AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

# Long-Read Sequencing Improves Recovery of Picoeukaryotic Genomes and Zooplankton Marker Genes from Marine Metagenomes

N. V. Patin,[a,b,c] K. D. Goodwin[a,c]

[a]Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, Miami, Florida, USA

[b]Cooperative Institute for Marine and Atmospheric Studies, Rosenstiel School of Marine, Atmospheric & Earth Science, University of Miami, Miami, Florida, USA

[c]Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, La Jolla, California, USA

**ABSTRACT** Long-read sequencing offers the potential to improve metagenome assemblies and provide more robust assessments of microbial community composition and function than short-read sequencing. We applied Pacific Biosciences (PacBio) CCS (circular consensus sequencing) HiFi shotgun sequencing to 14 marine water column samples and compared the results with those for short-read metagenomes from the corresponding environmental DNA samples. We found that long-read metagenomes varied widely in quality and biological information. The community compositions of the corresponding long- and short-read metagenomes were frequently dissimilar, suggesting higher stochasticity and/or bias associated with PacBio sequencing. Long reads provided few improvements to the assembly qualities, gene annotations, and prokaryotic metagenome-assembled genome (MAG) binning results. However, only long reads produced high-quality eukaryotic MAGs and contigs containing complete zooplankton marker gene sequences. These results suggest that high-quality long-read metagenomes can improve marine community composition analyses and provide important insight into eukaryotic phyto- and zooplankton genetics, but the benefits may be outweighed by the inconsistent data quality.

**IMPORTANCE** Ocean microbes provide critical ecosystem services, but most remain uncultivated. Their communities can be studied through shotgun metagenomic sequencing and bioinformatic analyses, including binning draft microbial genomes. However, most sequencing to date has been done using short-read technology, which rarely yields genome sequences of key microbes like SAR11. Long-read sequencing can improve metagenome assemblies but is hampered by technological shortcomings and high costs. In this study, we compared long- and short-read sequencing of marine metagenomes. We found a wide range of long-read metagenome qualities and minimal improvements to microbiome analyses. However, long reads generated draft genomes of eukaryotic algal species and provided full-length marker gene sequences of zooplankton species, including krill and copepods. These results suggest that long-read sequencing can provide greater genetic insight into the wide diversity of eukaryotic phyto- and zooplankton that interact as part of and with the marine microbiome.

**KEYWORDS** eDNA, long-read sequencing, marine microbiomes, metagenomics

Marine microbial ecology has benefited greatly from high-throughput sequencing, which allows community surveys of ecologically important habitats dominated by uncultured taxa (1–4). In particular, shotgun metagenomic sequencing and metagenome-assembled genome (MAG) binning have generated important insights into marine prokaryotic diversity and ecology (5, 6). Metagenomic analyses can expand the

**Editor** Holly Bik, University of California, Riverside



Address correspondence to N. V. Patin, nastassia.patin@noaa.gov.

The authors declare no conflict of interest.

**Received** 29 June 2022
**Accepted** 27 October 2022
**Published** 30 November 2022

tree of life (7), link taxonomy to biochemical function (e.g., see references 8 and 9), characterize the taxa driving global oceanographic processes like the biological pump (10), and generate testable hypotheses for understanding how microbes adapt to nutrient limitations (e.g., see reference 11).

To date, the vast majority of metagenomic sequencing has been done using short-read sequencing technology such as the Illumina MiSeq and HiSeq platforms. These instruments can generate billions of 150- to 600-bp reads at ever-decreasing costs; however, metagenomes generated from short-read sequencing have numerous shortcomings. Gene duplications or genomic regions with high levels of repeats, including ribosomal RNA (rRNA) genes, can prevent read assembly and MAG binning (12). Horizontally transferred genes can preclude the accurate reconstruction of population-level genomes as they may differ from the rest of the genome in nucleotide composition (13). Moreover, ecologically important but microdiverse lineages, including the ubiquitous SAR11 clade, have proven highly resistant to assembly and binning with short-read sequencing (14–16). Thus, important knowledge gaps remain in our understanding of marine metagenome composition and function.

The computational challenges associated with binning prokaryotic genomes are amplified for eukaryotic organisms. The larger genome sizes, the presence of introns, and compositional differences among chromosomes are just a few additional hurdles to the accurate recovery of eukaryotic MAGs (17). In the marine environment, these barriers have prevented the large-scale genomic characterization of single-celled eukaryotic phytoplankton (18, 19). Recent studies suggest a high level of unexplored diversity among these photosynthetic picoeukaryotes (20), and their estimated contribution to global carbon fixation is substantial (21–23). Improving our ability to generate high-quality eukaryotic plankton genome sequences thus remains an important target of metagenomic and bioinformatic studies.

Long-read sequencing can overcome some of the limitations of short-read technology. Platforms like the Pacific Biosciences (PacBio) and Oxford Nanopore Technologies platforms generate 5- to 50-kbp reads on average (24), which can be leveraged for marker gene surveys to improve the taxonomic resolution (25–28). While still hampered by high raw error rates and per-read costs compared to short-read sequencing, both factors have significantly improved over the last few years (24). In particular, PacBio circular consensus sequencing (CCS) performs multiple passes of a circularized template molecule, generating highly accurate HiFi long reads of $>$10 kbp in length (29). Recent comparisons of PacBio CCS with Illumina short-read amplicon sequencing showed that only long-read data accurately identified microbial taxa at the strain level (30, 31). Long reads have also been shown to improve complex metagenomic assemblies and lead to the recovery of higher-quality genome sequences (30, 32, 33). Furthermore, combining both short- and long-read data can provide abundance estimates through read recruitment to robust assemblies, harnessing the advantages of both technologies to improve both qualitative and quantitative metagenomic information (30, 34, 35).

In this study, we assessed the utility of PacBio CCS shotgun metagenomic sequencing of marine water column environmental DNA (eDNA) samples compared with Illumina short-read sequencing. We compared metagenome quality, recovery of taxonomic marker genes, and community composition (including strain-level taxonomic identifications) among short-read-only, long-read-only, and hybrid metagenomes. We further compared the numbers and qualities of MAGs generated across sequencing and assembly methods. We found that while prokaryotic community compositions and MAG recoveries were similar across short- and long-read-based metagenomes, only PacBio sequencing produced high-quality eukaryotic MAGs. Furthermore, long contigs produced by PacBio sequencing contained complete marker gene sequences from eukaryotic taxa, including copepods belonging to the genera *Calanus* and *Metridia* and the krill species *Euphausia pacifica*. We describe the implications of these results and discuss the potential for long-read sequencing to inform marine ecology studies.

## RESULTS

**Sample and sequencing results.** Fourteen samples collected from depths between 5 and 50 m off the coast of California were sequenced with both the Illumina (2 of 3 replicate filters) and PacBio (third replicate) platforms. The number of PacBio CCS reads ranged from 6,807 to 1,907,840 per metagenome (see Table S1 in the supplemental material). The numbers of Illumina reads from the corresponding samples ranged from 2,879,847 to 264,220,592. Four of the PacBio metagenomes were contaminated with reads from up to four human pathogen and/or common laboratory contamination taxa: *Comamonas acidovorans*, *Comamonas lacustris*, *Stenotrophomonas maltophilia*, and *Cutibacterium acnes* (Genome Taxonomy Database (GTDB) taxonomic designations). The average and maximum read lengths for each sample are provided in Table S1, along with the number of reads remaining in the four samples after decontamination.

**Assembly comparisons.** We compared the Illumina assemblies of the corresponding eDNA samples with two assemblies using only PacBio reads (hifiasm-meta and metaFlye) and hybrid assemblies with both Illumina and PacBio reads (hybridSPAdes). The longest contigs were longer in the assemblies generated with long reads only than in the short-read or hybrid assemblies (Fig. S1A). The metaFlye assembly program produced the longest contigs in 9 of the 14 assemblies, and in two cases, the longest metaFlye contig was nearly twice as long as the corresponding longest hifiasm-meta contig (Fig. S2A). Hifiasm-meta consistently performed the worst at generating contigs of >1 kbp in length (Fig. S2B). Open reading frames (ORFs) on the longest contig (>800 kbp from Las19c139_27m-3) were annotated as belonging to *Verrucomicrobiaceae* bacterium TMED86 according to NCBI taxonomy (GTDB taxonomy, SW10 sp002172625, phylum *Verrucomicrobiota*).

Other assembly metrics showed more variation among assemblers; the hybrid and short-read-only assemblies contained more contigs of >1 kbp than the long-read-only assemblies, although metaFlye performed comparably and exceeded other assemblers with numbers of contigs of >5 kbp (Fig. S1B). Hifiasm-meta assemblies had higher average $N_{50}$ values than the other assemblers (Fig. S1B). As expected, hybrid assemblies recruited a much higher proportion of short reads than the long-read-only assemblies, comparable to the proportions recruited by the short-read SPAdes assemblies (Fig. S1C). The mapping levels for the replicate short-read samples (Fig. S1C, black dots) were not as high as those for the short reads used in the hybrid assemblies (white dots) but were still much higher than those for the long-read-only assemblies. MetaFlye assemblies recruited slightly more reads on average than hifiasm-meta assemblies. Unmapped reads for the hifiasm-meta and metaFlye assemblies generally spanned all major taxa represented in both the long- and short-read metagenomes, with no clear taxonomic pattern for read recruitment (https://doi.org/10.6084/m9.figshare.21321225).

**Annotation comparisons.** We compared small-subunit (SSU) (16S and 18S) rRNA genes extracted from the Illumina, PacBio, and hybrid assemblies using hidden Markov models (HMMs) as well as from Illumina reads run through PhyloFlash (Fig. 1). The mean numbers of SSU rRNA sequences (~9) were comparable between the hybrid and short-read assemblies. Long-read (metaFlye) assemblies had both the highest mean (21) and lowest median (2) values, with the highest number of metagenomes containing no SSU genes (5) and 4 metagenomes with >25 SSU genes, while the other assemblies yielded a maximum of 22 SSU genes. Illumina short reads analyzed with PhyloFlash outperformed both the Illumina metaSPAdes and hybrid assemblies, followed by HMM analyses, with a mean of 19 SSU genes (Fig. 1A). SSU genes from the metaFlye assemblies were longer on average than those from any other source, although these assemblies also had the largest spread in mean lengths (Fig. 1B). PhyloFlash analysis of Illumina reads yielded genes with maximum lengths similar to those from all assemblies with shorter lengths (Fig. 1B), which likely contributed to similar confidence values in taxonomic assignments across all SSU sources (Fig. 1C). There were significant differences in confidence at multiple taxonomic levels between long- and short-read assemblies, between hybrid and short-read assemblies at two levels, and between the two short-read SSU extraction approaches at multiple levels (Table S2).

