Digital Microbe, a genome-informed data integration framework for team science on emerging model organisms

Iva Veseli[1,2,‡], Michelle A. DeMers[3,‡], Zachary S. Cooper[4,‡], Matthew S. Schechter[5], Samuel Miller[6], Laura Weber[7], Christa B. Smith[4], Lidimarie T. Rodriguez[8], William F. Schroer[4], Matthew R. McIlvin[7], Paloma Z. Lopez[7], Makoto Saito[7], Sonya Dyhrman[9], A. Murat Eren[1,2,6,10,11,†], Mary Ann Moran[4,†], Rogier Braakman[3,†]

[1]Helmholtz Institute for Functional Marine Biodiversity, 26129, Oldenburg, Germany
[2]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, 27570, Bremerhaven, Germany
[3]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Department of Marine Sciences, University of Georgia, Athens, GA 30602, USA
[5]Committee on Microbiology, The University of Chicago, Chicago, IL 60637, USA
[6]Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA 02543, USA
[7]Woods Hole Oceanographic Institution, Falmouth, MA 02543, USA
[8]Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611-0180, USA
[9]Lamont-Doherty Earth Observatory, and the Department of Earth and Environmental Sciences, Columbia University, New York, NY 10032, USA
[10]Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, Oldenburg, Germany
[11]Marine 'Omics Bridging Group, Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

† Corresponding authors: meren@mbl.edu; mmoran@uga.edu; braakman@mit.edu

‡ These authors contributed equally

Abstract

The remarkable pace of genomic data generation is rapidly transforming our understanding of life at the micron scale. Yet this data stream also creates challenges for team science. A single microbe can have multiple versions of genome architecture, functional gene annotations, and gene identifiers; additionally, the lack of mechanisms for collating and preserving advances in this knowledge raises barriers to community coalescence around shared datasets. "Digital Microbes" are frameworks for interoperable and reproducible collaborative science through open source, community-curated data packages built on a (pan)genomic foundation. Housed within an integrative software environment, Digital Microbes ensure real-time alignment of research efforts for collaborative teams and facilitate novel scientific insights as new layers of data are added. Here we describe two Digital Microbes: 1) the heterotrophic marine bacterium *Ruegeria pomeroyi* DSS-3 with >100 transcriptomic datasets from lab and field studies, and 2) the pangenome of the cosmopolitan marine heterotroph *Alteromonas* containing 339 genomes. Examples demonstrate how an integrated framework collating public (pan)genome-informed data can generate novel and reproducible findings.

Introduction

Expanded access to the genomic data of microbial organisms has been transforming the way we approach microbiology research. Genome sequences are subsequently enhanced with knowledge from experimental, modeling, and field studies (e.g., [1–4]) with the goal of yielding insights into microbial physiology, ecology, and biogeochemistry. Yet because different research teams independently consolidate and curate genome-related information via ad hoc solutions, these diverse data streams have created challenges for interoperable analyses, especially in collaborative work. More generally, the lack of a framework for establishing consensus versions of genome-linked reference data hinders community coalescence around shared datasets. To extend the impact of curated and collated microbial data beyond a single research group, requirements are: 1) an established reference dataset[5], which provides existing and updated knowledge in a standardized format; and 2) open access to these data, which allows multiple groups to collaboratively analyze and update the same genome and genome-linked information. The power of establishing a strategy for the open exchange of consensus microbial data linked to reference genomes for emerging model organisms, whether they are laboratory cultures or

those reconstructed from metagenomes, is increasing as team science takes on growing roles in environmental and life sciences research.

Contemporary software solutions for the analysis and exchange of microbial genomes and associated ‘omics survey data can be broadly characterized into three groups: 1) online portals that provide a centralized location for uploading or downloading genomes and/or ‘omics datasets; 2) online portals with embedded applications that allow the user to choose from pre-selected genomes and/or ‘omics datasets or, in some cases, upload their own data for analysis; and 3) downloadable tools that enable local analysis of genomes and/or ‘omics data (Table 1). While they provide important services for individual research groups, these solutions do not necessarily maximize the efficiency of collaborative team science efforts. Typically, datasets are provided either as raw data or as highly-polished summaries, and intermediate data products for coordination of downstream analyses are not maintained. Moreover, most existing solutions are centralized, in which case data curation and platform maintenance falls on a single entity vulnerable to loss of funding, while data format, updates, and accessibility are not fully under the control of researchers. An alternative solution that partially solves the data sharing needs of collaborative team science efforts is anvi’o[6] (https://anvio.org), an open-source software platform that can integrate a variety of data streams into interoperable, standalone SQL databases that can serve as collaborative data products[6]; however, anvi’o data products are not version-controlled. Inspired by the state-of-the-art technical opportunities offered by anvi’o, here we propose a general framework for the distribution and collaborative analysis of ‘omics datasets that is conducive to team science efforts. The ‘Digital Microbe’ (DM) concept describes features of a data product (#1-3) and a data implementation framework (#4-5) that:

1. Stores a genome sequence with sequence-linked information (e.g., curated gene calls, user-defined functional annotations, etc).
2. Supports additional layers of genome-associated data (e.g., genomic regions of particular interest, mutant strain availability, protein structures, etc).
3. Supports additional layers of experimental or environmental survey data, including intermediate analysis results of value to the research team (e.g., transcriptomic or proteomic activity across different experimental conditions, environmental distribution patterns through metagenomic or metatranscriptomic read recruitment analyses, etc).
4. Enables version-controlled addition of new data layers or curation of existing ones iteratively by any researcher.

5. Stores and enables the export of information in a universal format that is accessible to other programs and centralized or decentralized analysis platforms.

We developed the Digital Microbe concept and its implementation in the National Science Foundation (NSF) Science and Technology Center for Chemical Currencies of a Microbial Planet (C-CoMP; https://ccomp-stc.org) consisting of a research team geographically distributed across 12 institutions. Our construction of Digital Microbes enabled Center members to simultaneously access, analyze and update experimental and environmental datasets for the Center's two model marine bacterial species, *Ruegeria pomeroyi* DSS-3 and *Alteromonas macleodii* MIT1002, including diverse data types ranging from 'omics surveys and environmental parameters to metabolic models and metabolomes. Here we demonstrate the feasibility of the Digital Microbe concept as a solution addressing widespread needs in the microbiology community for reproducible, integrated data products and we describe Digital Microbe data packages for each of C-CoMP's model bacteria. The first Digital Microbe compiles knowledge of transcriptional response by *Ruegeria pomeroyi* DSS-3 gathered from 8 independent studies carried out between 2014 and 2023 (DOI: 10.5281/zenodo.7304959); the second describes an *Alteromonas* pangenome created by merging data from 339 isolate and metagenome-assembled genomes (DOI: 10.5281/zenodo.7430118).

Category	Examples	Limitations
Centralized databases (data upload/download)	NBCI RefSeq[95] DiatOmicBase[96] ECMDB[97] ProPortal[55]	Restricted data types Often specific to one type of organism Updates after the initial data deposit can be limited
Online portals (data upload + online analysis tools)	JGI IMG/M[43] KBase[98] Galaxy[99] PATRIC[100] CyVerse[101] RAST[102] SILVAngs[103] BV-BRC[104] Phycocosm[105]	May not enable sharing of data and results between users Analysis workflows are often black boxes without transparent or changeable parameters Dependence on the platform's computational resources can hinder high-throughput analyses May not facilitate automation of analysis workflows
Downloadable tools (local analysis with database snapshots)	Prokka[106] KofamScan[18] BLAST+[107] GToTree[108] Pathway Tools[109] DRAM[110]	May lack version control and/or easy mechanisms for updating database snapshot Can be difficult to share data/results between users with different local versions of the database snapshot Tool functionality is typically constrained to or tailored towards one kind of analysis

Table 1. Limitations of existing solutions for the sharing of 'omics information.

100 # Results and Discussion

Digital Microbe: Concept and Implementation

At its core, a Digital Microbe is a curated and versioned public data package that is (1) 'self-contained' (i.e., it can explain itself and its contents) and (2) 'extensible' (i.e., others can extend a Digital Microbe data package with additional layers of information coming from new experiments).
105 The package consists of multiple datasets organized and linked through reference to the genome of a single microbe or the pangenome of a group of microbes (Figure 1). Data collection consolidates information such as gene annotations, coverage and other read-mapping statistics,

and sample metadata. These data types can be flexible in scope and the extensibility of Digital Microbes via the programmatic addition of new ‘omics data types make them future-proof.

