Data in Brief 47 (2023) 108990

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset of 143 metagenome-assembled genomes from the Arctic and Atlantic Oceans, including 21 for eukaryotic organisms

Anthony Duncan[a], Kerrie Barry[b], Chris Daum[b], Emiley Eloe-Fadrosh[b], Simon Roux[b], Katrin Schmidt[c], Susannah G. Tringe[b], Klaus U. Valentin[d], Neha Varghese[b], Asaf Salamov[b], Igor V. Grigoriev[b], Richard M. Leggett[e], Vincent Moulton[a], Thomas Mock[c,*]

[a] School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, NR47TJ, UK
[b] US Department of Energy Joint Genome Institute, 1 Cyclotron Road, Berkeley, CA, 94720, USA
[c] School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, NR47TJ, UK
[d] Alfred-Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570, Bremerhaven, Germany
[e] Earlham Institute, Norwich Research Park, Norwich, NR4 7UG, UK

ARTICLE INFO

Article history:
Received 19 August 2022
Revised 18 January 2023
Accepted 9 February 2023
Available online 15 February 2023

Dataset link: Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data)
Dataset link: Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data)

ABSTRACT

This article presents metagenome-assembled genomes (MAGs) for both eukaryotic and prokaryotic organisms originating from the Arctic and Atlantic oceans, along with gene prediction and functional annotation for MAGs from both domains. Eleven samples from the chlorophyll-a maximum layer of the surface ocean were collected during two cruises in 2012; six from the Arctic in June-July on ARK-XXVII/1 (PS80), and five from the Atlantic in November on ANT-XXIX/1 (PS81). Sequencing and assembly was carried out by the Joint Genome Institute (JGI), who provide annotation of the assembled sequences, and 122 MAGs for prokaryotic organisms. A subsequent binning process identified 21 MAGs for eukaryotic organisms, mostly identified as Mamiellophyceae or Bacillariophyceae. The data for each MAG includes sequences in FASTA format, and tables of functional annotation of genes. For eukaryotic MAGs, transcript

* Corresponding author.
E-mail address: t.mock@uea.ac.uk (T. Mock).
Social media: @Th_Mock (T. Mock)

https://doi.org/10.1016/j.dib.2023.108990

Dataset link: Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data)
Dataset link: Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data)
Dataset link: Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data)
Dataset link: Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data)
Dataset link: Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data)
Dataset link: Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data)
Dataset link: Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data)
Dataset link: Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data)

Keyword:
Genome resolved metagenomics
Phycology
Marine algae
MAGs
Phytoplankton
Polar microbiomes

and protein sequences for predicted genes are available. A spreadsheet is provided summarising quality measures and taxonomic classifications for each MAG. These data provide draft genomes for uncultured marine microbes, including some of the first MAGs for polar eukaryotes, and can provide reference genetic data for these environments, or used in genomics-based comparison between environments.

Specifications Table

Subject	Microbiology: Microbiome		
Specific subject area	Surface ocean microbial communities		
Type of data	FASTA files Tables		
How the data were acquired	Seawater samples were sequenced using Illumina HiSeq platform, generating paired end 2×150bp reads. Reads from each sample were assembled using MEGAHIT. Sequencing and assembly performed by JGI.		
Data format	Raw and Analyzed		
Description of data collection	Samples were taken from seawater during cruises in 2012, six from the Arctic Polar Circle in June-July on ARK-XXVII/1 (PS80), and five from the tropical and sub-tropical Atlantic in November on ANT-XXIX/1 (PS81). Water samples were collected using 12L Niskin bottles, and seawater filtered onto 1.2-μm polycarbonate filters and frozen at -80 °C. DNA was extracted using EasyDNA Kit as described in Martin et al [1]. Samples were snap frozen in liquid nitrogen and stored at -80 °C until sequencing. Sequencing was performed by JGI using the Illumina HiSeq platform, generating paired end 2×150bp reads. Assembly, gene prediction and annotation were performed by JGI IMG pipeline [2]. This pipeline identified prokaryotic MAGs, but no eukaryotes. Eukaryotic bins were subsequently identified using EukRep [3] and MetaBat [4], and genes predicted by GeneMark-ES [5] and annotated using InterProScan [6].		
Data source location	Sample name	Latitude	Longitude
	P1	79.02	-9.52
	P2	78.87	-3.23
	P3a	78.87	8.11
	P3b	78.87	8.11
	P4	73.02	9.86
	P5	71.2	8.87
	P6	69.23	7.73
	NP1	34.88	-13.14
	NP2	26.05	-17.46
	NP3	15.25	-20.52
	NP4	2.41	-13.6
	NP5	-17.28	2.98
Data accessibility	Figshare: https://doi.org/10.6084/m9.figshare.c.5017517.v4 [7] Reads uploaded to NCBI SRA (https://www.ncbi.nlm.nih.gov/sra). BioProject accessions PRJNA365113, PRJNA365111, PRJNA330320, PRJNA365112, PRJNA406185, PRJNA406186, PRJNA365114, PRJNA366134, PRJNA366135, PRJNA365119, PRJNA365117, PRJNA365118.		
Related research article	Duncan, A., Barry, K., Daum, C. et al. Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans. Microbiome 10, 67 (2022). https://doi.org/10.1186/s40168-022-01254-7 [8]		

Value of the Data

- This data spans the Arctic Circle, enabling genomic comparison of surface ocean microbes across this strong polar-temperate environmental divide.
- The eukaryotic MAGs are among the first for ocean microbes, and can be used to expand the references genomes for this group of organisms beyond the small number sequenced from cultured species.
- Can be compared to MAGs from similar environmental conditions (i.e. Antarctic) to study evolutionary responses.
- Some MAGs closely related to known species can be included in pangenomic analyses.
- MAGs which appear to display high degrees of taxonomic and functional novelty (e.g. NP3_4P, Table 1)

Table 1
List summarizing the MAGs available in the dataset. For prokaryotes, taxonomy was generated by GTDB-Tk [13], here both the phylum and lowest rank with a non-placeholder name is given. For eukaryotes, taxonomy is based on placement in a phylogenomic tree including protist reference genomes. Two measures of functional novelty are given: the percentage of predicted genes which lack any functional annotation, and the percentage all the Pfam domains observed which were Domains of Unknown Function. The distance between each MAG and the closest reference genome in phylogenomic trees combining MAGs and reference is given as an estimate of taxonomic novelty. Trees for eukaryotes and prokaryotes were constructed separately as detailed in the related research article, so distances are not comparable between the two. Finally, the quality of MAGs is expressed through completeness and contamination; for eukaryotes this was generated by EukCC [14], and for prokaryotes using CheckM [15].