K-mer-based taxonomic composition comparisons between matched long- and short-read samples varied widely depending on the number of PacBio reads. On one

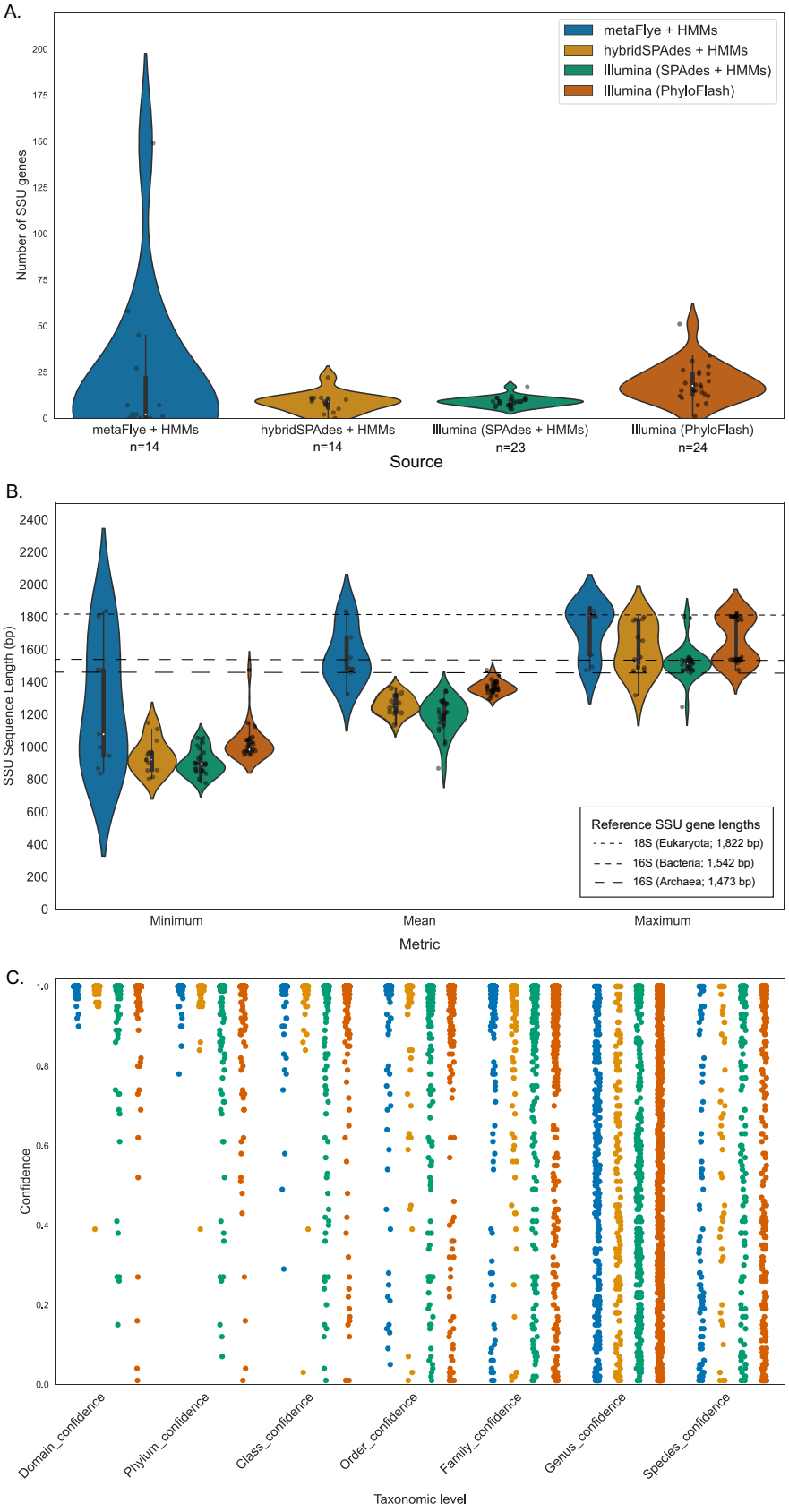



**FIG 1** Summary of the 16S and 18S small-subunit (SSU) rRNA genes extracted in four different ways: three assembly sets run against HMMs {long reads assembled with metaFlye ("metaFlye + HMMs"), long

(Continued on next page)

end of the spectrum, the PacBio metagenome 1903c111_10m-3 had only ~30,000 reads, 2% of which were taxonomically classified by sourmash and represented only 3 different species. The matching Illumina metagenomes had ~23 million and ~27 million reads, with taxonomic assignments of 10 and 8%, respectively, encompassing a diverse range of bacterial, archaeal, and eukaryotic taxa. In contrast, the PacBio metagenome with the most reads (nearly 2 million), 1903c119_11m, had reads classified to nearly twice as many bacterial and eukaryotic taxa as the corresponding Illumina metagenomes with over 18 million reads (182 versus 94 species-level classifications, respectively). The numbers of reads assigned to the genus *Pelagibacter*, a taxon of particular interest due to its ecological importance, ranged from 0 (8 samples) to 617,500 in PacBio metagenomes and from 282,371 to 23,779,853 in Illumina metagenomes (Table S3). The intragenus diversity varied widely between long- and short-read metagenome pairs and was inconsistent with regard to which method provided greater diversity. The taxonomic compositions of all samples at all levels of resolution can be found at https://doi.org/10.6084/m9.figshare.20883652.v1.

Alpha diversity analyses showed a much narrower range of values for Illumina and hybrid metagenomes than for PacBio metagenomes, whose average Shannon diversity values for both read-based (Fig. 2A) and assembly-based (Fig. 2C) analyses varied widely. Beta diversity analyses showed significant differences between long- and short-read samples (read-based permutational multivariate analysis of variance [PERMANOVA], $P = 0.001$ and $F = 11.062$; assembly-based PERMANOVA, $P = 0.016$ and $F = 3.29$) (Fig. 2), and paired samples did not have smaller Aitchison distances than unpaired samples (Fig. S3A). Larger Aitchison distances were linked to different sequencing platforms (Fig. S3B) but not different depth groupings (Fig. S3C). This separation was true for read-based (Fig. 2B) as well as assembly-based (Fig. 2D) taxonomic comparisons; in the latter, hybrid assemblies clustered closely with the Illumina metaSPAdes assemblies.

One of the major potential advantages of long-read sequencing is the recovery of complete protein-coding genes and the reconstruction of full biochemical gene pathways. We thus compared the total numbers of ORFs (annotated and unannotated), the fractions of ORFs with KEGG annotations, and the average numbers of annotated ORFs per contig among all assembly types (Fig. 3). We also compared the numbers of complete (>80% completion) KEGG modules (Table 1). The short-read and hybrid assemblies had the highest number of total ORFs (Fig. 3A), but the long-read assemblies had much higher fractions of annotated ORFs (Fig. 3B), thus leading to comparable numbers of total annotated ORFs among all assembly types (Fig. 3C). However, long-read assemblies performed the worst at reconstructing KEGG modules that were >80% complete (Table 1). Four metaFlye assemblies, which originated from metagenomes with <20,000 reads, produced no data for the KEGG module completeness output, while an additional four assemblies had five or fewer modules. Eleven of the hybrid assemblies had at least 10 KEGG modules meeting the completion criteria. In contrast, all 14 short-read SPAdes assemblies produced at least 30 KEGG modules with >80% completion. The average module completeness was highest for short-read assemblies, at 42%, while long-read assemblies averaged less than half that value (Table 1).

**Binning comparisons.** We compared the numbers and quality scores of MAGs generated from short-read, long-read, and hybrid assemblies using two different binning programs. Overall, hybrid assemblies binned with Vamb yielded the highest number of high-quality MAGs (18) and the highest average MAG quality score (40.3) (Table 2).

**FIG 1** Legend (Continued)
and short reads assembled with hybridSPAdes ("hybridSPAdes + HMMs"), and short reads assembled with metaSPAdes ["Illumina (SPAdes + HMMs)"]} as well as short reads run through PhyloFlash ["Illumina (PhyloFlash)"]. (A) Numbers of SSU genes extracted with each approach. One Illumina metagenome failed to assemble, which is why one more sample is included in the PhyloFlash analysis compared to the SPAdes assembly and HMM extraction. (B) Minimum, mean, and maximum lengths of the extracted SSU genes. The dashed lines show the lengths of reference 16S and 18S rRNA genes in all three domains of life. (C) Confidence values assigned to each taxonomic level of each extracted gene, from 0 to 1. Each level of taxonomic resolution featured a wider range of confidence values than the previous level.

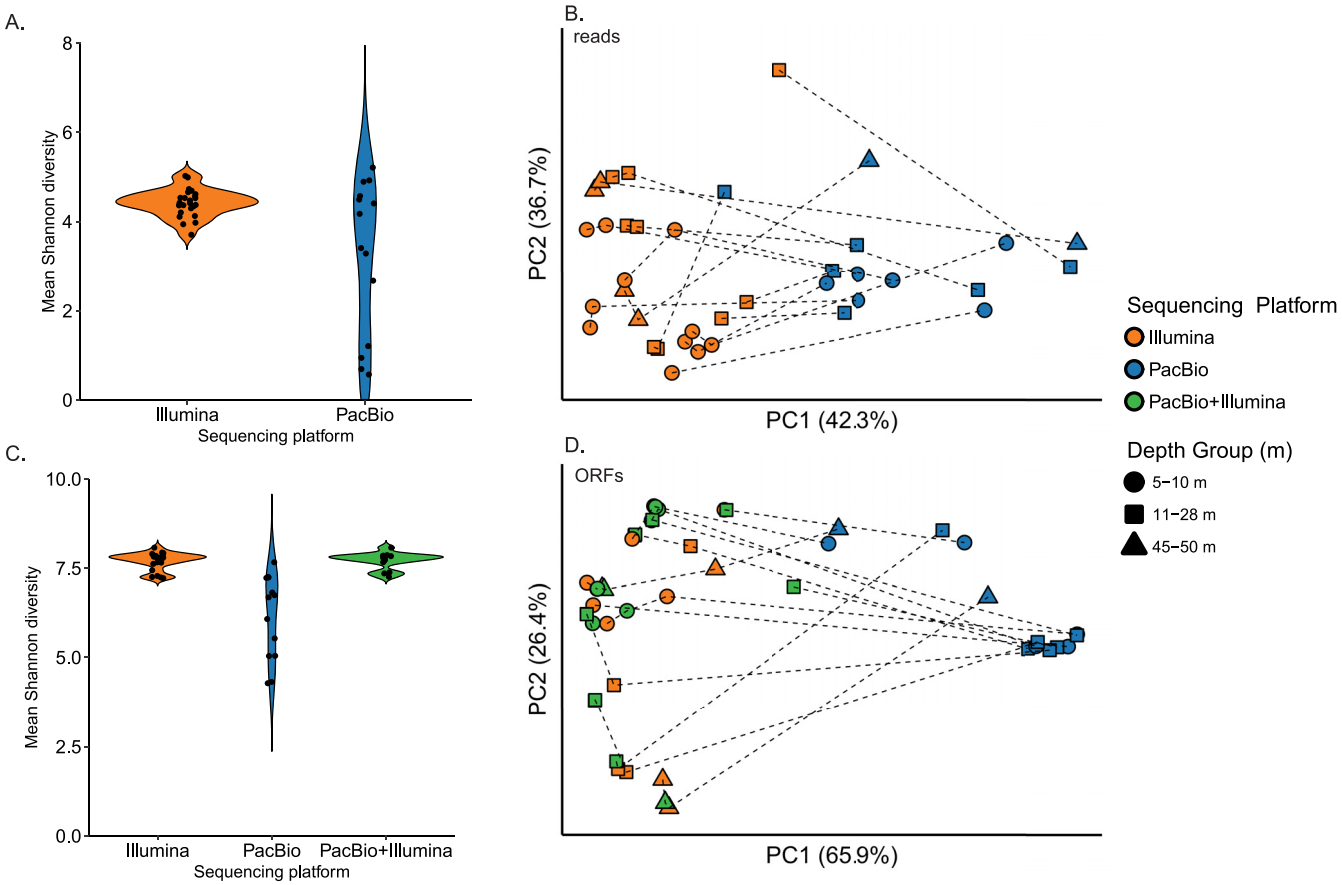



**FIG 2** Alpha and beta diversity analyses of (i) unassembled short and long reads and (ii) open reading frames (ORFs) extracted from long-read, short-read, and hybrid assemblies. (A) Shannon diversity values calculated from unassembled reads. (B) Principal-component analysis of metagenome taxonomic composition calculated from unassembled reads. (C) Shannon diversity values calculated from assembly ORFs. (D) Principal-component analysis of the metagenome taxonomic composition calculated from assembly ORFs. Metagenome types are denoted by colors, and sample depth groupings are denoted by shapes. Paired long-read, short-read, and hybrid metagenomes from the same eDNA sample are connected with dashed lines in the PCA plots.

None of the long-read assembly and binning combinations yielded more than three high-quality MAGs, and short-read assemblies produced the second- and third-highest-quality MAGs with MaxBin2 and Vamb, respectively (Table 2). Only the long-read assemblies yielded MAGs classified as *Pelagibacter* sp., although none of them had quality scores higher than zero.

Out of the 15 high-quality, dereplicated MAGs generated from all long-read and hybrid assemblies, 5 taxa (defined at the genus level) were not recovered from the corresponding short-read metagenomes (Table 3). These unique MAGs included members of the taxa SW10 (phylum *Verrucomicrobiota*), *Acidimicrobiales* TK06, and *Flavobacteriaceae*. Two of these MAGs likely benefited from the longer contigs provided by the PacBio assemblies, with longest MAG contigs of between 290,000 and 320,000 bp (Table 3). Nine out of 10 of the MAGs also recovered from short-read metagenomes were generated from hybrid assemblies, 8 of which were from the Vamb binning program.