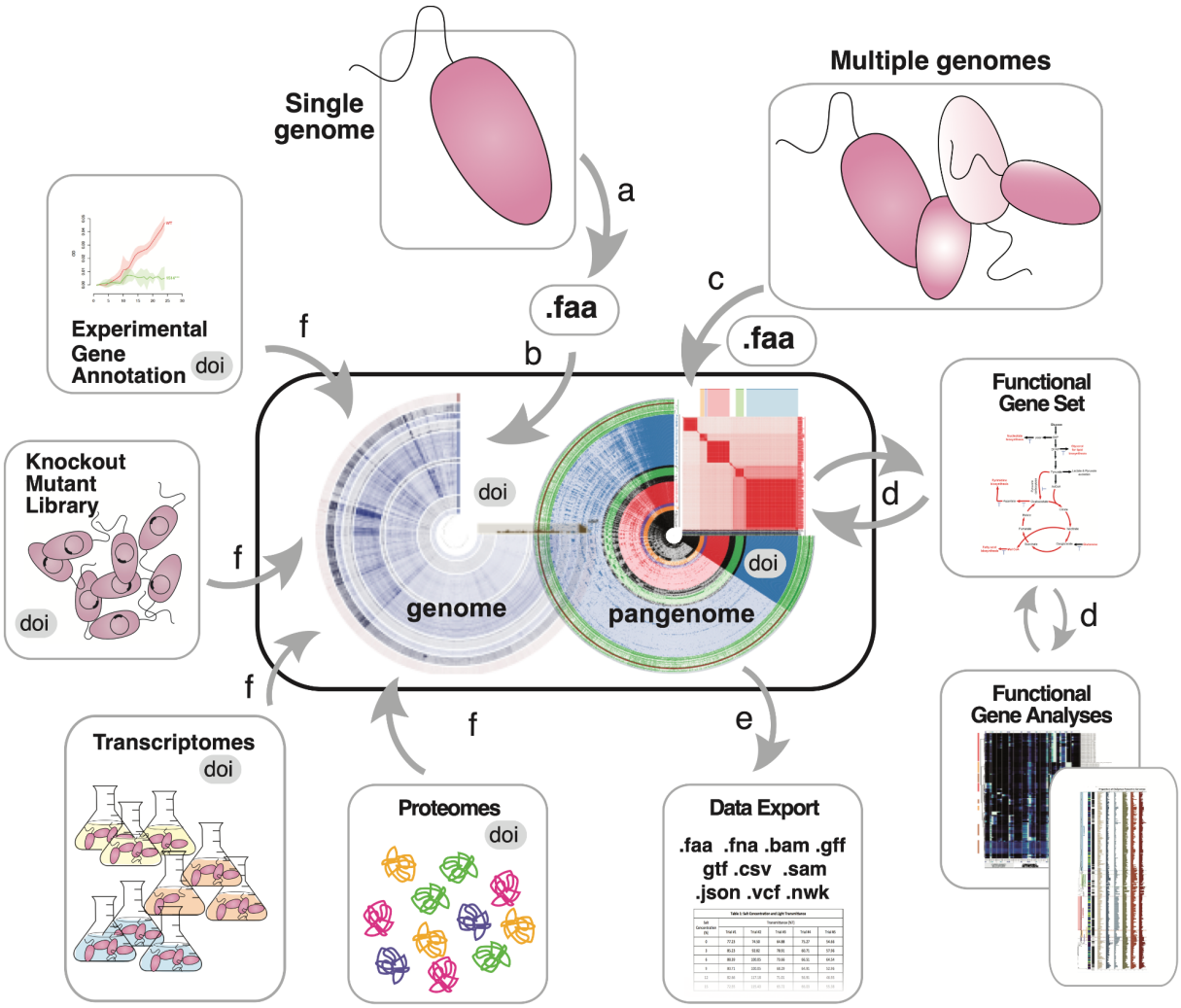


Figure 1. Architecture of a Digital Microbe. The genome of a model bacterium is (a) sequenced and (b) assembled and serves as the foundation of a Digital Microbe, a self-contained data package for a collaborative research team or a science community. (c) Alternatively, a pangenomic data package is assembled. (d) Intermediate datasets useful for downstream analyses are stored and reused, and (e) various data files and tables can be exported. (f) The Digital Microbe is iteratively populated with data layers referenced to individual genes, including mapped proteomes, transcriptomes, or gene-specific metadata types such as inventories of mutants or new annotations. Each Digital Microbe can be assigned a DOI (digital object identifier) and be versioned as new gene- or genome-referenced data are added.

The Digital Microbe framework utilizes a model organism’s genome or a clade’s pangenome as the foundation of a database file describing the DNA sequences (Figure 2). This database file is

hosted in a central data repository where it can be accessed by collaborators and community members. A software platform was needed for collaborative analyses, and we chose the open-source software platform anvi'o[6], which implements many of the Digital Microbe features described above (Figure 2). The concept behind the Digital Microbe framework, however, is independent of any one software platform. Similarly, C-CoMP hosts its Digital Microbe files on the data-sharing platform Zenodo (https://zenodo.org/), but other version-controlled storage solutions are available. As team science progresses, other genome- or gene-linked datasets (including both raw data and analysis results) can be added to the database by various groups, who update the publicly-hosted file to a new version that disseminates their data and findings to the team or community.

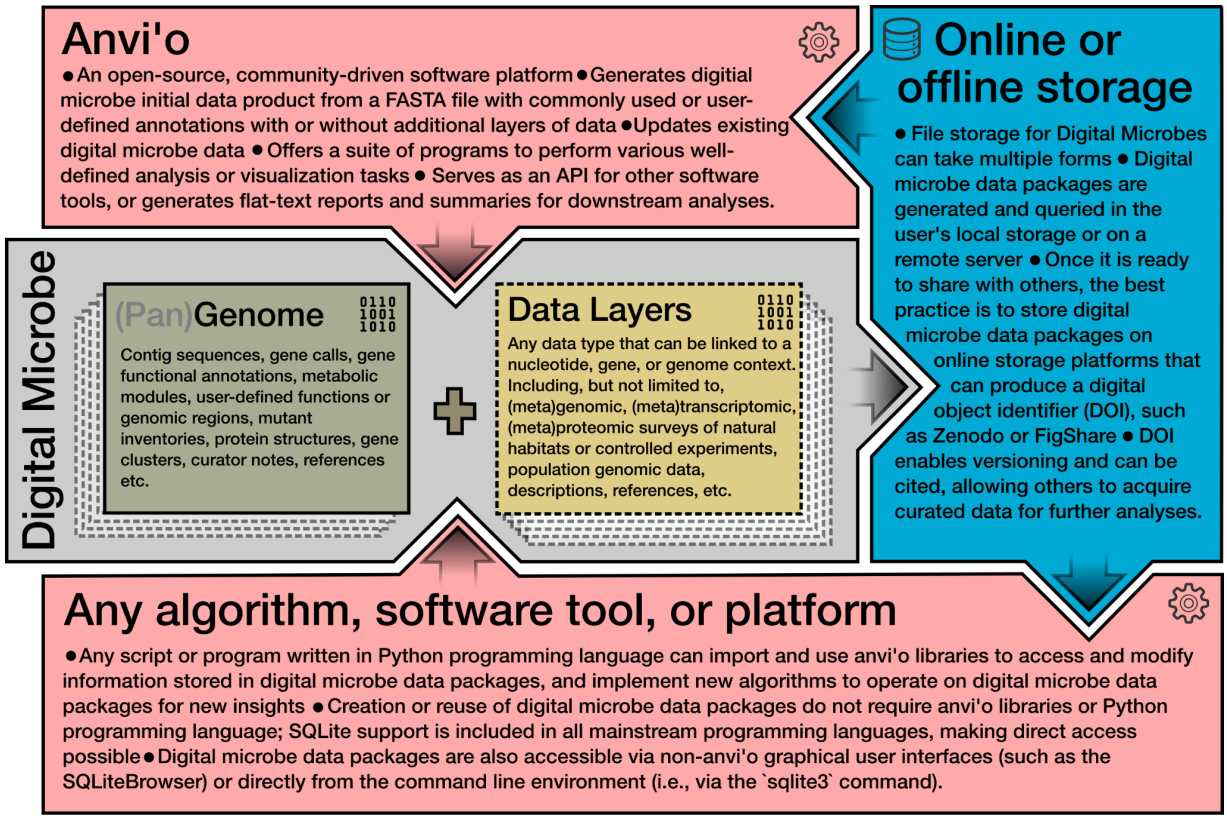


Figure 2. Situating the Digital Microbe concept in the existing computational environment. The Digital Microbe approach facilitates collaborative science by: establishing a version-controlled (pan)genomic reference; consolidating and cross-referencing collections of experimental and environmental data associated with a genome or pangenome; facilitating access to reusable intermediate analyses; and providing data export capabilities for transitioning to other programs or analysis software. While each of these features could be established by generating new software, we chose to use the existing open-source software platform anvi'o[6], which implements several aspects of a Digital Microbe via (pan)genomic data

storage in programmatically-queryable SQLite databases. The concept behind the Digital Microbe framework, however, is independent of any one software platform.

Here, we present two examples of Digital Microbes – one for the model organism *Ruegeria pomeroyi* and another for the pangenome of *Alteromonas* spp. – as well as case studies that exemplify how they can be used.

The *Ruegeria pomeroyi* Digital Microbe

Ruegeria pomeroyi DSS-3 is a representative of the Roseobacteraceae family, an important bacterial group in marine microbial communities[7] with its members among the most metabolically active bacterial cells in algal blooms and coastal environments[8]. *R. pomeroyi* has been well studied in the laboratory and field[9–11]; it grows well in both defined and rich media; and it is amenable to genetic alteration[12,13].

The *R. pomeroyi* Digital Microbe (Figure 3) is built on a well-curated genome assembly (*DM feature 1*) first annotated in 2004[14], reannotated in 2014[15], and enhanced with information from NCBI Clusters of Orthologous Groups (COG)[16], Pfam[17], and KEGG Kofam[18]. The Digital Microbe annotation is also continually updated (*DM feature 4*) with new experimental verifications of *R. pomeroyi* genes (e.g., [15,19–22]) that have not been captured in standardized genome annotation repositories (e.g., RefSeq GCF_000011965.2). The *R. pomeroyi* DSS-3 Digital Microbe is available on Zenodo[23].

