Title

Rokubacteria: Genomic Giants among the Uncultured Bacterial Phyla

Journal

Frontiers in Microbiology, 8(NOV)

ISSN

1664-302X

Authors

Becraft, Eric D

Woyke, Tanja

Jarett, Jessica

et al.

Publication Date

2017

DOI

10.3389/fmicb.2017.02264

Peer reviewed

frontiers in Microbiology

ORIGINAL RESEARCH
published: 28 November 2017
doi: 10.3389/fmicb.2017.02264

Rokubacteria: Genomic Giants among the Uncultured Bacterial Phyla

*Eric D. Becraft[1], Tanja Woyke[2], Jessica Jarett[2], Natalia Ivanova[2], Filipa Godoy-Vitorino[3], Nicole Poulton[1], Julia M. Brown[1], Joseph Brown[1], M. C. Y. Lau[4], Tullis Onstott[4], Jonathan A. Eisen[5], Duane Moser[6] and Ramunas Stepanauskas[1]**

[1] Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, United States, [2] Joint Genome Institute, Walnut Creek, CA, United States, [3] Department of Natural Sciences, Inter American University of Puerto Rico, San Juan, Puerto Rico, [4] Department of Geosciences, Princeton University, Princeton, NJ, United States, [5] College of Biological Sciences, Genome Center, University of California, Davis, Davis, CA, United States, [6] Desert Research Institute, Las Vegas, NV, United States

OPEN ACCESS

Edited by:
Frank T. Robb, University of Maryland, Baltimore, United States

Reviewed by:
David L. Bernick, University of California, Santa Cruz, United States
Brian P. Hedlund, University of Nevada, Las Vegas, United States
Marla Trindade, University of the Western Cape, South Africa

***Correspondence:**
Ramunas Stepanauskas
rstepanauskas@bigelow.org

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *14 August 2017*
Accepted: *02 November 2017*
Published: *28 November 2017*

Citation:
Becraft ED, Woyke T, Jarett J, Ivanova N, Godoy-Vitorino F, Poulton N, Brown JM, Brown J, Lau MCY, Onstott T, Eisen JA, Moser D and Stepanauskas R (2017) Rokubacteria: Genomic Giants among the Uncultured Bacterial Phyla. Front. Microbiol. 8:2264. doi: 10.3389/fmicb.2017.02264

Recent advances in single-cell genomic and metagenomic techniques have facilitated the discovery of numerous previously unknown, deep branches of the tree of life that lack cultured representatives. Many of these candidate phyla are composed of microorganisms with minimalistic, streamlined genomes lacking some core metabolic pathways, which may contribute to their resistance to growth in pure culture. Here we analyzed single-cell genomes and metagenome bins to show that the "Candidate phylum Rokubacteria," formerly known as SPAM, represents an interesting exception, by having large genomes (6–8 Mbps), high GC content (66–71%), and the potential for a versatile, mixotrophic metabolism. We also observed an unusually high genomic heterogeneity among individual Rokubacteria cells in the studied samples. These features may have contributed to the limited recovery of sequences of this candidate phylum in prior cultivation and metagenomic studies. Our analyses suggest that Rokubacteria are distributed globally in diverse terrestrial ecosystems, including soils, the rhizosphere, volcanic mud, oil wells, aquifers, and the deep subsurface, with no reports from marine environments to date.

Keywords: microbial ecology, microbial evolution, uncultivated bacteria, microbial dark matter, microbial genomics

INTRODUCTION

Technological innovations in single-cell genomics and metagenomics have led to a rapid improvement in our understanding of the genomic features, evolutionary histories and metabolic capabilities of tens of phylum-level branches of Archaea, Bacteria and Eukarya that lack cultured representatives (Yoon et al., 2011; Rinke et al., 2013; Becraft et al., 2015; Brown et al., 2015; Castelle et al., 2015). In these efforts, the subsurface has emerged as a bountiful reservoir of undiscovered, deeply branching microbial lineages that may hold clues to the emergence and evolution of life on our planet (Kallmeyer et al., 2012; Colwell and D'Hondt, 2013). Many of the recently discovered candidate phyla are composed of microorganisms with small genomes lacking some core metabolic pathways, which may be a factor contributing to the inability to obtain pure cultures of these organisms (Rinke et al., 2013; Becraft et al., 2015; Brown et al., 2015; Castelle et al., 2015). This apparent genomic reduction has given rise to hypotheses of genome-streamlining, parasitism,

symbiotic lifestyles, and large-scale community metabolic interdependence (Giovannoni et al., 2014; Castelle et al., 2015; Anantharaman et al., 2016).

Our preliminary findings from several subsurface environments indicated that the "Candidate phylum Spring Alpine Meadow" (SPAM) constitute an intriguing exception to genome streamlining in oligotrophic environments. The existence of this lineage was first suggested by several 16S rRNA gene sequences obtained in 2004 from an alpine soil from the Colorado Rocky Mountains (Lipson and Schmidt, 2004). Subsequently, related 16S rRNA gene sequences were identified on all continents except for Antarctica in environments such as crop soils (Hansel et al., 2008; Chen et al., 2012; Figuerola et al., 2015), copper mine soil (Rodrigues et al., 2014), the subsurface oxic sediments of Hanford Formation at Pacific Northwest National Laboratory (PNNL) (Lin et al., 2012), as well as volcanic mud and oil wells (unpublished). Our phylogenetic analyses suggested that the more recently described "Candidate phylum Rokubacteria," represented by several metagenome bins from the Rifle site, constitutes the same lineage as SPAM, although this is not noted in the original Rokubacteria publications (Anantharaman et al., 2016; Hug et al., 2016). Here we present genomic sequences from 19 individual Rokubacteria (formerly SPAM) cells from aquifers of different depths in Nevada, South Dakota and South Africa. We compare genomic data from these individual cells to Rokubacteria metagenome bins from Nevada groundwater, a *Tabebuia* rhizosphere in Puerto Rico, and a prior study of the Rifle DOE Scientific Focus Area (SFA) in Colorado (Anantharaman et al., 2016), where the first metagenome bins of these organisms were obtained. This first global 16S rRNA gene survey of Rokubacteria suggests that they comprise a monophyletic, phylum-level lineage that is most closely related to Nitrospirae. Different from the Nitrospirae, Rokubacteria genomes are consistently large, with high %GC and the potential for a mixotrophic metabolism, all packaged within small cells. Rokubacteria cells are also characterized by an unusually high genomic heterogeneity among individuals, with no environments identified to date with near-clonal populations. The unique combination of large genomes encoding the ability for a generalist metabolic strategy in oligotrophic environments contained within small cells is a rare observation among the recent explosion of candidate phyla characterization (Castelle et al., 2015; Anantharaman et al., 2016; Hug et al., 2016). High level of genetic heterogeneity among Rokubacteria individuals in studied samples is another intriguing feature that may present certain challenges to future studies.

MATERIALS AND METHODS

Field Sample Collection

Shallow aquifer water samples were collected from a groundwater evaluation well in Nye CO, Nevada, USA, named "Oasis Valley 2," hereafter referred to as "OV-2," on 14 December, 2014 (36.96°N, −116.72°W). OV-2 is a 4″ PVC-cased hole that was drilled in 2011 to a depth of 36.5 m in Tertiary and Quaternary age alluvial sand and gravel derived from nearby Tertiary volcanics. The well is screened (i.e., perforations were cut into the casing through which water can enter, but sand and other aquifer materials do not) over the interval from 9.1 to 27.4 m. Samples OV-2 P1, P2, and P3 were collected after removal of 1, 3, and 10 well volumes at a pumping rate of ∼300 L/min. Microbial biomass was collected on 0.2 μm polyethersulfone membrane filters (Millipore, Sterivex) from one, three, and five liters of samples at time points OV-2 P1, OV-2 P2, OV-2 P3, respectively.

Discharge water samples were collected from Crystal Spring, which is located adjacent to Death Valley, CA, USA, on 13 December, 2014 (36.42°N, −116.72°W). Producing ∼10,600 L per min, Crystal is the largest spring of the largest oasis of the Mojave Desert, Ash Meadows, Nye CO, NV, USA. It is located within the discharge zone for a regional aquifer hosted within the highly fractured Paleozoic carbonates of the Death Valley Regional Flow System (DVRFS) (Belcher et al., 2009).

