Digital Microbe: A Genome-Informed Data Integration Framework for Collaborative Research on Emerging Model Organisms

Iva Veseli[1,2,‡], Zachary S. Cooper[3,‡], Michelle A. DeMers[4,‡], Matthew S. Schechter[5], Samuel Miller[6], Laura Weber[7], Christa B. Smith[3], Lidimarie T. Rodriguez[8], William F. Schroer[3], 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[3,†], Rogier Braakman[4,†]

[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 Marine Sciences, University of Georgia, Athens, GA 30602, USA
[4]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, 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 focused on the physiology and ecology of microbes is rapidly transforming our understanding of life at the micron scale. Yet this data stream has also created challenges for finding interoperable and extensible modes of analysis. From our own experience, a single microbe often has multiple versions of its genome architecture, functional gene annotation, and gene naming system, without a straightforward mechanism for collating information and preserving crucial advances in annotation. These dispersed data sources raise barriers to collaborations, and more generally hinder community coalescence around shared datasets of model organisms. Here, we describe the “Digital Microbe” data product which provides a framework for interoperability, reproducibility, and collaborative microbial science. A Digital Microbe is an open source, community-curated data package built on a (pan)genome foundation, which is housed within an integrative software environment. Using Digital Microbes ensures real-time alignment of research efforts within collaborative teams, and, as new layers of 'omic, experimental, or modeling data are added, facilitates the generation of novel scientific insights. We describe two Digital Microbes, one for the model heterotrophic marine bacterium *Ruegeria pomeroyi* DSS-3 which includes >100 transcriptomic datasets from lab and field studies; and another for the pangenome of the cosmopolitan heterotrophic marine bacterial genus *Alteromonas* represented by 339 genomes. Examples are provided to demonstrate how an integrated framework that collates public (pan)genome-informed data can generate novel and reproducible findings.

Introduction

With a genome sequence as the price of admission, a growing number of cultured microbes are being developed as model organisms. Genome sequences are subsequently enhanced with knowledge from experimental, modeling, and field studies (e.g., Wetmore et al. 2015; Moore and Chisholm 1999; Sun et al. 2011; González et al. 2008) with the goal of yielding generalized knowledge relevant to microbial physiology, ecology, or biogeochemistry. To extend the impact of any microbial model beyond a single laboratory group, requirements are: 1) an established reference dataset (Leonelli 2013), which provides existing and newly-generated 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 model organism data is increasing as team science takes on growing roles in environmental and life sciences research.

In the NSF Center for Chemical Currencies of a Microbial Planet (C-CoMP; ccomp-stc.org), the necessity arose for a platform to house standardized data and enable iterative knowledge updates generated for the Center's two model marine bacterial species, *Ruegeria pomeroyi* DSS-3 and *Alteromonas macleodii* MIT1002. Collaborating Center members must simultaneously access and analyze the same experimental and modeling data types ranging from 'omics datasets and environmental parameters to metabolic models and metabolomes. As researchers are geographically distributed across 12 institutions, an established data sharing and curation strategy was particularly important. The Center's commitment to open science practices also required that the data platform be open-source and the data products be publicly accessible.

Several digital resources have been developed for the analysis and exchange of data focused on a model microbe or group of related microbes. These tools can be broadly characterized into three groups: 1) online portals that provide a centralized location for uploading or downloading 'omics datasets; 2) online portals that allow the user to choose from pre-selected 'omics datasets or, in some cases, upload their own data for analysis with embedded applications; and 3) downloadable tools that enable local analysis of 'omics data (Table 1). While they provide important services to many researchers, these platforms do not necessarily maximize a team science approach. Typically, datasets are provided as either raw or highly-polished results, but intermediate data products that ensure 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.

Most currently available resources therefore fell short in at least one feature that C-CoMP required from a microbe-centered team science data platform. These features are the ability to:

1. Host a standardized genome sequence that can be updated by team members with version control (e.g., modify gene annotations or ORF calls).
2. Iteratively add genome-linked data or metadata with version control (e.g., annotate genes with metabolic pathway membership or mutant strain availability).
3. Create, share, and modify intermediate analysis files of value to the research team (e.g., read-recruitment analyses or gene collections that group metabolically- or ecologically-related functions), including the ability to store and export information in universal formats accessible to other programs.
4. Analyze data using flexible workflows embedded in the platform (e.g., generate stoichiometric reaction networks for metabolic modeling or analyze the distribution of sequence variants in protein-coding genes).

We addressed these requirements using a 'Digital Microbes' solution. These are decentralized and data-centric bioinformatic structures for 'omics analysis and information exchange. Fundamentally, the Digital Microbe framework is a data packaging and sharing strategy for curated genome-linked information that supports collaborative research, with diverse options for implementation and exchange among scientists. Here, we describe Digital Microbe data packages for each of C-CoMP's model bacteria: one that 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); and one that creates an *Alteromonas* pangenome by merging data from 339 genomes (both isolates and metagenome-assembled genomes; DOI: 10.5281/zenodo.7430118). These taxa serve as laboratory models for identifying mechanisms underpinning microbial carbon cycling in the surface ocean.

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

Category	Examples	Limitations
Centralized databases (data upload/download)	NBCI RefSeq (O'Leary et al. 2016) DiatOmicBase (https://www.diatomicsbase.bio.ens.psl.eu/) ECMDB (Sajed et al. 2016) ProPortal (Kelly et al. 2012)	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 (Chen et al. 2023) KBase (Arkin et al. 2018) Galaxy (Galaxy Community 2022) PATRIC (Gillespie et al. 2011) CyVerse (Merchant et al. 2016) RAST (Aziz et al. 2008) SILVAngs (Quast et al. 2013) BV-BRC (Olson et al. 2023) Phycocosm (Grigoriev et al. 2021)	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 (Seemann 2014) KofamScan (Aramaki et al. 2020) BLAST+ (Camacho et al. 2009) GToTree (Lee 2019) Pathway Tools (Karp et al. 2015) DRAM (Shaffer et al. 2020)	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

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
95 a Digital Microbe data package with additional layers of information coming from new experiments). 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 are suitable for single microbial genomes as well as clade-level genomic collections. Further, the extensibility of Digital Microbes via the programmatic addition of new ‘omics data types make them future-proof.

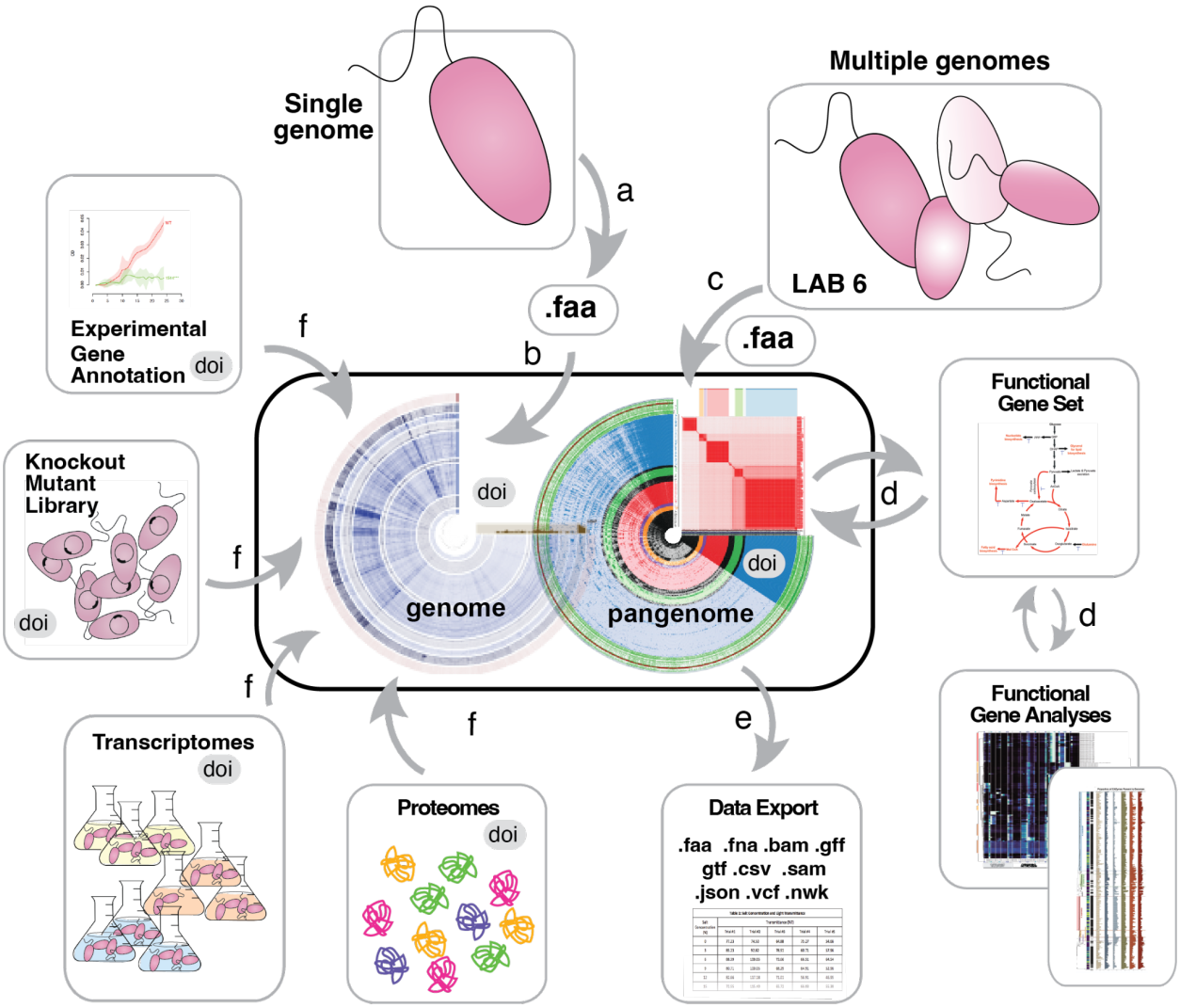


Figure 1. Architecture of a Digital Microbe. The genome of a model bacterium is (a) sequenced and (b) assembled and serves as the foundation of a Digital Microbe, a self-contained data package for a collaborative research team or a science community. (c) Alternatively, a pangenomic data package is assembled. (d) Intermediate datasets useful for downstream analyses are stored and reused (Fig. 4), 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-referenced data are added.