In addition, two eukaryotic MAGs were generated from hybrid assemblies. Following manual refinement (Fig. 4), taxonomic assignment of single-copy genes, and phylogenetic placement of DNA-dependent rRNA polymerase genes, they were identified as the picoeukaryotic taxa *Ostreococcus lucimarinus* and *Bathycoccus prasinos*. Phylogenies generated from the DNA-dependent RNA polymerase genes further validated the MAG classifications, with both a and b subunit genes falling within their associated clades (Fig. S4). The gene contents of the two MAGs were similar according to KEGG categories and subcategories (Fig. S5).

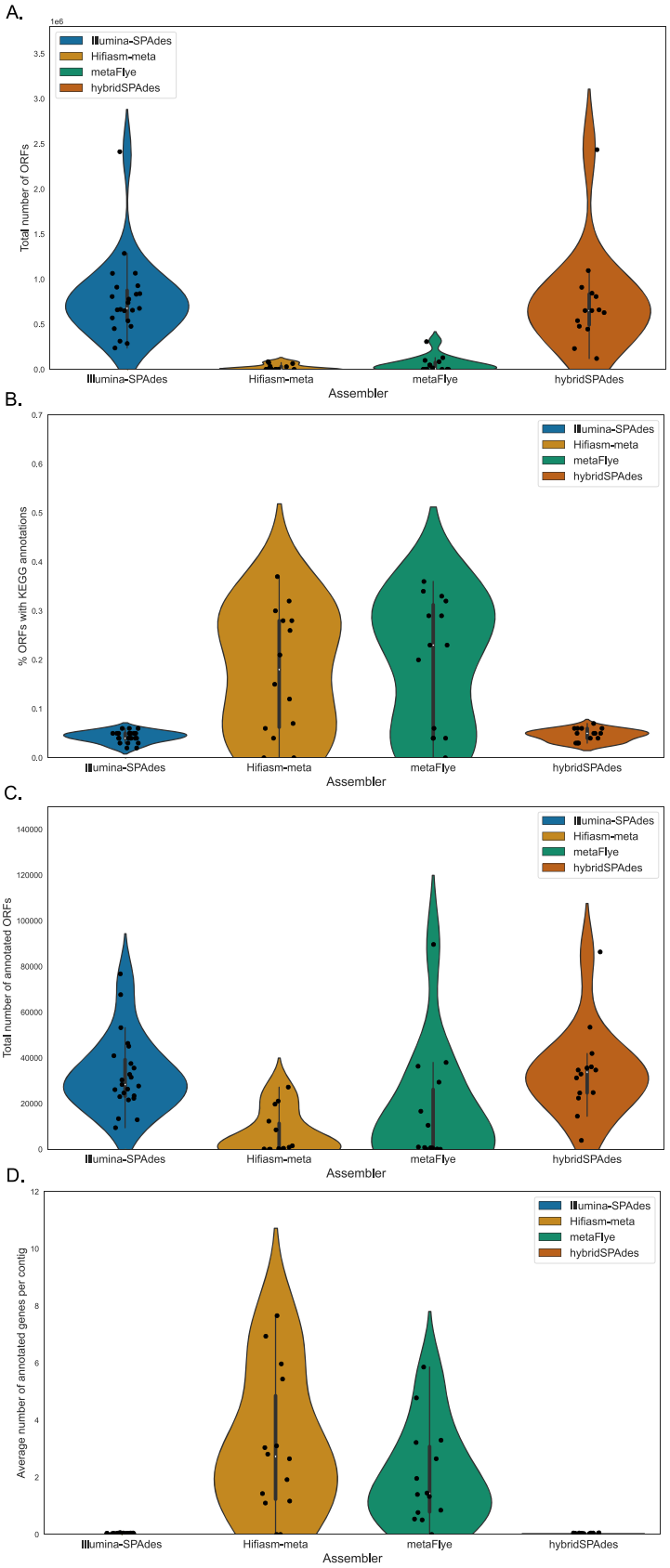



**FIG 3** Summary of annotated and unannotated open reading frames (ORFs) extracted from four assembly types: short read assemblies ("Illumina-SPAdes"), long reads assembled with hifiasm-meta ("Hifiasm-meta"),

(Continued on next page)

**TABLE 1** Average KEGG module completeness values for the metagenomes that generated results in MicrobeAnnotator[a]

| Sequencing platform(s) | Assembly type | No. of samples | Avg module completeness (%) |
|---|---|---|---|
| Illumina | metaSPAdes | 16 | 42.15 |
| PacBio | metaFlye | 14 | 20.29 |
| Illumina + PacBio | hybridSPAdes | 10 | 41.64 |

[a]Metagenomes are grouped by sequencing and assembly types.

All short-read metagenomes had reads that mapped onto all 15 high-quality, dereplicated MAGs from long-read and hybrid assemblies, including the 5 MAGs that were not recovered by short-read binning. Histograms of all mapped read distributions across all MAGs can be found at https://doi.org/10.6084/m9.figshare.21318000.v1.

**Eukaryotic rRNA gene sequences.** Nine contigs of >10,000 bp from five different metaFlye long-read assemblies contained full-length 18S rRNA gene sequences. All nine contigs also contained full or partial 28S rRNA genes. Of the nine 18S rRNA genes, two had top BLAST hits to the krill species *Euphausia pacifica*, and three had top hits to the copepod species *Calanus pacificus* or *Metridia* sp. The sequences matching *Metridia* sp. had coverage and identity values equal to those of *Metridia gerlachei*, *Metridia lucens*, and *Metridia pacifica*; phylogenies generated with genes from these and other *Metridia* species in GenBank show that our contig likely came from *Metridia pacifica* (Fig. S6). This result is also consistent with the known distributions of *Metridia* species in the northern Pacific Ocean (36, 37; K. Jacobson, personal communication).

The longest contigs for each of the three rRNA annotations were used as reference sequences for short-read mapping. Normalized coverages varied among samples for all three references, with sites J and N consistently having the lowest coverages (Fig. 5). Intergenic regions between 18S and 28S rRNA genes typically had low or no coverage, even when the rRNA genes recruited relatively high numbers of reads. Read recruitment plots showed that reads could be aligned at identity levels down to ~90%, with a relatively small fraction of mapped reads aligning at 100% sequence identity (Fig. S7) (for *Euphausia* plots, see https://doi.org/10.6084/m9.figshare.21308184.v3; for *Metridia* plots, see https://doi.org/10.6084/m9.figshare.21308175.v3; and for *Calanus* plots, see https://doi.org/10.6084/m9.figshare.21308169.v2).

## DISCUSSION

In this study, we investigated long-read PacBio sequencing as a way to improve our understanding of marine microbiomes. In particular, we assessed differences in metagenomic assemblies, community compositions, and MAG quantities and qualities. We found a few advantages of long reads for prokaryotic microbiome analyses. However, only hybrid metagenomic assemblies generated from both long and short reads yielded high-quality picoeukaryotic MAGs. Moreover, only long-read assemblies contained contigs with full 18S rRNA gene sequences of ecologically important zooplankton species. These data can help fill knowledge gaps in plankton diversity, biogeography, and marine food webs.

Assembly qualities were surprisingly comparable between short-read and long-read metagenomes (see Fig. S1 and S2 in the supplemental material). Moreover, quality (as determined by metrics such as the number of contigs of >1,000 bp in length) varied widely among samples, suggesting that there is no one-size-fits-all bioinformatic approach for metagenomic assemblies. Hybrid assemblies in general closely resembled short-read-only assemblies in terms of contig size distribution (Fig. S1A and B) and

**FIG 3** Legend (Continued)
long reads assembled with metaFlye ("metaFlye"), and short and long reads assembled together ("hybridSPAdes"). (A) Total numbers of ORFs for each assembly type. (B) Percentages of ORFs for each assembly type that were assigned a KEGG gene annotation. (C) Total numbers of annotated ORFs for each assembly type. (D) Average numbers of annotated ORFs per contig for each assembly type.

**TABLE 2** Numbers and qualities of MAGs generated from metagenomes from all combinations of sequencing, assembly, and binning approaches[a]

| Sequencing platform(s) | Assembly type | Binning approach | No. of metagenomes | Total no. of MAGs | Total no. of high-quality MAGs | Avg MAG quality | No. of *Pelagibacter* MAGs | *Pelagibacter* MAG quality |
|---|---|---|---|---|---|---|---|---|
| Illumina | SPAdes | MaxBin2 | 23 | 271 | 22 | −94.5 | 1 | −133.8 |
| Illumina | SPAdes | Vamb | 23 | 62 | 34 | 42.3 | 0 | NA |
| PacBio | hifiasm-meta | MaxBin2 | 14 | 147 | 1 | −59 | 1 | −156 |
| PacBio | metaFlye | MaxBin2 | 14 | 245 | 1 | −51 | 3 | 0, 0, −142 |
| PacBio | hifiasm-meta | Vamb | 14 | 82 | 3 | −128.4 | 3 | 0, 0, −52 |
| PacBio | metaFlye | Vamb | 14 | 39 | 2 | −4.8 | 0 | NA |
| Illumina + PacBio | hybridSPAdes | MaxBin2 | 14 | 160 | 7 | −90.8 | 1 | −102.8 |
| Illumina + PacBio | hybridSPAdes | Vamb | 14 | 37 | 18 | 40.3 | 0 | NA |

[a]Both Illumina replicates were included in binning analyses where possible. Values for MAGs in the genus *Pelagibacter* are also included due to their ecological importance and the known challenges associated with binning genomes from this taxon. NA, not applicable.

read mapping (Fig. S1C). Compared to long-read-only assemblies, hybrid assemblies recruited much higher levels of short reads from both the sample used for hybrid assembly as well as the replicate metagenome reads (Fig. S1C), highlighting the reliable replication of filter eDNA extraction and Illumina sequencing.

One potential advantage of long-read sequencing followed by metaFlye assembly is the number of extremely long (>100-kbp) contigs generated. Two metaFlye assemblies had contigs of ~800 kbp, one of which likely belonged to a bacterial genome in the *Verrucomicrobiota* phylum (although notably, this contig was absent from all of the high-quality MAGs, even before dereplication). It is unlikely that these long contigs were chimeric for different strains because the coverage of short reads from the corresponding samples was even across both contigs (https://doi.org/10.6084/m9.figshare.21320427.v1). These exceptionally long contigs may have contributed to the improved binning of prokaryotic genomes, as two of the three bacterial MAGs generated from only long-read assemblies contained contigs of >290,000 kbp (Table 3). However, they did not yield any high-quality SAR11 MAGs, which was one of the goals of this study. This may be because one of the obstacles to recovering microdiverse MAGs occurs at the read recruitment step of most binning algorithms, including MaxBin2 (38), in which short-read coverages are used to group contigs into likely genomes. We used short reads from the corresponding samples to perform binning on long-read assemblies, and the read diversity often varied widely between metagenome pairs (Fig. 2).

Eight PacBio metagenomes contained no reads classified as *Pelagibacter* (Table S3), and there were large discrepancies in intragenus diversity between matching PacBio and Illumina metagenomes, which may have prevented population-level read mapping to *Pelagibacter* contigs (see the Krona plots at https://doi.org/10.6084/m9.figshare.20883652.v1). It is possible that deeper long-read sequencing would allow long-read mapping in addition to the assembly and potential recovery of lineages like SAR11. However, this approach has yet to be tested and would likely require long-read sequencing an order of magnitude deeper than was done in this study.