			Genes		Pfams			Quality	
MAG	Phylum	Lowest Non-Placeholder Taxonomy	Number	Without Annotation	Number	of which DUF	Nearest Reference	Completeness	Contamination
NP5_1E	Bacillarophyta	Bacillarophyta	13678	16.3%	9468	3.5%	0.471	74.4%	0.0%
P1_1E	Bacillarophyta	Bacillarophyta	19182	24.5%	12443	2.4%	0.616	74.5%	3.9%
P1_2E	Bacillarophyta	Bacillarophyta	14003	18.2%	11113	2.4%	0.686	74.4%	0.0%
P1_4E	Bacillarophyta	Bacillarophyta	9447	20.8%	7763	2.3%	0.706	58.8%	0.0%
P1_5E	Bacillarophyta	Bacillarophyta	12446	25.2%	8718	2.5%	0.633	58.8%	5.9%
P2_3E	Bacillarophyta	Bacillarophyta	13298	16.6%	8736	2.2%	0.243	66.7%	3.9%
P3a_2E	Bacillarophyta	Bacillarophyta	11414	12.1%	9200	2.7%	0.618	62.8%	5.9%
P3a_4E	Bacillarophyta	Bacillarophyta	14484	21.3%	10613	2.4%	0.757	66.7%	0.0%
P6_1E	Bacillarophyta	Bacillarophyta	12916	19.7%	10227	2.3%	0.601	78.4%	2.0%
P6_2E	Bacillarophyta	Bacillarophyta	12096	23.7%	9242	2.5%	0.573	70.6%	7.8%
NP2_1E	Chlorophyta	Chlorophyta	5626	10.1%	5869	2.5%	0.318	64.1%	3.5%
NP2_2E	Chlorophyta	Chlorophyta	4808	7.0%	5661	2.8%	0.323	65.8%	0.9%
NP3_1E	Chlorophyta	Chlorophyta	5269	7.6%	6120	2.7%	0.425	70.5%	0.6%
P2_1E	Chlorophyta	Chlorophyta	11269	9.0%	11145	2.4%	0.122	94.0%	1.9%
P2_4E	Chlorophyta	Chlorophyta	12301	21.1%	9191	2.5%		81.9%	5.4%
P3a_3E	Chlorophyta	Chlorophyta	8289	11.9%	7059	2.8%	0.452	62.6%	3.8%
P5_1E	Chlorophyta	Chlorophyta	7595	12.6%	6793	2.8%	0.431	56.1%	1.8%
P6_3E	Chlorophyta	Chlorophyta	8488	11.2%	7373	2.6%	0.391	61.4%	1.8%
P3a_1E	Haptophyta	Haptophyta	29691	15.2%	20860	1.9%	0.538	68.9%	2.2%
P1_3E	Unknown	Unknown	13261	10.7%	12675	1.9%	0.817	48.7%	0.0%
P2_2E	Unknown	Unknown	16812	11.8%	15518	2.1%	0.820	64.1%	8.4%
NP1_22P	Actinobacteriota	MedAcidi-G3	2305	0.0%	2583	2.3%	0.802	85.5%	2.1%
NP2_25P	Actinobacteriota	MedAcidi-G1	1911	0.1%	2019	2.6%	0.999	76.9%	7.3%
NP2_26P	Actinobacteriota	Microtrichales	1990	0.1%	1729	2.5%		50.3%	9.0%
NP3_25P	Actinobacteriota	MedAcidi-G3	1920	18.1%	2187	2.1%	0.788	78.4%	2.1%
NP3_36P	Actinobacteriota	MedAcidi-G1	1743	21.4%	1928	2.5%	0.974	54.4%	6.8%
NP4_26P	Actinobacteriota	MedAcidi-G3	2325	0.0%	2689	2.2%	0.789	97.0%	3.0%
NP2_18P	Proteobacteria	Hyphomonas	2379	0.0%	2771	5.2%	0.267	75.2%	0.3%
NP3_20P	Proteobacteria	Brevundimonas	2813	24.3%	3124	4.8%	0.477	97.1%	3.6%
NP3_22P	Proteobacteria	Erythrobacter_A	2522	22.0%	2708	4.8%		75.2%	7.3%
NP4_22P	Proteobacteria	Rhodobacteraceae	2219	0.0%	2273	4.4%	0.796	50.7%	8.7%
NP4_8P	Proteobacteria	Pelagibaca bermudensis	4559	0.0%	5315	3.5%	0.108	92.5%	0.5%
NP5_12P	Proteobacteria	Micavibrionaceae	2377	30.9%	2553	3.3%	0.834	91.9%	3.7%
NP5_8P	Proteobacteria	Pelagibaca bermudensis	4053	19.7%	4444	3.4%	0.146	60.4%	1.9%
P1_23P	Proteobacteria	Loktanella	2988	18.9%	3324	3.5%		67.4%	1.4%
P1_24P	Proteobacteria	Planktomarina	2668	17.0%	3199	3.1%		92.7%	2.0%
P2_11P	Proteobacteria	Sulfitobacter_C	3134	20.6%	3580	3.4%	0.000	90.5%	1.1%
P3a_11P	Proteobacteria	Rhodobacteraceae	3669	20.6%	4135	2.8%	0.352	88.6%	1.2%
P3a_15P	Proteobacteria	Sulfitobacter_C	3214	21.8%	3651	3.5%	0.000	94.8%	3.5%
P3a_21P	Proteobacteria	Loktanella	2591	18.2%	2872	2.9%	0.338	58.9%	2.2%
P3b_2P	Proteobacteria	Sulfitobacter_C	3384	24.3%	3678	3.3%		94.8%	2.4%
NP1_23P	Bacteroidota	Croceibacter atlanticus	2086	0.1%	2232	5.9%		69.4%	0.7%
NP2_14P	Bacteroidota	Croceibacter atlanticus	2801	0.0%	2150	6.3%	0.001	63.5%	2.2%
NP3_30P	Bacteroidota	Croceibacter atlanticus	1688	34.7%	1632	5.5%	0.130	64.7%	9.3%

(*continued on next page*)

Table 1 (*continued*)