A Use Case: Exploring the Substrate Landscape of R. pomeroyi

Our team is leveraging *R. pomeroyi* as a whole-cell biosensor of labile components of the marine dissolved organic carbon (DOC) pool. A recent study using *R. pomeroyi* knockout mutants definitively identified the cognate substrate of 18 organic compound transporters[24]

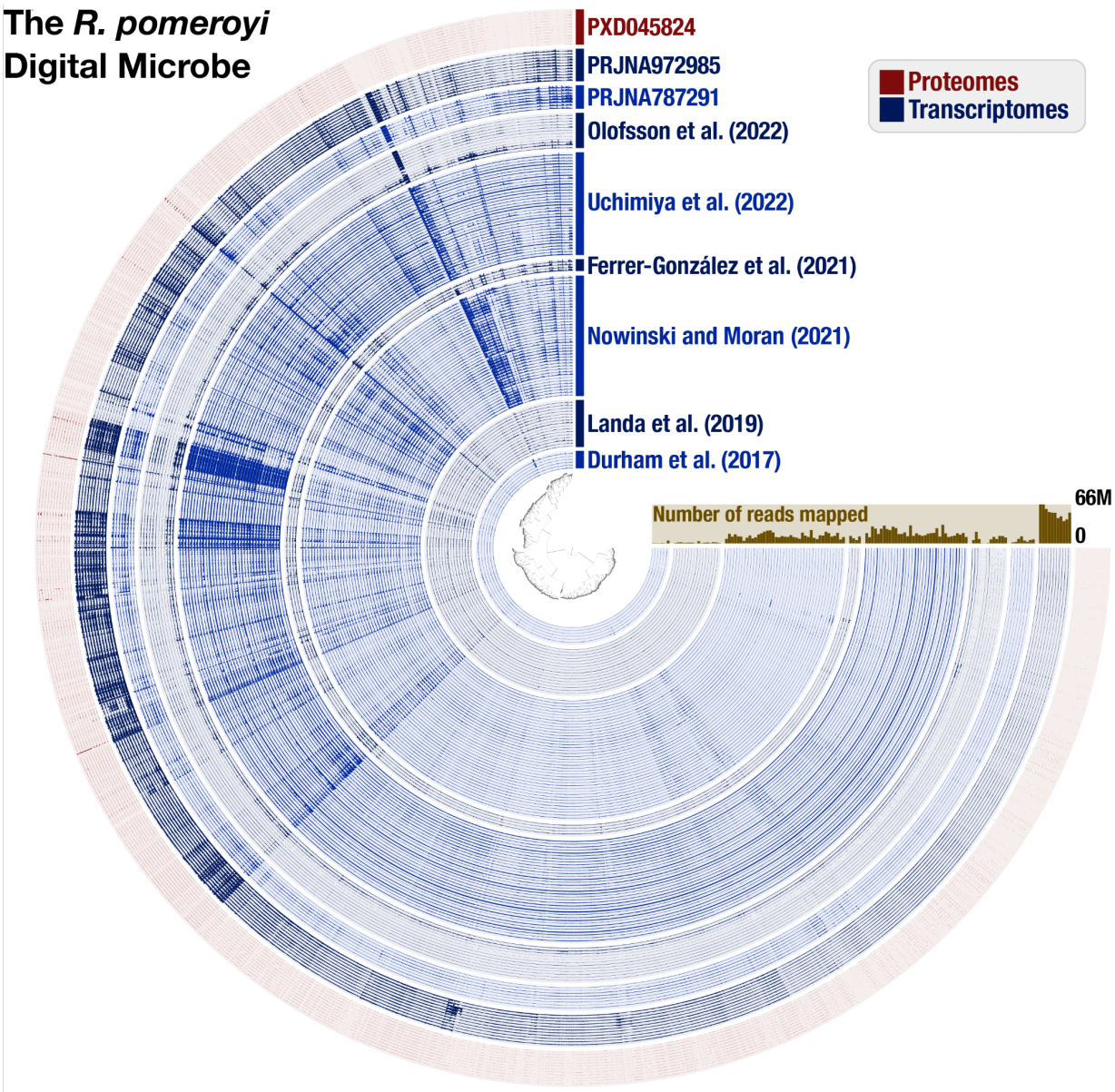


Figure 3. **Contents of the *R. pomeroyi* Digital Microbe**. As visualized in anvi'o 'gene mode', each item on the inner tree corresponds to one gene call in the *R. pomeroyi* genome, and the blue concentric circles display the coverage of each gene in a given transcriptome sample. The outermost red concentric circles correspond to normalized protein abundances from proteome samples (raw files available in the Proteomics Identifications Database (PRIDE) via Project PXD045824). Samples are grouped by their study of origin, with the data source indicated in text of the same color as the samples. The brown bar plot indicates the total number of reads that mapped from each transcriptome to the *R. pomeroyi* genome. This figure was generated from version 5.0 of the *R. pomeroyi* Digital Microbe databases on Zenodo.

which were added to the Digital Microbe (*DM feature 2*). Previous homology-based annotations of most of these transporter systems were either incorrect or vague, and therefore of minimal

ecological value. Although representing only a subset of the ~126 organic carbon influx transporter systems in the *R. pomeroyi* genome, the presence or expression of these 18 is unequivocally linked to a known metabolite. With the new annotations in hand, we undertook a meta-analysis of transporter expression across 133 previously sequenced *R. pomeroyi* transcriptomes from laboratory and field studies between 2014 and 2023 to gain insights into the availability of these 18 metabolites in marine environments.

We added transcriptomes of *R. pomeroyi* to the Digital Microbe by mapping them onto the genome as individual data layers (Figure 3, Table S1) (*DM feature 3*). Using the anvi'o interactive interface, we established a custom dataset that consisted of the 62 protein components of the 18 experimentally annotated transporters (*DM feature 2*). We normalized the read counts for each protein to transcripts per million (TPM) and clustered the resulting data (Euclidean distance and Ward clustering). To generate a heatmap of transporter expression, we extracted the data from anvi'o and visualized it using python (*DM feature 5*).

This meta-analysis captured responses by *R. pomeroyi* to available substrates under 43 different ecological conditions (Figure 4), including during co-culture growth with phytoplankton[25–28], on defined single or mixed substrates[20], and after introduction into a natural phytoplankton bloom[10]. At the broadest scale, the transporters enabling organic acid uptake (acetate, citrate, fumarate, and 3-hydroxybutyrate) had the highest relative expression across conditions, together accounting for an average of 48% (range: 9.7 – 86%) of the transcripts for transporters with confirmed substrates. Recent studies have indeed discovered that Roseobacteraceae members are organic acid catabolic specialists[29,30]. Transporter transcription patterns also revealed the differences in substrate availability across environments. Introduced into a natural dinoflagellate bloom[10], the

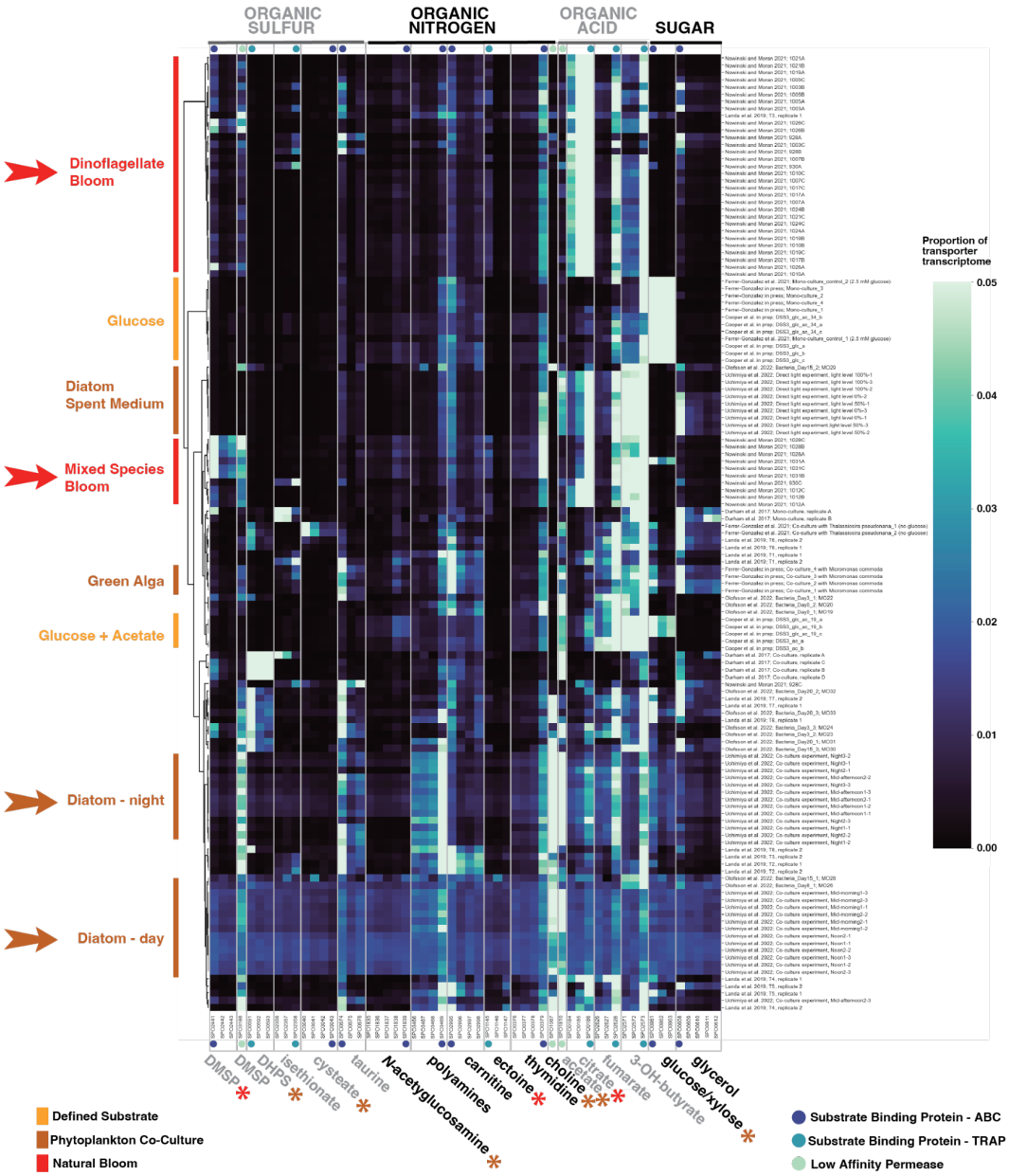


Figure 4. Clustered heatmap of relative gene expression for 18 experimentally annotated *R. pomeroyi* transporters compiled in a Digital Microbe. Each row represents a single transcriptome from the Digital Microbe dataset, and each column represents all transporter proteins with experimentally confirmed cognate substrates. Row labels indicate study and sample name (Table S1). Brighter colors indicate higher proportional expression (the scale maximum is ≥5% of the sum of the 18 transporter transcriptomes) while darker colors indicate lower. Arrows point out transcriptomes collected when substrates were derived from dinoflagellate-rich natural communities (red) or diatom co-cultures (brown); significant differences in transporter protein expression between these two substrate sources are indicated with asterisks colored red (enriched with dinoflagellates) or brown (with diatoms) (T-test, $p \leq 0.05$).

