doi.org/10.1126/science.1071698.
50. **Thorvaldsdottir H, Robinson JT, Mesirov JP.** 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform **14:**178–192. http://dx.doi.org/10.1093/bib/bbs017.
51. **Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, Hamamura N, Melendrez MC, Bateson MM, Ward DM, Heidelberg JF.** 2007. Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J **1:**703–713. http://dx.doi.org/10.1038/ismej.2007.46.