Subsurface water samples were collected from water at the Sanford Underground Research Facility (SURF, formerly the Homestake Mine) at 91.4 meters below land surface (mbls) in Lead, South Dakota, on 12 December, 2014 (44.35°N, −103.75°W). SURF samples were collected from perennial wall seeps associated with century-old horizontal legacy drifts in metamorphic rock. Subsurface water samples were also collected from a borehole at Finsch Mine at a depth of 857 mbls in South Africa on 11 November, 2012 (−28.38°S, 23.45°E).

All aquatic samples were collected aseptically from flowing pumped lines (submersible and peristaltic, at OV-2 and Crystal Spring, respectively) or directly from the source (SURF and Finsch). For single-cell genomics, one-milliliter aliquots were amended with 5% glycerol and 1x TE buffer (all final concentrations), frozen on dry ice in the field and stored at −80°C until further processing.

For metagenomics, the DNA from OV-2 samples was extracted from microbial biomass collected on 0.2 μm polyethersulfone membrane filters (Millipore, Sterivex) using the MO BIO PowerSoil DNA Isolation Kit (MO BIO Laboratories Inc., Carlsbad, CA) according to the manufacturer's protocol. An additional freeze/thaw cycle was included after the addition of solution C1 and immediately prior to the 10-min vortex step (30 min at −80°C followed by 10 min at 65°C). Additionally, a sample for metagenome sequencing was collected from a *Tabebuia* (*T. heterophylla*) rhizosphere in the serpentine area of Cabo Rojo Puerto Rico on 12 March, 2013. Three secondary roots from one tree, about 15–20 cm in length were collected, cut, stored in a 50 mL polyethylene centrifuge tube and transported on ice to the laboratory. The rhizosphere samples were obtained by washing the roots with 25 mL 1X PBS/Tween20 and shaking at 240 rpm horizontally for 1 h, and frozen at −80°C. The PBS/Tween20 solution with the rhizosphere was centrifuged at 9,000 × g for 20 min at 4°C. Genomic DNA was extracted from the resulting pellet using the MO BIO PowerSoil DNA Isolation Kit with bead tubes (Carlsbad, CA) following Earth Microbiome Project standard protocols (http://www.earthmicrobiome.org/protocols-and-standards/). Cells for single-cell genomics were not collected from the *Tabebuia* rhizosphere sample. Site images and the physicochemical characteristics of these field samples are reported in Supplemental Figure 2 and Supplemental Table 1.

Single-Cell Genomics

The generation, identification, sequencing and *de novo* assembly of single amplified genomes (SAGs) was performed at the Bigelow Laboratory Single-Cell Genomics Center (scgc.bigelow.org). The cryopreserved samples were thawed, pre-screened through a 40 μm mesh size cell strainer (Becton Dickinson) and incubated with the SYTO-9 DNA stain (Thermo Fisher Scientific) for 10–60 min. Fluorescence-activated cell sorting (FACS) was performed using a BD InFlux Mariner flow cytometer equipped with a 488 nm laser and a 70 μm nozzle orifice (Becton Dickinson, formerly Cytopeia). The cytometer was triggered on side scatter, and the "single-1 drop" mode was used for maximal sort purity. The sort gate was defined based on particle green fluorescence, light side scatter, and the ratio of green vs. red fluorescence (for improved discrimination of cells from detrital particles). For each sample, individual cells were deposited into 384-well plates containing 600 nL per well of 1x TE buffer and stored at −80°C until further processing. Of the 384 wells, 317 wells were dedicated for single particles, 64 wells were used as negative controls (no droplet deposition), and 3 wells received 10 particles each to serve as positive controls. Index sort data was collected using the BD FACS Sortware software. The DNA for each cell was amplified using WGA-X, as previously described in Stepanauskas et al. (2017). Cell diameters were determined using the FACS light forward scatter signal, which was calibrated against cells of microscopy-characterized laboratory cultures (Stepanauskas et al., 2017).

Illumina libraries were created, sequenced and assembled as previously described (Stepanauskas et al., 2017). This workflow was evaluated for assembly errors using three bacterial benchmark cultures with diverse genome complexity and %GC, indicating 60% average genome recovery, no non-target and undefined bases, and average frequencies of misassemblies, indels and mismatches per 100 kbp: 1.5, 3.0, and 5.0 (Stepanauskas et al., 2017). CheckM v1.0.6 (Parks et al., 2015) was used to calculate completeness of assemblies of environmental SAGs, which relies on single conserved marker genes, and genome size was estimated (assembly size divided by estimated genome completeness). We did not co-assemble SAGs due to the high genomic heterogeneity among individual cells. All SAGs were deposited in the Integrated Microbial Genomes database at the Joint Genome Institute (Supplemental Table 7).

The 16S rRNA gene sequences were aligned using SINA alignment software (Pruesse et al., 2012). Phylogenetic trees were inferred by MEGA 6.0 (Tamura et al., 2013) using the General TimeReversible (GTR) Model, with Gamma distribution with invariable sites (G+I), and 95% partial deletion for 1,000 replicate bootstraps. SAG assemblies were analyzed for protein-encoding regions using RAST (http://rast.nmpdr.org/) (Aziz et al., 2008), and genes (protein families) were annotated with Koala (KEGG) (http://www.kegg.jp/ghostkoala/) (Kanehisa et al., 2016) and InterProScan v5 (Jones et al., 2014). Average nucleotide identity (ANI) and average amino acid identity (AAI) of reciprocal hits were calculated using the online tools at the Kostas Lab website Environmental Microbial Genomics Laboratory (http://enve-omics.ce.gatech.edu/aai/) (Goris et al., 2007; Rodriguez and Konstantinidis, 2014). Synteny plots were produced using the Joint Genome Institute Integrated Microbial Genomes (IMG) system (https://img.jgi.doe.gov/) (Markowitz et al., 2014). Phage genes and transposases were identified as in Labonté et al. (2015b).

Metagenomic Sequencing and Analysis

For OV-2 samples, 1 ng of DNA was fragmented and adapter ligated using the Nextera XT kit (Illumina). The ligated DNA fragments were enriched with 12 cycles of PCR and purified using SPRI beads (Beckman Coulter). For the *Tabebuia* rhizosphere sample, 100 ng of DNA was sheared to 300 bp using the Covaris LE220 and size selected using SPRI beads (Beckman Coulter). The fragments were treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Inc) using the KAPA-Illumina library creation kit (KAPA biosystems). For both OV-2 and *Tabebuia* rhizosphere metagenomes, qPCR was used to determine the concentration of the libraries, and libraries were sequenced on an Illumina HiSeq. Metagenome reads were quality trimmed and filtered using rqcfilter tool from bbtools package (http://jgi.doe.gov/data-and-tools/bbtools/), which performs primer and adapter removal, trims reads to the quality of 10, and removes PhiX and human sequences. The resulting reads were error-corrected using BFC tool (https://github.com/lh3/bfc.git) (Li, 2015) with kmer length of 25 and removing reads containing unique kmers. The resulting filtered and error-corrected reads were assembled for each sample separately using SPAdes v.3.9.0 without error correction with kmers 27, 47, 67, 87, and 107 (Bankevich et al., 2012). Reads were mapped to the assemblies using Burrows-Wheel Aligner (BWA) v0.7.15 (Li and Durbin, 2010) and binned based on abundance patterns and kmer composition using MetaBAT v0.32.4 with minimum contig length of 3 kb and superspecific probability option (Kang et al., 2015). Differential coverage could not be utilized as there was little overlap between the 3 OV-2 samples (i.e., less than 10% of the reads from P1 and P2 could be mapped to P3, and vice versa). The bins corresponding to Rokubacteria were identified based on the presence of Rokubacteria 16S rRNA genes on contigs longer than 20 kb, as well as best BLAST hits to Rokubacteria SAG assemblies (Altschul et al., 1990). Additional Rokubacteria metagenome bins were identified by BLASTing annotated gene regions of SAGs against metagenome assemblies, and bins with ≥200 hits with ≤ 1e-50 e-value score were further analyzed with CheckM v1.0.6. Metagenome assemblies are deposited in the Integrated Microbial Genomes database at the Joint Genome Institute (Supplemental Table 7).

Recruitment of metagenome reads to single-amplified genomes (SAGs) was determined using in-house software and Burrows-Wheel Aligner (BWA) v0.7.15 (Li and Durbin, 2010) to map sequence reads to Rokubacteria SAG contigs that met the criteria of 100 bps overlap at ≥90% nucleotide identity. The relative abundance of SAG relatives was determined as the fraction of metagenome reads mapping per megabase of a reference genome.