To create a Digital Microbe, we first established the 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 found that the open-source software platform anvi'o (Eren et al. 2021) fulfilled the four resource features we required with minimal modification (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 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 entire team or community.

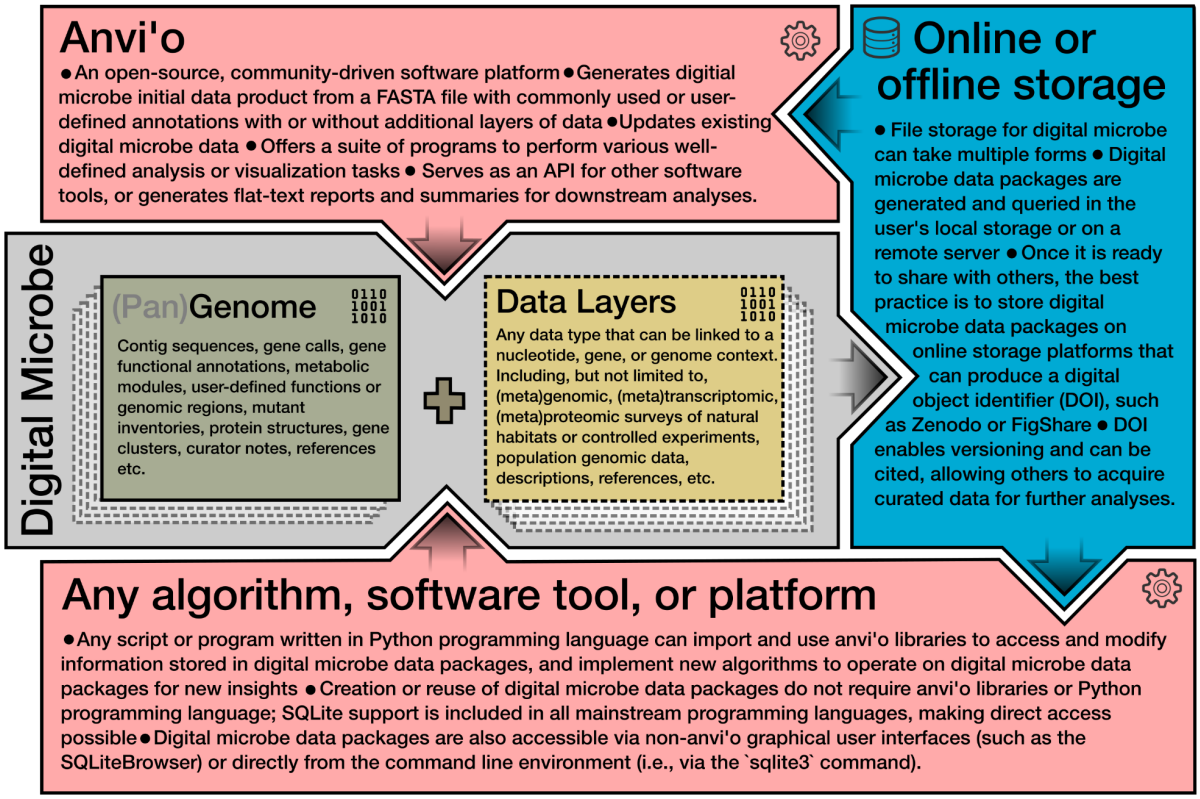


Figure 2. Situating the Digital Microbe concept in the existing computational environment. The Digital Microbe approach satisfies our four data science requirements by: establishing a version-controlled (pan)genomic reference; consolidating and cross-referencing collections of experimental and environmental data associated with a genome or pangenome; facilitating access to reusable intermediate analyses; and providing seamless data launch into a diversity of embedded analysis workflows. While

each requirement could be met by generating new software or using components of existing software, we found that the open-source software platform anvi'o (Eren et al. 2021) fulfilled feature requirements with minimal modification through a data storage solution in programmatically-queryable SQLite databases, interactive visualization capabilities, and integration with interoperable analysis programs enabling flexible workflows. 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 (Newton et al. 2010) with its members among the most metabolically active bacterial cells in algal blooms and coastal environments (Munson-McGee et al. 2022). *R. pomeroyi* has been well studied in the laboratory and field (Fu et al. 2020; Nowinski and Moran 2021; Schreier et al. 2023); it grows well in both defined and rich media; and it is amenable to genetic alteration (Howard et al. 2006; Mejia et al. 2022).

The *R. pomeroyi* Digital Microbe (Figure 3) is built on a well-curated genome assembly (*required feature 1*) first annotated in 2004 (Moran et al. 2004), reannotated in 2014 (Rivers, Smith, and Moran 2014), and enhanced with information from NCBI Clusters of Orthologous Groups (COG) (Galperin et al. 2021), Pfam (Mistry et al. 2021), and KEGG Kofam (Aramaki et al. 2020). The Digital Microbe annotation is also continually updated with new experimental verifications of *R. pomeroyi* genes (e.g., (Christie-Oleza, Miotello, and Armengaud 2012; Bullock et al. 2014; Rivers, Smith, and Moran 2014; Cunliffe 2016; Sharpe, Gifford, and Septer 2020) that have not been captured in genome annotations using standardized pipelines (e.g., RefSeq GCF_000011965.2). The *R. pomeroyi* DSS-3 Digital Microbe is available on Zenodo (Veseli and Cooper 2022).

