AMERICAN MUSEUM NOVITATES

Number 3975, 14 pp. August 25, 2021

The Bacterial Diversity Lurking in Protist Cell Cultures

AMERIS APONTE,[1, 2] YANGTSHO GYALTSHEN,[1] JOHN A. BURNS,[1, 3] AARON A. HEISS,[1] EUNSOO KIM,[1] AND SALLY D. WARRING[1]

ABSTRACT

Laboratory cultures of heterotrophic protists are often xenic, meaning that the culture contains more than one microbial organism. In this study, we analyzed genome-assembly data from cultures of four marine protist flagellates—the marine malawimonad *Imasa heleensis*, the undescribed mantamonad strain SRT-306, the discobid *Ophirina amphinema*, and the cryptist *Palpitomonas bilix*—specifically to search for genomes of cocultured bacteria. As no external bacteria have been added to the protist stock cultures, it is probable that the cocultured bacteria came from the original water samples from which the protists were isolated. At least some of these bacteria are consumed as a food source by the protists, all of which are obligate heterotrophs. From four separate metagenomic de novo assemblies for these mixed cultures, we identified 28 scaffolds, which BUSCO analyses suggest represent complete or near-complete bacterial genomes. These scaffolds range in length from 3,139,436 to 6,090,282 bp and encode 2873 to 5666 genes. Only eight of the 28 scaffolds corresponded to entries in the NCBI genome database, meaning that 20 of these scaffolds represent genomes from putatively novel bacterial species. Our findings highlight that data like these, which are often discarded or overlooked, can be a source of novel genomes and/or species.

[1] Sackler Institute for Comparative Genomics and Division of Invertebrate Zoology, American Museum of Natural History.

[2] Natural Science Faculty, Rio Piedras Campus, University of Puerto Rico.

[3] Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine.

ISSN 0003-0082

INTRODUCTION

As part of ongoing protist genome projects, we have generated large-scale genomic data from several recently discovered free-living heterotrophic flagellates, including the cryptist *Palipitomonas bilix* (Yabuki et al., 2010; 2014), the undescribed mantamonad strain SRT-306 (Glücksman et al., 2011; Brown et al., 2018), the phylogenetically deep-branching discobid *Ophirina amphinema* (Yabuki et al., 2018), and the marine malawimonad strain *Imasa heleensis* (Heiss et al., 2020). All of these protists have been recognized in the last decade as representing deeply branching lineages amongst all eukaryotes (Yabuki et al., 2010; Glücksman et al., 2011; Brown et al., 2018; Yabuki et al., 2018), and their genomic and cell structural data potentially have significant implications for better understanding early eukaryotic evolution, including the nature of the last eukaryotic common ancestor (LECA). Moreover, these microbes are phagotrophic nanoflagellates, an ecological group that plays an important role as grazers of bacteria (Collier and Rest, 2019). As each protist species investigated here is an obligate heterotroph, bacteria, presumably coisolated together with the protist species, are consumed by the protist as a necessary food source. However, the identity and community composition of the cocultured bacteria have been unknown.

Despite efforts to reduce bacterial load in DNA extraction, such as by physical separation using polycarbonate membrane filters before harvesting, the bacteria could not be completely removed, and their genomic sequences became a major source of contamination in the resulting data, often accounting for the majority of the generated reads. Therefore, the bacterial sequences had to be identified and separated from the eukaryotic reads during the process of protist genome assembly and annotation.

Studies in other systems have shown that there is value in nontarget sequencing data. For example, genomic and transcriptomic datasets generated from animal tissues can contain sequences originating from parasites and/or symbionts of those organisms. Classifying and identifying these sequences can shed light on the distribution of particular pathogens in animal populations (Lopes et al., 2017; Galen et al., 2020). In eukaryote genome projects like ours, these bacterial data, which could be valuable to the microbial research community, are often discarded without being assembled, annotated, and archived at a public data repository. However, these data may contain new genomes and/or species, and depending on sampling, culturing, and DNA extraction conditions, they could reveal population dynamics of the sampled habitat or the culture system. Here, we analyzed such "contaminating" bacterial data: from these four protist cultures, we assembled, annotated, and identified 28 scaffolds representing complete or near-complete bacterial genomes. Our analyses show that the majority of these bacterial scaffolds represent species that are not currently represented in GenBank's bacterial genome databases.

MATERIALS AND METHODS

Culture conditions: The culture strain of *Palpitomonas bilix* investigated in this study was established from a driftwood sample from Marcharchar Island, Palau, collected in June

2006 (Yabuki et al., 2010). The malawimonad *Imasa heleensis* and the discobid *Ophirina amphinema* were collected in September 2013 and March 2016, respectively, from a shallow lagoon on Nusa Lavata in the Hele island chain, in the Western Province of the Solomon Islands (Yabuki et al., 2018; Heiss et al. 2020). Mantamonad strain SRT-306 came from the scraped surface sample of a barracuda that was caught in a lagoon in Iriomote-jima, Okinawa Prefecture, Japan, in September 2013. All of the cultures grew well and were maintained in Erd-Schreiber medium (ESM: UTEX) or ESM fortified with 2.5% Cerophyl medium (ATCC 802) at 23° C, with the exception of SRT-306, which was maintained at 16° C.