MAG	Phylum	Lowest Non-Placeholder Taxonomy	Genes		Pfams		Nearest Reference	Quality	
			Number	Without Annotation	Number	of which DUF		Completeness	Contamination
NP4_11P	Bacteroidota	Croceibacter atlanticus	4431	0.0%	3066	5.8%	0.001	82.6%	3.9%
NP5_11P	Bacteroidota	Flavobacteriales	2280	28.2%	2377	3.9%	1.005	88.7%	7.7%
NP5_15P	Bacteroidota	Flavobacteriales	1784	23.1%	2010	3.4%	0.931	92.7%	0.5%
NP5_29P	Bacteroidota	Flavobacteriaceae	1148	17.2%	1341	5.1%	0.254	55.6%	4.7%
P1_15P	Bacteroidota	Saprospiraceae	3666	35.5%	3736	4.2%	0.618	81.6%	2.2%
P1_17P	Bacteroidota	Saprospiraceae	3454	34.2%	3575	4.7%		55.2%	1.7%
P1_21P	Bacteroidota	Croceibacter atlanticus	2922	25.3%	3295	6.3%	0.013	99.6%	2.8%
P1_34P	Bacteroidota	Flavobacteriaceae	1429	22.0%	1597	4.5%	0.002	64.0%	3.1%
P1_41P	Bacteroidota	Cryomorphaceae	1200	21.5%	1318	4.2%	0.010	69.1%	0.0%
P2_12P	Bacteroidota	Flavobacteriales	2445	27.0%	2673	4.4%	0.719	86.5%	4.9%
P2_23P	Bacteroidota	Cryomorphaceae	1591	44.8%	1169	3.7%	0.049	60.6%	7.5%
P2_25P	Bacteroidota	Bacteroidia	1358	28.1%	1350	4.6%	0.777	56.7%	0.0%
P3a_25P	Bacteroidota	Bacteroidia	1871	25.9%	2017	4.4%	0.768	82.1%	1.7%
P3a_27P	Bacteroidota	Flavobacteriaceae	1771	22.9%	1989	4.6%	0.008	80.1%	2.3%
P3a_30P	Bacteroidota	Cryomorphaceae	1661	21.7%	1866	3.5%	0.002	88.0%	5.4%
P3b_6P	Bacteroidota	Cryomorphaceae	1832	28.7%	1971	3.5%		92.1%	2.3%
P3b_8P	Bacteroidota	Flavobacteriaceae	1690	29.5%	1819	4.0%		78.3%	2.3%
P4_14P	Bacteroidota	Flavobacteriaceae	1064	23.5%	1263	3.9%	0.453	58.2%	2.0%
P4_19P	Bacteroidota	Flavobacteriales	975	23.2%	1150	4.2%	0.702	63.4%	1.1%
P5_11P	Bacteroidota	Cryomorphaceae	1934	29.7%	2073	3.5%	0.337	94.2%	5.0%
P6_12P	Bacteroidota	Saprospiraceae	2636	36.6%	2444	4.1%		77.9%	5.3%
P6_13P	Bacteroidota	Flavobacteriales	2162	26.2%	2393	3.9%	1.002	98.1%	0.5%
P6_15P	Bacteroidota	Flavobacteriaceae	1920	26.2%	2097	5.8%	0.461	80.2%	2.5%
P6_22P	Bacteroidota	Flavobacteriaceae	1651	20.8%	1937	6.4%		81.0%	0.0%
P6_35P	Bacteroidota	Flavobacteriaceae	1224	23.7%	1273	4.4%	0.022	50.7%	0.7%
P6_46P	Bacteroidota	Cryomorphaceae	1027	24.0%	1093	4.2%	0.337	51.1%	1.4%
NP2_9P	Proteobacteria	Pseudoalteromonas marina	3750	0.0%	4518	5.6%		72.5%	1.7%
NP3_13P	Proteobacteria	Pseudomonas_D sabulinigri	3226	18.2%	4151	5.4%	0.187	81.5%	1.7%
NP3_40P	Proteobacteria	Psychrobacter sp5	1396	18.4%	1522	5.0%		51.4%	0.6%
NP3_5P	Proteobacteria	Algiphilaceae	4273	19.4%	5299	5.3%	0.613	98.9%	3.6%
NP3_6P	Proteobacteria	Alteromonas macleodii	3996	20.1%	5024	5.1%	0.005	98.8%	3.1%
NP3_7P	Proteobacteria	Alcanivorax	4064	23.4%	4720	5.7%	0.097	96.7%	8.6%
NP4_10P	Proteobacteria	Alteromonas macleodii	3950	0.0%	4731	5.1%	0.005	86.6%	0.8%
NP4_18P	Proteobacteria	Alteromonas	2796	0.0%	3137	4.7%	0.012	50.3%	0.8%
NP4_41P	Proteobacteria	Legionellaceae	1615	0.1%	1803	2.2%		93.6%	1.2%
NP5_10P	Proteobacteria	Alcanivorax	3400	21.6%	3946	5.4%	0.097	93.6%	2.5%
NP5_19P	Proteobacteria	Halomonas	1705	14.7%	2058	4.2%		53.5%	0.0%
NP5_3P	Proteobacteria	Alteromonas	4250	23.6%	4855	5.2%		71.6%	1.7%
NP5_9P	Proteobacteria	Neptunomonas phycophila	3329	18.5%	4245	4.3%	0.355	89.4%	0.2%
P1_16P	Proteobacteria	Pseudoalteromonas marina	3941	22.5%	4798	5.3%	0.080	97.3%	2.0%
P1_20P	Proteobacteria	Bermanella	3272	25.4%	3793	4.5%	0.456	96.6%	6.4%
P1_25P	Proteobacteria	Nitrincolaceae	2421	14.5%	3098	3.1%		71.8%	0.0%
P1_26P	Proteobacteria	Porticoccaceae	2081	23.8%	2406	3.9%		55.2%	3.5%
NP4_11P	Bacteroidota	Croceibacter atlanticus	4431	0.0%	3066	5.8%	0.001	82.6%	3.9%
NP5_11P	Bacteroidota	Flavobacteriales	2280	28.2%	2377	3.9%	1.005	88.7%	7.7%
NP5_15P	Bacteroidota	Flavobacteriales	1784	23.1%	2010	3.4%	0.931	92.7%	0.5%
NP5_29P	Bacteroidota	Flavobacteriaceae	1148	17.2%	1341	5.1%	0.254	55.6%	4.7%
P1_15P	Bacteroidota	Saprospiraceae	3666	35.5%	3736	4.2%	0.618	81.6%	2.2%
P1_17P	Bacteroidota	Saprospiraceae	3454	34.2%	3575	4.7%		55.2%	1.7%
P1_21P	Bacteroidota	Croceibacter atlanticus	2922	25.3%	3295	6.3%	0.013	99.6%	2.8%
P1_34P	Bacteroidota	Flavobacteriaceae	1429	22.0%	1597	4.5%	0.002	64.0%	3.1%
P1_41P	Bacteroidota	Cryomorphaceae	1200	21.5%	1318	4.2%	0.010	69.1%	0.0%
P2_12P	Bacteroidota	Flavobacteriales	2445	27.0%	2673	4.4%	0.719	86.5%	4.9%
P2_23P	Bacteroidota	Cryomorphaceae	1591	44.8%	1169	3.7%	0.049	60.6%	7.5%
P2_25P	Bacteroidota	Bacteroidia	1358	28.1%	1350	4.6%	0.777	56.7%	0.0%

(*continued on next page*)

Table 1 (*continued*)