citrate transporter had the highest relative expression; in a diatom co-culture, the acetate transporter was the most highly expressed; co-cultured with a green alga, transporter genes indicated that taurine, glycerol, carnitine, and dimethylsulfoniopropionate (DMSP) were on the menu. The organic acid transporter that enables *R. pomeroyi* uptake of 3-hydroxybutyrate[24] was expressed across most growth conditions, yet this metabolite, also a precursor to the bacterium's storage polymer polyhydroxybutyrate (PHB), has not previously been identified as a relevant currency in bacterially-mediated carbon flux. The meta-analysis also showed a pattern in expression for transporters that contain a substrate binding protein gene (i.e., the ABC and TRAP transporter classes): the gene is expressed at consistently higher levels than other genes in the same transporter (i.e., higher than permeases and ATP-binding proteins) despite all having membership in the same operon. Additional layers of gene regulation are therefore occurring either as within-operon differential expression or as post-transcriptional selective degradation. Regardless, this regulatory strategy would benefit a bacterium in an environment where substrate acquisition is the growth-limiting step.

Three additional genome-linked data types recently added to the *R. pomeroyi* Digital Microbe include the locations of insertion sites of knockout mutants (covering 3,570 genes of 4,288 genes)[13,24]; proteomic data collected concurrently with one of the transcriptomic studies[31,32], and TnSeq mutant fitness measurements in synthetic microbial communities[11] (*DM feature 2*); these are enhancing collaborations among team members.

The *Alteromonas* Digital Microbe

Alteromonas is a genus of marine Gammaproteobacteria whose members associate with particles and can contribute significantly to heterotrophic activity of phytoplankton blooms, sometimes in the role of helper bacteria that provide benefits to the phytoplankton[33–35]. Bacteria in this genus are distinguished by genomes encoding an average of 4,000 genes that enable use of a broad spectrum of substrates[36], provide protection from reactive oxygen species to community members[33], and mediate polysaccharide degradation[37]. The type species of the genus is *Alteromonas macleodii*[34,38], with other notable species including *A. mediterranea*[39], *A. australica*[40], and *A. stellipolaris*[41]. While no single species has emerged as the primary model organism for this genus, the many genomes available for study provide an opportunity for pangenomic analysis to improve understanding of the evolution and diversity of this ubiquitous marine clade[42].

The assembled pangenome consists of 336 *Alteromonas* genomes with genes called and annotated using one standardized pipeline (*DM feature 1*). Of these, 78 are isolate genomes[43–45] and 258 are metagenome-assembled genomes (MAGs) obtained from a variety of marine environments in the global ocean[46]. Genomes represent members of the closely related 'surface' species *A. macleodii* (n=139) and 'deep' species *A. mediterranea* (n=25)[39], and the widely distributed *A. australica* (n=63)[47]. The 34,390 gene clusters of the pangenome are linked to an imported phylogenetic tree assembled from single-copy core genes (see Methods), annotated using NCBI COGs[16], KEGG KOfams[18], CAZyme HMMs[48] and orthology predictions from EggNOG-mapper[49–51] (DM *feature 2*), and assigned as core or accessory genes for the genus (*DM feature 3*) based on a Bayesian approach available in anvi'o[52]. The *Alteromonas* Digital Microbe with relevant pangenome and phylogeny files is accessible on Zenodo[53]. Examples of future versioned enhancements of this Digital Microbe might include additions of new *Alteromonas* genomes and improved annotations from culture studies and novel annotation programs.

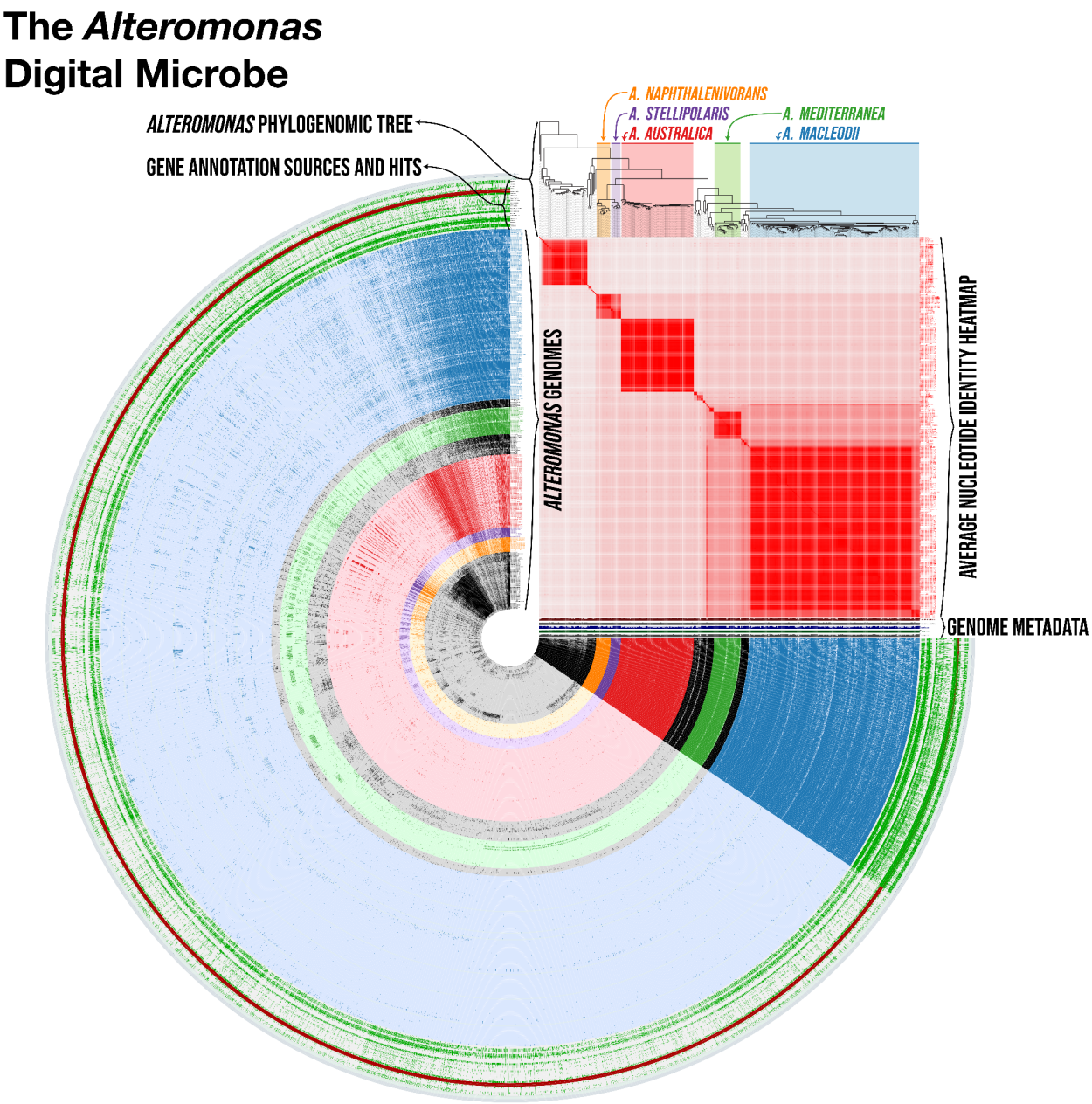


Figure 5. The *Alteromonas* Digital Microbe. Each concentric ring represents one *Alteromonas* genome, with colored rings identifying genomes from five clades of interest (*A. macleodii*, *A. mediterranea*, *A. austalica*, *A. stellioolaris*, and *A. naphthalenivorans*). The outermost green rings depict annotation sources applied to all genomes. Each spoke in the figure represents one gene cluster in the pangenome, with presence/absence denoted by darker/lighter colors, respectively. Genome metadata is shown next to each ring and includes total genome length, GC content, completion, number of genes per kbp, and number of gene clusters per genome. The red heatmap above the metadata shows the average nucleotide identity (ANI) percentage scores between genomes. The tree above the ANI heatmap shows the imported phylogenomic tree, with clades of interest color-referenced in the circular portion of the figure. This figure was generated using the anvi'o 'anvi-display-pan' from a version of the *Alteromonas* digital microbe without singleton genes, which is available on Zenodo under DOI:10.5281/zenodo.10421034.

A Use Case: Evolutionary Patterns of Carbohydrate Use

We leveraged the information contained within the Digital Microbe to examine diversity in the ability of this opportunistic marine genus to use poly-/oligosaccharides[36]. Sugars and sugar polymers are an abundant and diverse component of the ocean's dissolved organic carbon inventory[54], and differences in how microbes use them provide important clues on the evolutionary diversification of their roles in the oceanic carbon cycle. Moreover, the ability to annotate genes with the Carbohydrate-Active enZYme (CAZyme) Database[48] was recently added to anvi'o, allowing augmentation of the Digital Microbe with CAZyme annotations. The results indicate distinct CAZyme distributions across *Alteromonas* clades (Figure 6). For example, the *A. australica* and *A. stellipolaris* clades have more polysaccharide lyases than neighboring clades, while the *A. stellipolaris* clade is enriched in several other CAZyme categories as well. As patterns of diversity in CAZyme inventories are most distinct at the clade level compared to the within-clade level, carbohydrate utilization emerges as a potentially key driver of the large-scale niche partitioning of *Alteromonas* species.