DNA extraction and sequencing: For *I. heleensis* and *O. amphinema*, DNA was extracted on multiple occasions between January 2018 and May 2019. Cell cultures were scraped using a sterile cell scraper to lift adherent cells, and protist cells were collected on 0.6 µm and 0.4 µm Millipore polycarbonate filters, respectively, under partial vacuum, to reduce bacterial load. Cells were incubated on these filters for 3 hours at 56° C in lysis buffer from the Qiagen MagAttract HMW DNA Kit (Qiagen, Hilden, Germany). DNA was extracted from the lysates using this same kit according to the manufacturer's instructions. An aliquot of DNA (1 µg) from each culture was sent to the New York Genome Center for Illumina Paired End 2x150 bp sequencing on the Illumina HiSeqX platform (Illumina, San Diego, CA). Additional DNA (~1–4 µg) was prepared for sequencing on the Oxford Nanopore platform using either the SQK-LSK108 or SKK-LSK109 Genomic DNA by Ligation kit (Oxford Nanopore Technologies, Oxford, UK) according to manufacturer's instructions. Libraries were sequenced on the MinION platform using FLO-MIN106 SpotON R9 Flow Cells.

For *P. bilix* and mantamonad strain SRT-306, DNA was extracted using the Qiagen DNeasy Blood & Tissue Kit, following the manufacturer's instructions. Cells were collected onto 0.8 µm filters, as specified above, and washed three times each with 5–10 ml artificial seawater to reduce the bacterial load prior to DNA extraction. The plates on which SRT-306 were grown were scraped to lift adherent cells; no such procedure was required for the free-swimming *P. bilix*. DNA samples were sent to Cornell Sequencing Core and New York Genome Center for Illumina Nextera library preparations and Paired End 2x150 bp sequencing on the on the Illumina HiSeq2500 platform.

Base calling and assembly: Raw MinION FastQ files were base-called using Guppy v1.1 (Oxford Nanopore Technologies). Hybrid assemblies using both MinION and Illumina data from *I. heleensis* and *O. amphinema* were assembled using MaSuRCA v3.2.6 (Zimin et al., 2013). Illumina reads generated from cultures of *P. bilix* and SRT-306 were assembled using ALLPATHS-LG (Gnerre et al., 2011; Ribeiro et al., 2012).

Identification of bacterial genomes from metagenomic assemblies: The resulting assembled scaffolds and contigs were individually analyzed for completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO) v3 (Simão et al., 2015) with the "Bacteria odb9" dataset. The presence of 5S, 16S, and 23S rRNA genes was searched for using rnammer v1.2 (Lagesen et al., 2007). The scaffolds both (a) with a BUSCO completeness ≥80% and (b) that contained at least one complete rRNA operon were retained as "complete" bacterial scaffolds. The 80% completeness threshold was chosen because there was a sharp cutoff in completeness

TABLE 1. Features of the four metagenomic assemblies.

Name of Protist Culture	Sampling Location	Collection Year	DNA Extraction Year	Reference
SRT-306 (undescribed mantamonad)	Iriomote-jima, Japan	2013	2015	This study
Palpitomonas bilix	Macharchar Island, Republic of Palau	2006	2013–2014	Yabuki et al., 2010
Imasa heleensis	Nusa Lavata, Hele island chain, Solomon Islands	2013	2019	Heiss et al., 2020
Ophirina amphinema	Nusa Lavata, Hele island chain, Solomon Islands	2016	2019	Yabuki et al., 2018

for scaffolds below 80%, most of which were less than 60% complete and/or did not contain complete rRNA operons. In an attempt to circularize the complete scaffolds, two independent approaches were taken: MUMmer v3.0 (Kurtz et al., 2004) was used to look for overlaps at the ends of each scaffold, and Circlator (Hunt et al., 2015) was used to attempt circularization of each scaffold assembly.

Annotation and classification of bacterial genomes: Bacterial scaffolds were annotated for gene content using the rapid prokaryotic genome annotation tool, Prokka (Seemann, 2014). The scaffolds were classified into taxonomic groups using the Genome Taxonomy Database Tool Kit (GTDB-TK) version 0.3.1 (Chaumeil et al., 2019). Scaffolds were compared to each other using the online ANI calculator tool from the EZ BioCloud database (Yoon et al., 2017). The scaffolds are available on NCBI under the BioProject accession code PRJNA619388.

Abundance estimates: Relative abundance estimates were calculated for each complete genome by aligning the Illumina Paired End reads to the pooled set of complete bacterial genomes isolated from the same culture using the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009; Durbin, 2010), retaining only the uniquely mapping read pairs. Read pairs mapping to each scaffold were calculated using SAMtools version 1.9 (Li et al., 2009; Li, 2011). Final calculations and plots were made in R version 3.3.3 and ggplot2 version 3.3.2 (Wickham, 2016).

RESULTS

Identification of bacterial genomes from metagenomic assemblies: We screened mixed assemblies from four mono-isolate cultures of marine protists that were sequenced for the purpose of assembling high-quality reference genomes for the protist species (table 1). We identified 28 scaffolds that represent complete or near-complete bacterial genomes: four from SRT-306, three from *P. bilix*, six from *O. amphinema,* and 15 from *I. heleensis* (fig. 1, table 2). We attempted to circularize all genomes by looking for overlaps at the ends of the scaffolds by aligning each scaffold to itself and plotting the overlaps using MUMmer (Kurtz et al., 2004), and by running each scaffold through the Circlator assembly circularization pipeline (Hunt et al., 2015). MUMmer found no overlapping regions in any scaffold, and Circlator did not circularize any of the 28 scaffolds (data not shown). All but one of the 28 scaffolds contained at least one complete prokaryotic rRNA operon (i.e., with full-length 5S, 16S, and 23S rRNA; table 2). For scaffold CP051235, only one incomplete rRNA operon was annotated, which contained full-length 16S and 23S rRNA genes separated by two tRNA loci, but no downstream 5S rRNA. The 16S and 23S rRNA genes are located midscaffold on CP051235, and genes are annotated both upstream of the 16S and downstream of the 23S loci; thus, the lack of an annotated 5S rRNA is not due to an assembly truncation.