MAG	Phylum	Lowest Non-Placeholder Taxonomy	Genes		Pfams		Nearest Reference	Quality	
			Number	Without Annotation	Number	of which DUF		Completeness	Contamination
P3a_25P	Bacteroidota	Bacteroidia	1871	25.9%	2017	4.4%	0.768	82.1%	1.7%
P3a_27P	Bacteroidota	Flavobacteriaceae	1771	22.9%	1989	4.6%	0.008	80.1%	2.3%
P3a_30P	Bacteroidota	Cryomorphaceae	1661	21.7%	1866	3.5%	0.002	88.0%	5.4%
P3b_6P	Bacteroidota	Cryomorphaceae	1832	28.7%	1971	3.5%		92.1%	2.3%
P3b_8P	Bacteroidota	Flavobacteriaceae	1690	29.5%	1819	4.0%		78.3%	2.3%
P4_14P	Bacteroidota	Flavobacteriaceae	1064	23.5%	1263	3.9%	0.453	58.2%	2.0%
P4_19P	Bacteroidota	Flavobacteriales	975	23.2%	1150	4.2%	0.702	63.4%	1.1%
P5_11P	Bacteroidota	Cryomorphaceae	1934	29.7%	2073	3.5%	0.337	94.2%	5.0%
P6_12P	Bacteroidota	Saprospiraceae	2636	36.6%	2444	4.1%		77.9%	5.3%
P6_13P	Bacteroidota	Flavobacteriales	2162	26.2%	2393	3.9%	1.002	98.1%	0.5%
P6_15P	Bacteroidota	Flavobacteriaceae	1920	26.2%	2097	5.8%	0.461	80.2%	2.5%
P6_22P	Bacteroidota	Flavobacteriaceae	1651	20.8%	1937	6.4%		81.0%	0.0%
P6_35P	Bacteroidota	Flavobacteriaceae	1224	23.7%	1273	4.4%	0.022	50.7%	0.7%
P6_46P	Bacteroidota	Cryomorphaceae	1027	24.0%	1093	4.2%	0.337	51.1%	1.4%
NP2_9P	Proteobacteria	Pseudoalteromonas marina	3750	0.0%	4518	5.6%		72.5%	1.7%
NP3_13P	Proteobacteria	Pseudomonas_D sabulinigri	3226	18.2%	4151	5.4%	0.187	81.5%	1.7%
NP3_40P	Proteobacteria	Psychrobacter sp5	1396	18.4%	1522	5.0%		51.4%	0.6%
NP3_5P	Proteobacteria	Algiphilaceae	4273	19.4%	5299	5.3%	0.613	98.9%	3.6%
NP3_6P	Proteobacteria	Alteromonas macleodii	3996	20.1%	5024	5.1%	0.005	98.8%	3.1%
NP3_7P	Proteobacteria	Alcanivorax	4064	23.4%	4720	5.7%	0.097	96.7%	8.6%
NP4_10P	Proteobacteria	Alteromonas macleodii	3950	0.0%	4731	5.1%	0.005	86.6%	0.8%
NP4_18P	Proteobacteria	Alteromonas	2796	0.0%	3137	4.7%	0.012	50.3%	0.8%
NP4_41P	Proteobacteria	Legionellaceae	1615	0.1%	1803	2.2%		93.6%	1.2%
NP5_10P	Proteobacteria	Alcanivorax	3400	21.6%	3946	5.4%	0.097	93.6%	2.5%
NP5_19P	Proteobacteria	Halomonas	1705	14.7%	2058	4.2%		53.5%	0.0%
NP5_3P	Proteobacteria	Alteromonas	4250	23.6%	4855	5.2%		71.6%	1.7%
NP5_9P	Proteobacteria	Neptunomonas phycophila	3329	18.5%	4245	4.3%	0.355	89.4%	0.2%
P1_16P	Proteobacteria	Pseudoalteromonas marina	3941	22.5%	4798	5.3%	0.080	97.3%	2.0%
P1_20P	Proteobacteria	Bermanella	3272	25.4%	3793	4.5%	0.456	96.6%	6.4%
P1_25P	Proteobacteria	Nitrincolaceae	2421	14.5%	3098	3.1%		71.8%	0.0%
P1_26P	Proteobacteria	Porticoccaceae	2081	23.8%	2406	3.9%		55.2%	3.5%
NP1_17P	Verrucomicrobiota	Pedosphaerales	2893	0.1%	3325	8.1%	1.221	79.3%	5.4%
NP1_19P	Verrucomicrobiota	Opitutaceae	2612	0.1%	2533	9.5%	0.828	61.3%	3.1%
NP2_8P	Verrucomicrobiota	Roseibacillus	3503	0.0%	3725	12.8%	1.153	95.9%	0.5%
NP3_10P	Verrucomicrobiota	Roseibacillus	3044	27.6%	3224	11.4%	1.237	84.7%	0.0%
NP4_16P	Verrucomicrobiota	Roseibacillus	2981	0.0%	3170	10.9%	1.239	89.1%	1.0%
NP4_47P	Verrucomicrobiota	Pedosphaerales	1334	0.1%	1502	9.9%	1.143	52.7%	0.7%
NP5_7P	Verrucomicrobiota	Akkermansiaceae	3322	30.1%	3489	11.0%	1.100	90.7%	0.9%
P2_21P	Verrucomicrobiota	Puniceicoccaceae	1409	26.8%	1511	3.6%		50.0%	0.0%
P3a_28P	Verrucomicrobiota	Puniceicoccaceae	1588	22.7%	1846	3.7%	0.340	85.1%	0.7%
P5_21P	Verrucomicrobiota	Puniceicoccaceae	986	25.8%	1058	2.5%	0.368	52.4%	2.3%
P6_14P	Verrucomicrobiota	Puniceicoccaceae	2092	25.3%	2282	4.4%	0.332	78.6%	4.1%
P6_33P	Verrucomicrobiota	Puniceicoccaceae	1162	23.8%	1266	2.6%		53.7%	0.3%
NP2_12P	Poribacteria	Poribacteria	3178	0.1%	3013	4.2%	1.212	61.5%	7.1%

1. Objective

Ocean microbes are essential for marine life, they form the base of the ocean food web and play important roles in cycling of essential nutrients. A majority of marine microbes cannot be cultured, preventing access to their genomic information through isolate sequencing and assembly methods. Metagenomics has allowed insight into the genetic material of all members of these natural communities of microbes, but to fully understand the metabolic capability and roles of individual organisms from these communities, we need to place this sequence data back into a genomic context. Binning methods for recovering MAGs have been widely applied

to prokaryotes, but at the time of commencing our research we were aware of only 2 MAGs for eukaryotic marine microbes [9,10]. Our objective was to increase the range of marine eukaryotic microbes for which MAGs were available, to help better understand this environmentally significant unculturable majority. Here we describe in greater detail both the content of the repository containing MAGs and their annotation, and the methods used to produce the data.

2. Data Description

This data contains metagenome-assembled genomes, originating from samples collected in the Arctic Polar Circle and tropical and sub-tropical Atlantic Oceans. In total 143 MAGs were recovered, with 122 being prokaryotes, and 21 eukaryotes. Table 1 provides a list of all the MAGs available in the dataset. The sequence data for MAGs is the first archive making up this repository, and the annotation of the predicted genes the second. Fig. 1 shows the structure of these two components, showing directory and file structure, with more detail provided below.

143 FASTA files provide the DNA sequences for each MAG. For each eukaryotic MAG 3 files are given to describe functional annotation, and for each prokaryotic MAG 6 files are provided. Functional annotations are in different formats for eukaryotes and prokaryotes due to different tools being used for annotating them. Fig. 2 shows a summary of size, completion and taxonomy of these MAGs, with their potential functional novelty shown in Fig. 3.

For each prokaryote:

- 1 GFF file of predicted genes
- 5 tables giving annotation of genes with KEGG orthologs (KO), Enzyme Commission (EC) numbers, COG terms, Pfam domains, and a named gene product, each in tab-separated format.