One hypothesis on the advantages of long-read sequencing for microbiome analysis is that more SSU (16S and 18S) rRNA sequences would be captured in full, thereby improving the microbiome taxonomic resolution relative to short-read metagenomes. We found that the average numbers of extracted SSU rRNA genes were similar across all long- and short-read metagenomes (Fig. 1A). As with the assembly quality comparison, two long-read assemblies yielded an exceptionally high number of SSU rRNA genes, suggesting that the best long-read metagenomes can outperform the best short-read metagenomes in this respect. However, the overall pattern did not demonstrate a consistent advantage of long-read sequencing for generating high-quality SSU rRNA genes. Moreover, the average lengths of the extracted genes were also similar across assemblies (Fig. 1B), which was likely directly related to the similar confidence levels of taxonomic assignments (Fig. 1C).

**TABLE 3** Source, quality, metrics, and taxonomic assignment for the 15 high-quality, dereplicated MAGs binned from all long-read and hybrid assemblies[a]

| MAG ID | Assembly type | Binning method | Quality | Total length (bp) | Avg contig length (bp) | Longest contig (bp) | Domain | Phylum | Class | Order | Family | Genus | Species | Short-read metagenomes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S11C7135 | hybridSPAdes | Vamb | 94.4 | 2,037,740 | 7,125 | 36,998 | *Bacteria* | *Bacteroidota* | *Bacteroidia* | *Flavobacteriales* | *Flavobacteriaceae* | *Polaribacter* | None | X |
| S11C1624 | hybridSPAdes | Vamb | 90.1 | 1,815,185 | 7,439.3 | 42,585 | *Bacteria* | *Verrucomicrobiota* | *Verrucomicrobiae* | *Opitutales* | *Puniceicoccaceae* | BACL24 | BACL24 sp002292345 | X |
| S5C23781 | hybridSPAdes | Vamb | 77.5 | 1,154,357 | 29,598.90 | 92,412 | *Bacteria* | *Proteobacteria* | *Alphaproteobacteria* | *Rhodobacterales* | *Rhodobacteraceae* | *Amylibacter* | *Amylibacter* sp000153745 | X |
| S14C23837 | hybridSPAdes | Vamb | 74.7 | 628,173 | 29,913 | 97,354 | *Bacteria* | *Actinobacteriota* | *Acidimicrobiia* | TMED189 | TMED189 | TMED189 | TMED189 sp002683085 | X |
| S14C7553 | hybridSPAdes | Vamb | 71.1 | 1,003,441 | 5,067.9 | 24,945 | *Archaea* | *Thermoplasmatota* | *Poseidoniia* | *Poseidoniales* | *Thalassarchaeaceae* | MGIIb-O5 | MGIIb-O5 sp002726275 | X |
| S5C23599 | hybridSPAdes | Vamb | 71.0 | 1,785,992 | 24,465.60 | 143,545 | *Archaea* | *Thermoplasmatota* | *Poseidoniia* | *Poseidoniales* | *Poseidoniaceae* | MGIIa-L1 | MGIIa-L1 sp002697705 | X |
| S14C2330 | hybridSPAdes | Vamb | 67.1 | 1,090,873 | 5,896.60 | 26,986 | *Archaea* | *Thermoplasmatota* | *Poseidoniia* | *Poseidoniales* | *Thalassarchaeaceae* | MGIIb-O1 | None | X |
| S5C24179 | hybridSPAdes | Vamb | 47.4 | 1,475,824 | 56,762.50 | 144,120 | *Archaea* | *Thermoplasmatota* | *Poseidoniia* | *Poseidoniales* | *Poseidoniaceae* | MGIIa-L1 | MGIIa-L1 sp000246735 | X |
| 1903c122_28m_MaxBin2_016 | hybridSPAdes | MaxBin2 | 40.8 | 1,458,304 | 2,179.8 | 12,419 | *Bacteria* | *Proteobacteria* | *Gammaproteobacteria* | *Pseudomonadales* | *Halieaceae* | *Luminiphilus* | *Luminiphilus* sp000169115 | X |
| S14C286 | hifiasm-meta | Vamb | 57.8 | 1,594,643 | 46,901.30 | 299,336 | *Bacteria* | *Bacteroidota* | *Bacteroidia* | *Flavobacteriales* | *Flavobacteriaceae* | MS024-2A | MS024-2A sp002169865 | X |
| S8C188 | hifiasm-meta | Vamb | 46.5 | 1,184,433 | 26,320.70 | 71,595 | *Bacteria* | *Actinobacteriota* | *Acidimicrobiia* | *Acidimicrobiales* | TK06 | UBA9040 | UBA9040 sp003452475 | |
| Las19c135_5m_MaxBin2_001 | hifiasm-meta | MaxBin2 | 49.3 | 1,207,303 | 30,956.50 | 316,715 | *Bacteria* | *Bacteroidota* | *Bacteroidia* | *Flavobacteriales* | *Flavobacteriaceae* | Hel1-33-131 | Hel1-33-131 sp001735745 | |
| S14C2310 | metaFlye | Vamb | 70.4 | 2,673,573 | 53,471.50 | 292,039 | *Bacteria* | *Verrucomicrobiota* | *Verrucomicrobiae* | *Verrucomicrobiales* | *Akkermansiaceae* | SW10 | SW10 sp002323895 | |
| Las19c138_27m-1_MaxBin2_012-refined-v2 | hybridSPAdes | MaxBin2 | 85.5 | 12,229,370 | 3,054.30 | 33,976 | *Eukaryota* | Viridiplantae/ Chlorophyta | Mamiellophyceae | Mammiellales | Bathycoccaceae | *Bathycoccus* | *Bathycoccus prasinos* | |
| 1903c126_45m_MaxBin2_002-refined-v2 | hybridSPAdes | MaxBin2 | 84.3 | 11,803,494 | 7,991.50 | 54,512 | *Eukaryota* | Viridiplantae/ Chlorophyta | Mamiellophyceae | Mammiellales | Bathycoccaceae | *Ostreococcus* | *Ostreococcus lucimarinus* | |

[a]The two MAGs with "refined-v2" in the identifier were manually refined in anvi'o. The final column denotes whether or not a MAG classified to the same genus level was also recovered from any short-read metagenome.

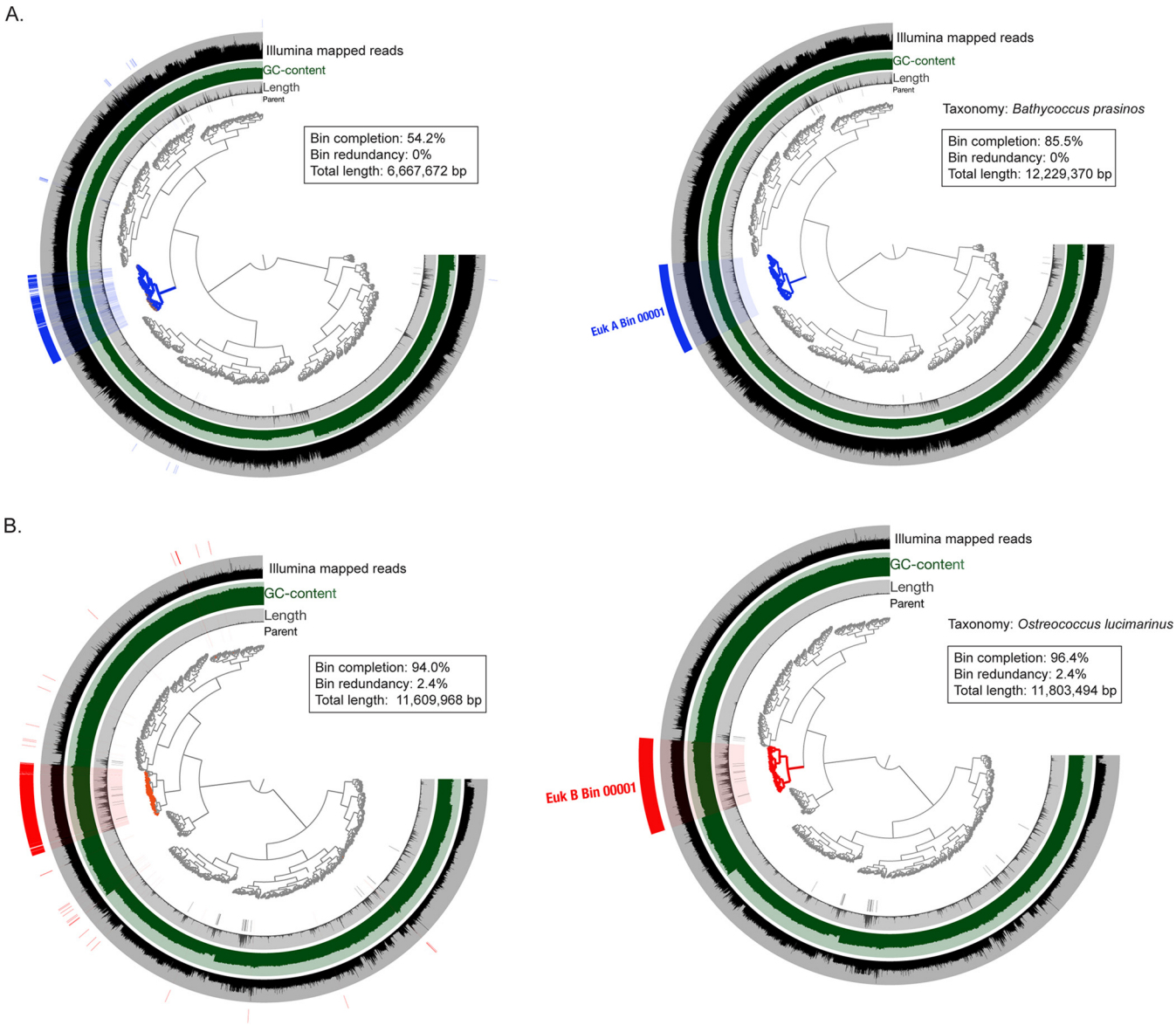
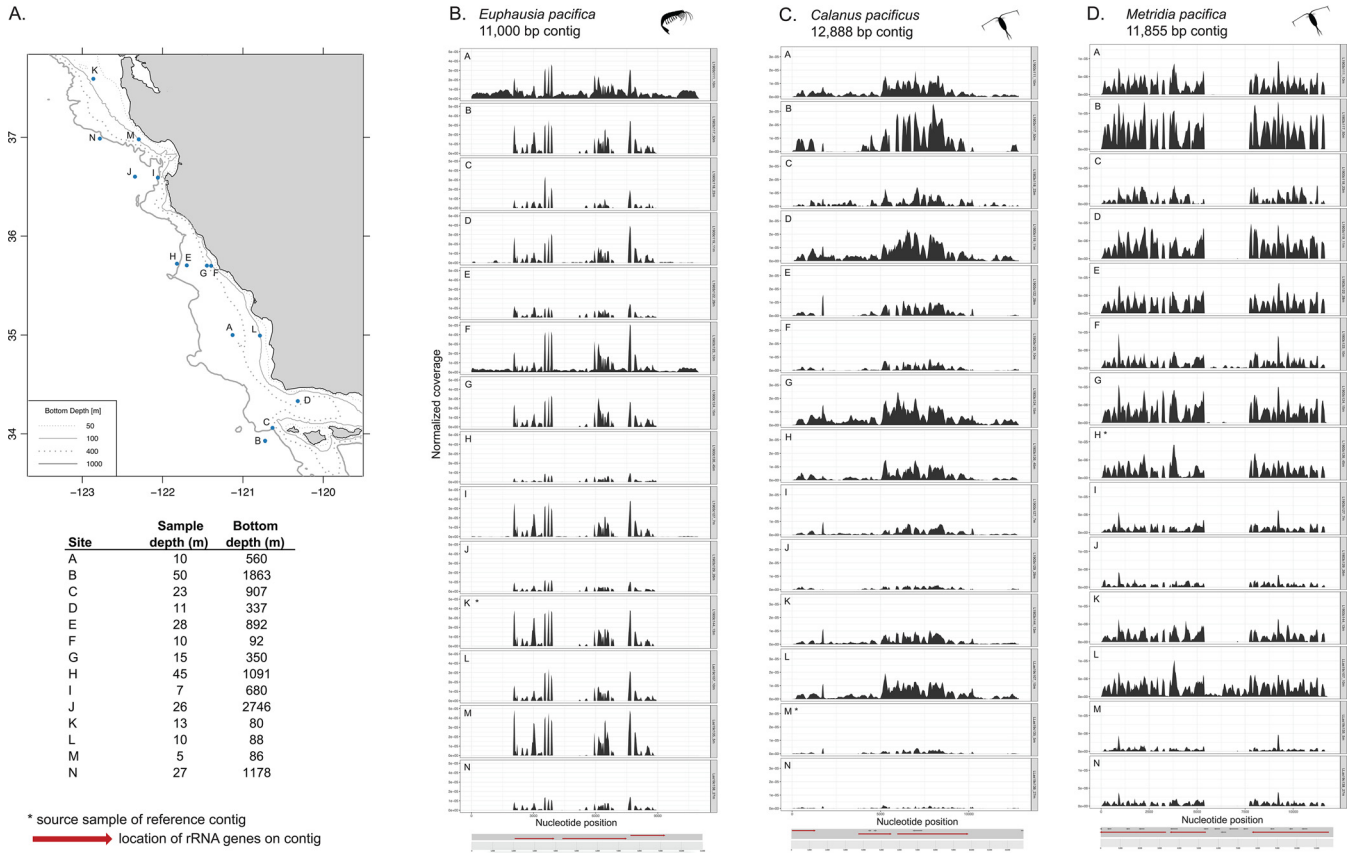



**FIG 4** Visual representation of the picoeukaryotic MAGs and their source metagenomes, before and after manual refinement, generated by anvi'o. The phylogram branches represent the metagenome contigs clustered by tetranucleotide frequency and coverage. On the left, results from the automated binning program show contigs belonging to the initial MAG. On the right, the final MAG contigs following manual refinement are shown. Completion, redundancy, and total MAG size are provided for each plot. The final taxonomic assignment is provided for the refined MAG. (A) *Bathycoccus prasinos* MAG before and after manual refinement. (B) *Ostreococcus lucimarinus* MAG before and after manual refinement.