Features and gene content of each bacterial scaffold: We used a rapid annotation tool, Prokka (Seemann, 2014), to annotate each bacterial scaffold for gene content and other basic genomic features (table 3). The 28 bacterial scaffolds ranged in length from 3.1 to 6.0 Mb, contained 2873 to 5666 annotated genes, and had an average of 44.5 tRNAs, congruent with what is typical of bacterial genomes (Land et al., 2015). GC content of these genomes ranged from 37.4 to 65.8%, within the range noted for bacterial genomes (Almpanis et al., 2018).

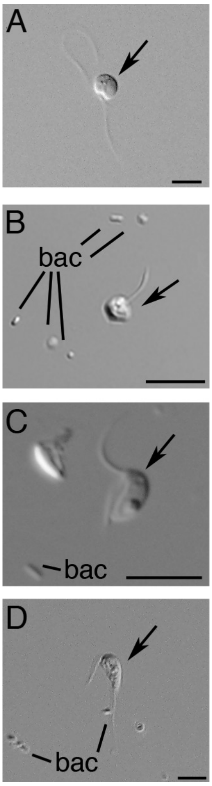

FIG. 1. Light microscopy images showing the four protist cells mentioned in this study. Arrow indicates protist cell, bac = bacteria. Scale bar = 5 µm. **A.** *Palpitomonas bilix*, **B.** mantamonad strain SRT-306; **C.** *Ophirina amphinema*; and **D.** *Imasa heleensis.*_

Taxonomic classification of the 28 scaffolds: We used the GTDB-TK tool to classify each genome taxonomically (Chaumeil et al., 2019). From the 28 scaffolds we found three phyla, five classes, eight orders, 10 families, and 21 genera represented (table 3). Eight of the scaffolds were classified to the species level. Two of these genomes represent the same taxon, *Alcanivorax* sp. DSM 26293 (GenBank Assembly Accession GCF_900107995.1), which is present in data from the cultures of *O. amphinema* and *I. heleensis*. These two scaffolds are close to identical at the sequence level, with an average nucleotide identity (ANI) of 99.99%. For the remaining 20 scaffolds, 19 were classified to the genus level and one (CP051235) was classified only to the family level. As such, these 20 scaffolds could represent either novel or known but unsequenced bacterial species. GTDB-TK establishes taxonomy using a combination of the placement of the query into a reference tree, the queries' Relative Evolutionary Distance (RED), and its ANI percent values compared to reference genomes (Matsen et al., 2010; Parks et al., 2018; Chaumeil et al., 2019). Queries placed on terminal branches of the reference tree (i.e., are sister to individual known species or strains) are placed in the genus of their sister taxon. A query is classified to species level if it is placed in an existing genus and its ANI value is within the circumscription radius (typically, 95%) of the closest species; otherwise the query is classified as a novel species in that genus (Chaumeil et al., 2019). For example, scaffold CP051248 had an ANI value of 99.94% and was classified as *Pseudooceanicola atlanticus*. In contrast, scaffold CP051241 was placed under the genus *Marinobacter*. Its closest ANI value, to the species *Marinobacter similis*, was 89.97%, suggesting that it is a novel species of *Marinobacter*. For queries that are placed on internal

TABLE 2. Features of each bacterial genome. Genomes are organized by protist culture of origin. Asterisks indicate different assemblies of the same bacterial strain.

Culture	Bacterial Assembly ID	Scaffold Length (base pairs)	GC Content (%)	Genes (no.)	tRNA (no.)	rRNA (no.)	BUSCO Completeness (%)	Coverage
Ophirina amphinema	CP051240	3961866	58.3	3465	44	9	98	64.19
	CP051241	4272096	57.5	3872	52	9	98.6	6995.73
	CP051246	4734807	37.4	4051	45	9	95.9	406.72
	CP051249	3255866	59.1	3101	42	3	87.2	36.68
	CP051253	3835116	58.5	3501	42	6	97.3	138.91
	CP051254	3573046	58.1	3409	42	3	96.6	40.01
Imasa heleensis	CP051228	4907657	62.93	4474	51	6	96.6	1042.67
	CP051229	3398407	64.06	3469	44	4	82.4	12.14
	CP051230	3829003	61.53	3650	42	3	97.3	169.94
	CP051232	4675334	65.78	4510	50	6	95.9	56.15
	CP051234	4470904	58.52	3937	50	3	95.3	1216.04
	CP051235	3307379	60.77	3227	47	2	96.6	32.3
	CP051237	6060643	59.7	5666	45	6	98	290.56
	CP051239	3139436	64.53	2873	36	6	88.5	120.15
	CP051242	3747071	38.28	3174	39	6	95.3	18.43
	CP051243	3439134	44.43	3006	36	6	94.6	1371.79
	CP051245	4052158	41.12	3702	39	6	95.3	108.24
	CP051247	6090282	57.87	4652	53	3	85.8	175.53
	CP051248	4533893	64.23	4245	47	6	95.9	151.18
	CP051250	4421565	57.26	4018	52	9	99.3	953.62
	CP051252	3837393	58.52	3500	43	6	97.3	62.26
SRT-306	CP051231	4309625	65.55	4215	43	5	95.9	95.3
	CP051233	3386480	62.39	3248	46	3	91.2	60.62
	CP051238	4744784	47.64	4176	46	3	85.1	150.07
	CP051251	3496013	64.62	3297	40	3	89.2	77.19
Palpitomonas bilix	CP051236	3707127	60.13	3496	49	7	97.3	351.93
	CP051244	3755467	43.86	3308	40	6	90.5	18.82
	CP051255	3958902	38.18	3177	37	6	91.9	20.98