For each eukaryote:

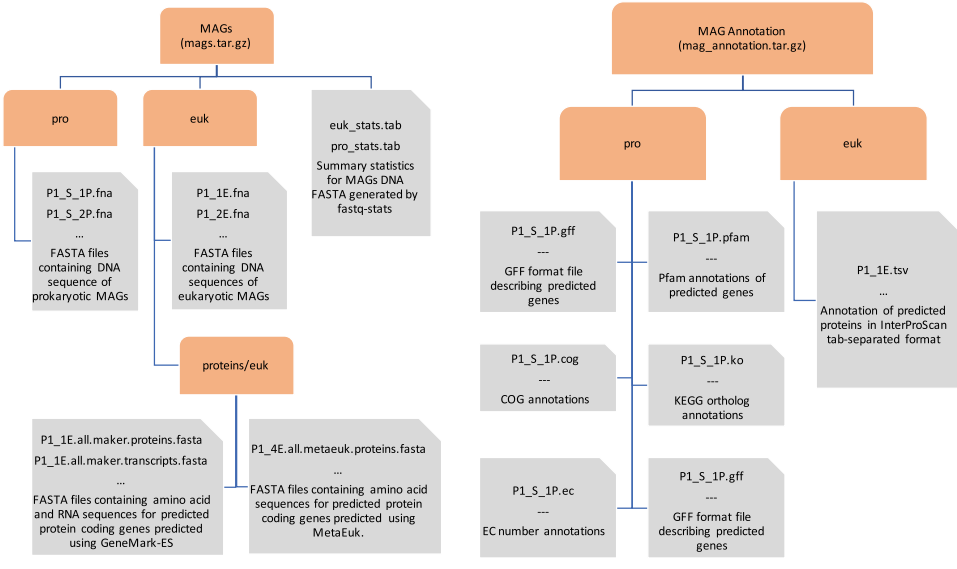

Fig. 1. Repository structure diagram. This describes two main archives which provide the sequence data and functional annotation for the eukaryotic and prokaryotic MAGs. Tan rounded corner nodes represent directories or compressed directories, and grey nodes files. Where ellipses are included in file descriptions, this indicates that there is one such file for each MAG.

Fig. 2. Summary of MAGs included in this dataset. Left column (red) shows MAGs recovered from non-polar assemblies, right column (blue) those within the Arctic Circle. These are further divided by domain, the top row shows eukaryotes, and the bottom row prokaryotes. Each point represents a MAG, with the size of point representing length of DNA sequence in the MAG, and the colour an estimated taxonomy. Each point is placed based MAG quality, with the horizontal axis being completeness and vertical axis contamination, assessed using EukCC [14] for eukaryotes, and CheckM [15] for prokaryotes.

- 1 FASTA files of predicted proteins amino acid sequences
- 1 FASTA file of predicted gene transcript RNA sequences for those annotated with GeneMark-ES [5] (all but MAGs P1_4E, P1_5E, and P2_4E)
- 1 table of InterProScan [6] output in tab-separated format

These files are assigned names indicating which sample they came from, the assembler used, a numeric identifier, and whether they are eukaryotic. For example, P1_S_2P originates from sample P1, the assembler used was SPAdes [11] (rather than MEGAHIT [12]), is the 2nd MAG from sample P1, and is given P for prokaryote (rather than E for eukaryote).

Hence for prokaryotes the file P1_S_2P.fna contains the contigs for this MAG, with P1_S_2P.gff the gene predictions, P1_S_2P.pfam the Pfam annotations of those genes, P1_S_2P.cog the COG annotations and so on for KOs, EC numbers, and gene products. For eukaryotes, P1_1E.fna again contains the contigs for the MAG, predicted genes are provided as their transcript and protein sequences in P1_1E.all.maker.transcripts.fasta and P1_1E.all.maker.proteins.fasta respectively, and annotation of these genes in P1_1E.tsv.

An Excel format spreadsheet contains summaries of sample and assembly details, and for MAGs their quality measures, taxonomic details, and associated metadata. The worksheets contained are:

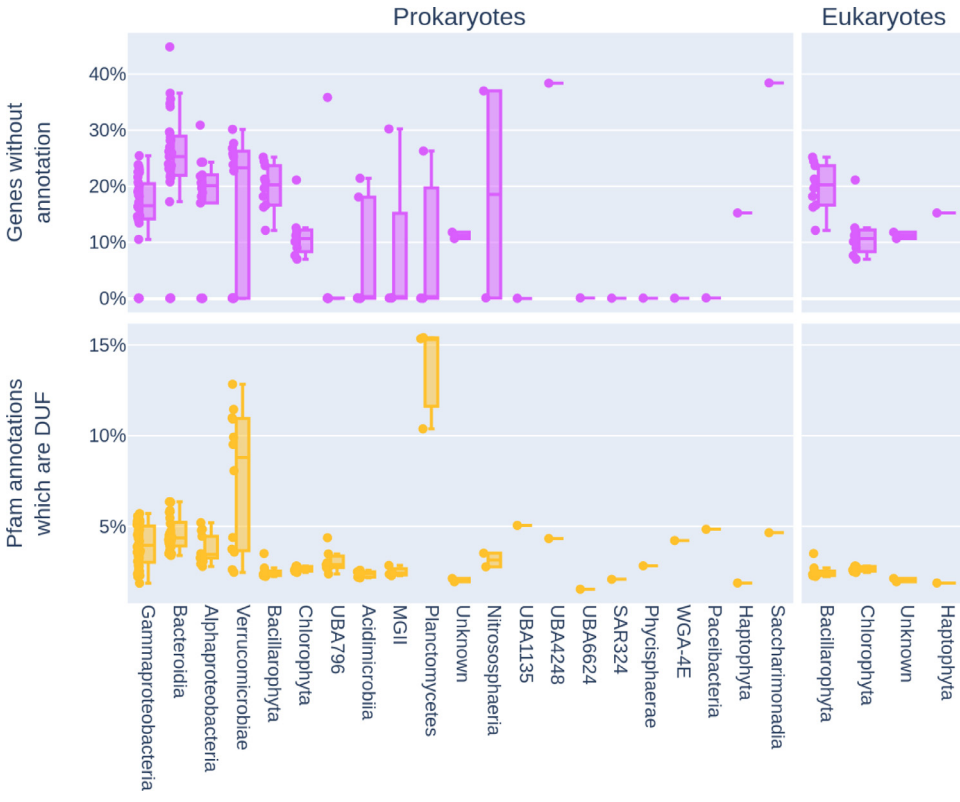

Fig. 3. Functional novelty by phylum. The top row in pink shows the proportion of predicted genes in a MAG which had no functional annotation, shown as both a box plot and points for each MAG to the left of the box. Some phyla, such as Bacteroidia, have a high level of unannotated genes with the potential to contain functional novelty. The bottom row in orange shows what proportion of all Pfam domains in predicted genes are Domains of Unknown Function (DUF), with Planctomycetes and Verrucomicrobiae showing higher proportions of DUF domains.