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

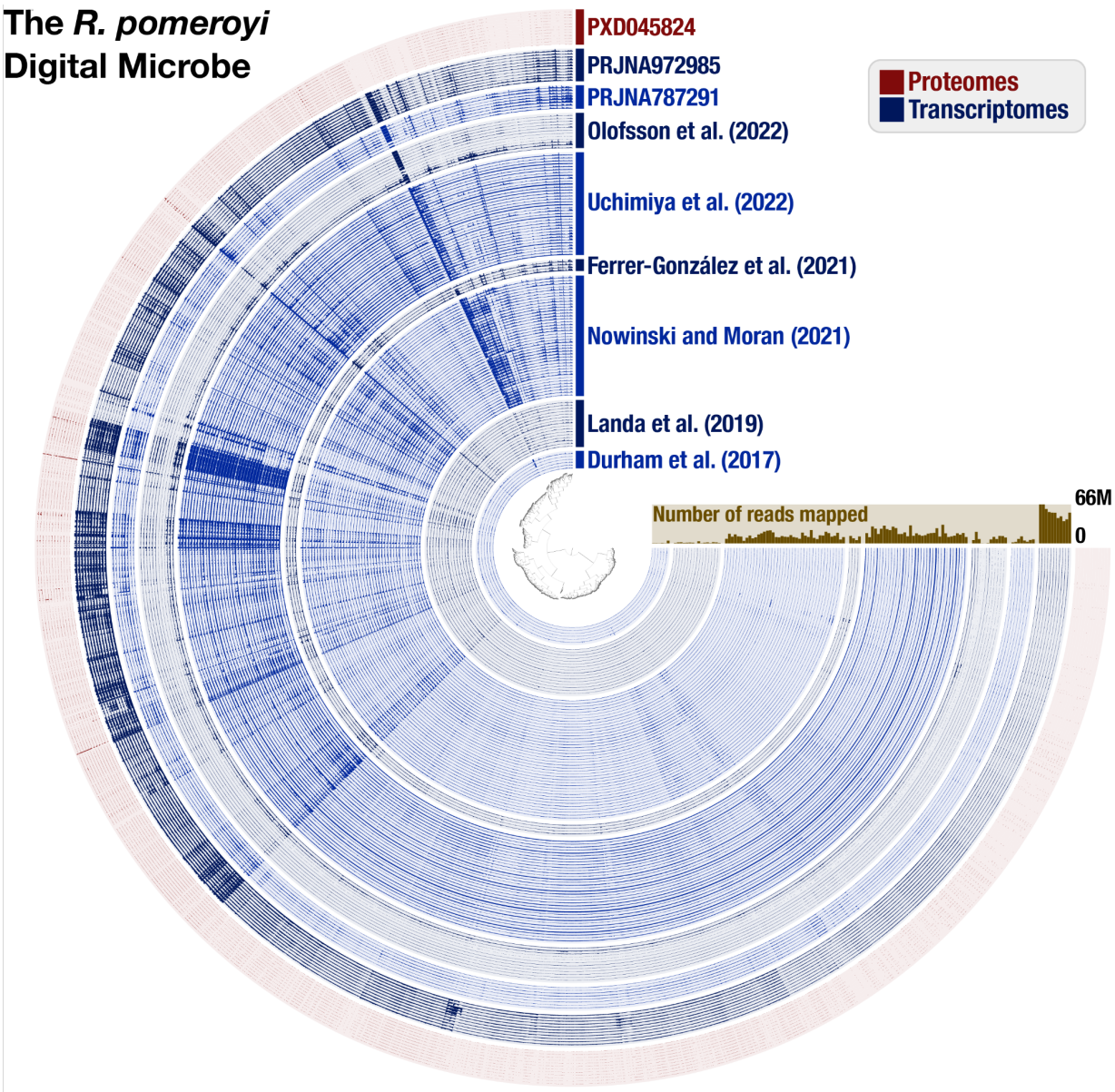


Figure 3. **Contents of the *R. pomeroyi* Digital Microbe**. As visualized in anvi'o 'gene mode', each item on the inner tree corresponds to one gene call in the *R. pomeroyi* genome, and the blue concentric circles display the coverage of each gene in a given transcriptome sample. Normalized protein abundances from proteome samples are shown in the red layers. Samples are grouped by their study of origin, with the data source indicated in text of the same color as the samples. The outermost group of transcriptomes correspond to the proteomes (yet unpublished but available on NCBI via the BioProject PRJNA972985). 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.

175 (Schroer et al. 2023) which were added to the Digital Microbe (*required 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
180 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
185 genome as individual data layers (Figure 3, Supplementary Table 1) (*required feature 2*). Using the anvi'o interactive interface, we established a custom dataset that consisted of the 62 protein components of the 18 experimentally annotated transporters (*required feature 3*). 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
190 expression, we extracted the data from anvi'o and then processed and visualized it using python (Figure 4) (*required feature 4*) (see https://github.com/C-CoMP-STC/digital-microbe/blob/main/rpom/rpom_dig_micro_transporter_expression_use_case.ipynb).

This meta-analysis captured responses by *R. pomeroyi* to available substrates under 43
195 different ecological conditions (Figure 4), including during co-culture growth with phytoplankton (Landa et al. 2017; Ferrer-González et al. 2023, 2021; Olofsson et al. 2022), on defined single or mixed substrates (Bullock et al. 2014), and after introduction into a natural phytoplankton bloom (Nowinski and Moran 2021). At the broadest scale, the transporters enabling organic acid uptake (acetate, citrate, fumarate, and 3-hydroxybutyrate) had the highest relative expression
200 across conditions, together accounting for an average of 48%, ranging from 9.7 – 86%, of the transcripts for transporters with confirmed substrates. Recent studies have indeed discovered that Roseobacteraceae members are organic acid catabolic specialists (Gralka, Pollak, and Cordero 2023; Forchielli, Sher, and Segrè 2022), and when *R. pomeroyi* is offered an organic acid/sugar mixture for growth, it first utilizes the acid fully (Cooper et al., unpublished).
205 Transporter transcription patterns also revealed the differences in substrate availability across environments. Introduced into a natural dinoflagellate bloom (Nowinski and Moran 2021), the

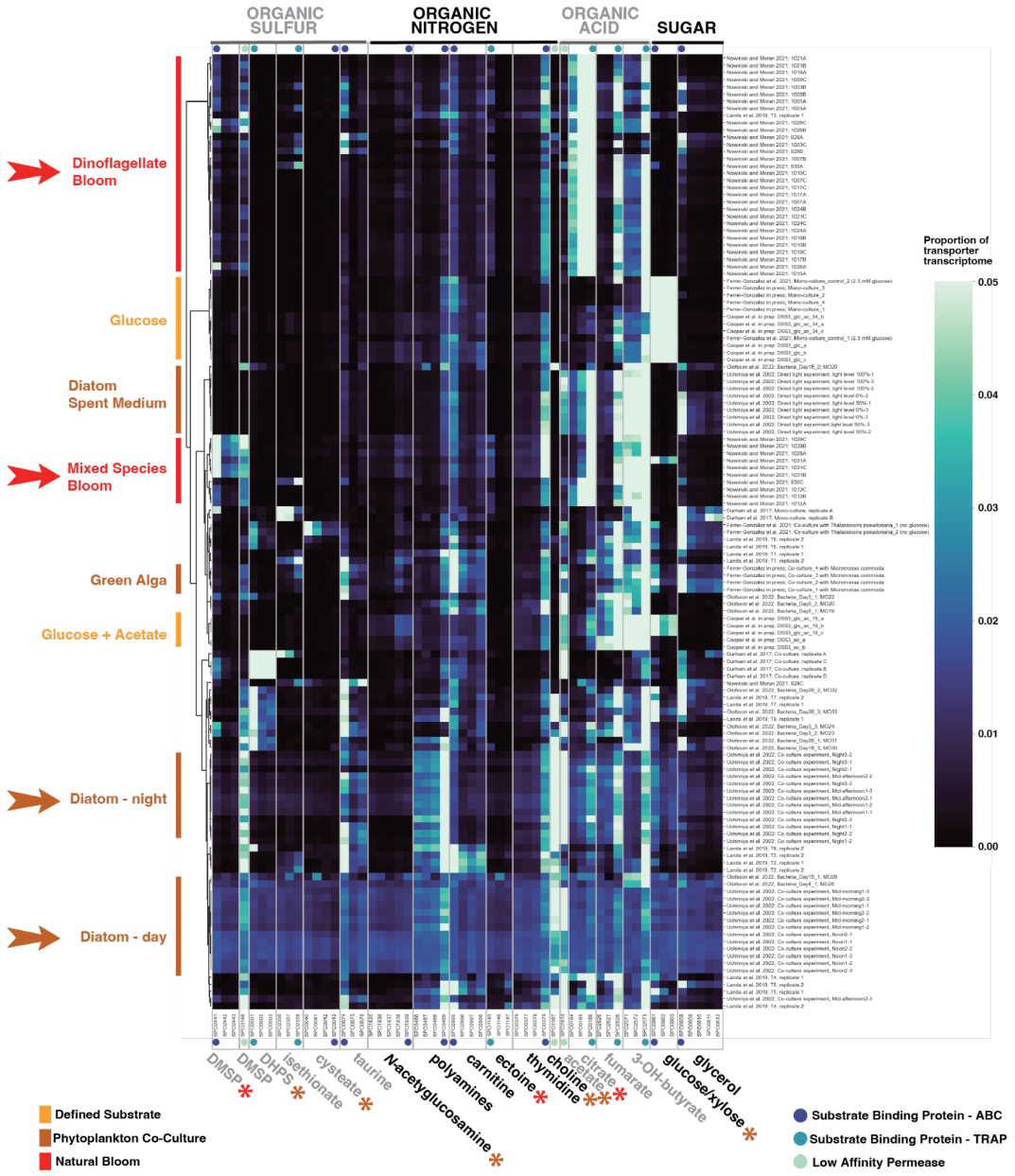


Figure 4. Clustered heatmap of relative gene expression for 18 experimentally annotated *R. pomeroyi* transporters compiled in a Digital Microbe. Each row represents a single transcriptome from the Digital Microbe dataset, and each column represents all transporter proteins with experimentally confirmed cognate substrates, likewise annotated in the Digital Microbe. Row labels indicate study and sample name (Supplementary Table 1). 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 to transcriptomes collected when substrates were derived from dinoflagellates (red) or diatoms (brown); significant differences 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 (Schroer et al. 2023) was expressed across most growth conditions, yet this metabolite has not previously been identified as a relevant currency in bacterially-driven 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; (Schroer et al. 2023; Mejia et al. 2022), proteomic data collected concurrently with one of the transcriptomic studies (Cooper et al., unpublished), and TnSeq mutant fitness measurements in synthetic microbial communities (Schreier et al. 2023) (*required feature 2*), and 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 (Hennon et al. 2017; López-Pérez et al. 2012; Henríquez-Castillo et al. 2022). Bacteria in this genus are distinguished by genomes encoding an average of 4,000 genes that enable use of a broad spectrum of substrates (Mikhailov, Romanenko, and Ivanova 2006), provide protection from reactive oxygen species to community members (Hennon et al. 2017), and mediate polysaccharide degradation (Koch et al. 2019). The type species of the genus is *Alteromonas macleodii* (Baumann et al. 1972; López-Pérez et al. 2012), with other notable species including *A. mediterranea* (Ivanova et al. 2015), *A. australica* (Ivanova et al. 2013), and *A. stellipolaris* (Van Trappen et al. 2004). 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 (López-Pérez and Rodriguez-Valera 2016).