RESULTS AND DISCUSSION

16S rRNA Gene Phylogeny and Biogeography

We used full-length 16S rRNA gene sequences of Rokubacteria SAGs as queries in BLASTn searches for related sequences in the NCBI nucleotide database that yielded 91 unique sequences with ≥85% nucleotide identity and ≥600 bps. A phylogenetic analysis of these sequences suggested that Rokubacteria form a strongly bootstrap-supported, monophyletic lineage (**Figure 1**). Nitrospirae was the most closely related phylum, sharing 79–83% 16S rRNA gene sequence identity with Rokubacteria. Some Rokubacteria 16S rRNA gene sequences were misclassified as Nitrospirae in public databases (green arrows in **Figure 1**, also see Supplemental Figure 3). Rokubacteria form a bootstrap-supported, monophyletic clade separate from previously described phyla, contain unifying genomic features (e.g., GC content), and fall below the median phylum-level 16S rRNA gene similarity threshold of 83.68% (range 81.6–85.93%) (Yarza et al., 2014). Therefore, we support the designation of Rokubacteria as a unique phylum-level lineage, as previously suggested from phylogenies based on ribosomal protein sequences, which agree with this phylogenetic placement (Anantharaman et al., 2016; Hug et al., 2016).

The Rokubacteria 16S rRNA gene sequences form two deeply branching sub-lineages that diverge from each other by ∼12–15%, i.e., at an operationally-defined class level (**Figure 1**; Supplemental Table 2) (Hugenholtz et al., 1998; Yarza et al., 2014). Apart from SAGs and PCR-derived sequences, one of the sub-lineages also included 16S rRNA genes from metagenome bins obtained from the Puerto Rican *Tabebuia* rhizosphere (light blue square in Figure 1) and from a previously published bin from the Rifle site, Colorado (orange square in **Figure 1**; Anantharaman et al., 2016). We propose naming this lineage, which encompassed the majority of Rokubacteria sequences originating from both soils and terrestrial subsurface environments to "Candidatus class Rokumicrobia." Another major lineage included a smaller set of sequences that originate exclusively from terrestrial subsurface sites. We propose naming this lineage "Candidatus class Infratellusbacteria," with reference to "infra" and "tellus" (Latin for "below" and "Earth"), hereby referred to as Infratellusbacteria, in order to reflect the predominant environment in which these microorganisms have been detected so far.

The sources of samples from which Rokubacteria 16S rRNA gene sequences were retrieved [25 in total; including 19 previously sampled sites (**Figure 1**)], suggest a cosmopolitan distribution in soils and terrestrial subsurface, with no evidence to date for presence in marine environments. Interestingly, Rokubacteria were low in abundance at almost every site where they were identified in this and prior studies (Lin et al., 2012; Figuerola et al., 2015), and often were represented by a single 16S rRNA gene sequence. An alternative analysis of Rokubacteria abundance in our study sites, by performing metagenome fragment recruitment on SAGs as references, provided further evidence that Rokubacteria comprised ∼1% of the microbial community in OV-2 (Supplemental Figure 4), similar to other samples (Lipson and Schmidt, 2004; Hansel et al., 2008; Chen et al., 2012; Lin et al., 2012; Rodrigues et al., 2014; Figuerola et al., 2015). A recent study identified Rokubacteria to constitute ∼10% of the microbial community in a grass root zone in the Angelo Coast Range Reserve, California, making it the most Rokubacteria-rich environment to date (Butterfield et al., 2016), though no 16S rRNA sequences were identified in the metagenome bins.

General Genome Features

The SAGs obtained from SURF, Finsch, OV-2 and Crystal Spring sites contained phylogenetically diverse representatives of both Rokubacteria classes, enabling us to explore their genomic content, metabolic potential and evolutionary histories. *De novo* genome assemblies of the 19 SAGs ranged from 0.05 to 2.86 Mbps (**Table 1**). The estimated Rokubacteria genome completeness ranged between 1 and 40% (average of 18.2%). This is significantly lower than the genome recovery from other SAGs using the same techniques in earlier studies, which averaged at around 50% (Rinke et al., 2013; Swan et al., 2013; Kashtan et al., 2014). Based on the presence of conserved single copy genes in the most complete SAG assemblies, we estimate that Rokubacteria complete genomes are 6–8 Mbps in length (average 6.8 Mbps; **Figure 2**, **Table 1**), which is slightly larger than estimates obtained from metagenome bins at the Rifle site (4–6 Mbps; Supplemental Table 3) (Anantharaman et al., 2016) and the Puerto Rican soil (**Table 1**). The CheckM-based genome size estimates from smaller SAG assemblies and contaminated metagenome bins were highly variable, while all estimates based on the more complete SAGs and metagenome bins converged on the average noted above (**Figure 2**). A relatively large fraction, between 8 and 17% of the Rokubacteria genomes, consists of nucleotides predicted to be non-coding. With a few intriguing exceptions (Sekiguchi et al., 2015), these features present a stark contrast to the predominantly small and streamlined genomes of most recently described bacterial and archaeal candidate phyla from diverse surface and subsurface environments, including the abundant and diverse candidate superphylum Patescibacteria (Rinke et al., 2013), which was later proposed to constitute an even larger evolutionary unit, the Candidate Phyla Radiation (CPR) (Brown et al., 2015; Castelle et al., 2015; Anantharaman et al., 2016).

The GC content of Rokubacteria SAG assemblies was at the high end of the reported spectrum for known organisms, ranging between 64 and 71%, with an average of 68% (**Figure 3**; **Table 1**). This is in agreement with the high % GC content of the Rokumicrobia metagenome bins reported by Anantharaman et al. (2016). The most closely related phylum to Rokubacteria, Nitrospirae, has a more variable GC content, ranging from 34% (*Thermodesulfovibrio islandicus*) to 62% (*Nitrospira moscoviensis*). The factors determining GC content remain unclear. The spontaneous mutations may favor nucleotide shifts to A and T (Hershberg and Petrov, 2010; Hildebrand et al., 2010), and the lower nitrogen content of AT may provide a selective advantage to low GC organisms in N-limited environments (Giovannoni et al., 2014). Yet, high %GC is present in a wide range of lineages and habitats (Hershberg