- station_details: Information on the stations and sampling, including location, date, sampling depth, in-situ metadata including temperature, salinity, and nutrient measurements. Includes JGI and NCBI accessions for the samples.
- read_fastq_stats: Summary statistics for reads from each sample, generated by fastq-stats (length, mean quality, base frequency etc.)
- all_assembly: Summary of assembly quality for all (MEGAHIT and SPAdes) assemblies, provided by JGI.
- assembly: Same as worksheet all_assembly, but restricted to only the MEGAHIT assemblies used for eukaryotic binning
- euk_summary: The size of data at each step of eukaryotic binning. Each step gives the number of read or contigs, and the length in base pairs, for instance reads and reads_bp is the number or reads and total length of read respectively. The contigs columns give the number and length of contigs in the assembly, eukrep columns the number and length of contigs predicted as eukaryotic by EukRep, the binned columns the number and length of contigs placed in bins by MetaBat, and the mqbinned columns the number and length of contigs in medium quality bins as assessed by BUSCO.
- eukrep: Summary statistics of the predicted eukaryotic contigs, generated by BBMap.
- eukbinned: Details of the medium quality eukaryotic MAGs. Summary of sequence statistics generated by BBMap are indicated by blue columns, quality as assessed by BUSCO by

red columns, and quality assessed by EukCC by yellow columns. The estimated phylum and number of predicted proteins are also given.
- pro_summary: The size of data at each step of prokaryotic binning. Columns are the same as the worksheet euk_summary, but the eukrep and binned columns are blank. EukRep was not used for prokaryote binning, and bins below medium quality were discarded by the IMG pipeline and so their size is unknown.
- pro_binned: Details of the medium quality or higher prokaryotic MAGs. Identifiers for the MAG are provided, both the name used in the repository and the Bin ID used by IMG. The column 'Bin Quality' contains either MQ for medium quality, or HQ for high quality. The columns in red are the quality and lineage estimated by CheckM; the usually more specific lineage from GTDB-Tk is also provided. Number and length of contigs, and number of predicted genes, are also given.
- probinned_bbmapstats: Summary statistics of the nucleic acid sequences for each MAG, generated by BBMap (number of contigs, N50, GC% etc.)
- pro_assembledby: Indicates which assembly was used for prokaryotic binning, MH being MEGAHIT, and SP SPAdes.

3. Experimental Design, Materials and Methods

Elven samples in total were collected for metagenome sequencing during two RV Polarstern expeditions in 2012 [1]. Samples were taken from six stations within the Arctic Polar Circle (ARK-XXVII/1 (PS80), 17th June to 9th July), and five from the tropical and subtropical Atlantic (ANT-XXIX/1 (PS81), 1st to 24th November). Two filtering steps were carried out, samples were first pre-filtered with a 100μm mesh to remove larger zooplankton, then filtered onto 1.2μm Nucleopore membrane filters. These were stored at -80°C. To extract DNA, the EasyDNA Kit was used with modifications. Pre-heated (65 °C) solution A was used to wash cells off the filter, and the supernatant transferred into a new tub with a small spoon of glass beads (425–600 μm, acid-washed) (Sigma-Aldrich, USA). Samples were vortexed three times in intervals of 3s. RNAse A was added to the samples and incubated for 30 min at 65 °C. The supernatant was transferred into a new tube, and solution B from the kit was added followed by a chloroform phase separation and an ethanol precipitation. DNA was pelleted by centrifugation and washed several times with isopropanol, air-dried, and suspended in 100 μL TE buffer. DNA concentration was measured with a Nanodrop (Thermo Fisher Scientific, Waltman, MA, USA), samples snap-frozen in liquid nitrogen and stored at -80 °C until sequencing.

Sequencing was carried out by the Joint Genome Institute, with assembly and annotation performed by their Integrated Microbial Genomes & Microbiomes (IMG/M) pipeline. The processes making up these pipelines have been published [2,16], and summarized below here.

Sequencing using the Illumina HiSeq platform generated 2×150bp paired-end reads. Illumina adapters were removed using BBDuk (v35.87) [17]. Subsequently reads were trimmed and filtered again using BBDuk. First read ends with quality less than 12 were trimmed. Any read pair with either three or more N characters, average quality score across the read less than 3, or length less than 51bp after trimming were discarded. Reads which map to the human HG19 genome with greater than 93% identity were also discarded, a standard part of the JGI QC pipeline. After quality control, a total approximately 629Gbp reads remained.

Quality controlled reads were assembled using MEGAHIT (v1.0.3) [12] with default parameters and a range of k-mers 23, 43, 63, 83, 103, 123. MEGAHIT assemblies contain approximately 26Gbp in 42 million contigs. The quality-controlled reads were mapped back to the assemblies to generate coverage using seal [18].

Reads from six samples (P1, P2, P3a, P6, NP3, NP5) were later reassembled using SPAdes (3.10.0-dev) [11]. This assembly used the raw unfiltered reads, which were corrected using bfc (r181) and a k-mer size of 21, then assembled using SPAdes with the meta option and range of k-mers 21, 33, 55, 77, 99, 127. The SPAdes assemblies total approximately 10Gbp and 18 million contigs. In general, the SPAdes assemblies are smaller than their MEGAHIT counterparts, but

with longer mean contig lengths. Reads were mapped back to the assembly to generate overage using bwa-mem (version 0.7.15-r1142-dirty) [19] with default parameters.

Genes were predicted for each of these assemblies using an ensemble of gene prediction tools: prokaryotic GeneMark.hmm (v2.8), Prodigal (v2.6.3), MetaGeneAnnotator (August 2008), and FragGeneScan (v1.1.6) [20–23]. tRNA were predicted using INFERNAL (v1.1.1) [24], and rRNA with HMMER (3.1b2) [25]; both of these need the domain as a parameter, so are run three times. Predictions from these tools are combined based on a majority consensus, with ties broken based on the predicting tool in the order they were listed above. A set of rules are applied to resolve conflicts between protein coding genes and other features (e.g. tRNA) [16]. Protein coding genes shorter than 32 amino acids are discarded.

Protein coding genes are functionally annotated with COGs, Pfams, KEGG orthologs, and EC numbers. COGs are assigned using RPS-BLAST (v2.2.31) to search against the CDD database [26,27], with an e-value cutoff of 0.1; Pfams are assigned based on search against profile HMMs using HMMER (v3.1b2) and the model specific cutoffs; KOs are assigned from LAST (737+) [28] search results against the IMG database of isolate reference genomes, and EC number based on mapping between KO and EC numbers. The best LAST hit is used to assign taxonomy to the gene, and the taxonomy of contig is the lowest common ancestor of all the genes on the contig, where 30% or greater of the genes have any LAST hits. A total of approximately 50 million genes were predicted.

The binning process incorporated into the IMG/M pipeline identified 122 prokaryotic MAGs. Each assembly was binned individually using MetaBat (v2.12.1) [4] using a minimum contig size of 3000bp, coverage of the contigs in samples other than the one the assembly was generated from was not used. Quality of bins were assessed using CheckM (v1.0.12) [15], and only medium quality bins were retained ($\geq$50% completeness, $\leq$10% contamination). Taxonomy of MAGs was assessed with GTDB-Tk (v0.2.2, database release 86) [13]. These MAGs are available both in this repository, and on the IMG website using the bin identifiers included in the summary spreadsheet.