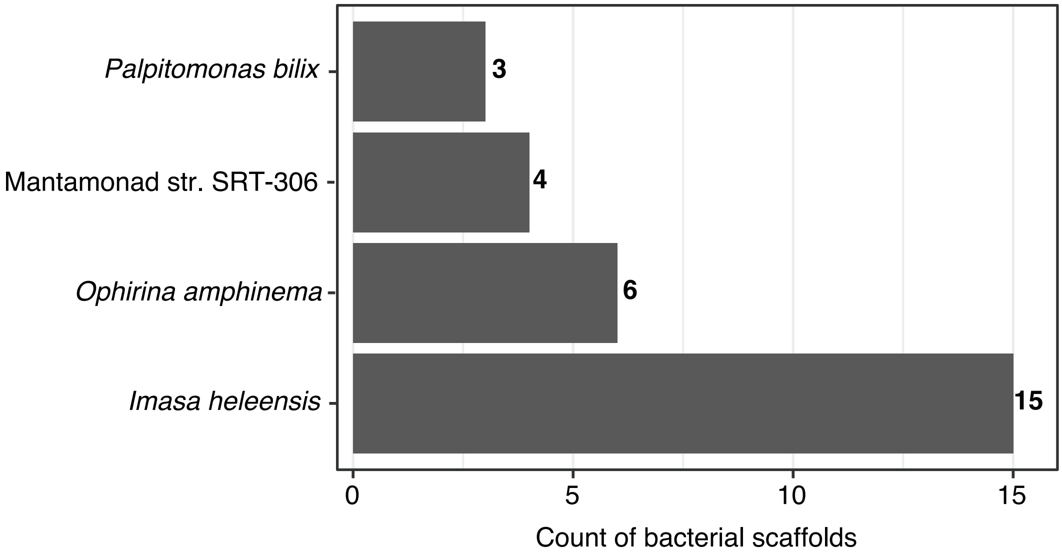

FIG. 2. Count of scaffolds in each metagenome that are ≥80% complete or more, as determined by BUSCO.

branches, taxonomy is instead determined by placement and/or RED (Parks et al., 2018; Chaumeil et al., 2019). Only one of our queries, scaffold CP051235 (which was the scaffold lacking the 5S rRNA gene), was not placed in an existing genus. This scaffold was placed in the family *Hyphomicrobiaceae*, and potentially represents a new genus therein.

Relative abundance of bacterial species in each culture: We next evaluated the relative abundance of the 28 bacterial scaffolds by calculating the number of Illumina read pairs that aligned uniquely to each genome from the four sequencing libraries (fig. 2). Two of the metagenomes, from *O. amphinema* and *P. bilix* cultures, were each dominated by a single bacterial taxon, *Marinobacter* and *Nitratireductor* respectively, that accounted for ~90% of the bacterial data from each library. The metagenomic data from the *I. heleensis* and SRT-306 cultures were not dominated by any one species (fig. 2).

DISCUSSION

In this study, we investigated bacterial data that are often discarded in the course of sequencing genomes from xenic protist cultures. From four metagenomic protist genome assemblies, we identified 28 assembly scaffolds that represent complete or near-complete bacterial chromosomes.

We were able to classify eight of the bacterial scaffolds to species level, 19 to genus level, and one to family level. Only two of these 28 bacterial scaffolds represented the same species, *Alcanivorax* sp. DSM 26293, which was present in the *O. amphinema* and *I. heleensis* cell culture metagenomes, making a total of 27 unique bacterial scaffolds from our dataset. *Ophirina amphinema* and *I. heleensis* were originally isolated from the same lagoon, so this *Alcanivorax* bacterium is possibly from this original source.

TABLE 3. Taxonomic classification of bacterial genomes. Results given by GTDB-TK database. "NA" indicates scaffolds for which no closest hit was given. Asterisks (*) indicate placeholder family and genus names provided by GTDB-TK when no family and/or genus name is present in NCBI database.

Bacterial Assembly ID	Family	Genus	Species	Species Name/ID, Closest Hit (% ANI)
CP051228	GCA-2696645 *	GCA-2696645 *	–	NA
CP051229	*Hyphomonadaceae*	UBA7672 *	–	NA
CP051230	*Hyphomonadaceae*	*Hyphomonas*	–	*Hyphomonas atlantica/* GCA_000682715.1 (83.4)
CP051231	*Rhodobacteraceae*	*Oceanicola*	–	*Oceanicola litoreus/* GCA_900142295.1 (85.71)
CP051232	*Rhodobacteraceae*	*Maritimibacter*	–	*Maritimibacter sp./* GCA_002701395.1 (87.84)
CP051233	*Rhodobacteraceae*	*Thalassobius*	–	*Rhodobacteraceae bacterium/* GCA_002708925.1 (80.76)
CP051234	*Hyphomicrobiaceae*	*Filomicrobium*	–	*Alphaproteobacteria* bacterium BRH_c36/ GCA_001516065.1 (78.08)
CP051235	*Hyphomicrobiaceae*	–	–	NA
CP051236	*Phyllobacteriaceae*	*Nitratireductor*	–	*Nitratireductor* sp./ GCA_002697745.1 (89.84)
CP051237	*Rhizobiaceae*	*Pararhizobium*	–	*Pararhizobium haloflavum/* GCA_002750855.1 (77.4)
CP051238	*Alteromonadaceae*	*Aestuariibacter*	–	*Aestuariibacter aggregatus/* GCA_900129565.1(87.13)
CP051239	*Alcanivoracaceae*	*Alcanivorax*	–	*Alcanivorax* sp. UBA2685/ GCA_002354605.1 (94.97)
CP051240	*Alcanivoracaceae*	*Alcanivorax*	–	*Alcanivorax nanhaiticus/* GCA_000756665.1 (84.13)
CP051241	*Alteromonadaceae*	*Marinobacter*	–	*Marinobacter similis/* GCA_000830985.1 (89.97)
CP051242	*Balneolaceae*	UBA7797 *	–	*Balneolaceae* bacterium UBA7797/ GCA_002480645.1 (76.65)
CP051243	*Cryomorphaceae*	UBA7878 *	–	*Flavobacteriales* bacterium UBA7878/ GCA_002501205.1 (78.02)
CP051244	*Flavobacteriaceae*	*Muricauda*	–	NA
CP051245	*Flavobacteriaceae*	*Muricauda*	–	NA
CP051246	*Flavobacteriaceae*	*Salegentibacter*	–	NA
CP051247	*Planctomycetaceae*	UBA9033 *	–	*Planctomycetaceae* bacterium UBA2671/ GCA_002359185.1 (76.63)
CP051248	*Rhodobacteraceae*	*Pseudooceanicola*	*Pseudooceanicola atlanticus*	*Pseudooceanicola atlanticus/* GCA_000768315.1 (99.94)
CP051249	*Rhodobacteraceae*	*Epibacterium*	*Epibacterium mobile*	*Epibacterium mobile/* GCA_001681715.1 (96.59)