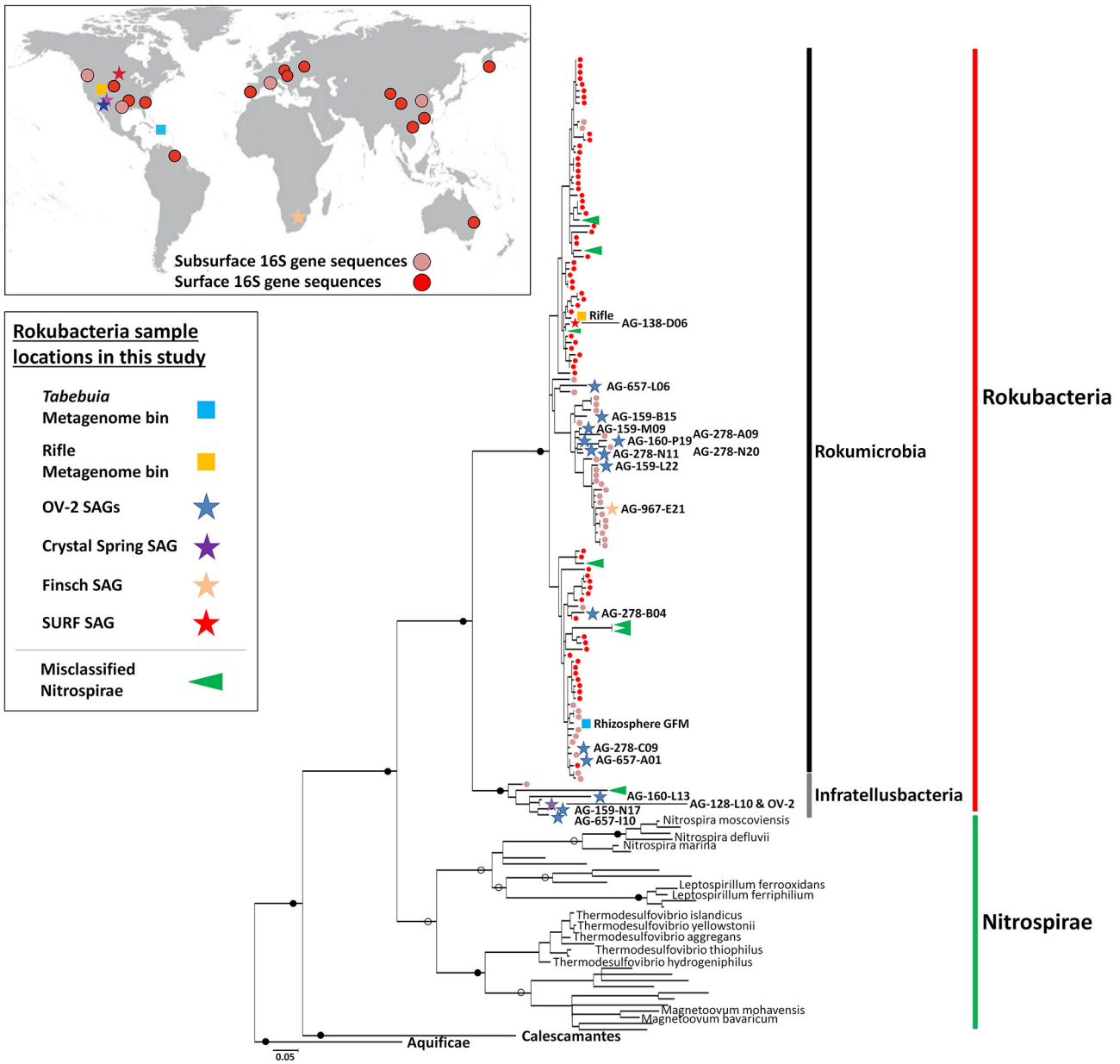

FIGURE 1 | Maximum-likelihood phylogeny of the Rokubacteria (red line), based on partial 16S rRNA gene sequences (~600 bps in length). Included are NCBI sequences with ≥85% nucleotide identity to Rokubacteria SAGs. The Rokumicrobia are demarcated by a black line, and the Infratellusbacteria are demarcated by a grey line. Nitrospirae genome 16S rRNA gene sequences representing classified genera and sequences misclassified as Nitrospirae are demarcated by a green vertical bar and green arrows, respectively. 16S rRNA gene sequence from previous surface (red circles), and subsurface (tan circles) studies are also indicated at the terminal branch of each sequence. Map insert (upper left) shows geographic distribution of reported Rokubacteria 16S rRNA gene sequences from past surface (red circles), and subsurface (tan circles) studies. Sequence identifiers are reported in Supplemental Figure 3. Previously sampled subsurface sites from Puerto Rico and Nevada where Rokubacteria 16S rRNA genes sequences were identified are not shown in Figure 1 insert due to space constants. Stars indicate Rokubacteria SAGs (blue = OV-2, purple = Crystal Spring; tan = Finsch mine; and red = SURF) and squares indicate metagenome bins (light blue = *Tabebuia*; orange = Rifle site), and color corresponds to site SAGs were isolated from (left; also see Supplemental Figure 2). All SAGs and metagenome bins that contained a 16S rRNA gene are included in the phylogeny. Full circles indicate bootstrap values >90%; open circles indicate boot-strap values >70%. Scale bar represents 0.05 nucleotide substitutions per site.

and Petrov, 2010). Factors determining high %GC remain controversial, with some studies suggesting the importance of temperature and solar radiation as selective variables (Foerstner et al., 2005; Hildebrand et al., 2010), while other reports refute these findings (Lassalle et al., 2015; Li et al., 2015). Furthermore, while some studies suggest GC content is evolutionarily

TABLE 1 | Rokubacteria genome assembly statistics and predicted completeness.

Rokubacteria SAGs	Class	Site	Raw PE reads (millions)	Assembly (Mbps)	GC%	16S rRNA+	Estimated Genome Completeness (%)	Predicted Genome Size-Mbps	% Contamination[d]	Genome quality[e]
*AD-967-E21	Rokumicrobia	Finsch	13.7	1.52	66	Yes	30	5.1	<1	Low
*AG-128-L10	Infratellusbacteria	CS[a]	13.2	1.89	70	Yes	24	7.9	<1	Low
AG-138-D06	Rokumicrobia	SURF[b]	12.0	0.29	67	Yes	6	4.9	0	Low
AG-657-A01	Rokumicrobia	OV-2	12.2	0.40	68	Yes	<1	NA	0	Low
*AG-657-I10	Infratellusbacteria	OV-2	10.8	0.93	71	Yes	<1	NA	0	Low
*AG-657-L06	Rokumicrobia	OV-2	7.7	1.03	65	Yes	<1	NA	0	Low
AG-159-B15	Rokumicrobia	OV-2	8.8	0.05	64	Yes	<1	NA	0	Low
AG-159-G23	Rokumicrobia	OV-2	7.6	0.40	64	No	5	8.1	0	Low
*AG-159-L22	Rokumicrobia	OV-2	8.4	0.69	65	Yes	16	4.3	0	Low
*AG-159-M09	Rokumicrobia	OV-2	7.5	0.93	65	Yes	4	22.2	0	Low
AG-159-N17	Infratellusbacteria	OV-2	9.8	0.41	67	Yes	<1	NA	0	Low
AG-159-P01	Rokumicrobia	OV-2	5.1	0.07	65	No	<1	NA	0	Low
AG-160-L13	Infratellusbacteria	OV-2	7.0	0.41	64	Yes	3	13.7	0	Low
*AG-160-P19	Rokumicrobia	OV-2	8.2	1.38	64	Yes	27	5.1	<1	Low
AG-278-A09	Rokumicrobia	OV-2	0.08	0.24	65	Yes	4	6.1	0	Low
*AG-278-B04	Rokumicrobia	OV-2	8.4	2.61	69	Yes	32	8.2	<1	Low
*AG-278-C09	Rokumicrobia	OV-2	8.5	1.68	68	Yes	18	9.21	0	Low
*AG-278-N11	Rokumicrobia	OV-2	7.4	1.15	67	Yes	15	7.8	0	Low
*AG-278-N20	Rokumicrobia	OV-2	0.16	2.86	67	Yes	40	7.2	0	Low
Rokubacteria metagenome bins										
*OV-2 bin8[c]	Infratellusbacteria	OV-2	-	5.69	72	Yes	89	6.4	2	Medium
*Rhizosphere bin[c]	Rokumicrobia	PR	-	4.01	69	Yes	63	6.4	1	Low
OV-2 bin1	Unknown	OV-2	-	11.84	62	No	100	11.8	332	NA
OV-2 bin2	Unknown	OV-2	-	11.68	69	No	100	11.6	175	NA
OV-2 bin6	Unknown	OV-2	-	6.06	69	No	73	8.3	59	NA
OV-2 bin9	Unknown	OV-2	-	5.05	71	No	78	6.5	34	NA
OV-2 bin11	Unknown	OV-2	-	4.68	68	No	98	4.8	153	NA
OV-2 bin43	Unknown	OV-2	-	1.32	72	No	14	9.5	1	Low

*Asterisks indicate more complete genome assemblies that were used in **Figure 4**.*
[a,b]Crystal Spring, Nevada and Sanford Underground Research Facility (300 m).
[c]Metagenome bins from OV-2 and Tabebuia rhizosphere sample taken in Puerto Rico.
[d]Estimated with CheckM.
[e]Genome quality reported according to Bowers et al. (2017).

conserved within lineages (Lassalle et al., 2015; Reichenberger et al., 2015), others show large GC variation among lineages that were thought to be exclusively high in GC, such as the Actinobacteria phylum (Ghai et al., 2012; Swan et al., 2013) and the Roseobacter clade of the Alphaproteobacteria (Swan et al., 2013; Zhang et al., 2016). The high %GC of Rokubacteria contrasts low %GC in most of the major, uncultured branches of Bacteria and Archaea explored with single-cell genomics (Rinke et al., 2013) and metagenome binning (Anantharaman et al., 2016; Hug et al., 2016) to date (**Figure 3**). It remains to be understood what evolutionary processes are involved in the emergence and maintenance of high %GC, and to what extent the discovery of novel microbial lineages with high %GC has been hampered by biases in DNA amplification (Stepanauskas et al., 2017) and sequencing techniques (Chen et al., 2013).