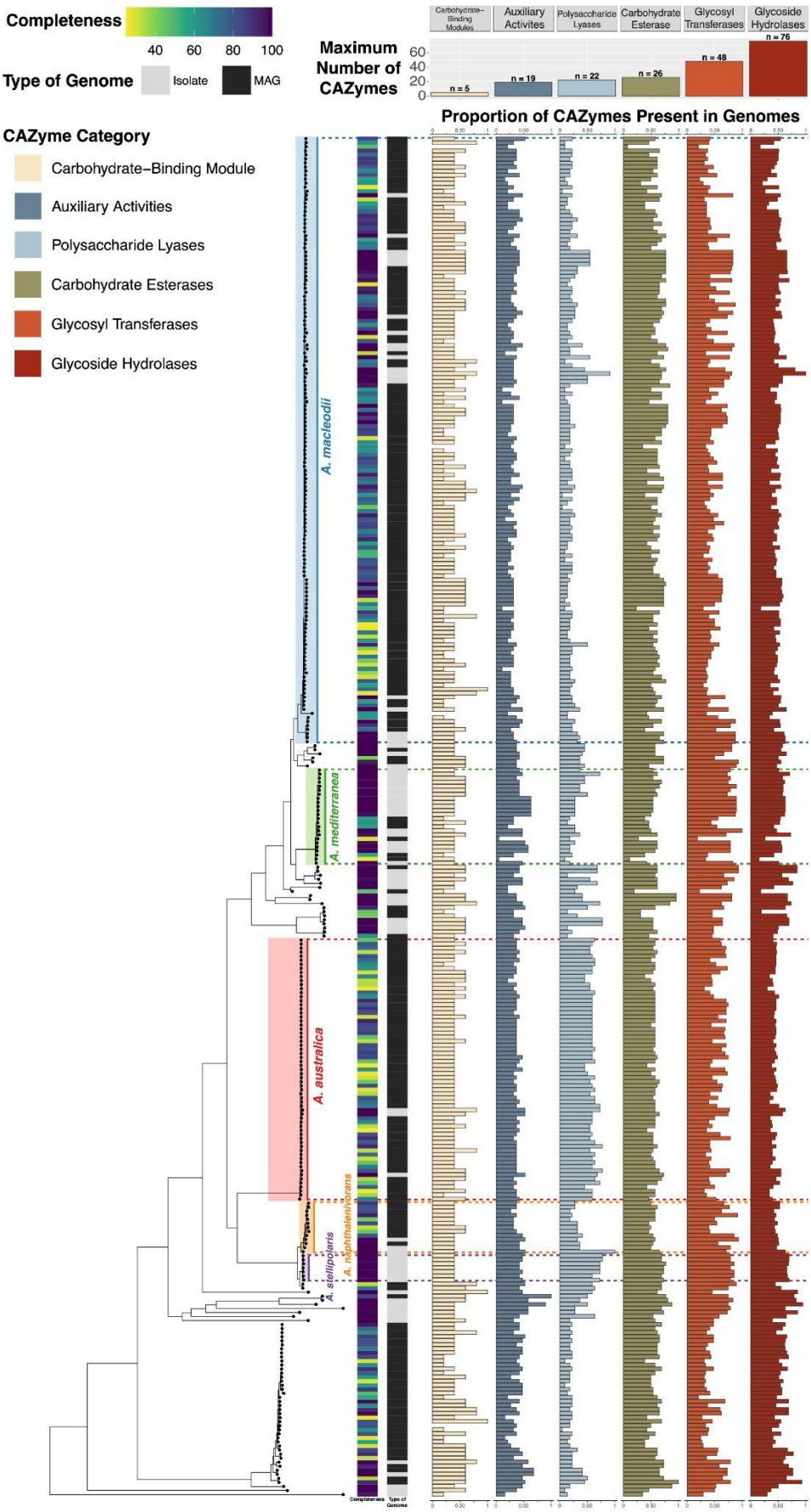

Completeness
40
60
80
100
Type of Genome
Isolate
MAG
Maximum Number of CAZymes
60
40
20
0
Carbohydrate-Binding Modules
Auxiliary Activites
Polysaccharide Lyases
Carbohydrate Esterase
Glycosyl Transferases
Glycoside Hydrolases
n = 5
n = 19
n = 22
n = 26
n = 48
n = 76
Proportion of CAZymes Present in Genomes
CAZyme Category
Carbohydrate-Binding Module
Auxiliary Activities
Polysaccharide Lyases
Carbohydrate Esterases
Glycosyl Transferases
Glycoside Hydrolases
A. macleodii
A. mediterranea
A. australica
A. naphthalenivorans
A. stellipolaris
Completeness
Type of Genome

Figure 6. Distribution of CAZYme annotations across a phylogeny of 336 isolate and MAG genomes from the genus *Alteromonas*. The phylogeny of the genus is displayed on the left side of the figure, with genomes represented by points on the tree and five of the clades (*A. macleodii*, *A. mediterranea*, *A. australica*, *A. napthalenivorans*, and *A. stellipolaris*) highlighted. Each row on the right side of the figure represents one genome. Completeness and type of genome are shown in the two heatmaps to the right of the phylogeny. The horizontal bar plots of different colors show the proportion of CAZymes in each genome relative to the maximum number of all categories of CAZymes as indicated in the legend in the inset at the upper left. The maximum number for each CAZyme category is represented by the vertical bar plot at the top of the figure.

We also gained insights into how carbohydrate usage has shaped *Alteromonas* evolution and ecology from gene phylogenies of selected CAZymes (Supplementary Figure 1). The topology of several CAZyme phylogenies broadly recapitulates the topology of the genome phylogeny built from single-copy core genes (Supplementary Figure 2), suggesting that vertical descent has dominated the evolution of these genes. However, the topologies of other CAZyme phylogenies have significant discordance with the genome phylogeny (Supplementary Figure 2), suggesting that horizontal transfer has also had an important role in the evolution of carbohydrate utilization strategies in *Alteromonas*. The divergent evolutionary trajectories of different CAZymes highlight selective pressures acting on the metabolic diversification of *Alteromonas*, and may offer clues on how this diversification has in turn impacted the evolution of carbon cycling in the ocean.

Future directions

Digital Microbe data packages furnish an architecture for reproducible, open, and extensible collaborative work in microbiology and its many derivative fields. While we present here a specific architecture tailored to our research focus, it is only one manifestation of the broader digital microbe concept: that decentralized taxon-specific databases are key mechanisms for capturing knowledge accumulating from genome-informed data that are now so vast and distributed as to be intractable to synthesize[55]. Digital Microbe packages allow one-stop shopping for data spread across multiple public archives, allow coordinated selection and documentation of genome structure and annotations within and between research teams, and are extensible to new data types. While the framework presented here is designed for bacterial and archaeal data, the development of digital microbes for eukaryotic model organisms is an important future application[56]. One enhancement under development by C-CoMP is an integrated toolkit for metabolic modeling, but the nature and scope of future applications can be defined by any research group that uses a digital microbe framework for their research. Organized and open access to taxon-explicit data is an essential foundation for modern microbiology.

Methods

Both Digital Microbes were generated and analyzed using v7.1-dev or later versions of anvi'o[6].

Creation of the *Ruegeria pomeroyi* Digital Microbe

We created the *Ruegeria pomeroyi* Digital Microbe from the *R. pomeroyi* DSS-3 complete genome and megaplasmid sequences[14] and (meta)transcriptome samples from [10,27,28,32,57–59]. We generated a contigs database from the genome and megaplasmid sequences with `anvi-gen-contigs-database` and annotated the resulting Prodigal[60] gene calls with *de novo* annotations from NCBI Clusters of Orthologous Genes (COGs)[16], KEGG KOfams[18], and Pfam[17] via the associated anvi'o program for each annotation source. We also identified single-copy core genes using `anvi-run-hmms` and associated these genes with taxonomy data from the GTDB[61] using `anvi-run-scg-taxonomy`. We imported manually curated gene annotations, including annotations indicating which genes have available mutants[13], using the program `anvi-import-functions`.

To process the (meta)transcriptomes, we quality-filtered the samples using FASTX-toolkit[62] with the parameters described in [25]. We mapped the reads to the DSS-3 genome using Bowtie 2[63] and samtools[64]. Each sample's read mapping data were converted into an anvi'o profile database using `anvi-profile`, and all samples were merged into a single database with `anvi-merge`. To add proteomic data[31], we normalized spectral abundance counts with a normalized spectral abundance factor to make data comparable across all proteomes. We generated a 'genes database' to store gene-level information by running `anvi-interactive` on the established contigs and profile databases with the `--gene-mode` flag, and we imported the normalized abundances for each gene into the genes database using the program `anvi-import-misc-data`. We also used this program to import fitness data associated with gene mutants from [11] into the same genes database.

Transporter expression analysis for *Ruegeria pomeroyi*

The genes database in the *R. pomeroyi* Digital Microbe contains gene-level transcript coverage information from >100 samples. To assess the proportional expression of substrate-confirmed transporter genes, we used the anvi'o interactive interface to create a bin containing the transporter genes, and generated a static summary page with the "init-gene-coverages" box checked to export annotation and coverage data for each contig region where our genes of interest were located. After reading the exported data into dataframes using python v3.7.8 and pandas[65,66], we extracted the coverage data for our specific genes of interest, normalized the

coverages to TPM using the total number of reads per sample, and relativized these data to represent the proportional expression of each gene. We visualized these data as a clustermap using the seaborn package[67] and assessed statistical differences in the mean gene expression using the a t-test implemented in the scipy stats package[68].