CP051250	*Alteromonadaceae*	*Marinobacter*	*Marinobacter adhaerens*	*Marinobacter adhaerens*/ GCA_001717765.1 (97.23)
CP051251	*Rhodobacteraceae*	*Pseudooceanicola*	*Pseudooceanicola nitratireducens*	*Pseudooceanicola nitratireducens*/ GCA_900109195.1 (97.09)
CP051252	*Alcanivoracaceae*	*Alcanivorax*	*Alcanivorax* sp. DSM 26293	*Alcanivorax* sp. DSM 26293/ GCA_900107995.1 *(98.96)*
CP051253	*Alcanivoracaceae*	*Alcanivorax*	*Alcanivorax* sp. DSM 26293	*Alcanivorax* sp. DSM 26293/ GCA_900107995.1 (98.96)
CP051254	*Hyphomonadaceae*	*Hyphomonas*	*Hyphomonas atlantica*	*Hyphomonas atlantica*/ GCA_000682715.1 (98.11)
CP051255	*Balneolaceae*	*Balneola*	*Balneola* sp. EhC07	*Balneola* sp. EhC07/ GCA_001650905.1 (96.92)

A literature search was done of the top taxonomic hits for each genome to ascertain the typical environmental ranges of the source organisms. *Salegentibacter* is a genus found in hypersaline lakes, on the surfaces of marine fauna, and in marine sediments (McCammon and Bowman, 2000; Bowman, 2016). *Hyphomonas* (Abraham, 2020), *Epibacterium* (Wirth and Whitman, 2020), *Maritimibacter* (Lee et al., 2007), *Aestuariibacter* (Yi et al., 2004), *Thalassobius* (Arahal et al., 2005), *Muricauda* (Bruns et al., 2001; Bruns and Berthe-Corti, 2015), *Nitratireductor* (Singh et al. 2012), *Balnenola* (Urios et al., 2006), *Oceanicola* (Cho and Giovannoni, 2004), *Pseudoceanicola* (Lai et al., 2015), *Planctomycetaceae* (Ward, 2015), and *Balneolaceae* (Xia et al., 2016) represent groups of widely dispersed marine bacteria isolated from seawater and/or marine sediments or surfaces. *Alcanivorax* (Yakimov et al., 1998; Golyshin et al., 2015), *Marinobacter* (Gauthier et al., 1992; Bowman and McMeekin, 2015), *Hyphomicrobiaceae* (Kesy et al. 2019), and *Filomicrobium* (Schlesner, 1987; 2015) are marine groups previously isolated from seawater and seawater sediments enriched with crude oil and/or hydrocarbons. *Nitratireductor* is a nitrate-reducing genus found in various marine habitats (Labbe et al., 2004), members of which have previously been isolated from diatom cultures (Jang et al., 2011). The family *Cryomorphaceae* is present within a wide range of nonextreme ecosystems, both marine and terrestrial (Bowman, 2015). Lastly, *Pararhizobium* is a nitrogen-fixing genus that associates with plant roots, which has a worldwide distribution (Mousavi et al., 2015).

In summary, all but one of the described groups (*Pararhizobium*) are characteristically found in marine habitats, consistent with the fact that the cultures are from marine environments. These cocultured bacteria therefore are likely from the source samples from which each protist was isolated. New microbial genomic data generated from little-sampled habitats has the potential to be highly valuable to the research community, as studies have shown that marine and other infrequently sampled niches may contain more novel microbial diversity than habitats that are much closer at hand, like the human gut or local soils (Bech et al., 2020; Tessler et al., 2017). However, we cannot rule out that they are environmental contaminants, considering that most of the taxa have global marine distributions, the seawater culture medium would select for marine species, and our lab in Manhattan is surrounded by seawater and bacteria are commonly dispersed through the air and on surfaces (Mayol, 2017).