Rokubacteria SAG assemblies shared between 36 and 922 orthologous protein-encoding genes (average of 308 reciprocal orthologous protein hits). The average amino acid identity (AAI) was 46.2% (range from 34.6 to 64.2%; **Figure 4**), demonstrating high cell-to-cell genome divergence. Interestingly, SAGs originating from the OV-2 sample shared roughly the same proportion of protein-coding genes as SAGs from geographically distant sites. The most divergent Rokubacteria SAGs were obtained from the same OV-2 site (**Figure 1**), both within and between class-level lineages. Genomes were mostly non-syntenic on larger scales. However, many shared proteins of related function were located in small islands of synteny in the six least fragmented SAG assemblies (Supplemental Figure 5). Causes for the unusually variable genome content among cells in each study site remain unclear. Dispersal of dormant cells to the sampling

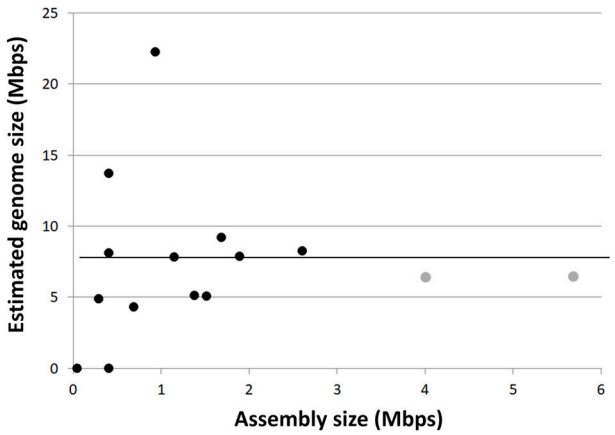

FIGURE 2 | Single-cell genome (black circles) and metagenome bin (gray circles) assembly sizes compared to CheckM estimated genome sizes. Horizontal line indicates average estimated genome size (~6.8 Mbps).

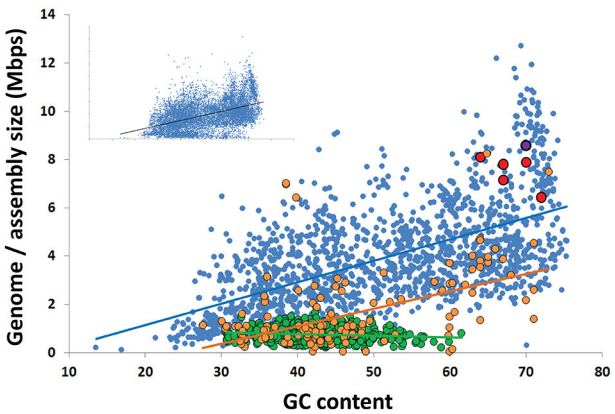

FIGURE 3 | Correlation between the G+C content and genome size among all finished bacterial genomes in IMG (blue; $R^2 = 0.35$), the 4 most complete Rokubacteria SAGs from OV-2 (red), metagenome bins from OV-2 bin 8 and Puerto Rican *Tabebuia* rhizosphere that contain a 16S rRNA gene (purple), and the Rifle metagenome bin that contains a 16S rRNA gene (black) (also see Supplemental Table 3). Also displayed are the estimated genome sizes for Candidate Phyla Radiation (CPR) genomes (green; $R^2 = 0.02$) and candidate phyla genomes not a part of the CPR lineage (orange; $R^2 = 0.25$). The insert contains all genomes in IMG, including all partial SAGs and metagenome bins from all bacterial phyla ($R^2 = 0.35$).

sites from a multitude of evolutionarily distant populations is one plausible explanation. An alternative explanation may be the accumulation of point mutations, gene acquisitions, gene loss and genome rearrangements at a rate that outpaces cell division. The latter possibility is highly speculative and contradicts conventional models of microbial evolution, but should be viewed in the context of bacterial generation times potentially ranging in hundreds and even thousands of years in some low-energy, subsurface environments (Labonté et al., 2015a).

Rokubacteria genomes contain numerous transposases and integrases (4–60 per SAG assembly; Supplemental Table 4). Genes of potential viral origin and CRISPR regions were also identified in most Rokubacteria SAGs (Supplemental Table 4). The contigs that contained phage-like genes were never found to be entirely viral, indicating prophage integration into host chromosomes. These observations are similar to the recent finding of abundant transposable prophages in Firmicutes in the deep subsurface of the Witwatersrand Basin (Labonté et al., 2015a) and indicate a potentially important role of viruses as vectors of horizontal gene transfer in low-energy, subsurface environments.

Predicted Phenotype and Energy Production

We employed forward light scatter (FSC) signals from FACS, which where calibrated against a series of benchmark cultures, to estimate approximate diameters of the cells from which SAGs were generated (Stepanauskas et al., 2017). This indicated that Rokubacteria cell diameters ranged between 0.3 and 0.4 μm (**Figure 5**). While this estimate is greater than the 0.15–0.20 μm diameter reported for some of the CPR cells (Luef et al., 2015), and the ~0.2–0.3 μm average diameter of the most abundant marine bacterioplankton lineage *Pelagibacter* (Giovannoni et al., 2005, 2014), it is approaching the theoretical lower limit for cell sizes (NRC, 1999). Such small cells, including the Patescibacteria, *Pelagibacter, Mycoplasma*, ultrasmall Actinomycetes, and *Prochlorococcus*, have small, streamlined genomes that range between 0.8 and 2.5 Mbps (Biller et al., 2014; Nakai et al., 2016; Parrott et al., 2016). In the case of Rokubacteria, the presence of large genomes in small cells may imply extensive DNA packaging or dormancy. In partial support of this hypothesis, a variety of DNA packaging and super coiling proteins were annotated in the Rokubacteria SAGs and metagenome bins (Supplementary Table 5). Further experimental work is required to confirm these predictions.

Rokubacteria contain numerous genes that are typical of Gram-negative (diderm) organisms, including the majority of genes involved in the production and transport of lipids across the cytoplasmic membrane for outer membrane and LPS assembly (Sutcliffe, 2010; Supplemental Figure 1 and Supplemental Table 5), which is consistent with their phylogenetic affiliation with the Gram-negative Nitrospirae. We identified multiple genes involved in twitching motility in 11 Rokumicrobia SAGs, 4 Infratellusbacteria SAGs, and both metagenome bins, possibly indicating a conserved mechanism of pili motility in the Rokubacteria (Supplemental Section 1 and Supplemental Table 5). We also identified genes in 3 Rokumicrobia SAGs and Infratellusbacteria OV-2 bin 8 that are predicted to encode flagella structural proteins, while propeller filament genes were absent in all SAG assemblies and metagenome bins (Supplemental Figure 1). While OV-2 bin 8 contained genes involved in flagella assembly, Infratellusbacteria SAGs lacked genes required for the assembly of flagella, though gene absence could be due to fewer and less complete SAG assemblies. Furthermore, putative genes were identified in the majority of assemblies for methyl-accepting chemotaxis proteins, two-component sensor kinases, ATP motor proteins,