MAGs identified by the IMG/M pipeline were all prokaryotic, prompting a separate binning effort to recover eukaryotes. Only the MEGAHIT assemblies were used for eukaryotic binning. Eukaryotic contigs in were identified in each assembly using EukRep (v0.6.5) [3] with default parameters, producing a total of approximately 4Gbp and 2 million eukaryotic contigs. To estimate the coverage of these eukaryotic contigs in all samples, reads from each sample were pseudo-aligned to each of the 12 sets of eukaryotic contigs using the Kallisto (v0.44.0) [29] kallisto-quant command with default parameters. The estimated mean coverage of each contig was taken to be the number of reads estimated to originate from that contig multiplied by the read length (150bp) divided by the length of the contig. This was formatted into a table for each set of eukaryotic contigs, with the contig as rows, set of reads as columns, and each entry the estimated coverage. Binning was performed for each set of eukaryotic contigs with MetaBat (v2.12.1) with this coverage information as input and a minimum contig size of 1500bp, and otherwise default parameters. This produced 59 bins; to match the prokaryotes the quality of these bins was assessed using BUSCO (v3.0.2) [30] and the eukaryota_odb9 set of genes, and only the 18 MAGs which were medium quality or better retained.

Although genes had been predicted on all contigs by the IMG/M pipeline, this had been using tools which were not adapted to the more complex gene structure of eukaryotes. Hence, genes were predicted for these 18 eukaryotic MAGs using MAKER (v2) [31] and GeneMark-ES (v4.38) [5] in self-training mode. A GeneMark-ES model was trained using gmes_petap.pl command with the MAG contigs as input with a minimum contig length of 5000bp. The resulting model was used by MAKER with otherwise default parameters. GeneMark-ES has the assumption that all contigs originate from a single genome, so gene prediction had to be carried out after binning for these eukaryotic MAGs.

After this initial eukaryotic binning effort, colleagues at JGI identified 3 additional eukaryotic bins (P1_4E, P1_5E, and P2_4E) using alternative methods. Starting with the assemblies, contigs were searched against the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) [32] database using MMSeqs2 [33] to filter for eukaryotic contigs. These were each

binned using Metabat (v2.12.1), and resulting bins checked for taxonomic consistency using the MMSeqs2 results. Any bin with 50% or greater contigs from a single phylum and total length 5Mbp or greater was retained, and filtered to remove contigs from other taxa. This resulted in three additional MAGs, for which genes were predicted using MetaEuk [34] with NR [35] used as reference database. These three additional MAGs were added to the repository.

Completeness and contamination of these 21 eukaryotic MAGs was assessed using EukCC (v0.2) [14] to obtain lineage specific estimates of quality. For these eukaryotic MAGs, the predicted proteins were annotated using InterProScan (v5.37-75.0) [6] with default parameters.

Ethics Statements

The authors have consulted the publishers Ethics in Publishing standards, and believe the manuscript meets these standards.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data) (NCBI BioProject).
Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data) (NCBI BioProject).
Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data) (NCBI BioProject).
Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data) (NCBI BioProject).
Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data) (NCBI BioProject).
Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data) (NCBI BioProject).
Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data) (NCBI BioProject).
Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data) (NCBI BioProject).
Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data) (NCBI BioProject).
Metagenome Assembled Genomes of 21 eukaryotic and 122 prokaryotic phytoplankton, with predicted proteins and functional annotation (Original data) (NCBI BioProject).

CRediT Author Statement

Anthony Duncan: Investigation, Formal analysis, Writing – original draft, Visualization; **Kerrie Barry:** Investigation; **Chris Daum:** Investigation; **Emiley Eloe-Fadrosh:** Investigation; **Simon Roux:** Investigation, Formal analysis, Writing – review & editing; **Katrin Schmidt:** Investigation; **Susannah G. Tringe:** Investigation, Formal analysis; **Neha Varghese:** Investigation, Formal analysis; **Asaf Salamov:** Investigation, Formal analysis, Writing – review & editing; **Igor V. Grigoriev:** Investigation, Formal analysis, Writing – review & editing; **Richard M. Leggett:** Conceptualization, Writing – review & editing, Supervision; **Vincent Moulton:** Conceptualization,

Writing – review & editing, Supervision; **Thomas Mock:** Conceptualization, Writing – review & editing, Supervision.

Acknowledgments

The authors would like to thank the following collaborators from the Joint Genome Institute: A. Clum, A. Copeland, B. Foster, Br. Foster, M. Huntemann, N. N. Ivanova, N. C. Kyrpides, E. Lindquist, S. Mukherjee, K. Palaniappan and T.B.K. Reddy.

This work was supported by the Natural Environmental Research Council [grant number NE/N012070/1]. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231.

References

[1] K. Martin, K. Schmidt, A. Toseland, C.A. Boulton, K. Barry, B. Beszteri, C.P.D. Brussaard, A. Clum, C.G. Daum, E. Eloe-Fadrosh, A. Fong, B. Foster, B. Foster, M. Ginzburg, M. Huntemann, N.N. Ivanova, N.C. Kyrpides, E. Lindquist, S. Mukherjee, K. Palaniappan, T.B.K. Reddy, M.R. Rizkallah, S. Roux, K. Timmermans, S.G. Tringe, W.H. van de Poll, N. Varghese, K.U. Valentin, T.M. Lenton, I.V. Grigoriev, R.M. Leggett, V. Moulton, T. Mock, The biogeographic differentiation of algal microbiomes in the upper ocean from pole to pole, Nat. Commun. 12 (2021) 5483, doi:10.1038/s41467-021-25646-9.

[2] I.-M.A. Chen, K. Chu, K. Palaniappan, M. Pillay, A. Ratner, J. Huang, M. Huntemann, N. Varghese, J.R. White, R. Seshadri, T. Smirnova, E. Kirton, S.P. Jungbluth, T. Woyke, E.A. Eloe-Fadrosh, N.N. Ivanova, N.C. Kyrpides, IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes, Nucl. Acids Res. 47 (2019) D666–D677, doi:10.1093/nar/gky901.

[3] P.T. West, A.J. Probst, I.V. Grigoriev, B.C. Thomas, J.F. Banfield, Genome-reconstruction for eukaryotes from complex natural microbial communities, Genome Res. 28 (2018) 569–580, doi:10.1101/gr.228429.117.

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

[5] V. Ter-Hovhannisyan, A. Lomsadze, Y.O. Chernoff, M. Borodovsky, Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training, Genome Res. 18 (2008) 1979–1990, doi:10.1101/gr.081612.108.

[6] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A.F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, S. Hunter, InterProScan 5: genome-scale protein function classification, Bioinformatics 30 (2014) 1236–1240, doi:10.1093/bioinformatics/btu031.

[7] A. Duncan, Metagenome-assembled genomes of phytoplankton communities across the Arctic Circle, (2020). doi:10.6084/m9.figshare.c.5017517.

[8] A. Duncan, K. Barry, C. Daum, E. Eloe-Fadrosh, S. Roux, K. Schmidt, S.G. Tringe, K.U. Valentin, N. Varghese, A. Salamov, I.V. Grigoriev, R.M. Leggett, V. Moulton, T. Mock, Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans, Microbiome 10 (2022) 67, doi:10.1186/s40168-022-01254-7.