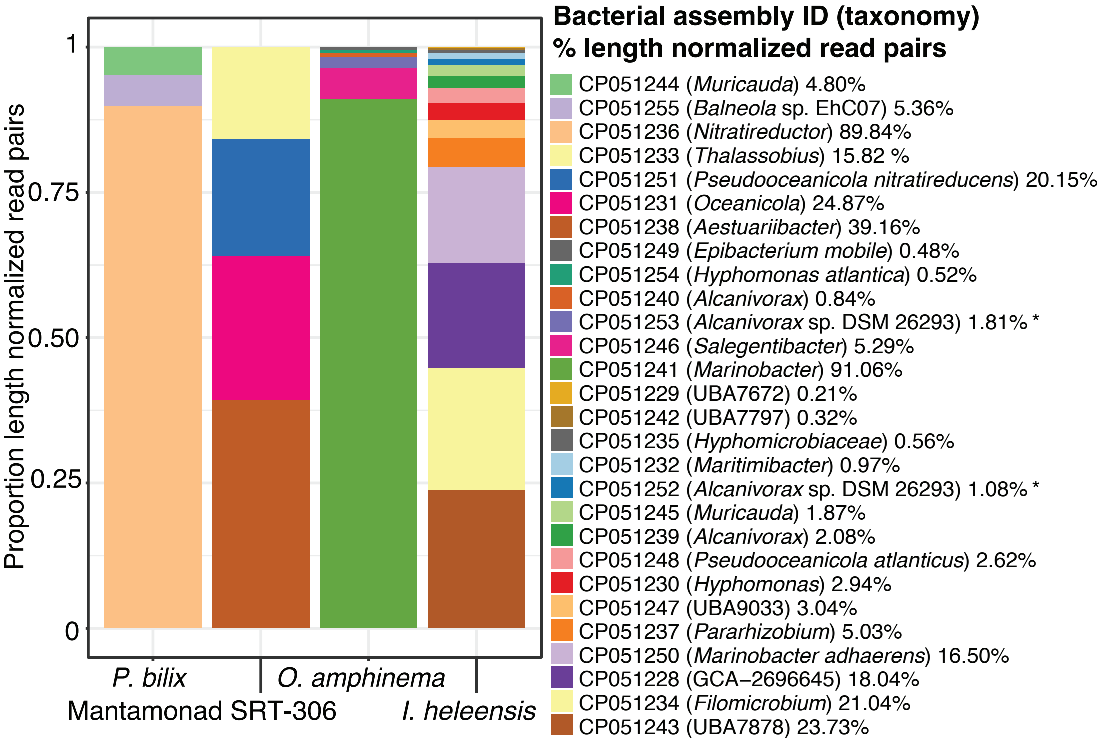

FIG. 3. Relative abundance of each bacterial scaffold in Illumina sequencing libraries, represented as proportion of length normalized read pairs for the portion of each sequencing library mapping to the bacterial scaffolds. The two *Alcanivorax* sp. DSM 26293 scaffolds are marked with asterisks (*).

When we calculated the relative abundance of each bacterial scaffold across the four cultures, we found that the species are unevenly represented in the sequencing libraries, with both *Ophirina amphinema* and *Palpitomonas bilix* having one dominant bacterial taxon, *Marinobacter* and *Nitratireductor* respectively, while the novel malawimonad and mantamonad strains each had a more even distribution of bacterial taxa (fig. 2). This distribution may represent population patterns in the source cultures possibly caused by varying bacterial diversity in the original samples or by differing grazing preferences of each isolated protist in culture. However, each culture was preprocessed using a polycarbonate membrane filter prior to DNA extraction, which may have influenced the distribution of each species, by retaining both larger bacterial cells and bacterial flocs. Therefore, the bacterial abundances presented in figure 2 are not necessarily representative of the bacterial community of each culture.

Importantly, 20 of the 28 scaffolds we identified represent taxa that are not yet present in bacterial genome databases. As mentioned above, these are data that we usually discard from our protist genomics projects, but we find that they can also be a source of novel bacterial genomes and species. We note as well that we obtained these unidentified species from long-term cultures grown under constant environmental conditions and containing relatively few different organisms. This suggests that establishing axenic cultures of at least some of these novel bacteria might be practicable, thereby allowing for their formal description. However, the existence of axenic cultures may not remain the only requirement for formal description. A recent consensus statement (Murray, 2020) suggests mechanisms for the establishment of new prokaryote taxa based on genomic data. Should this consensus be formally accepted, all of our new taxa should be candidates for formal description. In either case, we find that the prokaryotic genomes generated during eukaryote sequencing projects need not be disposable by-products. Instead, we suggest they are a key source for potentially important new discoveries.

ACKNOWLEDGMENTS

This work was supported by a grant from the Simons Foundation (SF-382790) awarded to E.K., internship funding from the National Science Foundation (DBI 1358465) awarded to A.A., and postdoctoral support from the Gerstner Fellowship and the Lerner-Gray Fund awarded to A.A.H. The authors thank Ashley Yang and Apurva Narechania for their assistance with cell culturing and computational work, respectively.

REFERENCES

Abraham, W.R. 2020. *Hyphomonas*. *In* M.E. Trujillo et al. (editors), Bergey's manual of systematics of archaea and bacteria: 1–14. Hoboken, NJ: Wiley.

Almpanis, A., M. Swain, D. Gatherer, and N. McEwan. 2018. Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. Microbial Genomics 4 (4).

Arahal, D.R., M.C. Macian, E. Garay, and M.J. Pujalte. 2005. *Thalassobius mediterraneus* gen. nov., sp. nov, and reclassification of *Ruegeria gelatinovorans* as *Thalassobius gelatinovorus* comb. nov. International Journal of Systematic and Evolutionary Microbiology 55 (6): 2371–2376.

Bech, P.K., K.L. Lysdal, L. Gram, M. Bentzon-Tilia, and M.L. Strube. 2020. Marine sediments hold an untapped potential for novel taxonomic and bioactive bacterial diversity. mSystems Sep 2020: 5 (5).

Bowman, J.P. 2015. *Cryomorphaceae*. *In* W.B. Whitman et al. (editors), Bergey's manual of systematics of archaea and bacteria: 1–7. Hoboken, NJ: Wiley.