Creation of the *Alteromonas* Digital Microbe

To create the *Alteromonas* Digital Microbe, we collected 78 isolate genomes and 258 MAGs from the Joint Genome Institute's Integrated Microbial Genomes (IMG) project[43], NCBI[69], and [46]. We converted each genome into an anvi'o contigs database using `anvi-gen-contigs-database`. For the genomes from IMG and NCBI, we determined completion and contamination statistics using CheckM v1.0.18[70]; for the MAGs that were taken from [46], we used the mean completeness and mean contamination statistics reported in that publication. We annotated the genes in each contigs database with the NCBI Clusters of Orthologous Genes (COGs)[16], KEGG KOfams[18], and Carbohydrate-Active enZYme (CAZyme) HMMs[48] via the associated anvi'o program for each annotation source, and imported externally-run annotations from EggNOG-mapper[49–51] and KEGG GhostKOALA[71] into the databases using `anvi-import-functions`.

We ran `anvi-pan-genome` to create the pangenome and computed the average nucleotide identity (ANI) between all pairs of genomes using `anvi-compute-genome-similarity`. To extract the core genome from the pangenome (i.e., genes found in all genomes), we used a Bayesian statistical method[52] implemented in `anvi-script-compute-bayesian-pan-core`. This method employs mOTUpan.py to determine the gene clusters likely to be core based on individual genome completeness scores.

Phylogenomic Analysis of the *Alteromonas* genomes

To build the phylogeny of *Alteromonas*, we aligned and concatenated the sequences from 110 single-copy core gene clusters using `anvi-get-sequences-for-gene-clusters`. We imported these sequences into the tree building software RAxML, version 8.2.12[72], and built the tree under the "PROTGAMMAAUTO" model setting. We used FigTree v1.4.4[73] to midpoint root the tree and save it in newick file format. To incorporate the tree into the pangenome, we imported the newick tree with the program `anvi-import-misc-data`. For the phylogenomic CAZyme analysis, we used `anvi-split` to subset gene clusters with known CAZyme functions into a new pangenome database and ran `anvi-summarize` on this smaller pangenome to count the number of CAZymes per category, per genome. We visualized these data as a function of the previously-determined phylogeny in R

385 v4.1.1[74] using the packages aplot v0.1.9[75], BiocManager v1.30.20[76], dplyr v1.1.0[77], ggnewscale v0.4.8[78], ggplot2 v3.4.1[79], ggstance v0.3.6.9000[80], ggtree v3.7.1.003[81–85], ggtreeExtra v1.9.1.992[81,86], nationalparkcolors v0.1.0[87], plyr v1.8.8[88], RColorBrewer v1.1-3[89], scales v1.2.1[90], and tidyr v1.3.0[91].

We then repeated the initial steps above to generate a phylogeny for the subset of isolate
390 genomes (n=78), which resulted in a tree built from 111 single-copy core gene clusters. After subsetting the gene clusters with known CAZyme annotations into a smaller pangenome, we identified eight CAZyme-related gene clusters that were part of the single-copy core genome. We then generated an individual phylogeny from each of these CAZymes. We used R to compare the CAZyme phylogenies with the overall core genome phylogeny for these isolate genomes, with
395 the packages listed above in addition to colorBlindness v0.1.9[92], easyalluvial v0.3.1[93], and gridExtra v2.3[94].

Data Availability

The *Ruegeria pomeroyi* Digital Microbe is available via DOI:10.5281/zenodo.7304959 and the *Alteromonas* Digital Microbe is available via DOI:10.5281/zenodo.7430118. The raw proteomics
400 data included in the *Ruegeria pomeroyi* Digital Microbe is available on the Proteomics Identifications Database (PRIDE) project PXD045824 with accompanying metadata and processed data available in Biological and Chemical Oceanography Data Management Office (BCO-DMO) dataset 927507 via DOI: 10.26008/1912/bco-dmo.927507.1. The accompanying raw transcriptomic expression data to the proteomics data is available under the National Center for
405 Biotechnology Information (NCBI) BioProject PRJNA972985 with metadata available in BCO-DMO dataset 916134 via DOI:10.26008/1912/bco-dmo.916134.1.

Code Availability

Reproducible workflows for the generation of the Digital Microbes and the analyses described in this work can be accessed at https://github.com/C-CoMP-STC/digital-microbe. In particular, the
410 Jupyter notebook for the *Ruegeria pomeroyi* use-case analysis can be found at https://github.com/C-CoMP-STC/digital-microbe/blob/main/rpom/rpom_dig_micro_transporter_expression_use_case.ipynb and the workflow for the *Alteromonas* use-case analysis can be found at the following link:

https://github.com/C-CoMP-STC/digital-microbe/blob/main/alteromonas/useCase/alteromonasUseCases.md.

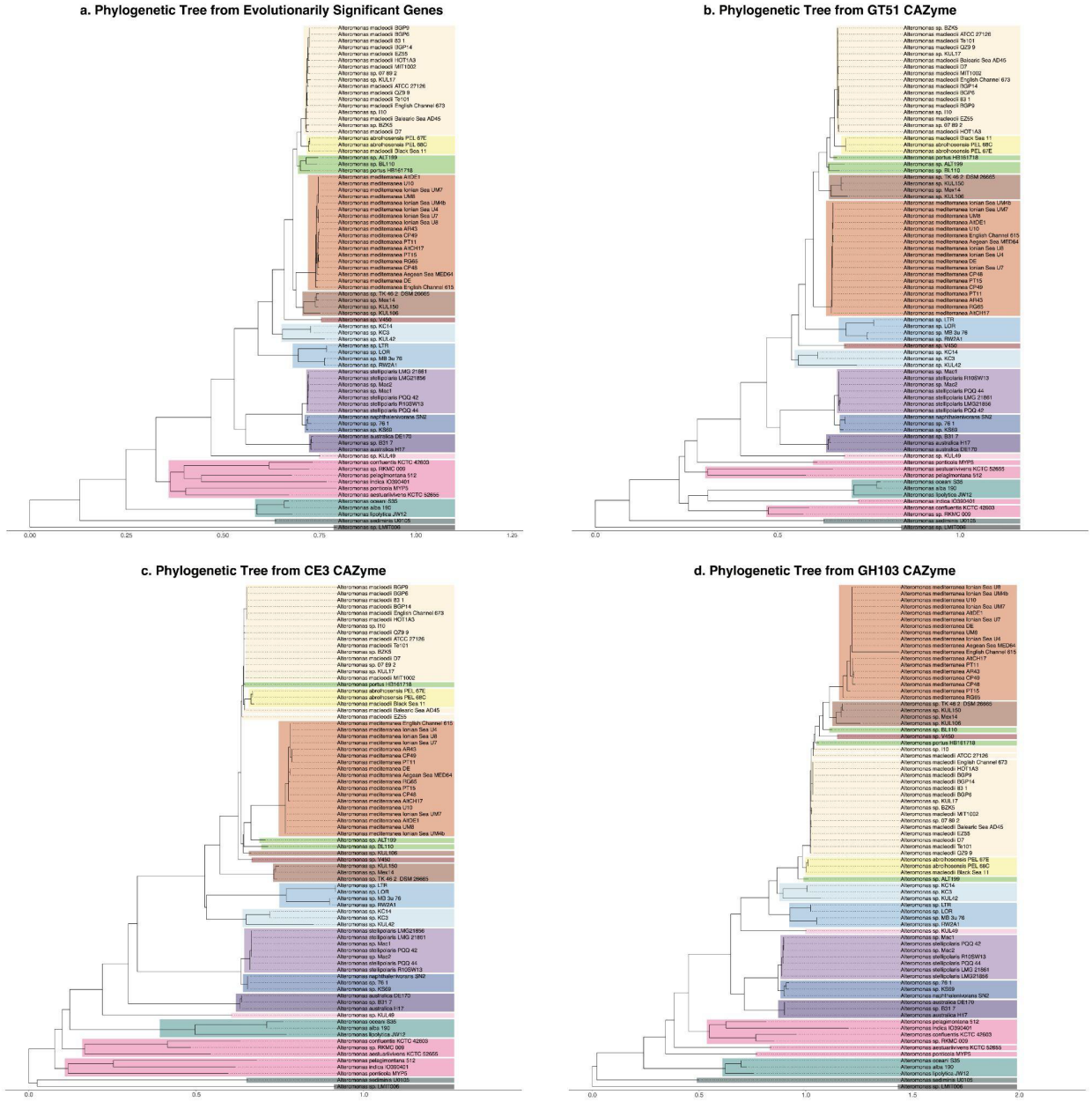
Acknowledgements

This work was supported under NSF grant OCE-2019589 to the Center for Chemical Currencies of a Microbial Planet. IV acknowledges support by the National Science Foundation Graduate Research Fellowship under Grant No. 1746045. MAM acknowledges support by Simons Foundation Grant 542391 within the Principles of Microbial Ecosystems Collaborative. This is C-CoMP publication #024.