The discrepancies between matching short- and long-read metagenomes suggested that community richness, evenness, and composition were not well replicated with the two sequencing technologies (Fig. 2). The wider range of alpha diversity values and differences in beta diversity were true for both read-based (Fig. 2A and B) and assembly-based (Fig. 2C and D) analyses. Thus, despite the loss of information from short reads to assembled contigs, the assembly process did not lead to more comparable results between short- and long-read metagenomes. The variability among long-read metagenomes may have been due in part to the metagenome size; samples with relatively few (<50,000) long reads could be linked to only a few taxa, while larger metagenomes showed a wide range of marine taxa (see the Krona plots at https://doi.org/10.6084/m9.figshare.20883652.v1). There was no strong linear correlation between the numbers of metagenome reads and Shannon diversity values for the read-based (Illumina $R^2 = 0.184$; PacBio $R^2 = 0.318$) or assembly-based (Illumina $R^2 = 0.0072$; PacBio $R^2 = 0.241$) analyses,




| Site | Sample depth (m) | Bottom depth (m) |
|---|---|---|
| A | 10 | 560 |
| B | 50 | 1863 |
| C | 23 | 907 |
| D | 11 | 337 |
| E | 28 | 892 |
| F | 10 | 92 |
| G | 15 | 350 |
| H | 45 | 1091 |
| I | 7 | 680 |
| J | 26 | 2746 |
| K | 13 | 80 |
| L | 10 | 88 |
| M | 5 | 86 |
| N | 27 | 1178 |

**FIG 5** Short-read metagenome coverage for three contigs generated from long-read assemblies. Each contig was annotated as a different zooplankton species according to 18S and 28S rRNA gene sequences located on the contig. Coverage values were normalized to metagenome sizes, and the *y* axes for all panels within each plot are set to the same scale. (A) Map showing sample collection sites and table with associated bottom depths. (B) The contig annotated as the krill species *Euphausia pacifica* recruited short reads across the entire contig length from sites A and J. Short reads from all other metagenomes mapped only to the rRNA gene regions. (C) The contig annotated as the copepod species *Calanus pacificus* recruited short reads across the entire contig length from about half of the sites, while reads from the other half mapped only to the rRNA gene regions. (D) The contig annotated as the copepod species *Metridia pacifica* recruited short reads across the entire contig length from sites F and L. Short reads from all other metagenomes mapped only to the rRNA gene regions.

although there was a stronger logarithmic correlation with the assembly-based PacBio analysis ($R^2 = 0.405$). These results suggest much more stochasticity or, possibly, taxonomic bias in PacBio long-read sequencing technology relative to short-read Illumina platforms. Alternatively, the DNA extraction protocols used for the different sample sets may have yielded DNA enriched in different taxa.

Paradoxically, the total numbers of annotated ORFs were comparable among all assembly types because the higher total numbers of ORFs in the short-read and hybrid assemblies were balanced out by the smaller fractions of annotated ORFs (Fig. 3). The larger contigs from the long-read assemblies likely contributed to the higher levels of annotations. However, the similar total numbers of annotations did not lead to similar alpha diversity values (Fig. 2C) and were not reflected in taxonomic composition similarities (Fig. 2D). Moreover, the lower levels of KEGG module completeness (Table 1) suggest that a higher fraction of annotated genes were not grouped into larger gene pathways and originated from different fragments of DNA, while short-read and hybrid assemblies performed better at assembling sequences from longer genomic regions. Thus, one of the main potential advantages of long-read sequencing was not observed in the functional annotation of genes and KEGG modules.

Two-thirds of the prokaryotic MAGs recovered from the long-read and hybrid assemblies were also generated from the matching short-read metagenomes (Table 3), suggesting a moderate advantage of long-read sequencing for assembling high-quality microbial genomes. The MAGs unique to the long-read and hybrid metagenomes were

generally well represented in the short-read metagenomes, suggesting that these populations were present in the replicate filter samples but benefited from long-read sequencing for binning (see the read mapping histograms at https://doi.org/10.6084/m9.figshare.21318000). One possible reason for the additional MAG recovery is that the long reads were assembled with two different programs, while the short reads were run through metaSPAdes only (albeit with duplicate metagenomes compared to a single long-read metagenome). Assemblers are known to vary in assembly output (which is also the binning input), and having a second, different long-read assembly for each sample may have been an important source of additional MAGs. However, the average length of the longest contig for the five unique long-read MAGs was >150,000 bp, versus 92,000 bp for the MAGs also recovered from short-read metagenomes (Table 3). The higher number of long contigs from long-read assemblies may thus have contributed to the higher-quality MAGs retained in the final count. One of the long-read-only MAGs belonged to the species SW10 sp002323895 (phylum *Verrucomicrobiota*), which was also the taxonomy of the longest long-read contig (802,392 bp from the sample Las19c138_27m-3 metaFlye assembly); surprisingly, however, the MAG did not contain this ultralong contig. This omission may have been due to flawed binning programs or because the 800-kbp contig belonged to a subpopulation of SW10 sp002323895 distinct from the MAG. Regardless of the cause, this taxon may be an example of the benefits of long-read sequencing due to genome characteristics that prevent assembly from short reads alone.

One important advantage of supplementing short-read sequences with long reads was the recovery of two eukaryotic MAGs from hybrid long- and short-read assemblies. Following manual refinement, the MAGs were classified as belonging to *Bathycoccus prasinos* and *Ostreococcus lucimarinus*, respectively (Fig. 4; Fig. S4). *Bathycoccus* and *Ostreococcus* are two of the three major cosmopolitan marine picoeukaryotic genera (the third being *Micromonas*), which comprise basal lineages of the phylum Chlorophyta. These single-celled algae are characterized by tiny cell sizes (~1 $\mu$m in diameter) and genomes (10 to 15 Mbp), which can provide valuable information on the minimum gene requirements for free-living eukaryotic phytoplankton as well as the subsequent evolution of plant lineages (39, 40). There are genome sequences available from cultured strains of both *Ostreococcus* and *Bathycoccus*, including *O. lucimarinus* isolated from coastal Southern California (40), and recent studies have presented picoeukaryotic MAGs from the Atlantic, South Pacific, Indian, and Arctic Oceans as well as the Mediterranean and Red Seas (41, 42). To our knowledge, there are no *Bathycoccus* whole genomes or MAGs from the California Current. These MAGs, and potentially many more that can be generated from long-read sequencing, may thus help fill in knowledge gaps on phytoplankton biogeography and evolution (43). While an extensive metabolic characterization is beyond the scope of this study, the gene contents of the two MAGs appear similar at the level of KEGG categorization, with a few exceptions in specific glycan and lipid metabolism gene categories (Fig. S5). We also note that the genomes of the *B. prasinos* isolates feature two "outlier" chromosomes with different GC contents (potentially influenced by horizontal gene transfer) (39) that we may not have captured in our analysis.

We identified several long contigs containing entire 18S rRNA gene sequences belonging to ecologically important zooplankton species from metaFlye assemblies. In particular, contigs annotated as the krill species *Euphausia pacifica* and two copepod species, *Calanus pacificus* and *Metridia pacifica*, were used as reference contigs for short-read mapping to compare coverages at all sampling sites (*Euphausia* recruitment plots: https://doi.org/10.6084/m9.figshare.21308184, *Metridia* recruitment plots: https://doi.org/10.6084/m9.figshare.21308175, *Calanus* recruitment plots: https://doi.org/10.6084/m9.figshare.21308169). Krill and copepod species have been linked to large-scale oceanographic processes, including El Niño-Southern Oscillation (ENSO) events and warm water anomalies (44, 45); improving eDNA detection methods would thus assist ecosystem monitoring efforts. The read coverage was generally low for all samples, reflecting the relatively small proportion of eukaryotic DNA in the eDNA pool. Site J in particular featured almost no reads mapping to any of the eukaryotic reference contigs (Fig. 5);

this was also the site with the deepest bottom depth, suggesting that bathymetry may influence surface zooplankton communities in ways that can be detected with eDNA. However, because it is challenging to link gene copy numbers with biomass in multicellular organisms, we do not make any inferences regarding the relative abundances of copepods or krill in our samples. Additionally, the extent of genetic variability among the krill and copepod species in the California Current is poorly understood (37); coverage values include reads with <100% sequence identity (Fig. S7; see also the read recruitment plots cited above), which may not represent the same species as that of the reference contig. Moreover, short reads from metagenomes corresponding to the long-read metagenomes from which the reference contigs were generated did not show complete coverage, possibly because the zooplankton DNA was absent from the biological eDNA replicates. This observation suggests a level of stochasticity in the sampling and/or sequencing processes that precludes conclusions about organism genetic relatedness from metagenomes. Differences in coverage between genic and intergenic regions, along with the percent identities of mapped reads, show a range of mapping fidelities among samples (Fig. S7; see also the read recruitment plots cited above). Despite these sources of uncertainty, however, the coverage patterns across the contig sequence were generally conserved across sites. The rRNA marker genes showed similar peaks and valleys in recruitment levels, showing which gene regions are more or less conserved across a spatial range in the California Current. Data linking zooplankton genetics with biogeography are rare but potentially invaluable, as modeling planktonic food webs is highly challenging and suffers from a dearth of environmental data (46). A better understanding of the genetic variability among closely related zooplankton species will improve the interpretation of coverage results and provide valuable data on biogeography and population connectivity.

In conclusion, we found that long-read metagenomes of marine eDNA samples did not reliably reproduce the results on microbial community composition and function from replicate short-read metagenomes. This may be due to the inconsistent quality of long-read sequencing in this study. There may also have been biological differences between the replicate eDNA samples due to the different DNA extraction protocols used on the filters destined for short- versus long-read sequencing. Additionally, assembly qualities, the extraction of SSU rRNA genes, and the recovery of prokaryotic MAGs were only marginally improved by long-read sequencing. However, only long-read metagenomes provided high-quality eukaryotic MAGs. These genomes fill valuable gaps in knowledge on an understudied component of marine planktonic communities in the California Current. Moreover, long-read contigs provided full sequences of marker genes from ecologically important zooplankton species, including krill and copepods. These results suggest that a combination of short- and long-read metagenomes can generate a more complete overview of the planktonic community than amplicon sequencing or short-read metagenomic sequencing alone. However, to go beyond presence/absence information for eukaryotic organisms, a better understanding of genetic variability (e.g., the level of conservation of intergenic SSU rRNA sequences) among species is necessary. This study provides a baseline for investigating these questions and leveraging eDNA sequence data to complement morphology-based surveys in marine ecosystems.