Bowman, J.P. 2016. *Salegentibacter*. *In* W.B. Whitman et al. (editors), Bergey's manual of systematics of archaea and bacteria: 1–11. Hoboken, NJ: Wiley.

Bowman, J.P., and T.A. McMeekin. 2015. *Marinobacter*. *In* W.B. Whitman et al. (editors), Bergey's manual of systematics of archaea and bacteria: 1–6. Hoboken, NJ: Wiley.

Brown, M.W., et al. 2018. Phylogenomics places orphan protistan lineages in a novel eukaryotic supergroup. Genome Biology and Evolution 10 (2): 427–433.

Bruns, A., and L. Berthe-Corti. 2015. *Muricauda*. *In* W.B. Whitman et al. (editors), Bergey's manual of systematics of archaea and bacteria: 1–8. Hoboken, NJ: Wiley.

Bruns, A., M. Rohde, and L. Berthe-Corti. 2001. *Muricauda ruestringensis* gen. nov., sp. nov., a facultatively anaerobic, appendaged bacterium from German North Sea intertidal sediment. International Journal of Systematic and Evolutionary Microbiology 51 (6): 1997–2006.

Chaumeil, P.A., A.J. Mussig, P. Hugenholtz, and D.H. Parks. 2019. Gtdb-tk: A toolkit to classify genomes with the genome taxonomy database. Bioinformatics.

Cho, J.C., and S.J. Giovannoni. 2004. *Oceanicola granulosus* gen. nov., sp. nov. and *Oceanicola batsensis* sp. nov., poly-beta-hydroxybutyrate-producing marine bacteria in the order '*Rhodobacterales.*' International Journal of Systematic and Evolutionary Microbiology 54 (4): 1129–1136.

Collier, J.L., and J.S. Rest. 2019. Swimming, gliding, and rolling toward the mainstream: cell biology of marine protists. Molecular Biology of the Cell 30 (11): 1245–1248.

Galen, S.C., J. Borner, S.L. Perkins, and J.D. Weckstein. 2020. Phylogenomics from transcriptomic "bycatch" clarify the origins and diversity of avian trypanosomes in North America. PLoS ONE 15 (10).

Gauthier, M.J., et al. 1992. *Marinobacter hydrocarbonoclasticus* gen. nov., sp. nov., a new, extremely halotolerant, hydrocarbon-degrading marine bacterium. International Journal of Systematic Bacteriology 42 (4): 568–576.

Glücksman, E., et al. 2011. The novel marine gliding zooflagellate genus *Mantamonas* (Mantamonadida ord. n.: Apusozoa). Protist 162 (2): 207–221.

Gnerre, S., et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America 108 (4): 1513–1518.

Golyshin, P.N., S. Harayama, K.N. Timmis, and M.M. Yakimov. 2015. *Alcanivorax*. *In* W.B. Whitman et al. (editors), Bergey's manual of systematics of archaea and bacteria: 1–7. Hoboken, NJ: Wiley.

Heiss, A. A., et al. 2020. Description of *Imasa heleensis*, gen. nov., sp. nov. (Imasidae, fam. nov.), a deep-branching marine malawimonad and possible key taxon in understanding early eukaryotic evolution. Journal of Eukaryotic Microbiology e12837.

Hunt, M., et al. 2015. Circlator: Automated circularization of genome assemblies using long sequencing reads. Genome Biology 16: 294.

Jang, G.I., C.Y. Hwang, and B.C. Cho. 2011. *Nitratireductor aquimarinus* sp. nov., isolated from a culture of the diatom *Skeletonema costatum*, and emended description of the genus *Nitratireductor*. International Journal of Systematic and Evolutionary Microbiology 61 (11): 2676–2681.

Kurtz, S., et al. 2004. Versatile and open software for comparing large genomes. Genome Biology 5 (2): R12.

Labbe, N., S. Parent, and R. Villemur. 2004. *Nitratireductor aquibiodomus* gen. nov., sp. nov., a novel alpha-proteobacterium from the marine denitrification system of the Montreal biodome (Canada). International Journal of Systematic and Evolutionary Microbiology 54 (1): 269–273.

Lagesen, K., et al. 2007. Rnammer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research 35 (9): 3100–3108.

Lai, Q., et al. 2015. *Pseudooceanicola atlanticus* gen. nov. sp. nov., isolated from surface seawater of the Atlantic Ocean and reclassification of *Oceanicola batsensis*, *Oceanicola marinus*, *Oceanicola nitratireducens*, *Oceanicola nanhaiensis*, *Oceanicola antarcticus* and *Oceanicola flagellatus*, as *Pseudooceanicola batsensis* comb. nov., *Pseudooceanicola marinus* comb. nov., *Pseudooceanicola nitratireducens* comb. nov., *Pseudooceanicola nanhaiensis* comb. nov., *Pseudooceanicola antarcticus* comb. nov., and *Pseudooceanicola flagellatus* comb. nov. Antonie Van Leeuwenhoek 107 (4): 1065–1074.

Land, M., et al. 2015. Insights from 20 years of bacterial genome sequencing. Functional & Integrative Genomics 15 (2): 141–161.

Lee, K., Y.J. Choo, S.J. Giovannoni, and J.C. Cho. 2007. *Maritimibacter alkaliphilus* gen. nov., sp. nov., a genome-sequenced marine bacterium of the *Roseobacter* clade in the order *Rhodobacterales*. International Journal of Systematic and Evolutionary Microbiology 57 (7): 1653–1658.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27 (21): 2987–2993.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25 (14): 1754–1760.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26 (5): 589–595.