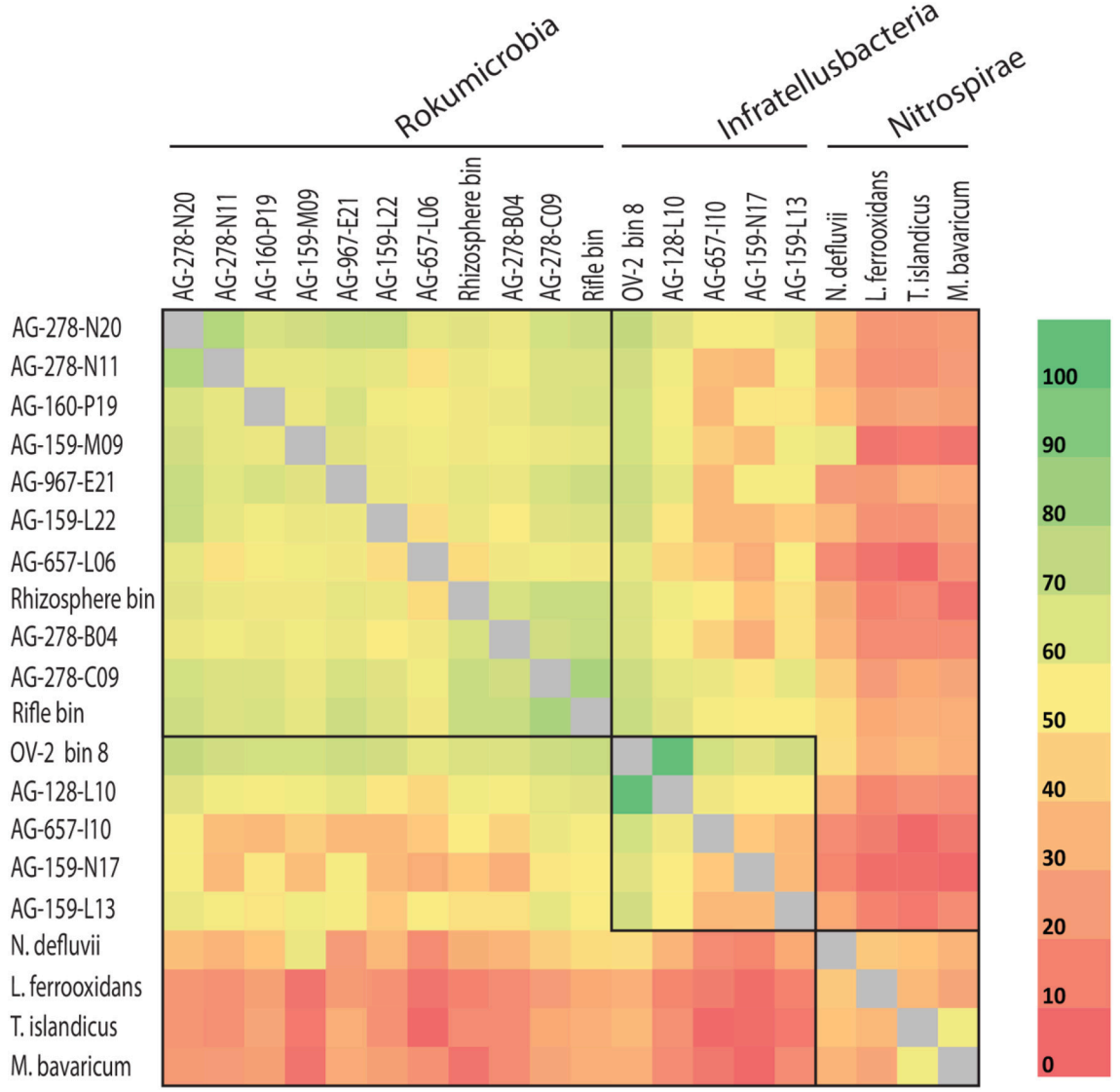

FIGURE 4 | Average amino acid identity (AAI) for shared proteins among more complete Rokubacteria single amplified genomes (SAGs) (>0.5 Mbps) and Rokubacteria metagenome bin 8 and rhizosphere bin (indicated by asterisks in Table 1), and select genomes from the Nitrospirae phylum. Boxes indicate phylogenetically defined class-level lineages (also see **Figure 1** and Supplemental Table 2).

and sensor proteins for nitrogen, oxygen, zinc/lead, and acetoacetate, indicating that Rokubacteria can respond to a broad range of chemical stimuli. Rokubacteria encode multiple carbon transport proteins, including those specializing in lipids, peptides and sugars. Rokubacteria encode for glycolysis, the TCA cycle, electron transport complexes required for aerobic respiration, and fermentative metabolisms (Supplemental Section 1 and Supplemental Figure 1). Rokumicrobia SAGs also contain genes involved in nitrogen respiration that could act as electron acceptors during anaerobic conditions, and nitrite oxidoreductases, which are universally conserved nitrification proteins in the Nitrospirae lineage (Supplemental Figure 6). These findings indicate that Rokubacteria can utilize diverse electron donors and acceptors under aerobic and anaerobic conditions (see Supplemental Section I for detailed metabolic predictions and discussion).

Critical Analysis of Single-Cell Genomes and Metagenome Bins

The availability of several partial genomes of Rokubacteria from single cells and metagenomes from this and prior studies (**Table 1** and Supplemental Table 3) offered an opportunity to compare the type and quality of information that can be extracted using these two approaches. Genome completeness is one important quality metric of *de novo* assemblies. The ratio of the number of single copy marker genes that are found vs. expected in an assembly is the most commonly used proxy for genome completeness and is

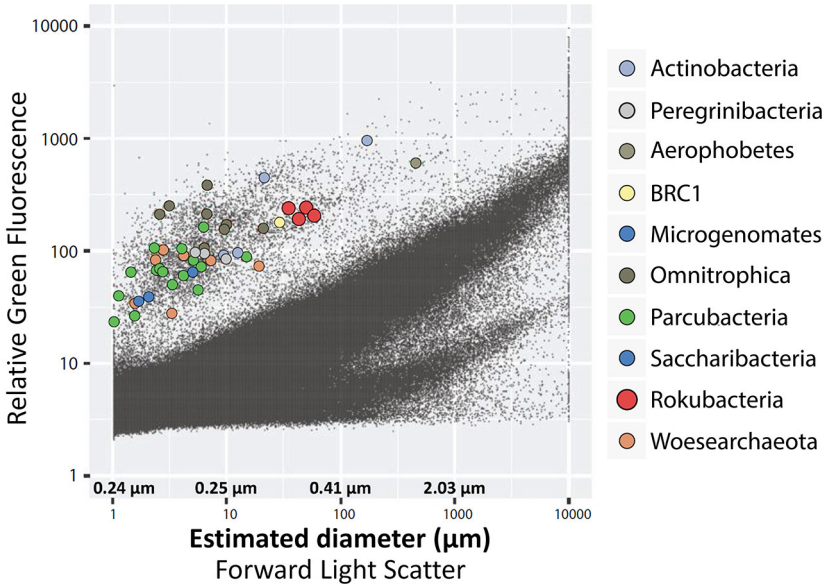

FIGURE 5 | Optical properties and estimated diameters of cells sorted from the OV-2 sample that contained the largest number of cells identified as Rokubacteria. Colored dots indicate cells that were successfully identified by their 16S rRNA gene. Black dots indicate all particles detected by the fluorescence-activated cell sorter.

implemented in popular computational tools, such as CheckM (Parks et al., 2015). In our study, CheckM-based estimates of assembly completeness of Rokubacteria SAGs and metagenome bins ranged between 1–40 and 14–100%, respectively, suggesting that individual SAG assemblies tended to be less complete than metagenome bins (**Table 1**). However, the higher average completeness of metagenome bins came with important caveats: high estimated contamination in five out of eight bins (**Table 1**), absence of rRNA genes in six out of eight bins, and the lack of knowledge of the number and genetic diversity of cells that contributed DNA sequences to each bin. These caveats may limit the interpretability of metagenome bins in the context of microbial ecology and evolution.

While the CheckM-based estimates of SAG contamination were always below 1%, they ranged between 1 and 332% (average 95%) in our OV-2 for metagenome bins (**Table 1**) and between 2 and 14% in bins of an earlier study (Anantharaman et al., 2016) (Supplemental Table 3), suggesting quality limitations of most bins (**Table 1**). These observations are in general agreement with the recent benchmarking effort employing > 1,000 previously sequenced strains of microorganisms and mobile genetic elements, which found that the performance of metagenome assembly and binning is impaired by the presence of related strains in a sample (Sczyrba, 2017).

The CheckM estimates of contamination are based on the phylogenetic placement of the assembly's single copy marker genes against a built-in database (Parks et al., 2015), which lacks many uncultured lineages, including Rokubacteria. To the best of our knowledge, the ability of CheckM to detect contamination that originates from lineages that are absent from its database has never been evaluated. To address this question, we created pairwise combinations of assemblies of each Infratellusbacteria SAG with each Rokumirobia SAG. The CheckM-estimated contamination in these combined assemblies was significantly smaller than the real, cross-class contamination (**Figure 6**), suggesting that CheckM may fail detecting contamination from lineages not represented in the CheckM database. Strikingly, the majority of our artificial combinations of SAGs from different phylogenetic classes would be considered "high quality" genomes according to the recently proposed genome standards for SAGs and metagenome bins (Bowers et al., 2017). In order to assess whether similar, cross-class contamination may be affecting our metagenome bins, we analyzed AAI among Rokubacteria SAGs and the only OV-2 metagenome bin that contained a rRNA gene (bin 8). While the rRNA gene placed this bin firmly among the Infratellusbacteria (**Figure 1**), its AAI suggested affiliation with Rokumicrobia (**Figure 4**). Furthermore, the best BLASTn hits to bin 8 genes consisted of SAGs from both class-level lineages, including multiple near full-length alignments at >95% nucleotide identity with Rokumicrobia SAGs (Supplemental Table 6). This indicates that the CheckM-based estimate of 2% contamination for this bin may be a major underestimate. These findings imply that improvements are urgently needed in the quality control of genome assemblies originating from uncultured microbial groups and in the validation of the performance of QC software.