## MATERIALS AND METHODS

**Sample collection.** Samples were collected during the 2019 Rockfish Recruitment and Ecosystem Assessment Survey (RREAS) aboard the NOAA Ship *Reuben Lasker* between 30 May and 7 June 2019. At each site, water was collected from three depths using Niskin bottles mounted on a Conductivity, Temperature, and Depth (CTD) rosette. Three bottles each were triggered at 5 m, 100 m, and the deep chlorophyll maximum, which varied among sites. Triplicate 1-L samples from each depth were collected, with each triplicate containing approximately 1/3 L water from each of the three bottles from the sampled depths. Water was mixed in 2-L Whirl-Pak bags before filtration onto 0.22-$\mu$m, 47-mm polyvinylidene difluoride (PVDF) membrane filters using vacuum pumps. Filters were stored in 2-mL cryotubes at −80°C until the end of the survey, at which point they were transported on dry ice to the Monterey Bay Aquarium Research Institute (MBARI), where they were stored at −80°C.

**DNA extraction, library preparation, and sequencing.** Two of the triplicate filters were shipped on dry ice to Oregon State University's Center for Quantitative Life Sciences Core Facilities for DNA

extractions in preparation for short-read sequencing. Extractions were performed using the Omega Bio-Tek Mag-Bind blood and tissue DNA HDQ 96 kit according to the manufacturer's instructions and a KingFisher instrument to automate extractions. DNA was shipped to CosmosID on dry ice for library preparation and sequencing.

DNA extractions for long-read sequencing were performed at CosmosID. The third replicate filters from each sample were shipped on dry ice to the facility, and DNA was extracted using the OLX/DNAexpress sample kit for the isolation of DNA from marine filters (Claremont Biosolutions, Upland, CA) according to the manufacturer's instructions. DNA was quantified using a Qubit fluorometer (Invitrogen, Waltham, MA).

Both short- and long-read library preparation and metagenomic sequencing were performed at CosmosID. Short-read metagenome libraries were prepared using the Nextera XT protocol, and sequencing was performed on an Illumina HiSeq platform using a 2× 150-bp paired-end (PE) protocol. The long-read sequencing library was prepared for PacBio sequencing according to the manufacturer's instructions for preparing HiFi SMRTbell libraries from ultralow DNA input (https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-SMRTbell-Libraries-from-Ultra-Low-DNA-Input-.pdf). The library was sequenced on one 8M SMRT cell on a PacBio Sequel II platform to generate circular consensus sequencing (CCS) data.

**Short-read bioinformatics. (i) Sequence quality control and decontamination.** Each set of 14 matching Illumina short-read metagenomes was processed for comparison with long reads. Raw Illumina reads were quality filtered with fastp (47) using the parameters –cut_tail -g -l 100 -W 3 -M 30 -w 16 –adapter_fasta adapters.fa; the adapter Fasta file is provided in the manuscript FigShare repository (https://doi.org/10.6084/m9.figshare.21554595.v1). Human DNA was removed from the quality-controlled sequences using BMTagger by identifying reads matching the latest human genome sequence (GRCh38_latest_genomic). All further analyses were performed using quality-controlled, BMTagger-filtered reads.

**(ii) Assembly and binning.** Metagenome assemblies were generated using the quality-controlled, BMTagger-filtered reads in metaSPAdes (48). Metagenomes that failed to assemble were run through BBNorm (https://sourceforge.net/projects/bbmap/) with the parameters target=100 and min=5 (see Table S1 in the supplemental material), and the normalized reads were rerun through metaSPAdes. Assembly qualities were assessed using Quast (49).

Binning was performed using two different programs: MaxBin2 (38) and Vamb (50). The assemblies of all short-read metagenomes (including both replicates) were run individually through MaxBin2 with a minimum contig length of 1,000 bp and concatenated for input to Vamb according to the instructions provided in the GitHub repository (https://github.com/RasmussenLab/vamb). The resulting MAGs from both binning programs were assessed for quality using CheckM (51) and anvi'o (52) (anvi-gen-contigs-database, anvi-run-hmms, and anvi-estimate-genome-completeness commands); quality scores were generated by calculating completion − (5× contamination) (CheckM) and completion − (5× redundancy) (anvi'o), and MAGs were considered to be of high quality if both the CheckM and anvi'o quality scores were above 40. High-quality MAGs were dereplicated using dRep (53) with a 0.95 average nucleotide identity (ANI) threshold.

**(iii) Taxonomic and functional annotation.** The taxonomic compositions of the metagenomes were assessed in three ways: (i) with unassembled reads, (ii) with assembled open reading frames (ORFs), and (iii) with extracted SSU rRNA sequences.

For the read-based analysis, reads were first interleaved using the mergepe command in seqtk (https://github.com/lh3/seqtk). Metagenome signatures were generated with the sourmash sketch dna command (54, 55), and the resulting signatures were run against the GTDB (56) and Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP) (57) preprepared databases (https://osf.io/wxf9z/) using the sourmash gather command (58, bioRxiv). Finally, the gather output was matched to the taxonomy using sourmash tax metagenome and the GTDB and MMETSP taxonomy reference sheets (https://osf.io/wxf9z/). Both Krona and CSV output formats were generated, and a Krona plot was generated using ktimporttext in a Krona Conda environment. All Krona taxonomy summaries and plots can be found at https://doi.org/10.6084/m9.figshare.20883652.v1.

For taxonomic annotation of the assemblies, ORFs were generated with Prodigal (59) and run against the NCBI-nr database using DIAMOND (60) with BLASTp and "very-sensitive" parameters. DIAMOND output files were parsed using a custom Python script (parse_DIAMOND_output_for_qiime.ipynb [https://doi.org/10.6084/m9.figshare.21320475.v1]), and taxonomy was assigned using taxonkit with the lineage command (60). In addition, ORFs were run through MicrobeAnnotator (61) for estimates of KEGG module numbers and completeness per metagenome.

SSU rRNA sequences were extracted in two ways: (i) with PhyloFlash (62) for 16S and 18S rRNA genes and (ii) using rRNA HMM databases in anvi'o for 16S, 18S, 23S, and 28S rRNA genes (anvi-run-hmms and anvi-get-sequences-for-hmm-hits commands) (52). The 16S and 18S rRNA genes extracted using both approaches were run against the SILVA (64) and PR2 (65) reference databases, respectively, using the vsearch sintax taxonomy classifier with no confidence cutoff (66). The RESCRIPt-formatted SILVA database was downloaded from https://docs.qiime2.org/2021.2/data-resources/#silva-16s-18s-rrna and formatted for sintax according to the vsearch instructions at https://github.com/torognes/vsearch/issues/438. The RESCRIPt-formatted PR2 database was downloaded from https://github.com/pr2database/pr2database/releases.

The vsearch output was extracted, and confidence levels at each taxonomic level were tested for significant differences using one-way analysis of variance (ANOVA) followed by Tukey's honestly significant difference (HSD) test using the Python libraries SciPy and bioinfokit. Analysis details can be found in the parse_vsearch_sintax_output_for_taxa_confidences.ipynb Jupyter notebook at https://doi.org/10.6084/m9.figshare.21320475.v1. SSU genes *Escherichia coli*, *Halobacterium salinarum*, and *Carcharadon*

*carcharias* were queried in NCBI GenBank and their lengths were used as references for bacterial, archaeal, and eukaryotic domains in Figure 1, respectively.

**Long-read bioinformatics. (i) Sequence decontamination.** Unlike the short-read metagenomes, the PacBio metagenomes were not found to contain human DNA sequences. However, four PacBio samples were filtered for likely laboratory microbial contamination belonging to the nonmarine taxa *Delftia acidovorans*, *Delftia lacustris*, *Cutibacterium acnes*, and *Stenotrophomonas maltophilia*, identified from taxonomic annotation results using sourmash and GTDB (see below for full taxonomic annotation methods). These four reference genomes were downloaded from GenBank, and filtered PacBio reads were aligned to the genome Fasta files using minimap2 (67). Unmapped (decontaminated) reads were then used for all downstream analyses.

**(ii) Assembly, binning, and read mapping.** The 10 raw and 4 decontaminated samples were assembled using three different approaches: hifiasm-meta (68) of long reads only, metaFlye (69) of long reads only, and hybridSPAdes (70) using both long PacBio and short Illumina reads. For the latter, the corresponding duplicate Illumina metagenomes were compared, and the sample with the higher number of reads was chosen for the hybrid assembly. Assembly qualities were assessed using Quast.

Binning (MaxBin2 and Vamb), MAG quality assessment, and MAG dereplication were performed as described above for the short-read assemblies. The short-read metagenomes used for long-read and hybrid binning are shown in Table S1. Annotation with MicrobeAnnotator and rRNA HMMs were also performed as described above.

Reads from all short-read metagenomes were competitively mapped to all 15 high-quality, dereplicated MAGs from long-read and hybrid assemblies using Magic BLAST (71). The results were filtered for a minimum read length of 70 bp and a minimum alignment length/read length value of 0.7 using the Python script 01c_MagicBlast_ShortRead_Filter.py, found at https://github.com/rotheconrad/00_in-situ_GeneCoverage, and read recruitment plots were generated using the Recplot_4 R package (https://github.com/KGerhardt/Recplot_4). Plots show reads aligned at a 0.5% depth resolution and a 100-bp read length resolution.

Additionally, two metaFlye assemblies (Las19c135_5m-3 and Las19c138_27m-3) each were found to have one exceptionally long contig (~800 kbp). Reads from the corresponding short-read metagenomes (Las19c135_5m-1, Las19c135_5m-2, and Las19c138_27m-1) were mapped to the contigs to assess the coverage depth across the contig sequence. Reads were mapped for the two short-read metagenomes mapped against the contig from Las19c135_5m-3 and the single short-read metagenome mapped against the contig from Las19c138_27m-3. Read mapping, filtering, and plot generation were performed as described above for the MAGs.

**(iii) Taxonomic and functional annotation.** The taxonomic composition of long-read metagenomes was assessed using the same three approaches as the ones described above for the short-read metagenomes: running sourmash (54) and FracMinHash (Irber et al., 2022, bioRxiv) against GTDB (56) and MMETSP (57), annotating the assembled ORFs with DIAMOND (72) against the NCBI-nr database, and running the extracted 16S and 18S rRNA genes against SILVA (64) and PR2, respectively. Reads were not interleaved since they were not generated with paired-end chemistry. PhyloFlash (63) is designed for short reads and thus was not used for extracting SSU rRNA genes; only rRNA HMM databases were used in anvi'o to extract 16S, 18S, 23S, and 28S rRNA genes.

**Analyses applied to both short- and long-read metagenomes. (i) Assembly read mapping.** Short reads from both Illumina metagenome replicates (when available) were mapped to all four sets of assemblies: short reads assembled with SPAdes, long reads assembled with hifiasm-meta, long reads assembled with metaFlye, and short and long reads assembled with hybridSPAdes. minimap2 (63) was used for read mapping, and the resulting bam files were parsed with SAMtools (73) to generate counts for mapped and unmapped reads. The unmapped reads were extracted with SAMtools and run through sourmash with the GTDB and MMETSP databases as described above to assess the taxonomic composition.

**(ii) Alpha diversity.** Shannon diversity was estimated with the R package vegan (74) for Illumina and PacBio metagenomes using sourmash taxonomic composition results. The same estimation was applied to Illumina (SPAdes), PacBio (metaFlye), and hybrid assemblies using ORF taxonomic annotation counts as the taxonomy table input. Shannon diversity values were plotted against metagenome read numbers for both read-based and assembly-based analyses, and correlation coefficients were calculated for linear and logarithmic trend lines in Excel. The code for alpha diversity analyses is contained in the sourmash_alpha_beta_diversity.R and ORF_alpha_beta_diversity.R scripts at the GitHub repository (https://doi.org/10.6084/m9.figshare.21320475.v1).