The assembled pangenome consists of 336 *Alteromonas* genomes with genes called and annotated using one standardized pipeline *(required feature 1)*. Of these, 78 are isolate genomes (Chen et al. 2023; Mukherjee et al. 2023; Sayers et al. 2021) and 258 are metagenome-assembled genomes (MAGs) obtained from a variety of marine environments in the global ocean (Paoli et al. 2022). Genomes represent members of the closely related 'surface' species *A. macleodii* (n=139) and 'deep' species *A. mediterranea* (n=25) (Ivanova et al. 2015), and the widely distributed *A. australica* (n=63) (López-Pérez et al. 2014). The 34,390 gene clusters of the pangenome are linked to an imported phylogenetic tree assembled from single-copy core genes (see Methods), are annotated using NCBI COGs (Galperin et al. 2021), KEGG KOfams (Aramaki et al. 2020), CAZyme HMMs (Yin et al. 2012) and orthology predictions from EggNOG-mapper (Cantalapiedra et al. 2021; Huerta-Cepas et al. 2019; Buchfink, Reuter, and Drost 2021) (r*equired feature 2*), and are assigned as core or accessory genes for the genus (*required feature 3*) based on a Bayesian approach available in anvi'o (Buck, Mehrshad, and Bertilsson 2022) (*required feature 4*). The *Alteromonas* Digital Microbe with relevant (pan)genome and phylogeny files is accessible on Zenodo (DeMers and Braakman 2023). Examples of future versioned enhancements of this Digital Microbe might include additions of new *Alteromonas* genomes and improved (pan)genome annotations from culture studies and novel annotation programs.

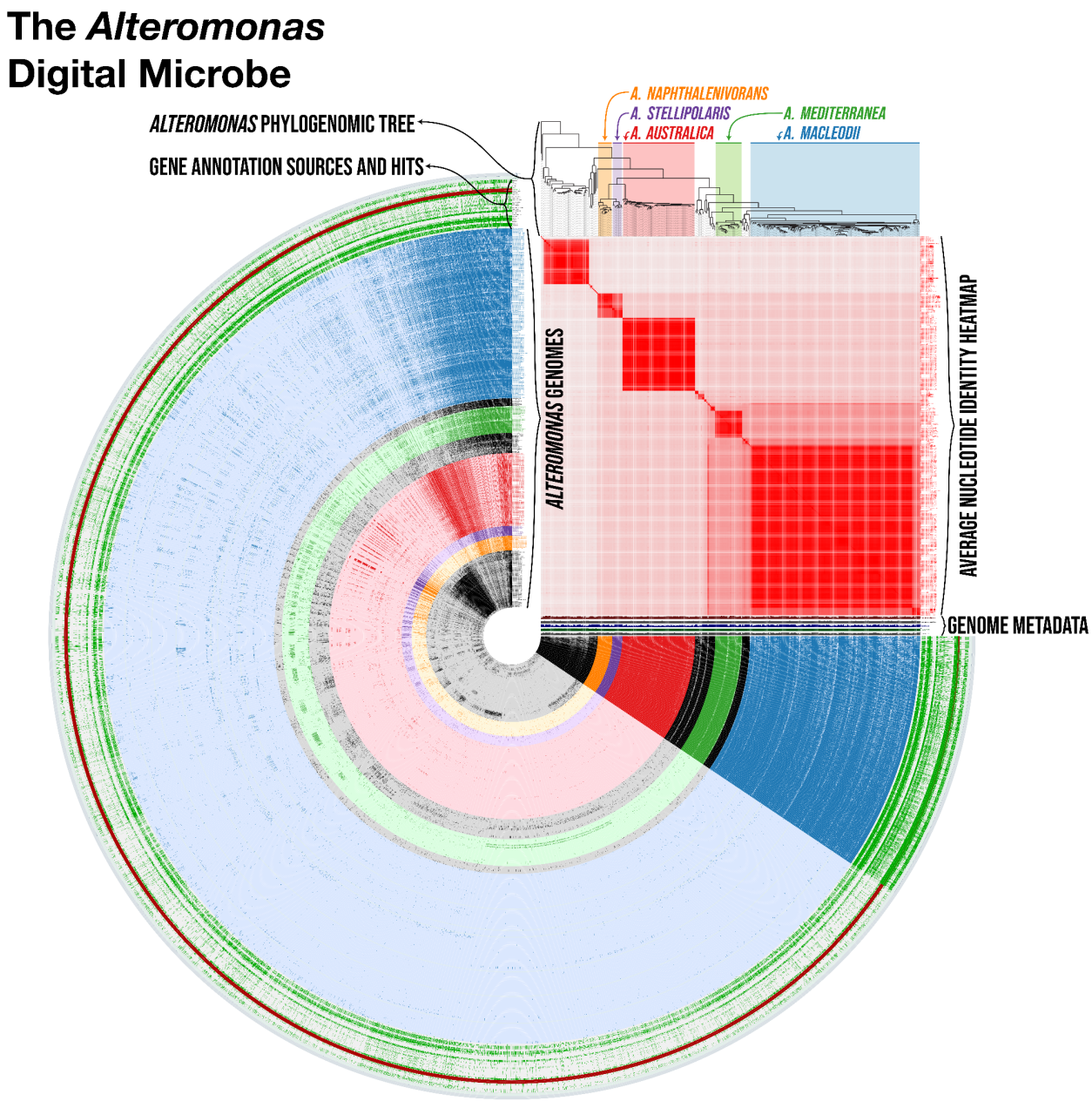


Figure 5. The Alteromonas Digital Microbe. Each concentric ring represents one Alteromonas genome, with colored rings identifying genomes form five clades of interest (*A. macleodii*, *A. mediterranea*, *A. austalica*, *A. stellioolaris*, and *A. naphthalenivorans*). The outermost green rings depict annotation sources added for all genomes in the digital microbe. 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*

285 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 and amino sugars
290 (Mikhailov, Romanenko, and Ivanova 2006). Sugars and sugar polymers are an abundant and diverse component of the ocean's dissolved organic carbon inventory (Repeta 2015), 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 (Yin et al. 2012) was recently added to
295 anvi'o, bringing this analysis within easy reach (*required feature 4*). We therefore augmented the Digital Microbe with CAZyme annotations to explore the evolutionary patterns of these proteins. The results indicate distinct CAZymes 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. mediterranea* and *A. stellipolaris* clades have larger total
300 CAZyme inventories that suggest highly diverse carbohydrate utilization capabilities. As patterns of diversity in CAZyme inventories are most distinct at the clade level as compared to the within-clade level, carbohydrate utilization emerges as a potentially key driver of the large-scale niche partitioning of *Alteromonas* species.