[9] T.O. Delmont, C. Quince, A. Shaiber, Ö.C. Esen, S.T. Lee, M.S. Rappé, S.L. McLellan, S. Lücker, A.M. Eren, Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes, Nat. Microbiol. 3 (2018) 804, doi:10.1038/s41564-018-0176-9.

[10] N. Joli, A. Monier, R. Logares, C. Lovejoy, Seasonal patterns in Arctic prasinophytes and inferred ecology of Bathycoccus unveiled in an Arctic winter metagenome, ISME J 11 (2017) 1372–1385, doi:10.1038/ismej.2017.7.

[11] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, P.A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477, doi:10.1089/cmb.2012.0021.

[12] D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, Bioinformatics 31 (2015) 1674–1676, doi:10.1093/bioinformatics/btv033.

[13] P.-A. Chaumeil, A.J. Mussig, P. Hugenholtz, D.H. Parks, GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database, Bioinformatics 36 (2020) 1925–1927, doi:10.1093/bioinformatics/btz848.

[14] P. Saary, A.L. Mitchell, R.D. Finn, Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC, Genome Biol. 21 (2020) 244, doi:10.1186/s13059-020-02155-4.

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

[16] M. Huntemann, N.N. Ivanova, K. Mavromatis, H.J. Tripp, D. Paez-Espino, K. Tennessen, K. Palaniappan, E. Szeto, M. Pillay, I.-M.A. Chen, A. Pati, T. Nielsen, V.M. Markowitz, N.C. Kyrpides, The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4), Stand. Genom. Sci. 11 (2016) 17, doi:10.1186/s40793-016-0138-x.

[17] B. Bushnell, BBTools software package, URL Http://Sourceforge.Net/Projects/Bbmap. (2014).

[18] L. Pireddu, S. Leo, G. Zanetti, SEAL: a distributed short read mapping and duplicate removal tool, Bioinformatics 27 (2011) 2159–2160, doi:10.1093/bioinformatics/btr325.
[19] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, ArXiv 1303 (2013) 3997 [q-Bio] http://arxiv.org/abs/1303.3997 . (accessed April 9, 2019).
[20] A.V. Lukashin, M. Borodovsky, GeneMark.hmm: New solutions for gene finding, Nucl. Acids Res. 26 (1998) 1107–1115, doi:10.1093/nar/26.4.1107.
[21] D. Hyatt, G.-L. Chen, P.F. LoCascio, M.L. Land, F.W. Larimer, L.J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinformatics 11 (2010) 119, doi:10.1186/1471-2105-11-119.
[22] H. Noguchi, T. Taniguchi, T. Itoh, MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes, DNA Res. 15 (2008) 387–396, doi:10.1093/dnares/dsn027.
[23] M. Rho, H. Tang, Y. Ye, FragGeneScan: predicting genes in short and error-prone reads, Nucl. Acids Res. 38 (2010) e191–e191, doi:10.1093/nar/gkq747.
[24] E.P. Nawrocki, S.R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches, Bioinformatics 29 (2013) 2933–2935, doi:10.1093/bioinformatics/btt509.
[25] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucl. Acids Res. 39 (2011) W29–W37, doi:10.1093/nar/gkr367.
[26] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucl. Acids Res. 25 (1997) 3389–3402, doi:10.1093/nar/25.17.3389.
[27] A. Marchler-Bauer, J.B. Anderson, M.K. Derbyshire, C. DeWeese-Scott, N.R. Gonzales, M. Gwadz, L. Hao, S. He, D.I. Hurwitz, J.D. Jackson, Z. Ke, D. Krylov, C.J. Lanczycki, C.A. Liebert, C. Liu, F. Lu, S. Lu, G.H. Marchler, M. Mullokandov, J.S. Song, N. Thanki, R.A. Yamashita, J.J. Yin, D. Zhang, S.H. Bryant, CDD: a conserved domain database for interactive domain family analysis, Nucl. Acids Res. 35 (2007) D237–D240, doi:10.1093/nar/gkl951.
[28] S.M. Kiełbasa, R. Wan, K. Sato, P. Horton, M.C. Frith, Adaptive seeds tame genomic sequence comparison, Genome Res. 21 (2011) 487–493, doi:10.1101/gr.113985.110.
[29] N.L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification, Nat. Biotechnol. 34 (2016) 525–527, doi:10.1038/nbt.3519.
[30] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212, doi:10.1093/bioinformatics/btv351.
[31] C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects, BMC . 12 (2011) 491, doi:10.1186/1471-2105-12-491.
[32] P.J. Keeling, F. Burki, H.M. Wilcox, B. Allam, E.E. Allen, L.A. Amaral-Zettler, E.V. Armbrust, J.M. Archibald, A.K. Bharti, C.J. Bell, B. Beszteri, K.D. Bidle, C.T. Cameron, L. Campbell, D.A. Caron, R.A. Cattolico, J.L. Collier, K. Coyne, S.K. Davy, P. Deschamps, S.T. Dyhrman, B. Edvardsen, R.D. Gates, C.J. Gobler, S.J. Greenwood, S.M. Guida, J.L. Jacobi, K.S. Jakobsen, E.R. James, B. Jenkins, U. John, M.D. Johnson, A.R. Juhl, A. Kamp, L.A. Katz, R. Kiene, A. Kudryavtsev, B.S. Leander, S. Lin, C. Lovejoy, D. Lynn, A. Marchetti, G. McManus, A.M. Nedelcu, S. Menden-Deuer, C. Miceli, T. Mock, M. Montresor, M.A. Moran, S. Murray, G. Nadathur, S. Nagai, P.B. Ngam, B. Palenik, J. Pawlowski, G. Petroni, G. Piganeau, M.C. Posewitz, K. Rengefors, G. Romano, M.E. Rumpho, T. Rynearson, K.B. Schilling, D.C. Schroeder, A.G.B. Simpson, C.H. Slamovits, D.R. Smith, G.J. Smith, S.R. Smith, H.M. Sosik, P. Stief, E. Theriot, S.N. Twary, P.E. Umale, D. Vaulot, B. Wawrik, G.L. Wheeler, W.H. Wilson, Y. Xu, A. Zingone, A.Z. Worden, The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing, PLOS Biol. 12 (2014) e1001889, doi:10.1371/journal.pbio.1001889.
[33] M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, Nat. Biotechnol. 35 (2017) 1026–1028, doi:10.1038/nbt.3988.
[34] E. Levy Karin, M. Mirdita, J. Söding, MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics, Microbiome 8 (2020) 48, doi:10.1186/s40168-020-00808-x.
[35] N.A. O'Leary, M.W. Wright, J.R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C.M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V.S. Joardar, V.K. Kodali, W. Li, D. Maglott, P. Masterson, K.M. McGarvey, M.R. Murphy, K. O'Neill, S. Pujar, S.H. Rangwala, D. Rausch, L.D. Riddick, C. Schoch, A. Shkeda, S.S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R.E. Tully, A.R. Vatsan, C. Wallin, D. Webb, W. Wu, M.J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T.D. Murphy, K.D. Pruitt, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, Nucl. Acids Res. 44 (2016) D733–D745, doi:10.1093/nar/gkv1189.