**(iii) Beta diversity.** Aitchison distances among all metagenomes (PacBio and Illumina samples) were calculated using DEICODE for a robust Aitchison principal-component analysis (PCA) (75) for both read-based (sourmash) and assembly-based (DIAMOND) taxonomic composition results. The DEICODE input for the read-based analysis was a BIOM-formatted taxonomic composition table of the number of base pairs assigned to each species by sourmash (Python script merge_sourmash_results.ipynb at the GitHub URL mentioned above), and for the assembly-based analysis, it was a BIOM-formatted count table generated for the number of ORFs assigned to each taxon by the DIAMOND analysis described above. The DEICODE rpca command was run using default parameters for the read-based analysis and with a min-feature-frequency 40 parameter for the assembly-based analysis. The resulting distance and ordination files were imported into the R package qiime2R (76) and used to generate a PCA plot using ggplot2 (77). PERMANOVAs were run using the adonis2 function in the vegan package (74) to test for significant differences among depth groups and sequencing platforms. Based on these results (differences were significant between sequencing platforms but not among depth groups), Aitchison distances between all pairs of samples were plotted as a histogram, and differences between paired and unpaired samples were tested by analysis of variance (ANOVA) using the aov function in the R vegan package. The code for the read-based and assembly-based beta diversity

analyses can be found in the sourmash_alpha_beta_diversity.R and ORFs_alpha_beta_diversity.R scripts, respectively, at GitHub (https://doi.org/10.6084/m9.figshare.21320475.v1).

**(iv) MAG refinement and annotation.** One smaller MAG with an ambiguous taxonomic assignment was examined using the anvi'o interactive interface and determined to be highly contaminated. Manual refinement did not improve the quality or taxonomic assignment confidence with GTDB-Tk v2.0.2 (78) (data not shown). The MAG was removed from the final count of high-quality, dereplicated MAGs.

Two large (>10-Mb) MAGs with taxonomic assignments limited to the phylum level were refined using the anvi'o interactive interface (Fig. 4). MAGs were annotated with BUSCO (79), and the resulting single-copy genes were run against the NCBI-nr database using DIAMOND (72). DNA-dependent RNA polymerase genes from both MAGs were identified using HMM databases for the a and b enzyme subunits from Delmont et al. (2022). The extracted genes for both subunits were aligned with the corresponding database sequences belonging to the genera *Micromonas*, *Ostreococcus*, *Bathycoccus*, and *Pycnococcus* using MAFFT (v.7.490) (80). Phylogenies of the two alignments were built using RAxML (v.8.2.12) (81) with 100 bootstraps and the PROTGAMMAAUTO parameter to automatically determine the best model. The resulting trees were visualized with iTOL (82), and colors and fonts were edited in Adobe Illustrator.

**(v) Eukaryotic rRNA annotation and coverage.** To identify potentially ecologically important eukaryotic organisms in metagenomes, metaFlye, hybridSPAdes, and Illumina assemblies were run against HMM databases in anvi'o to extract 5S, 12S, 18S, 23S, and 28S rRNA genes. The resulting 18S rRNA hits were run against the NCBI-nr database with BLAST for taxonomic annotation. Based on these results, the longest contigs containing genes annotated as *Calanus* sp., *Metridia* sp., and *Euphausia pacifica* genes, respectively, were used as reference contigs for coverage calculations.

To calculate krill and copepod 18S rRNA gene coverages in each short-read metagenome, reads from all matching Illumina samples were aligned against the assemblies using minimap2 (67). Assemblies and mapping results were visualized in anvi'o by merging individual profiles, and the coverage across the contig of interest was examined by calculating split coverages (anvi-get-split-coverages), manually normalizing the reported coverages by the short-read metagenome size, and visualizing the normalized results (anvi-script-visualize-split-coverages).

For the fine-scale visualization of read mapping levels, the same matching Illumina sample reads were aligned with the reference contig of interest using Magic BLAST (71). The results were filtered to retain hits with a minimum read length of 70 bp and a ratio of the alignment length to the read length of 0.7 using the Python script 01c_MagicBlast_ShortRead_Filter.py found at https://github.com/rotheconrad/00_in-situ_GeneCoverage. Read recruitment plots were generated with the filtered Magic BLAST results using the Recplot_4 package in R (https://github.com/KGerhardt/Recplot_4). Plots show reads aligned at a 0.5% depth resolution and a 100-bp read length resolution.

The extracted *Metridia* sp. 18S and 28S rRNA gene sequences were aligned with 18S and 28S rRNA genes from several species of *Metridia* from GenBank. Alignments were performed with MUSCLE (83), and maximum likelihood phylogenies were built using MEGA (84).

**Data availability.** Raw PacBio and Illumina sequencing reads are available in the NCBI Sequence Read Archive (BioProject accession number PRJNA853328). All high-quality, dereplicated MAGs generated from PacBio sequence data, additional data generated from the analyses described in this article, and code for statistical analyses and data visualizations are available at https://figshare.com/projects/PacBio_Marine_Metagenomes/144459.

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**FIG S1**, PDF file, 0.8 MB.
**FIG S2**, PDF file, 0.6 MB.
**FIG S3**, PDF file, 0.9 MB.
**FIG S4**, PDF file, 0.3 MB.
**FIG S5**, PDF file, 2.1 MB.
**FIG S6**, PDF file, 0.8 MB.
**FIG S7**, PDF file, 1.1 MB.
**TABLE S1**, PDF file, 0.02 MB.
**TABLE S2**, PDF file, 0.02 MB.
**TABLE S3**, PDF file, 0.01 MB.

## ACKNOWLEDGMENTS

We thank the captain and crew of the NOAA Ship *Reuben Lasker* for facilitating safe and effective water sampling in conjunction with the fisheries survey. We are also grateful to Francisco Chavez, Kobun Truelove, and Katherine Pitz at MBARI for arranging the DNA extractions at Oregon State University and providing valuable suggestions for marine water column filter DNA extraction protocols for long-read sequencing. CosmosID provided both long-read and short-read sequencing services, and we are grateful for scientific discussions

with CosmosID staff regarding this study. Finally, we thank Leocadio Blanco-Bercial and Kym Jacobson for their guidance on zooplankton genetic variability.

This work was made possible by a grant from the National Oceanographic Partnership Program (NOPP) and the NOAA/OAR 'Omics Program and was carried out in part under the auspices of the Cooperative Institute of Marine and Atmospheric Studies (CIMAS), a Cooperative Institute of the University of Miami and the National Oceanic and Atmospheric Administration, under cooperative agreement number NA20OAR4320472.

## REFERENCES

1. Martinez-Gutierrez CA, Aylward FO. 2022. Evolutionary genomics of marine bacteria and archaea, p 327–354. *In* Stal LJ, Cretoiu MS (ed), The marine microbiome, 2nd ed. Springer International Publishing, Cham, Switzerland.
2. Needham DM, Fichot EB, Wang E, Berdjeb L, Cram JA, Fichot CG, Fuhrman JA. 2018. Dynamics and interactions of highly resolved marine plankton via automated high-frequency sampling. ISME J 12:2417–2432. https://doi.org/10.1038/s41396-018-0169-y.
3. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-H, Smith HO. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74. https://doi.org/10.1126/science.1093857.
4. Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, Welch DBM, Martiny JBH, Sogin M, Boetius A, Ramette A. 2011. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. PLoS One 6:e24570. https://doi.org/10.1371/journal.pone.0024570.
5. Anderson RE, Reveillaud J, Reddington E, Delmont TO, Eren AM, McDermott JM, Seewald JS, Huber JA. 2017. Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents. Nat Commun 8:1114. https://doi.org/10.1038/s41467-017-01228-6.
6. Colatriano D, Tran PQ, Guéguen C, Williams WJ, Lovejoy C, Walsh DA. 2018. Genomic evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean Chloroflexi bacteria. Commun Biol 1:90. https://doi.org/10.1038/s42003-018-0086-7.
7. Castelle CJ, Banfield JF. 2018. Major new microbial groups expand diversity and alter our understanding of the tree of life. Cell 172:1181–1197. https://doi.org/10.1016/j.cell.2018.02.016.
8. Aylward FO, Santoro AE. 2020. Heterotrophic Thaumarchaea with small genomes are widespread in the dark ocean. mSystems 5(3):e00415-20. https://doi.org/10.1128/mSystems.00415-20.
9. Sun X, Ward BB. 2021. Novel metagenome-assembled genomes involved in the nitrogen cycle from a Pacific oxygen minimum zone. ISME Commun 1:26. https://doi.org/10.1038/s43705-021-00030-2.
10. Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, Darzi Y, Audic S, Berline L, Brum J, Coelho LP, Espinoza JCI, Malviya S, Sunagawa S, Dimier C, Kandels-Lewis S, Picheral M, Poulain J, Searson S, Tara Oceans Coordinators, Stemmann L, Not F, Hingamp P, Speich S, Follows M, Karp-Boss L, Boss E, Ogata H, Pesant S, Weissenbach J, Wincker P, Acinas SG, Bork P, de Vargas C, Iudicone D, Sullivan MB, Raes J, Karsenti E, Bowler C, Gorsky G. 2016. Plankton networks driving carbon export in the oligotrophic ocean. Nature 532:465–470. https://doi.org/10.1038/nature16942.
11. Hogle SL, Hackl T, Bundy RM, Park J, Satinsky B, Hiltunen T, Biller S, Berube PM, Chisholm SW. 2022. Siderophores as an iron source for picocyanobacteria in deep chlorophyll maximum layers of the oligotrophic ocean. ISME J 16:1636–1646. https://doi.org/10.1038/s41396-022-01215-w.
12. Ghurye JS, Cepeda-Espinoza V, Pop M. 2016. Metagenomic assembly: overview, challenges and applications. Yale J Biol Med 89:353–362.
13. Douglas GM, Langille MGI. 2019. Current and promising approaches to identify horizontal gene transfer events in metagenomes. Genome Biol Evol 11:2750–2766. https://doi.org/10.1093/gbe/evz184.
14. Delmont TO, Kiefl E, Kilinc O, Esen OC, Uysal I, Rappé MS, Giovannoni S, Eren AM. 2019. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. Elife 8:e46497. https://doi.org/10.7554/eLife.46497.
15. Delmont TO, Quince C, Shaiber A, Esen ÖC, Lee ST, Rappé MS, McLellan SL, Lücker S, Eren AM. 2018. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat Microbiol 3:804–813. https://doi.org/10.1038/s41564-018-0176-9.
16. Tully BJ, Graham ED, Heidelberg JF. 2018. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 5:170203. https://doi.org/10.1038/sdata.2017.203.
17. Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB. 2009. Protists are microbes too: a perspective. ISME J 3:4–12. https://doi.org/10.1038/ismej.2008.101.
18. Caron DA, Alexander H, Allen AE, Archibald JM, Armbrust EV, Bachy C, Bell CJ, Bharti A, Dyhrman ST, Guida SM, Heidelberg KB, Kaye JZ, Metzner J, Smith SR, Worden AZ. 2017. Probing the evolution, ecology and physiology of marine protists using transcriptomics. Nat Rev Microbiol 15:6–20. https://doi.org/10.1038/nrmicro.2016.160.
19. del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I. 2014. The others: our biased perspective of eukaryotic genomes. Trends Ecol Evol 29:252–259. https://doi.org/10.1016/j.tree.2014.03.006.
20. de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury J-M, Bittner L, Chaffron S, Dunthorn M, Engelen S, Flegontova O, Guidi L, Horák A, Jaillon O, Lima-Mendez G, Lukeš J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Tara Oceans Coordinators, Acinas SG, Bork P, Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E. 2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science 348:1261605. https://doi.org/10.1126/science.1261605.
21. Kirkham AR, Lepère C, Jardillier LE, Not F, Bouman H, Mead A, Scanlan DJ. 2013. A global perspective on marine photosynthetic picoeukaryote community structure. ISME J 7:922–936. https://doi.org/10.1038/ismej.2012.166.
22. Rii YM, Duhamel S, Bidigare RR, Karl DM, Repeta DJ, Church MJ. 2016. Diversity and productivity of photosynthetic picoeukaryotes in biogeochemically distinct regions of the South East Pacific Ocean. Limnol Oceanogr 61:806–824. https://doi.org/10.1002/lno.10255.
23. Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ. 2015. Environmental science. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. Science 347:1257594. https://doi.org/10.1126/science.1257594.
24. Tedersoo L, Albertsen M, Anslan S, Callahan B. 2021. Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. Appl Environ Microbiol 87:e00626-21. https://doi.org/10.1128/AEM.00626-21.
25. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK. 2019. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res 47:e103. https://doi.org/10.1093/nar/gkz569.
26. Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, Ehrlich RL, Palmer JN, Workman AD, Blasetti M, Sen B, Hammond J, Cohen NA, Ehrlich GD, Mell JC. 2018. Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. Microbiome 6:190. https://doi.org/10.1186/s40168-018-0569-2.
27. Heeger F, Bourne EC, Baschien C, Yurkov A, Bunk B, Spröer C, Overmann J, Mazzoni CJ, Monaghan MT. 2018. Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. Mol Ecol Resour 18:1500–1514. https://doi.org/10.1111/1755-0998.12937.
28. Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK. 2016. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ 4:e1869. https://doi.org/10.7717/peerj.1869.
29. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C-S, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S,