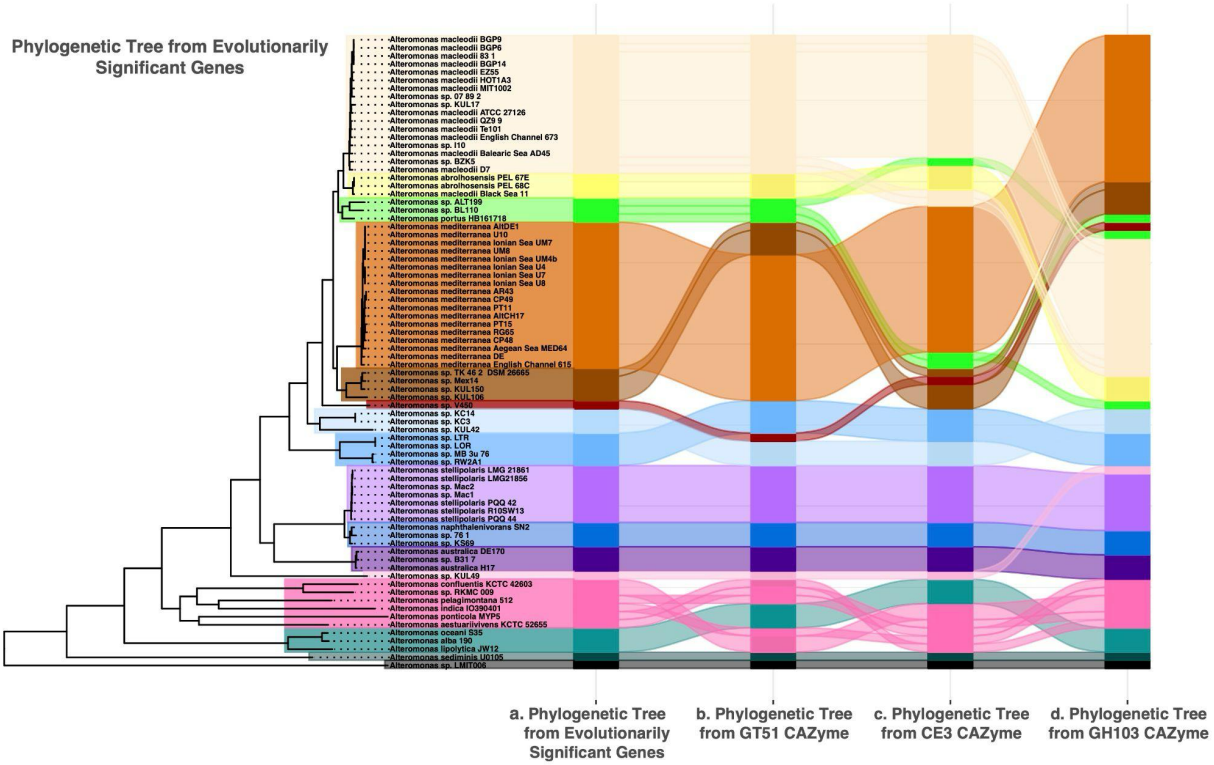
Supplementary Tables

Table S1. Transcriptome samples included in the *Ruegeria pomeroyi* Digital Microbe.

Supplementary Figures



Supplementary Figure 1. Individual phylogenies of core genes (a) and CAZymes (b, c, d) from the *Alteromonas* pangenome. The core genes (n=111) used to build the accepted phylogeny in (a) consist of those found in all 78 isolate genomes as single copy genes ('--max-num-genes-from-each-genome' set to 1), with the '--max-functional-homogeneity-index' set to 0.9 and '--min-geometric-homogeneity-index' set to 0.925. The CAZyme phylogenies (b, c, d) are each built from the alignment of one gene - GT51, CE3, and GH103, respectively. Clades are highlighted based on color assignments in Supplementary Figure 2. The horizontal axis provides a means for interpreting branch length as a measure of genetic change.



Supplementary Figure 2. Comparison of core gene (a) and CAZyme (b, c, d) phylogenies for the *Alteromonas* pangenome. The core genes (n=111) used to build the accepted phylogeny in (a) consist of those found in all 78 isolate genomes as single copy genes ('--max-num-genes-from-each-genome' set to 1), with the '--max-functional-homogeneity-index' set to 0.9 and '--min-geometric-homogeneity-index' set to 0.925. The CAZyme phylogenies (b, c, d) are built from the alignment of one gene each. The alluvial plot displays the movement of clades within the tree. Further branch definition of each tree is shown in Supplementary Figure 1.

References

1. Wetmore, K. M. *et al.* Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio* **6**, e00306–15 (2015).
2. Moore, L. R. & Chisholm, S. W. Photophysiology of the marine cyanobacterium Prochlorococcus: Ecotypic differences among cultured isolates. *Limnol. Oceanogr.* **44**, 628–638 (1999).
3. Sun, J. *et al.* One carbon metabolism in SAR11 pelagic marine bacteria. *PLoS One* **6**,

e23973 (2011).

4. González, J. M. *et al.* Genome analysis of the proteorhodopsin-containing marine bacterium *Polaribacter* sp. MED152 (Flavobacteria). *Proc. Natl. Acad. Sci. U. S. A.* **105**, 8724–8729 (2008).

5. Leonelli, S. Model Organism. in *Encyclopedia of Systems Biology* (eds. Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H.) 1398–1401 (Springer New York, 2013).

6. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* **6**, 3–6 (2021).

7. Newton, R. J. *et al.* Genome characteristics of a generalist marine bacterial lineage. *ISME J.* **4**, 784–798 (2010).

8. Munson-McGee, J. H. *et al.* Decoupling of respiration rates and abundance in marine prokaryoplankton. *Nature* **612**, 764–770 (2022).

9. Fu, H., Uchimiya, M., Gore, J. & Moran, M. A. Ecological drivers of bacterial community assembly in synthetic phycospheres. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 3656–3662 (2020).

10. Nowinski, B. & Moran, M. A. Niche dimensions of a marine bacterium are identified using invasion studies in coastal seawater. *Nat Microbiol* **6**, 524–532 (2021).

11. Schreier, J. E., Smith, C. B., Ioerger, T. R. & Moran, M. A. A mutant fitness assay identifies bacterial interactions in a model ocean hot spot. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2217200120 (2023).

12. Howard, E. C. *et al.* Bacterial taxa that limit sulfur flux from the ocean. *Science* **314**, 649–652 (2006).

13. Mejia, C. *et al.* An arrayed transposon library of *Ruegeria pomeroyi* DSS-3. *bioRxiv* 2022.09.11.507510 (2022) doi:10.1101/2022.09.11.507510.

14. Moran, M. A. *et al.* Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* **432**, 910–913 (2004).

15. Rivers, A. R., Smith, C. B. & Moran, M. A. An updated genome annotation for the model marine bacterium *Ruegeria pomeroyi* DSS-3. *Stand. Genomic Sci.* **9**, 11 (2014).
16. Galperin, M. Y. *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).
17. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
18. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
19. Christie-Oleza, J. A., Miotello, G. & Armengaud, J. High-throughput proteogenomics of *Ruegeria pomeroyi*: seeding a better genomic annotation for the whole marine Roseobacter clade. *BMC Genomics* **13**, 73 (2012).
20. Bullock, H. A., Reisch, C. R., Burns, A. S., Moran, M. A. & Whitman, W. B. Regulatory and functional diversity of methylmercaptopropionate coenzyme A ligases from the dimethylsulfoniopropionate demethylation pathway in *Ruegeria pomeroyi* DSS-3 and other proteobacteria. *J. Bacteriol.* **196**, 1275–1285 (2014).
21. Cunliffe, M. Purine catabolic pathway revealed by transcriptomics in the model marine bacterium *Ruegeria pomeroyi* DSS-3. *FEMS Microbiol. Ecol.* **92**, (2016).
22. Sharpe, G. C., Gifford, S. M. & Septer, A. N. A model Roseobacter, *Ruegeria pomeroyi* DSS-3, employs a diffusible killing mechanism to eliminate competitors. *mSystems* **5**, (2020).
23. Veseli, I. & Cooper, Z. *Ruegeria pomeroyi* digital microbe databases. (2022) doi:10.5281/zenodo.7439166.
24. Schroer, W. F. *et al.* Functional annotation and importance of marine bacterial transporters of plankton exometabolites. *ISME Communications* **3**, 1–10 (2023).
25. Landa, M., Burns, A. S., Roth, S. J. & Moran, M. A. Bacterial transcriptome remodeling during sequential co-culture with a marine dinoflagellate and diatom. *ISME J.* **11**, 2677–

500 2690 (2017).

26. Ferrer-González, F. X. *et al.* Bacterial transcriptional response to labile exometabolites from photosynthetic picoeukaryote *Micromonas commoda*. *ISME Commun* **3**, 5 (2023).

27. Ferrer-González, F. X. *et al.* Resource partitioning of phytoplankton metabolites that support bacterial heterotrophy. *ISME J.* **15**, 762–773 (2021).

505 28. Olofsson, M. *et al.* Growth-stage-related shifts in diatom endometabolome composition set the stage for bacterial heterotrophy. *ISME Communications* **2**, 1–9 (2022).

29. Gralka, M., Pollak, S. & Cordero, O. X. Genome content predicts the carbon catabolic preferences of heterotrophic bacteria. *Nat Microbiol* **8**, 1799–1808 (2023).

30. Forchielli, E., Sher, D. & Segrè, D. Metabolic phenotyping of marine heterotrophs on

510 refactored media reveals diverse metabolic adaptations and lifestyle strategies. *mSystems* **7**, e0007022 (2022).

31. Cooper, Z. S. *et al.* Normalized protein abundance data and protein annotations for proteomic data from laboratory cultures of *Ruegeria pomeroyi* DSS-3 and *Alteromonas macleodii* MIT1002 in 2022. (2024) doi:10.26008/1912/bco-dmo.927507.1.

515 32. Cooper, Z. S., Gray, L., Rauch, S. & Moran, M. A. Metadata for transcriptomic expression data from cultures of *Ruegeria pomeroyi* DSS-3 and *Alteromonas macleodii* MIT1002 grown in defined culture media with either glucose, acetate, or a mix of both as carbon substrates. (2023) doi:10.26008/1912/bco-dmo.916134.1.

33. Hennon, G. M. M. *et al.* The impact of elevated CO_2 on *Prochlorococcus* and microbial

520 interactions with 'helper' bacterium *Alteromonas*. *ISME J.* **12**, 520–531 (2017).