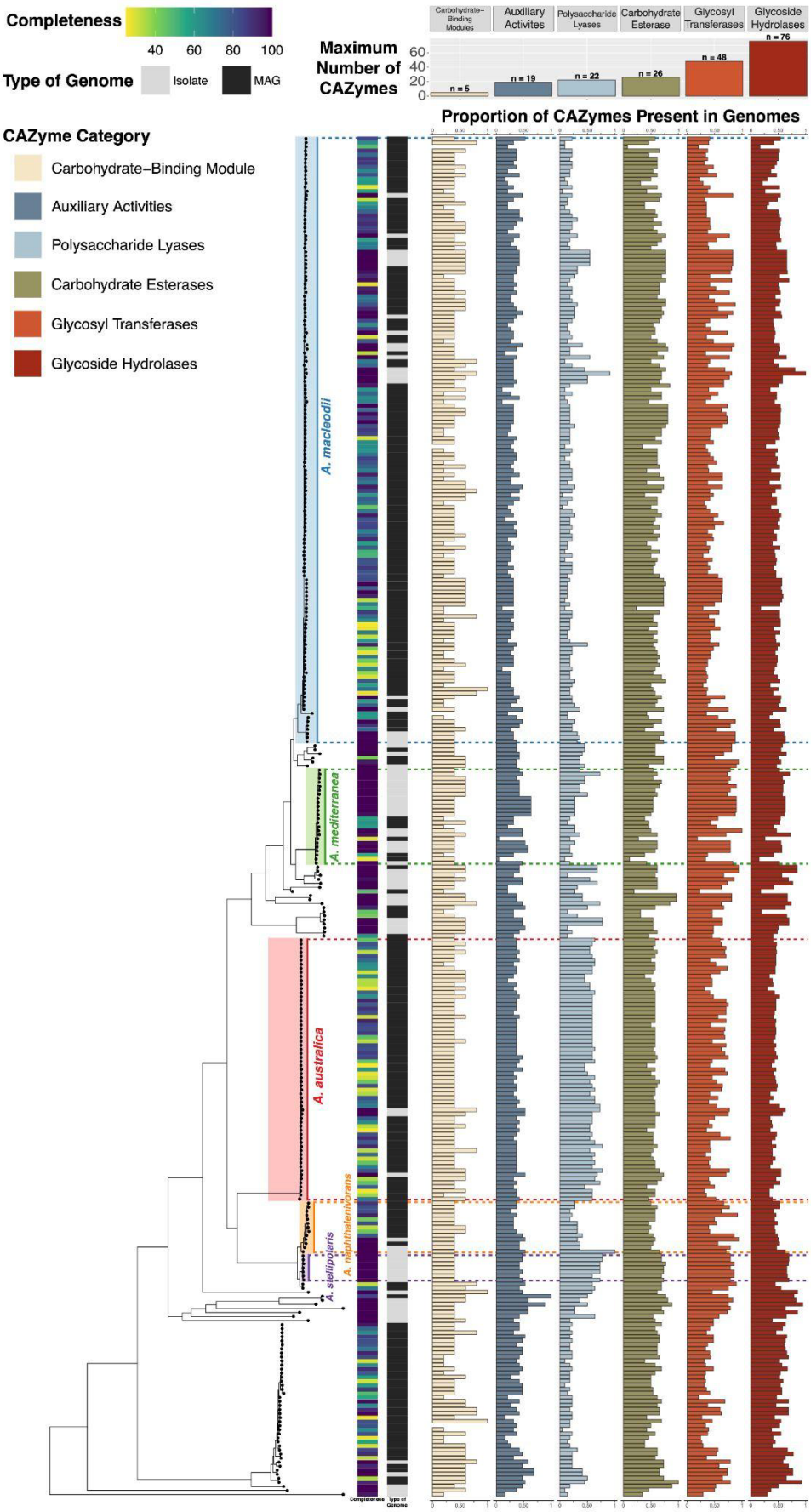

Completeness
40
60
80
100
Type of Genome
Isolate
MAG
Maximum Number of CAZymes
60
40
20
0
Carbohydrate-Binding Modules
Auxiliary Activites
Polysaccharide Lyases
Carbohydrate Esterase
Glycosyl Transferases
Glycoside Hydrolases
n = 5
n = 19
n = 22
n = 26
n = 48
n = 76
Proportion of CAZymes Present in Genomes
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.

Conclusion

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 (Kelly et al. 2012). 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 (Groussman et al. 2023). Another enhancement under ongoing 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 adopts digital microbes for their work. Organized and democratized access to taxon-explicit data is emerging as a cornerstone of modern microbiology.

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. 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.

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. ZSC, CBS, WFS, and MAM acknowledge support by Simons Foundation Grant 542391 within the Principles of Microbial Ecosystems Collaborative. This is C-CoMP publication #024.

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.

Methods

Both Digital Microbes were generated and analyzed using v7.1-dev or later versions of anvi'o (Eren et al. 2021).

365 **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 (Moran et al. 2004) and its associated (meta)transcriptome samples (Durham et al. 2015; Landa et al. 2019; Ferrer-González et al. 2021; Nowinski and Moran 2021; Olofsson et al. 2022; Uchimiya et al. 2022; Cooper et al., unpublished). We generated a
370 contigs database from the genome and megaplasmid sequences with `anvi-gen-contigs-database` and annotated the resulting Prodigal (Hyatt et al. 2010) gene calls with *de novo* annotations from NCBI Clusters of Orthologous Genes (COGs) (Galperin et al. 2021), KEGG KOfams (Aramaki et al. 2020), and Pfam (Mistry et al. 2021) via the associated anvi'o program for each annotation source. We also identified single-copy core genes using `anvi-run-hmms` and
375 associated these genes with taxonomy data from the GTDB (Parks et al. 2022) using `anvi-run-scg-taxonomy`. We imported manually curated gene annotations, including annotations indicating which genes have available TnSeq mutants (Mejia et al. 2022), using the program `anvi-import-functions`.

380 To process the (meta)transcriptomes, these samples were quality-filtered using FASTX-toolkit (Gordon and Hannon 2010) with the parameters described in (Landa et al. 2017). We mapped the reads to the DSS-3 genome using Bowtie 2 (Langmead and Salzberg 2012) and samtools (Danecek et al. 2021). Each sample's read mapping data was converted into an anvi'o profile database using `anvi-profile`, and all samples were merged into one database with `anvi-merge`.
385 To add proteomic data (Cooper et al., unpublished), we normalized spectral abundance counts with a normalized spectral abundance factor to make these 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
390 `anvi-import-misc-data`. We also used this program to import fitness data associated with gene mutants from (Schreier et al. 2023) into the same genes database.

Transporter analysis for *Ruegeria pomeroyi*

The genes database in the *R. pomeroyi* Digital Microbe contains coverage information for each gene in each sample associated with it. To assess the proportional expression of substrate-
395 confirmed transporter genes, we used the anvi'o interactive interface. There, we created a bin containing these 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 ('split')

where our genes of interest were located. After reading the exported data into dataframes using python v3.7.8 and pandas (McKinney 2010; The pandas development team 2023), we extracted the coverage data for our specific genes of interest, normalized the coverages to TPM using the total number of reads for each sample, and then relativized these data to represent the proportional expression of each gene. We visualized these data as a clustermap using the seaborn package (Waskom 2021) and assessed statistical differences in the mean gene expression using the a t-test implemented in the scipy stats package (Virtanen et al. 2020).

Creation of the Alteromonas Digital Microbe

To create the *Alteromonas* Digital Microbe, we collected a total of 78 isolate genomes and 258 MAGs from the Joint Genome Institute's Integrated Microbial Genomes (IMG) project (Chen et al. 2023), NCBI (Sayers et al. 2022), and (Paoli et al. 2022). 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 (Parks et al. 2015); for the MAGs that were taken from (Paoli et al. 2022), 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) (Galperin et al. 2021), KEGG KOfams (Aramaki et al. 2020), and Carbohydrate-Active enZYme (CAZyme) HMMs (Yin et al. 2012) via the associated anvi'o program for each annotation source, and we imported externally-run annotations from EggNOG-mapper (Cantalapiedra et al. 2021; Huerta-Cepas et al. 2019; Buchfink, Reuter, and Drost 2021) and KEGG GhostKOALA (Kanehisa, Sato, and Morishima 2016) 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 (Buck, Mehrshad, and Bertilsson 2022) 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 (Stamatakis 2014), and

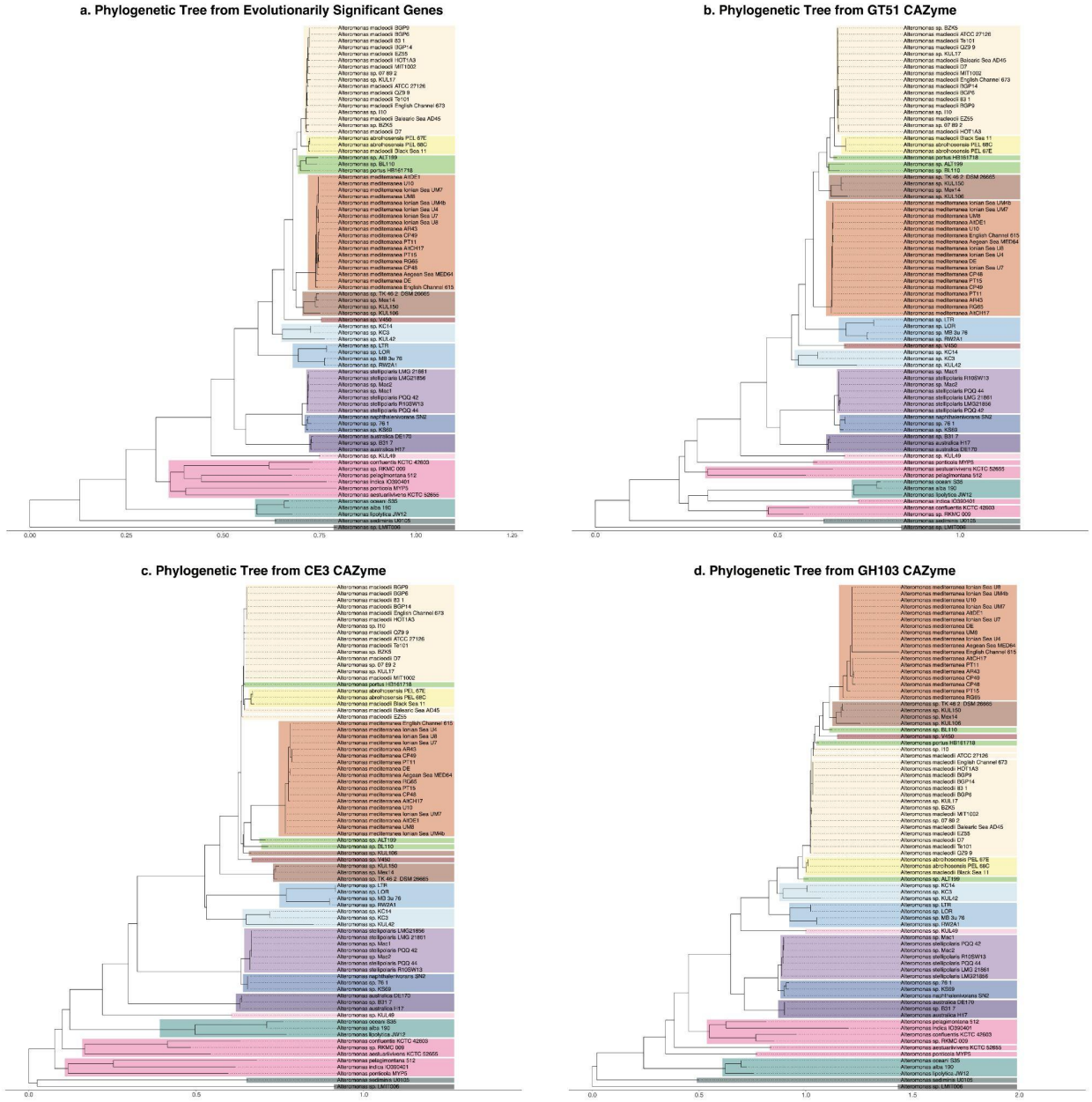
430 built the tree under the “PROTGAMMAAUTO” model setting. We used FigTree v1.4.4 (Rambaut n.d.) to midpoint root the tree and save it in newick file format. To incorporate the tree into the pangenome, we imported the newick tree into the pangenome database with the program `anvi-import-misc-data`. 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-
435 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 v4.1.1 (R Core Team 2018) using the packages aplot v0.1.9 (Yu 2023), BiocManager v1.30.20 (Morgan 2023), dplyr v1.1.0 (Wickham et al. 2023), ggnewscale v0.4.8 (Campitelli 2022), ggplot2 v3.4.1 (Wickham 2016), ggstance v0.3.6.9000 (Henry, Wickham, and Chang 2022),
440 ggtree v3.7.1.003 (Yu 2022; Xu et al. 2022; Yu 2020; Yu et al. 2018, 2017), ggtreeExtra v1.9.1.992 (Yu 2022; Xu et al. 2021), nationalparkcolors v0.1.0 (Jolly n.d.), plyr v1.8.8, RColorBrewer v1.1-3 (Neuwirth 2022), scales v1.2.1 (Wickham and Seidel 2022), and tidyr v1.3.0 (Wickham, Vaughan, and Girlich 2023).