Li, H., et al. 2009. The sequence alignment/map format and samtools. Bioinformatics 25 (16): 2078–2079.

Lopes, R.L., M.M. Mérida, and M. Carneiro, M. 2017. Unleashing the potential of public genomic resources to find parasite genetic data. Trends in Parasitology 33 (10): 750–753.

Matsen, F.A., R.B. Kodner, and E.V. Armbrust. 2010. Pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11: 538.

Mayol, E., et al. 2017. Long-range transport of airborne microbes over the global tropical and subtropical ocean. Nature Communications 8 (1): 201.

McCammon, S.A., and J.P. Bowman. 2000. Taxonomy of antarctic flavobacterium species: description of *Flavobacterium gillisiae* sp. nov., *Flavobacterium tegetincola* sp. nov., and *Flavobacterium xanthum* sp. nov., nom. rev. and reclassification of [*Flavobacterium*] *salegens* as *Salegentibacter salegens* gen. nov., comb. nov. International Journal of Systematic and Evolutionary Microbiology 50 (3): 1055–1063.

Mousavi, S.A., A. Willems, X. Nesme, P. de Lajudie, and K. Lindstrom. 2015. Revised phylogeny of Rhizobiaceae: proposal of the delineation of *Pararhizobium* gen. nov., and 13 new species combinations. Systematic and Applied Microbiology 38 (2): 84–90.

Murray, A.E., et al. 2020. Roadmap for naming uncultivated archaea and bacteria. Nature Microbiology 5 (8): 987–994.

Parks, D.H., et al. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology 36 (10): 996–1004.

Ribeiro, F.J., et al. 2012. Finished bacterial genomes from shotgun sequence data. Genome Research 22 (11): 2270–2277.

Schlesner, H. 1987. *Filomicrobium fusiforme* gen. nov., sp. nov., a slender budding, hyphal bacterium from brackish water. Systematic and Applied Microbiology 10 (1): 63–67.

Schlesner, H. 2015. *Filomicrobium*. *In* W.B. Whitman et al. (editors), Bergey's manual of systematics of archaea and bacteria: 1–6. Hoboken, NJ: Wiley.

Seemann, T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30 (14): 2068–2069.

Simão, F.A., R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, and E.M. Zdobnov. 2015. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31 (19): 3210–3212.

Tessler, M., et al. 2017. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Scientific Reports 7: 6589.

Urios, L., H. Agogue, F. Lesongeur, E. Stackebrandt, and P. Lebaron. 2006. *Balneola vulgaris* gen. nov., sp. nov., a member of the phylum *Bacteroidetes* from the north-western Mediterranean sea. International Journal of Systematic and Evolutionary Microbiology 56 (8): 1883–1887.

Ward, N.L. 2015. *Planctomycetaceae*. *In* W.B. Whitman et al. (editors), Bergey's manual of systematics of archaea and bacteria: 1–2. Hoboken, NJ: Wiley.

Wickham, H. 2016. Ggplot2: elegant graphics for data analysis. New York: Springer-Verlag.

Wirth, J.S., and W.B. Whitman. 2020. *Epibacterium*. *In* W.B. Whitman (editors), Bergey's manual of systematics of archaea and bacteria: 1–13. Hoboken, NJ: Wiley.

Xia, J., S.K. Ling, X.Q. Wang, G.J. Chen, and Z.J. Du. 2016. *Aliifodinibius halophilus* sp. nov., a moderately halophilic member of the genus *Aliifodinibius*, and proposal of *Balneolaceae* fam. nov. International Journal of Systematic and Evolutionary Microbiology 66 (6): 2225–2233.

Yabuki, A., Y. Inagaki, and K. Ishida. 2010. *Palpitomonas bilix* gen. et sp. nov.: a novel deep-branching heterotroph possibly related to *Archaeplastida* or *Hacrobia*. Protist 161 (4): 523–538.

Yabuki, A., et al. 2014. *Palpitomonas bilix* represents a basal cryptist lineage: Insight into the character evolution in *Cryptista*. Scientific Reports 4: 4641.

Yabuki, A., Y. Gyaltshen, A.A. Heiss, K. Fujikura, and E. Kim. 2018. *Ophirina amphinema* n. gen., n. sp., a new deeply branching discobid with phylogenetic affinity to jakobids. Scientific Reports 8 (1): 16219.

Yakimov, M.M., et al. 1998. *Alcanivorax borkumensis* gen. nov., sp. nov., a new, hydrocarbon-degrading and surfactant-producing marine bacterium. International Journal of Systematic Bacteriology 48 (2): 339–348.

Yi, H., K.S. Bae, and J. Chun. 2004. *Aestuariibacter salexigens* gen. nov., sp. nov. And *Aestuariibacter halophilus* sp. nov., isolated from tidal flat sediment, and emended description of *Alteromonas macleodii*. International Journal of Systematic and Evolutionary Microbiology 54 (2): 571–576.

Yoon, S.H., et al. 2017. Introducing ezbiocloud: a taxonomically united database of 16s rRNA gene sequences and whole-genome assemblies. International Journal of Systematic and Evolutionary Microbiology 67 (5): 1613–1617.

Zimin, A.V., et al. 2013. The MaSuRCA genome assembler. Bioinformatics 29 (21): 2669–2677.

All issues of *Novitates* and *Bulletin* are available on the web (http://digitallibrary.amnh.org/dspace). Order printed copies on the web from:
http://shop.amnh.org/a701/shop-by-category/books/scientific-publications.html

or via standard mail from:

American Museum of Natural History—Scientific Publications
Central Park West at 79th Street
New York, NY 10024

∞ This paper meets the requirements of ANSI/NISO Z39.48-1992 (permanence of paper).