The comparison of SAGs and metagenome bins demonstrates that the two approaches provide two fundamentally different types of data and should be interpreted accordingly. While SAG assemblies represent fragments of discrete genomes from individual cells, the metagenome bins are fragments of consensus sequences derived from a multitude of genetically non-identical organisms. The consistency of certain general features between Rokubacteria SAGs and bins (e.g., high %GC, large estimated

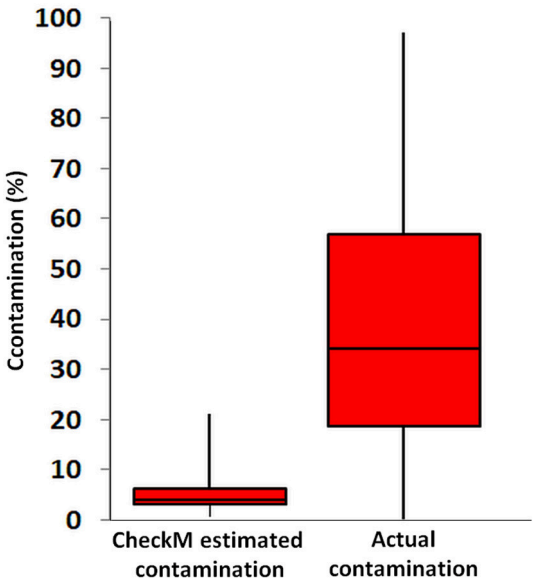

FIGURE 6 | Contamination predicted by CheckM software from all pairwise SAG combinations across the class-level lineages Rokumicrobia and Infratellusbacteria (15 Rokumicrobia and 4 Infratellusbacteria SAGs; 60 combinations total), compared to actual contamination calculated from all artificially combined SAGs. Two-sample *t*-test assuming equal variances was significant ($p = <000.1$).

genome size, and many shared metabolic pathways) suggests that metagenome bins provide useful consensus information about this candidate phylum (**Figure 3** and Supplemental Figure 1). However, consensus sequences appear to mask extensive genetic diversity among Rokubacteria cells in the studied environments. On a more fundamental level, metagenome assembly and binning relies on the assumption that microbial communities are composed of near-clonal populations. An increasing body of evidence shows that this assumption is not valid in many microorganismal lineages and environments, with genomic rearrangements and horizontal gene transfer being more prevalent than previously thought (Ochman et al., 2000; Feldgarden et al., 2003; Shapiro, 2010; Kashtan et al., 2014; Labonté et al., 2015a). By recovering data from the most fundamental units of biological organization, single-cell genomics does not rely on the assumption of clonality, offers an opportunity to improve our understanding of microbial microevolutionary processes (Garrity and Lyons, 2003; Engel et al., 2014; Kashtan et al., 2014), and helps calibrating the performance and interpretation of metagenomics tools when working with complex, natural microbiomes.

CONCLUDING REMARKS

Recent discoveries of many novel phyla and superphyla of microorganisms are revolutionizing our understanding of the genealogy and current diversity of life. Here, a focused analysis of the single-cell genomic and metagenome sequences of Rokubacteria (formerly known as SPAM) suggests that they constitute a monophyletic, phylum-level lineage that is most closely related to Nitrospirae among the currently described phyla. Large genomes, high %GC, and a global presence at low abundance in soils and terrestrial subsurface environments appear to be general features of this candidate phylum. Rokubacteria genomes predict didermy, mixotrophy, motility, and versatile DNA packaging mechanisms. It is plausible that the latter feature interferes with gDNA amplification, in part explaining the difficulty of recovering high quality genomes from Rokubacteria single cells. Furthermore, large cell-to-cell genomic heterogenetity and low relative abundance in most environments studied to date may be among the factors contributing to their limited recovery in metagenome bins. Our analysis also demonstrates major differences in the quality of genomic data obtained from SAGs and metagenome bins: while assemblies with greatest estimated genome recovery were obtained by metagenome binning, SAGs delivered contamination-free data from discrete biological units, making them easier to interpret and revealing significant genomic diversity within this candidate phylum, including a split into two class-level lineages.

AUTHOR CONTRIBUTIONS

EB: Project leader, primary author and primary data analyst. RS: PI of project, data creation and secondary author. TW: data creation and scientific correspondent. JE: phylogenetics and scientific correspondent. TO: Sample collection and scientific correspondent. DM: Sample collection and scientific correspondent. ML: Sample collection and scientific correspondent. JMB, JB, and NI: bioinformatics and data analysis. FG-V: Sample collection and scientific correspondent. JJ: Project leader and scientific correspondent. NP: sample processing. Co-PI: TO, DM, TW, and JE.

ACKNOWLEDGMENTS

We thank the staff of the Bigelow Laboratory Single-Cell Genomics Center and the Joint Genome Institute for the generation of single-cell and metagenomic data. We are grateful to Olukayode Kuloyo, Borja Linage, Sarah Hendrickson, Cara Magnabosco, Melody Lindsay, Petra Diamonds and the management and staff for Finsch diamond mine, South Africa for their help obtaining the samples. We are also grateful to Laura Vann for her help obtaining Nevada field samples. Thanks also to Darrell Lacy, Levi Kryder, and John Klenke of the Nevada Nuclear Waste Repository Program Office (NWRPO) and the Nature Conservancy for access to and sampling assistance with well OV-2. We are grateful to Refuge Manager, Annji Bagozzi, and the U.S. Fish and Wildlife service for permitting and guidance related to the sampling of Crystal Spring at Ash Meadows National Wildlife Refuge. We are also grateful to Jaret Heise, Tom Regan, Kathy Hart, and many others at the Sanford Underground Research Facility for safe underground access and sampling assistance. Special thanks to Brittany Kruger, Joshua Sackett, and Scott Hamilton-Brehm of the Moser Lab for sample and metadata collection. This work was supported by the U.S.

National Science Foundation grants DEB-1441717 and OCE-1335810. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02264/full#supplementary-material

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., et al. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* 7:13219. doi: 10.1038/ncomms13219

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Becraft, E. D., Dodsworth, J. A., Murugapiran, S. K., Ohlsson, J. I., Briggs, B. R., Kanbar, J., et al. (2015). Single-cell-genomics-facilitated read binning of candidate phylum EM19 genomes from geothermal spring metagenomes. *Appl. Environ. Microbiol.* 82, 992–1003. doi: 10.1128/AEM.03140-15

Belcher, C. M., Finch, P., Collinson, M. E., Scott, A. C., and Grassineau, N. V. (2009). Geochemical evidence for combustion of hydrocarbons during the K-T impact event. *Proc. Natl. Acad. Sci. U.S.A.* 106, 4112–4117. doi: 10.1073/pnas.0813117106

Biller, S. J., Berube, P. M., Berta-Thompson, J. W., Kelly, L., Roggensack, S. E., Awad, L., et al. (2014). Genomes of diverse isolates of the marine cyanobacterium Prochlorococcus. *Sci. Data* 1:140034. doi: 10.1038/sdata.2014.34

Bowers, R., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35, 725–731. doi: 10.1038/nbt.3893

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

Butterfield, C. N., Li, Z., Andeer, P. F., Spaulding, S., Thomas, B. C., Singh, A., et al. (2016). Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ* 4:e2687. doi: 10.7717/peerj.2687

Castelle, C. J., Wrighton, K. C., Thomas, B. C., Hug, L. A., Brown, C. T., Wilkins, M. J., et al. (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* 25, 690–701. doi: 10.1016/j.cub.2015.01.014

Chen, X., Su, Y., He, X., Wei, Y., Wei, W., and Wu, J. (2012). Soil bacterial community composition and diversity respond to cultivation in Karst ecosystems. *World J. Microbiol. Biotechnol.* 28, 205–213. doi: 10.1007/s11274-011-0809-0

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., and Hwang, C. C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS ONE* 8:e62856. doi: 10.1371/journal.pone.0062856