We then repeated the initial steps above to generate a phylogeny for the subset of isolate
445 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 8 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,
450 with the packages listed above in addition to colorBlindness v0.1.9 (Ou 2021), easyalluvial v0.3.1 (Oettinghaus 2022), and gridExtra v2.3 (Auguie 2017).

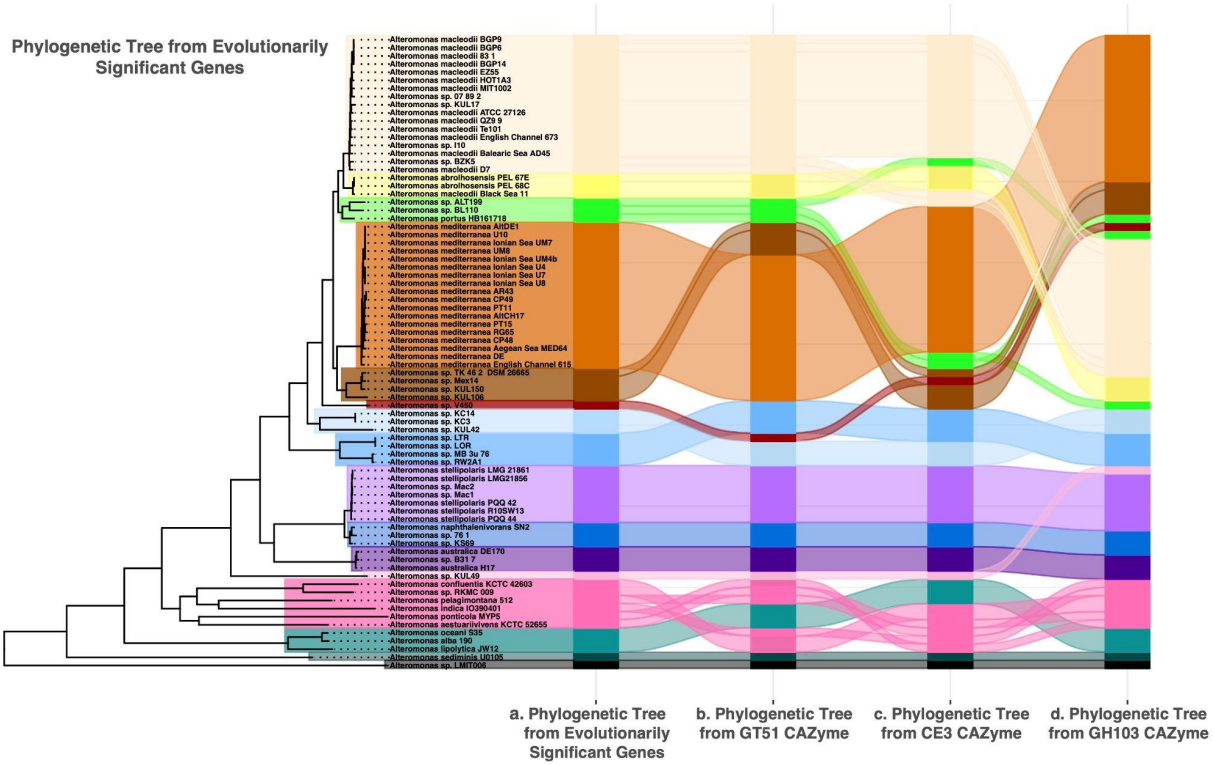
Supplementary Tables

Supplementary Table 1. 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.



465 **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
470 plot displays the movement of clades within the tree. Further branch definition of each tree is shown in Supplementary Figure 1.

References

Aramaki, Takuya, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. 2020. "KofamKOALA: KEGG Ortholog Assignment
475 Based on Profile HMM and Adaptive Score Threshold." *Bioinformatics* 36 (7): 2251–52.

Arkin, Adam P., Robert W. Cottingham, Christopher S. Henry, Nomi L. Harris, Rick L. Stevens, Sergei Maslov, Paramvir Dehal, et al. 2018. "KBase: The United States Department of Energy Systems Biology Knowledgebase." *Nature Biotechnology* 36 (7): 566–69.

Auguie, Baptiste. 2017. *Miscellaneous Functions for "Grid" Graphics [R Package gridExtra*
480 *Version 2.3]*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=gridExtra.

Aziz, Ramy K., Daniela Bartels, Aaron A. Best, Matthew DeJongh, Terrence Disz, Robert A. Edwards, Kevin Formsma, et al. 2008. "The RAST Server: Rapid Annotations Using

Subsystems Technology." *BMC Genomics* 9 (February): 75.
Baumann, Linda, Paul Baumann, M. Mandel, and Richard D. Allen. 1972. "Taxonomy of Aerobic Marine Eubacteria." *Journal of Bacteriology* 110 (1): 402–29.
Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. "Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND." *Nature Methods* 18 (4): 366–68.
Buck, Moritz, Maliheh Mehrshad, and Stefan Bertilsson. 2022. "mOTUpan: A Robust Bayesian Approach to Leverage Metagenome-Assembled Genomes for Core-Genome Estimation." *NAR Genomics and Bioinformatics* 4 (3): lqac060.
Bullock, Hannah A., Chris R. Reisch, Andrew S. Burns, Mary Ann Moran, and William B. Whitman. 2014. "Regulatory and Functional Diversity of Methylmercaptopropionate Coenzyme A Ligases from the Dimethylsulfoniopropionate Demethylation Pathway in Ruegeria Pomeroyi DSS-3 and Other Proteobacteria." *Journal of Bacteriology* 196 (6): 1275–85.
Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (December): 421.
Campitelli, Elio. 2022. *Multiple Fill and Colour Scales in "ggplot2" [R Package Ggnewscale Version 0.4.8]* (version 0.4.8). Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=ggnewscale.
Cantalapiedra, Carlos P., Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021. "eggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale." *Molecular Biology and Evolution* 38 (12): 5825–29.
Chen, I-Min A., Ken Chu, Krishnaveni Palaniappan, Anna Ratner, Jinghua Huang, Marcel Huntemann, Patrick Hajek, et al. 2023. "The IMG/M Data Management and Analysis System v.7: Content Updates and New Features." *Nucleic Acids Research* 51 (D1): D723–32.
Christie-Oleza, Joseph A., Guylaine Miotello, and Jean Armengaud. 2012. "High-Throughput Proteogenomics of Ruegeria Pomeroyi: Seeding a Better Genomic Annotation for the Whole Marine Roseobacter Clade." *BMC Genomics* 13 (February): 73.
Cooper, Z.S., Schroer, W.F., Garfield, C.M., Scott, H., Yang, M., Smith. C.B., Uchimiya, M., Lopez, P.Z., McIlvin, M., Saito, M., Covert, M., Segrè, D., Moran, M.A. 2024. "Dynamic metabolic efficiency of substrate utilization by copiotrophic marine bacteria". *Manuscript in preparation.*
Cunliffe, Michael. 2016. "Purine Catabolic Pathway Revealed by Transcriptomics in the Model Marine Bacterium Ruegeria Pomeroyi DSS-3." *FEMS Microbiology Ecology* 92 (1). https://doi.org/10.1093/femsec/fiv150.
Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab008.
DeMers, Michelle, and Rogier Braakman. 2023. "Alteromonas Digital Organism Databases." https://doi.org/10.5281/zenodo.7430118.
Durham, Bryndan P., Shalabh Sharma, Haiwei Luo, Christa B. Smith, Shady A. Amin, Sara J. Bender, Stephen P. Dearth, et al. 2015. "Cryptic Carbon and Sulfur Cycling between Surface Ocean Plankton." *Proceedings of the National Academy of Sciences of the United States of America* 112 (2): 453–57.
Eren, A. Murat, Evan Kiefl, Alon Shaiber, Iva Veseli, Samuel E. Miller, Matthew S. Schechter, Isaac Fink, et al. 2021. "Community-Led, Integrated, Reproducible Multi-Omics with Anvi'o." *Nature Microbiology* 6 (1): 3–6.
Ferrer-González, Frank X., Maria Hamilton, Christa B. Smith, Jeremy E. Schreier, Malin Olofsson, and Mary Ann Moran. 2023. "Bacterial Transcriptional Response to Labile