Carroll A, Rank DR, Hunkapiller MW. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37:1155–1162. https://doi.org/10.1038/s41587-019-0217-9.
30. Gehrig JL, Portik DM, Driscoll MD, Jackson E, Chakraborty S, Gratalo D, Ashby M, Valladares R. 2022. Finding the right fit: evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. Microb Genom 8:e000794. https://doi.org/10.1099/mgen.0.000794.
31. Martijn J, Lind AE, Schön ME, Spiertz I, Juzokaite L, Bunikis I, Pettersson OV, Ettema TJG. 2019. Confident phylogenetic identification of uncultured prokaryotes through long read amplicon sequencing of the 16S-ITS-23S rRNA operon. Environ Microbiol 21:2485–2498. https://doi.org/10.1111/1462-2920.14636.
32. Overholt WA, Hölzer M, Geesink P, Diezel C, Marz M, Küsel K. 2020. Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. Environ Microbiol 22:4000–4013. https://doi.org/10.1111/1462-2920.15186.
33. Somerville V, Lutz S, Schmid M, Frei D, Moser A, Irmler S, Frey JE, Ahrens CH. 2019. Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. BMC Microbiol 19:143. https://doi.org/10.1186/s12866-019-1500-0.
34. Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VGH, McHardy AC, Nederbragt AJ, Pope PB. 2016. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci Rep 6:25373. https://doi.org/10.1038/srep25373.
35. White RA, III, Bottos EM, Roy Chowdhury T, Zucker JD, Brislawn CJ, Nicora CD, Fansler SJ, Glaesemann KR, Glass K, Jansson JK. 2016. Moleculo long-read sequencing facilitates assembly and genomic binning from complex soil metagenomes. mSystems 1(3):e00045-16. https://doi.org/10.1128/mSystems.00045-16.
36. Batchelder HP. 1985. Seasonal abundance, vertical distribution, and life history of *Metridia pacifica* (Copepoda: Calanoida) in the oceanic subarctic Pacific. Oceanogr Res Pap 32:949–964. https://doi.org/10.1016/0198-0149(85)90038-X.
37. Harvey JBJ, Johnson SB, Fisher JL, Peterson WT, Vrijenhoek RC. 2017. Comparison of morphological and next generation DNA sequencing methods for assessing zooplankton assemblages. J Exp Mar Biol Ecol 487:113–126. https://doi.org/10.1016/j.jembe.2016.12.002.
38. Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605–607. https://doi.org/10.1093/bioinformatics/btv638.
39. Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N, Van Bel M, Poulain J, Katinka M, Hohmann-Marriott MF, Piganeau G, Rouzé P, Da Silva C, Wincker P, Van de Peer Y, Vandepoele K. 2012. Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. Genome Biol 13:R74. https://doi.org/10.1186/gb-2012-13-8-r74.
40. Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, Zhou K, Otillar R, Merchant SS, Podell S, Gaasterland T, Napoli C, Gendler K, Manuell A, Tai V, Vallon O, Piganeau G, Jancek S, Heijde M, Jabbari K, Bowler C, Lohr M, Robbens S, Werner G, Dubchak I, Pazour GJ, Ren Q, Paulsen I, Delwiche C, Schmutz J, Rokhsar D, Van de Peer Y, Moreau H, Grigoriev IV. 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci U S A 104:7705–7710. https://doi.org/10.1073/pnas.0611046104.
41. Duncan A, Barry K, Daum C, Eloe-Fadrosh E, Roux S, Schmidt K, Tringe SG, Valentin KU, Varghese N, Salamov A, Grigoriev IV, Leggett RM, Moulton V, Mock T. 2022. Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans. Microbiome 10:67. https://doi.org/10.1186/s40168-022-01254-7.
42. Delmont TO, Gaia M, Hinsinger DD, Frémont P, Vanni C, Fernandez-Guerra A, Eren AM, Kourlaiev A, d'Agata L, Clayssen Q, Villar E, Labadie K, Cruaud C, Poulain J, Da Silva C, Wessner M, Noel B, Aury J-M, Sunagawa S, Acinas SG, Bork P, Karsenti E, Bowler C, Sardet C, Stemmann L, de Vargas C, Wincker P, Lescot M, Babin M, Gorsky G, Grimsley N, Guidi L, Hingamp P, Jaillon O, Kandels S, Iudicone D, Ogata H, Pesant S, Sullivan MB, Not F, Lee K-B, Boss E, Cochrane G, Follows M, Poulton N, Raes J, Sieracki M, Speich S. 2022. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. Cell Genomics 2:100123. https://doi.org/10.1016/j.xgen.2022.100123.
43. Sibbald SJ, Archibald JM. 2017. More protist genomes needed. Nat Ecol Evol 1:145. https://doi.org/10.1038/s41559-017-0145.
44. Fisher JL, Peterson WT, Rykaczewski RR. 2015. The impact of El Niño events on the pelagic food chain in the northern California Current. Glob Chang Biol 21:4401–4414. https://doi.org/10.1111/gcb.13054.
45. Lilly LE, Ohman MD. 2021. Euphausiid spatial displacements and habitat shifts in the southern California Current System in response to El Niño variability. Prog Oceanogr 193:102544. https://doi.org/10.1016/j.pocean.2021.102544.
46. D'Alelio D, Eveillard D, Coles VJ, Caputi L, d'Alcalà MR, Iudicone D. 2019. Modelling the complexity of plankton communities exploiting omics potential: from present challenges to an integrative pipeline. Curr Opin Syst Biol 13:68–74. https://doi.org/10.1016/j.coisb.2018.10.003.
47. Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.
48. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834. https://doi.org/10.1101/gr.213959.116.
49. Mikheenko A, Saveliev V, Gurevich A. 2016. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32:1088–1090. https://doi.org/10.1093/bioinformatics/btv697.
50. Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, Jensen LJ, Nielsen HB, Petersen TN, Winther O, Rasmussen S. 2021. Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol 39:555–560. https://doi.org/10.1038/s41587-020-00777-4.
51. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.
52. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, Pan JN, Yousef M, Fogarty EC, Trigodet F, Watson AR, Esen ÖC, Moore RM, Clayssen Q, Lee MD, Kivenson V, Graham ED, Merrill BD, Karkman A, Blankenberg D, Eppley JM, Sjödin A, Scott JJ, Vázquez-Campos X, McKay LJ, McDaniel EA, Stevens SLR, Anderson RE, Fuessel J, Fernandez-Guerra A, Maignien L, Delmont TO, Willis AD. 2021. Community-led, integrated, reproducible multi-omics with anvi'o. Nat Microbiol 6:3–6. https://doi.org/10.1038/s41564-020-00834-3.
53. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. DRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 11:2864–2868. https://doi.org/10.1038/ismej.2017.126.
54. Pierce NT, Irber L, Reiter T, Brooks P, Brown CT. 2019. Large-scale sequence comparisons with sourmash. F1000Res 8:1006. https://doi.org/10.12688/f1000research.19675.1.
55. Brown CT, Irber L. 2016. sourmash: a library for MinHash sketching of DNA. J Open Source Softw 1:27. https://doi.org/10.21105/joss.00027.
56. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res 50:D785–D794. https://doi.org/10.1093/nar/gkab776.
57. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, Beszteri B, Bidle KD, Cameron CT, Campbell L, Caron DA, Cattolico RA, Collier JL, Coyne K, Davy SK, Deschamps P, Dyhrman ST, Edvardsen B, Gates RD, Gobler CJ, Greenwood SJ, Guida SM, Jacobi JL, Jakobsen KS, James ER, Jenkins B, John U, Johnson MD, Juhl AR, Kamp A, Katz LA, Kiene R, Kudryavtsev A, Leander BS, Lin S, Lovejoy C, Lynn D, Marchetti A, McManus G, Nedelcu AM, Menden-Deuer S, Miceli C, Mock T, Montresor M, Moran MA, Murray S, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol 12:e1001889. https://doi.org/10.1371/journal.pbio.1001889.
58. Irber L, Brooks PT, Reiter T, Pierce-Ward NT, Hera MR, Koslicki D, Brown CT. 2022. Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. bioRxiv 11:475838. https://doi.org/10.1101/2022.01.11.475838.
59. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
60. Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368. https://doi.org/10.1038/s41592-021-01101-x.

61. Shen W, Ren H. 2021. TaxonKit: a practical and efficient NCBI taxonomy toolkit. J Genet Genomics 48:844–850. https://doi.org/10.1016/j.jgg.2021.03.006.
62. Ruiz-Perez CA, Conrad RE, Konstantinidis KT. 2021. MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. BMC Bioinformatics 22:11. https://doi.org/10.1186/s12859-020-03940-5.
63. Gruber-Vodicka HR, Seah BKB, Pruesse E. 2020. phyloFlash: rapid small-subunit rRNA profiling and targeted assembly from metagenomes. mSystems 5(5):e00920-20. https://doi.org/10.1128/mSystems.00920-20.
64. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and Web-based tools. Nucleic Acids Res 41:D590–D596. https://doi.org/10.1093/nar/gks1219.
65. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, de Vargas C, Decelle J, Del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Logares R, Mahé F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet A-L, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R. 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res 41:D597–D604. https://doi.org/10.1093/nar/gks1160.
66. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584. https://doi.org/10.7717/peerj.2584.
67. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
68. Feng X, Cheng H, Portik D, Li H. 2022. Metagenome assembly of high-fidelity long reads with hifiasm-meta. Nat Methods 19:671–674. https://doi.org/10.1038/s41592-022-01478-3.
69. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, Pevzner PA. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods 17:1103–1110. https://doi.org/10.1038/s41592-020-00971-x.
70. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. 2016. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics 32:1009–1015. https://doi.org/10.1093/bioinformatics/btv688.
71. Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL. 2019. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinformatics 20:405. https://doi.org/10.1186/s12859-019-2996-x.
72. Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368. https://doi.org/10.1038/s41592-021-01101-x.
73. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. Gigascience 10:giab008. https://doi.org/10.1093/gigascience/giab008.
74. Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P, Stevens M, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, De Caceres M, Durand S, Evangelista H, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill M, Lahti L, McGlinn D, Ouellette M, Ribeiro Cunha E, Smith T, Stier A, Ter Braak C, Weedon J. 2022. _vegan: Community Ecology Package_. R package version 2.6-2. https://CRAN.R-project.org/package=vegan.
75. Martino C, Morton JT, Marotz CA, Thompson LR, Tripathi A, Knight R, Zengler K. 2019. A novel sparse compositional technique reveals microbial perturbations. mSystems 4:e00016-19. https://doi.org/10.1128/mSystems.00016-19.
76. Bisanz JE. 2018. qiime2R: importing QIIME2 artifacts and associated data into R sessions. https://github.com/jbisanz/qiime2R.
77. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY.
78. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36:1925–1927. https://doi.org/10.1093/bioinformatics/btz848.
79. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38:4647–4654. https://doi.org/10.1093/molbev/msab199.
80. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
81. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
82. Letunic I, Bork P. 2019. Interactive Tree of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.
83. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
84. Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis version 11. Mol Biol Evol 38:3022–3027. https://doi.org/10.1093/molbev/msab120.