34. López-Pérez, M. *et al.* Genomes of surface isolates of *Alteromonas macleodii*: the life of a widespread marine opportunistic copiotroph. *Sci. Rep.* **2**, 696 (2012).

35. Henríquez-Castillo, C. *et al.* Metaomics unveils the contribution of *Alteromonas* bacteria to carbon cycling in marine oxygen minimum zones. *Frontiers in Marine Science* **9**, (2022).

525 36. Mikhailov, V. V., Romanenko, L. A. & Ivanova, E. P. The genus *Alteromonas* and related

Proteobacteria. in *The Prokaryotes: A Handbook on the Biology of Bacteria Volume 6: Proteobacteria: Gamma Subclass* (eds. Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H. & Stackebrandt, E.) 597–645 (Springer New York, 2006).

37. Koch, H. *et al.* Biphasic cellular adaptations and ecological implications of *Alteromonas macleodii* degrading a mixture of algal polysaccharides. *ISME J.* **13**, 92–103 (2019).
38. Baumann, L., Baumann, P., Mandel, M. & Allen, R. D. Taxonomy of aerobic marine eubacteria. *J. Bacteriol.* **110**, 402–429 (1972).
39. Ivanova, E. P. *et al.* Ecophysiological diversity of a novel member of the genus *Alteromonas*, and description of *Alteromonas mediterranea* sp. nov. *Antonie Van Leeuwenhoek* **107**, 119–132 (2015).
40. Ivanova, E. P. *et al. Alteromonas australica* sp. nov., isolated from the Tasman Sea. *Antonie Van Leeuwenhoek* **103**, 877–884 (2013).
41. Van Trappen, S., Tan, T.-L., Yang, J., Mergaert, J. & Swings, J. *Alteromonas stellipolaris* sp. nov., a novel, budding, prosthecate bacterium from Antarctic seas, and emended description of the genus *Alteromonas*. *Int. J. Syst. Evol. Microbiol.* **54**, 1157–1163 (2004).
42. López-Pérez, M. & Rodriguez-Valera, F. Pangenome evolution in the marine bacterium *Alteromonas*. *Genome Biol. Evol.* **8**, 1556–1570 (2016).
43. Chen, I.-M. A. *et al.* The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.* **51**, D723–D732 (2023).
44. Mukherjee, S. *et al.* Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v.9. *Nucleic Acids Res.* **51**, D957–D963 (2023).
45. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **49**, D10–D17 (2021).
46. Paoli, L. *et al.* Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
47. López-Pérez, M., Gonzaga, A., Ivanova, E. P. & Rodriguez-Valera, F. Genomes of

Alteromonas australica, a world apart. *BMC Genomics* **15**, 483 (2014).

48. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, W445–51 (2012).
49. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
50. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
51. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
52. Buck, M., Mehrshad, M. & Bertilsson, S. mOTUpan: a robust Bayesian approach to leverage metagenome-assembled genomes for core-genome estimation. *NAR Genom Bioinform* **4**, lqac060 (2022).
53. DeMers, M. & Braakman, R. *Alteromonas* Digital Organism Databases. (2023) doi:10.5281/zenodo.7517480.
54. Repeta, D. J. Chapter 2 - Chemical characterization and cycling of dissolved organic matter. in *Biogeochemistry of Marine Dissolved Organic Matter (Second Edition)* (eds. Hansell, D. A. & Carlson, C. A.) 21–63 (Academic Press, 2015).
55. Kelly, L., Huang, K. H., Ding, H. & Chisholm, S. W. ProPortal: a resource for integrated systems biology of *Prochlorococcus* and its phage. *Nucleic Acids Res.* **40**, D632–40 (2012).
56. Groussman, R. D., Blaskowski, S., Coesel, S. N. & Armbrust, E. V. MarFERReT, an open-source, version-controlled reference library of marine microbial eukaryote functional genes. *Sci Data* **10**, 926 (2023).
57. Durham, B. P. *et al.* Cryptic carbon and sulfur cycling between surface ocean plankton. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 453–457 (2015).

58. Landa, M. *et al.* Sulfur metabolites that facilitate oceanic phytoplankton–bacteria carbon flux. *ISME J.* **13**, 2536–2550 (2019).

580 59. Uchimiya, M., Schroer, W., Olofsson, M., Edison, A. S. & Moran, M. A. Diel investments in metabolite production and consumption in a model microbial system. *ISME J.* **16**, 1306–1317 (2022).

60. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

585 61. Parks, D. H. *et al.* 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 (2022).

62. Gordon, A. & Hannon, G. J. *Fastx-toolkit*. (2010).

63. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*

590 **9**, 357–359 (2012).

64. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).

65. McKinney, W. Data structures for statistical computing in Python. in *Proceedings of the 9th Python in Science Conference* (SciPy, 2010). doi:10.25080/majora-92bf1922-00a.

66. The pandas development team. *Pandas-dev/pandas: Pandas*. (2023).

595 doi:10.5281/zenodo.7794821.

67. Waskom, M. *Seaborn: statistical data visualization*. (The Open Journal, 2021). doi:10.21105/joss.03021.

68. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

600 69. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **50**, D20–D26 (2022).

70. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and

metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

605 71. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).

72. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

610 73. Rambaut, A. *FigTree*.

74. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, Vienna, Austria., 2018).

75. Yu, G. *Decorate a 'ggplot' with associated information [R package aplot version 0.1.10]*. (Comprehensive R Archive Network (CRAN), 2023).

615 76. Morgan, M. *Access the Bioconductor project package repository [R package BiocManager version 1.30.20]*. (Comprehensive R Archive Network (CRAN), 2023).

77. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. *A grammar of data manipulation [R package dplyr version 1.1.2]*. (Comprehensive R Archive Network (CRAN), 2023).

620 78. Campitelli, E. *Multiple fill and colour scales in 'ggplot2' [R package ggnewscale version 0.4.8]*. (Comprehensive R Archive Network (CRAN), 2022).

79. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

80. Henry, L., Wickham, H. & Chang, W. *Ggstance: horizontal ggplot2 components*. (Github,

625 2022).

81. Yu, G. *Data Integration, Manipulation and Visualization of Phylogenetic Trees*. (CRC Press, Taylor & Francis Group, 2022).

82. Xu, S. *et al. Ggtree*: a serialized data object for visualization of a phylogenetic tree and annotation data. (2022) doi:10.1002/imt2.56.

630 83. Yu, G. Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020).

84. Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Mol. Biol. Evol.* **35**, 3041–3043 (2018).

85. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. Ggtree : An R package for
635 visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

86. Xu, S. *et al.* ggtreeExtra: Compact visualization of richly annotated phylogenetic data. (2021) doi:10.1093/molbev/msab166.

87. Jolly, K. *Nationalparkcolors: R package for color themes inspired by National Parks*.
640 (Github).

88. Wickham, H. The Split-Apply-Combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–29 (2011).

89. Neuwirth, E. *RColorBrewer: ColorBrewer palettes. [R package RColorBrewer version 1.1-3]*. (Comprehensive R Archive Network (CRAN), 2022).

645 90. Wickham, H. & Seidel, D. *Scale functions for visualization [R package scales version 1.2.1]*. (Comprehensive R Archive Network (CRAN), 2022).

91. Wickham, H., Vaughan, D. & Girlich, M. *Tidyr: Tidy messy data [R package tidyr version 1.3.0]*. (Comprehensive R Archive Network (CRAN), 2023).

92. Ou, J. *Safe color set for color blindness [R package colorBlindness version 0.1.9]*.
650 (Comprehensive R Archive Network (CRAN), 2021).

93. Oettinghaus, B. *Easyalluvial: Generate alluvial plots with a single line of code [R package version 0.3.1]*. (Github, 2022).

94. Auguie, B. *Miscellaneous functions for 'grid' graphics [R package gridExtra version 2.3]*. (Comprehensive R Archive Network (CRAN), 2017).

655 95. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status,

taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).

96. DiatOmicBase. https://www.diatomicsbase.bio.ens.psl.eu/.

97. Sajed, T. *et al.* ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. *Nucleic Acids Res.* **44**, D495–501 (2016).

98. Arkin, A. P. *et al.* KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* **36**, 566–569 (2018).

99. Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* **50**, W345–W351 (2022).

100. Gillespie, J. J. *et al.* PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.* **79**, 4286–4298 (2011).

101. Merchant, N. *et al.* The iPlant Collaborative: Cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol.* **14**, e1002342 (2016).

102. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).

103. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).

104. Olson, R. D. *et al.* Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res.* **51**, D678–D689 (2023).

105. Grigoriev, I. V. *et al.* PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res.* **49**, D1004–D1011 (2021).

106. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

107. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

108. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* **35**, 4162–

4164 (2019).

109. Karp, P. D. *et al.* Pathway Tools version 24.0: Integrated software for Pathway/genome informatics and systems biology. *arXiv [q-bio.GN]* (2015).

110. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).

Author Contributions

ZSC, MSS, LTR, SD, AME, MAM, and RB conceptualized the study. ZSC, MAD, IV, SM, CBS, LTR, WFS, MRM, PZL, and MS curated data. ZSC, MAD, MAM, and RB performed formal analyses. ZSC, MAD, CBS, LTR, WFS, MRM, PZL, and MS conducted investigations. IV, MSS, SM, and AME developed methodology. LW and AME administered the project. AME and MAM provided resources. IV, MSS, SM, and AME developed software. LW, AME, MAM, and RB supervised the project. ZSC, MAD, and IV validated the results. ZSC, MAD, LW, MAM, and RB worked on visualization. ZSC, MAD, IV, MSS, SM, LW, SD, AME, MAM, and RB wrote the paper with critical input from all authors.

Competing Interests

The authors declare no competing interests.