535 Exometabolites from Photosynthetic Picoeukaryote Micromonas Commoda." *ISME Communications* 3 (1): 5.
Ferrer-González, Frank X., Brittany Widner, Nicole R. Holderman, John Glushka, Arthur S. Edison, Elizabeth B. Kujawinski, and Mary Ann Moran. 2021. "Resource Partitioning of Phytoplankton Metabolites That Support Bacterial Heterotrophy." *The ISME Journal* 15 (3):
540 762–73.
Forchielli, Elena, Daniel Sher, and Daniel Segrè. 2022. "Metabolic Phenotyping of Marine Heterotrophs on Refactored Media Reveals Diverse Metabolic Adaptations and Lifestyle Strategies." *mSystems* 7 (4): e0007022.
Fu, He, Mario Uchimiya, Jeff Gore, and Mary Ann Moran. 2020. "Ecological Drivers of Bacterial
545 Community Assembly in Synthetic Phycospheres." *Proceedings of the National Academy of Sciences of the United States of America* 117 (7): 3656–62.
Galaxy Community. 2022. "The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2022 Update." *Nucleic Acids Research* 50 (W1): W345–51.
Galperin, Michael Y., Yuri I. Wolf, Kira S. Makarova, Roberto Vera Alvarez, David Landsman,
550 and Eugene V. Koonin. 2021. "COG Database Update: Focus on Microbial Diversity, Model Organisms, and Widespread Pathogens." *Nucleic Acids Research* 49 (D1): D274–81.
Gillespie, Joseph J., Alice R. Wattam, Stephen A. Cammer, Joseph L. Gabbard, Maulik P. Shukla, Oral Dalay, Timothy Driscoll, et al. 2011. "PATRIC: The Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species." *Infection and*
555 *Immunity* 79 (11): 4286–98.
González, José M., Beatriz Fernández-Gómez, Antoni Fernàndez-Guerra, Laura Gómez-Consarnau, Olga Sánchez, Montserrat Coll-Lladó, Javier Del Campo, et al. 2008. "Genome Analysis of the Proteorhodopsin-Containing Marine Bacterium *Polaribacter* Sp. MED152 (Flavobacteria)." *Proceedings of the National Academy of Sciences of the United States of*
560 *America* 105 (25): 8724–29.
Gordon, A., and G. J. Hannon. 2010. *Fastx-Toolkit* (version 0.0.13). http://hannonlab.cshl.edu/fastx_toolkit/.
Gralka, Matti, Shaul Pollak, and Otto X. Cordero. 2023. "Genome Content Predicts the Carbon Catabolic Preferences of Heterotrophic Bacteria." *Nature Microbiology* 8 (10): 1799–1808.
565 Grigoriev, Igor V., Richard D. Hayes, Sara Calhoun, Bishoy Kamel, Alice Wang, Steven Ahrendt, Sergey Dusheyko, et al. 2021. "PhycoCosm, a Comparative Algal Genomics Resource." *Nucleic Acids Research* 49 (D1): D1004–11.
Groussman, R. D., S. Blaskowski, S. N. Coesel, and E. V. Armbrust. 2023. "MarFERReT, an Open-Source, Version-Controlled Reference Library of Marine Microbial Eukaryote
570 Functional Genes." *Scientific Data* 10 (1): 926.
Hennon, Gwenn M. M., J. Jeffrey Morris, Sheean T. Haley, Erik R. Zinser, Alexander R. Durrant, Elizabeth Entwistle, Terje Dokland, and Sonya T. Dyhrman. 2017. "The Impact of Elevated CO2 on Prochlorococcus and Microbial Interactions with 'Helper' Bacterium Alteromonas." *The ISME Journal* 12 (2): 520–31.
575 Henríquez-Castillo, Carlos, Alvaro M. Plominsky, Salvador Ramírez-Flandes, Anthony D. Bertagnolli, Frank J. Stewart, and Osvaldo Ulloa. 2022. "Metaomics Unveils the Contribution of Alteromonas Bacteria to Carbon Cycling in Marine Oxygen Minimum Zones." *Frontiers in Marine Science* 9. https://doi.org/10.3389/fmars.2022.993667.
Henry, Lionel, Hadley Wickham, and Winston Chang. 2022. *Ggstance: Horizontal ggplot2*
580 *Components* (version 0.3.6.9). Github. https://github.com/lionel-/ggstance.
Howard, Erinn C., James R. Henriksen, Alison Buchan, Chris R. Reisch, Helmut Bürgmann, Rory Welsh, Wenying Ye, et al. 2006. "Bacterial Taxa That Limit Sulfur Flux from the Ocean." *Science* 314 (5799): 649–52.
Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K.
585 Forslund, Helen Cook, Daniel R. Mende, et al. 2019. "eggNOG 5.0: A Hierarchical,

Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses." *Nucleic Acids Research* 47 (D1): D309–14.
Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (March): 119.
Ivanova, Elena P., Mario López-Pérez, Mila Zabalos, Song Ha Nguyen, Hayden K. Webb, Jason Ryan, Kiril Lagutin, Mikhail Vyssotski, Russell J. Crawford, and Francisco Rodriguez-Valera. 2015. "Ecophysiological Diversity of a Novel Member of the Genus Alteromonas, and Description of Alteromonas Mediterranea Sp. Nov." *Antonie van Leeuwenhoek* 107 (1): 119–32.
Ivanova, Elena P., Hooi Jun Ng, Hayden K. Webb, Valeriya V. Kurilenko, Natalia V. Zhukova, Valery V. Mikhailov, Olga N. Ponamoreva, and Russell J. Crawford. 2013. "Alteromonas Australica Sp. Nov., Isolated from the Tasman Sea." *Antonie van Leeuwenhoek* 103 (4): 877–84.
Jolly, Katie. n.d. *Nationalparkcolors: R Package for Color Themes Inspired by National Parks* (version 0.1.0). Github. Accessed May 2, 2023. https://github.com/katiejolly/nationalparkcolors.
Kanehisa, Minoru, Yoko Sato, and Kanae Morishima. 2016. "BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences." *Journal of Molecular Biology* 428 (4): 726–31.
Karp, Peter D., Suzanne M. Paley, Peter E. Midford, Markus Krummenacker, Richard Billington, Anamika Kothari, Wai Kit Ong, Pallavi Subhraveti, Ingrid M. Keseler, and Ron Caspi. 2015. "Pathway Tools Version 24.0: Integrated Software for Pathway/genome Informatics and Systems Biology." *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1510.03964.
Kelly, Libusha, Katherine H. Huang, Huiming Ding, and Sallie W. Chisholm. 2012. "ProPortal: A Resource for Integrated Systems Biology of Prochlorococcus and Its Phage." *Nucleic Acids Research* 40 (Database issue): D632–40.
Koch, Hanna, Alexandra Dürwald, Thomas Schweder, Beatriz Noriega-Ortega, Silvia Vidal-Melgosa, Jan-Hendrik Hehemann, Thorsten Dittmar, et al. 2019. "Biphasic Cellular Adaptations and Ecological Implications of Alteromonas Macleodii Degrading a Mixture of Algal Polysaccharides." *The ISME Journal* 13 (1): 92–103.
Landa, Marine, Andrew S. Burns, Bryndan P. Durham, Kaitlin Esson, Brent Nowinski, Shalabh Sharma, Alexey Vorobev, Torben Nielsen, Ronald P. Kiene, and Mary Ann Moran. 2019. "Sulfur Metabolites That Facilitate Oceanic Phytoplankton–bacteria Carbon Flux." *The ISME Journal* 13 (10): 2536–50.
Landa, Marine, Andrew S. Burns, Selena J. Roth, and Mary Ann Moran. 2017. "Bacterial Transcriptome Remodeling during Sequential Co-Culture with a Marine Dinoflagellate and Diatom." *The ISME Journal* 11 (12): 2677–90.
Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.
Lee, Michael D. 2019. "GToTree: A User-Friendly Workflow for Phylogenomics." *Bioinformatics* 35 (20): 4162–64.
Leonelli, Sabina. 2013. "Model Organism." In *Encyclopedia of Systems Biology*, edited by Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, 1398–1401. New York, NY: Springer New York.
López-Pérez, Mario, Aitor Gonzaga, Elena P. Ivanova, and Francisco Rodriguez-Valera. 2014. "Genomes of Alteromonas Australica, a World Apart." *BMC Genomics* 15 (1): 483.
López-Pérez, Mario, Aitor Gonzaga, Ana-Belen Martin-Cuadrado, Olga Onyshchenko, Akbar Ghavidel, Rohit Ghai, and Francisco Rodriguez-Valera. 2012. "Genomes of Surface Isolates of Alteromonas Macleodii: The Life of a Widespread Marine Opportunistic Copiotroph." *Scientific Reports* 2 (September): 696.