Colwell, F. S., and D'Hondt, S. (2013). Nature and extent of the deep biosphere. *Rev. Mineral. Geochem.* 75, 547–574. doi: 10.2138/rmg.2013.75.17

Engel, P., Stepanauskas, R., and Moran, N. A. (2014). Hidden diversity in honey bee gut symbionts detected by single-cell genomics. *PLoS Genet.* 10:e1004596. doi: 10.1371/journal.pgen.1004596

Feldgarden, M., Byrd, N., and Cohan, F. M. (2003). Gradual evolution in bacteria: evidence from Bacillus systematics. *Microbiology* 149, 3565–3573. doi: 10.1099/mic.0.26457-0

Figuerola, E. L., Guerrero, L. D., Turkowsky, D., Wall, L. G., and Erijman, L. (2015). Crop monoculture rather than agriculture reduces the spatial turnover of soil bacterial communities at a regional scale. *Environ. Microbiol.* 17, 678–688. doi: 10.1111/1462-2920.12497

Foerstner, K. U., von Mering, C., Hooper, S. D., and Bork, P. (2005). Environments shape the nucleotide composition of genomes. *EMBO Rep.* 6, 1208–1213. doi: 10.1038/sj.embor.7400538

Garrity, G. M., and Lyons, C. (2003). Future-proofing biological nomenclature. *OMICS* 7, 31–33. doi: 10.1089/153623103322006562

Ghai, R., McMahon, K. D., and Rodriguez-Valera, F. (2012). Breaking a paradigm: cosmopolitan and abundant freshwater actinobacteria are low GC. *Environ. Microbiol. Rep.* 4, 29–35. doi: 10.1111/j.1758-2229.2011.00274.x

Giovannoni, S. J., Cameron Thrash, J., and Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *ISME J.* 8, 1553–1565. doi: 10.1038/ismej.2014.60

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., et al. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309, 1242–1245. doi: 10.1126/science.1114057

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. System. Evol. Microbiol.* 57, 81–91. doi: 10.1099/ijs.0.64483-0

Hansel, C. M., Fendorf, S., Jardine, P. M., and Francis, C. A. (2008). Changes in bacterial and archaeal community structure and functional diversity along a geochemically variable soil profile. *Appl. Environ. Microbiol.* 74, 1620–1633. doi: 10.1128/AEM.01787-07

Hershberg, R., and Petrov, D. A. (2010). Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6:e1001115. doi: 10.1371/journal.pgen.1001115

Hildebrand, F., Meyer, A., and Eyre-Walker, A. (2010). Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 6:e1001107. doi: 10.1371/journal.pgen.1001107

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nat. Microbiol.* 1:16048. doi: 10.1038/nmicrobiol.2016.48

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

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., and D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16213–16216. doi: 10.1073/pnas.1203849109

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* 428, 726–731. doi: 10.1016/j.jmb.2015.11.006

Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165. doi: 10.7717/peerj.1165

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science* 344, 416–420. doi: 10.1126/science.1248575

Labonté, J. M., Field, E. K., Lau, M., Chivian, D., Van Heerden, E., Wommack, K. E., et al. (2015a). Single cell genomics indicates horizontal gene transfer and viral infections in a deep subsurface Firmicutes population. *Front. Microbiol.* 6:349. doi: 10.3389/fmicb.2015.00349

Labonté, J. M., Swan, B. K., Poulos, B., Luo, H., Koren, S., Hallam, S. J., et al. (2015b). Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *ISME J.* 9, 2386–2399. doi: 10.1038/ismej.2015.48

Lassalle, F., Perian, S., Bataillon, T., Nesme, X., Duret, L., and Daubin, V. (2015). GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet.* 11:e1004941. doi: 10.1371/journal.pgen.1004941

Li, H. (2015). BFC: correcting illumina sequencing errors. *Bioinformatics* 31, 2885–2887. doi: 10.1093/bioinformatics/btv290

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, J., Zhou, J., Wu, Y., Yang, S., and Tian, D. (2015). GC-content of synonymous codons profoundly influences amino acid usage. *G3 (Bethesda)* 5, 2027–2036. doi: 10.1534/g3.115.019877

Lin, X., Kennedy, D., Fredrickson, J., Bjornstad, B., and Konopka, A. (2012). Vertical stratification of subsurface microbial community composition across geological formations at the Hanford Site. *Environ. Microbiol.* 14, 414–425. doi: 10.1111/j.1462-2920.2011.02659.x

Lipson, D. A., and Schmidt, S. K. (2004). Seasonal changes in an alpine soil bacterial community in the colorado rocky mountains. *Appl. Environ. Microbiol.* 70, 2867–2879. doi: 10.1128/AEM.70.5.2867-2879.2004

Luef, B., Frischkorn, K. R., Wrighton, K. C., Holman, H. Y., Birarda, G., Thomas, B. C., et al. (2015). Diverse uncultivated ultra-small bacterial cells in groundwater. *Nat. Commun.* 6:6372. doi: 10.1038/ncomms7372

Markowitz, V. M., Chen, I. M., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., et al. (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* 42, D560–D567. doi: 10.1093/nar/gkt963

Nakai, R., Fujisawa, T., Nakamura, Y., Nishide, H., Uchiyama, I., Baba, T., et al. (2016). Complete genome sequence of Aurantimicrobium minutum type Strain KNCT, a planktonic ultramicrobacterium isolated from river water. *Genome Announc.* 4:e00616-16. doi: 10.1128/genomeA.00616-16

NRC (1999). "Size limits of very small microorganisms," in *Proceedings of a Workshop, Steering Group for the Workshop on Size Limits of Very Small Microorganisms* (Washington, DC: National Academy Press).

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304. doi: 10.1038/35012500

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

Parrott, G. L., Kinjo, T., and Fujita, J. (2016). A compendium for *Mycoplasma pneumoniae*. *Front. Microbiol.* 7:513. doi: 10.3389/fmicb.2016.00513

Pruesse, E., Peplies, J., and Glockner, F. O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28, 1823–1829. doi: 10.1093/bioinformatics/bts252

Reichenberger, E. R., Rosen, G., Hershberg, U., and Hershberg, R. (2015). Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. *Genome Biol. Evol.* 7, 1380–1389. doi: 10.1093/gbe/evv063

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437. doi: 10.1038/nature12352

Rodrigues, V. D., Torres, T. T., and Ottoboni, L. M. (2014). Bacterial diversity assessment in soil of an active Brazilian copper mine using high-throughput sequencing of 16S rDNA amplicons. *Antonie Van Leeuwenhoek* 106, 879–890. doi: 10.1007/s10482-014-0257-6

Rodriguez, R., and Konstantinidis, K. (2014). Bypassing cultivation to identify bacterial species. *Microbe*. 9, 111–118. doi: 10.1128/microbe.9.111.1

Sczyrba, A. (2017). Critical assessment of metagenome interpretation – a benchmark of computational metagenomics software. *bioRxiv*. doi: 10.1101/099127

Sekiguchi, Y., Ohashi, A., Parks, D. H., Yamauchi, T., Tyson, G. W., and Hugenholtz, P. (2015). First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking. *PeerJ* 3:e740. doi: 10.7717/peerj.740

Shapiro, J. A. (2010). Mobile DNA and evolution in the 21st century. *Mob. DNA* 1:4. doi: 10.1186/1759-8753-1-4

Stepanauskas, R., Fergusson, E. A., Brown, J., Poulton, N. J., Tupper, B., Labonté, J. M., et al. (2017). Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat. Commun.* 8:84. doi: 10.1038/s41467-017-00128-z

Sutcliffe, I. C. (2010). A phylum level perspective on bacterial cell envelope architecture. *Trends Microbiol.* 18, 464–470. doi: 10.1016/j.tim.2010.06.005

Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., Gonzalez, J. M., et al. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci. U.S.A.* 110, 11463–11468. doi: 10.1073/pnas.1304246110

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K. H., et al. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* 12, 635–645. doi: 10.1038/nrmicro3330

Yoon, H. S., Price, D. C., Stepanauskas, R., Rajah, V. D., Sieracki, M. E., Wilson, W. H., et al. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* 332, 714–717. doi: 10.1126/science.1203163

Zhang, Y., Sun, Y., Jiao, N., Stepanauskas, R., and Luo, H. (2016). Ecological genomics of the uncultivated marine roseobacter lineage CHAB-I-5. *Appl. Environ. Microbiol.* 82, 2100–2111. doi: 10.1128/AEM.03678-15

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.