López-Pérez, Mario, and Francisco Rodriguez-Valera. 2016. “Pangenome Evolution in the Marine Bacterium Alteromonas.” *Genome Biology and Evolution* 8 (5): 1556–70.
McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” In *Proceedings of the 9th Python in Science Conference*. SciPy. https://doi.org/10.25080/majora-92bf1922-00a.
Mejia, Catalina, Lidimarie Trujillo Rodriguez, Ravin Poudel, Adam Ellington, Adam R. Rivers, and Christopher R. Reisch. 2022. “An Arrayed Transposon Library of Ruegeria Pomeroyi DSS-3.” *bioRxiv*. https://doi.org/10.1101/2022.09.11.507510.
Merchant, Nirav, Eric Lyons, Stephen Goff, Matthew Vaughn, Doreen Ware, David Micklos, and Parker Antin. 2016. “The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences.” *PLoS Biology* 14 (1): e1002342.
Mikhailov, Valery V., Lyudmila A. Romanenko, and Elena P. Ivanova. 2006. “The Genus Alteromonas and Related Proteobacteria.” In *The Prokaryotes: A Handbook on the Biology of Bacteria Volume 6: Proteobacteria: Gamma Subclass*, edited by Martin Dworkin, Stanley Falkow, Eugene Rosenberg, Karl-Heinz Schleifer, and Erko Stackebrandt, 597–645. New York, NY: Springer New York.
Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, et al. 2021. “Pfam: The Protein Families Database in 2021.” *Nucleic Acids Research* 49 (D1): D412–19.
Moore, Lisa R., and Sallie W. Chisholm. 1999. “Photophysiology of the Marine Cyanobacterium Prochlorococcus: Ecotypic Differences among Cultured Isolates.” *Limnology and Oceanography* 44 (3): 628–38.
Moran, Mary Ann, Alison Buchan, José M. González, John F. Heidelberg, William B. Whitman, Ronald P. Kiene, James R. Henriksen, et al. 2004. “Genome Sequence of Silicibacter Pomeroyi Reveals Adaptations to the Marine Environment.” *Nature* 432 (7019): 910–13.
Morgan, Martin. 2023. *Access the Bioconductor Project Package Repository [R Package BiocManager Version 1.30.20]* (version 1.30.20). Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/BiocManager/index.html.
Mukherjee, Supratim, Dimitri Stamatis, Cindy Tianqing Li, Galina Ovchinnikova, Jon Bertsch, Jagadish Chandrabose Sundaramurthi, Mahathi Kandimalla, et al. 2023. “Twenty-Five Years of Genomes OnLine Database (GOLD): Data Updates and New Features in v.9.” *Nucleic Acids Research* 51 (D1): D957–63.
Munson-McGee, Jacob H., Melody R. Lindsay, Eva Sintes, Julia M. Brown, Timothy D’Angelo, Joe Brown, Laura C. Lubelczyk, et al. 2022. “Decoupling of Respiration Rates and Abundance in Marine Prokaryoplankton.” *Nature* 612 (7941): 764–70.
Neuwirth, Erich. 2022. *RColorBrewer: ColorBrewer Palettes. [R Package RColorBrewer Version 1.1-3]* (version 1.1-3). Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=RColorBrewer.
Newton, Ryan J., Laura E. Griffin, Kathy M. Bowles, Christof Meile, Scott Gifford, Carrie E. Givens, Erinn C. Howard, et al. 2010. “Genome Characteristics of a Generalist Marine Bacterial Lineage.” *The ISME Journal* 4 (6): 784–98.
Nowinski, Brent, and Mary Ann Moran. 2021. “Niche Dimensions of a Marine Bacterium Are Identified Using Invasion Studies in Coastal Seawater.” *Nature Microbiology* 6 (4): 524–32.
Oettinghaus, Björn. 2022. *Easyalluvial: Generate Alluvial Plots with a Single Line of Code [R Package Version 0.3.1]* (version 0.3.1). Github. https://github.com/erblast/easyalluvial.
O’Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. “Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation.” *Nucleic Acids Research* 44 (D1): D733–45.
Olofsson, Malin, Frank X. Ferrer-González, Mario Uchimiya, Jeremy E. Schreier, Nicole R. Holderman, Christa B. Smith, Arthur S. Edison, and Mary Ann Moran. 2022. “Growth-

Stage-Related Shifts in Diatom Endometabolome Composition Set the Stage for Bacterial Heterotrophy." *ISME Communications* 2 (1): 1–9.
Olson, Robert D., Rida Assaf, Thomas Brettin, Neal Conrad, Clark Cucinell, James J. Davis, Donald M. Dempsey, et al. 2023. "Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): A Resource Combining PATRIC, IRD and ViPR." *Nucleic Acids Research* 51 (D1): D678–89.
Ou, Jianhong. 2021. *Safe Color Set for Color Blindness [R Package colorBlindness Version 0.1.9]* (version 0.1.19). Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=colorBlindness.
Paoli, Lucas, Hans-Joachim Ruscheweyh, Clarissa C. Forneris, Florian Hubrich, Satria Kautsar, Agneya Bhushan, Alessandro Lotti, et al. 2022. "Biosynthetic Potential of the Global Ocean Microbiome." *Nature* 607 (7917): 111–18.
Parks, Donovan H., Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2022. "GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy." *Nucleic Acids Research* 50 (D1): D785–94.
Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. "CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes." *Genome Research* 25 (7): 1043–55.
Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (Database issue): D590–96.
Rambaut, Andrew. n.d. *FigTree* (version 1.4.4). Accessed May 2, 2023. http://tree.bio.ed.ac.uk/software/figtree/.
R Core Team. 2018. *R: A Language and Environment for Statistical Computing* (version 4.1.1). R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Repeta, Daniel J. 2015. "Chapter 2 - Chemical Characterization and Cycling of Dissolved Organic Matter." In *Biogeochemistry of Marine Dissolved Organic Matter (Second Edition)*, edited by Dennis A. Hansell and Craig A. Carlson, 21–63. Boston: Academic Press.
Rivers, Adam R., Christa B. Smith, and Mary Ann Moran. 2014. "An Updated Genome Annotation for the Model Marine Bacterium Ruegeria Pomeroyi DSS-3." *Standards in Genomic Sciences* 9 (December): 11.
Sajed, Tanvir, Ana Marcu, Miguel Ramirez, Allison Pon, An Chi Guo, Craig Knox, Michael Wilson, Jason R. Grant, Yannick Djoumbou, and David S. Wishart. 2016. "ECMDB 2.0: A Richer Resource for Understanding the Biochemistry of E. Coli." *Nucleic Acids Research* 44 (D1): D495–501.
Sayers, Eric W., Jeffrey Beck, Evan E. Bolton, Devon Bourexis, James R. Brister, Kathi Canese, Donald C. Comeau, et al. 2021. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 49 (D1): D10–17.
Sayers, Eric W., Evan E. Bolton, J. Rodney Brister, Kathi Canese, Jessica Chan, Donald C. Comeau, Ryan Connor, et al. 2022. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 50 (D1): D20–26.
Schreier, Jeremy E., Christa B. Smith, Thomas R. Ioerger, and Mary Ann Moran. 2023. "A Mutant Fitness Assay Identifies Bacterial Interactions in a Model Ocean Hot Spot." *Proceedings of the National Academy of Sciences of the United States of America* 120 (12): e2217200120.
Schroer, William F., Hannah E. Kepner, Mario Uchimiya, Catalina Mejia, Lidimarie Trujillo Rodriguez, Christopher R. Reisch, and Mary Ann Moran. 2023. "Functional Annotation and Importance of Marine Bacterial Transporters of Plankton Exometabolites." *ISME Communications* 3 (1): 1–10.

Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69.
Shaffer, Michael, Mikayla A. Borton, Bridget B. McGivern, Ahmed A. Zayed, Sabina Leanti La Rosa, Lindsey M. Solden, Pengfei Liu, et al. 2020. “DRAM for Distilling Microbial Metabolism to Automate the Curation of Microbiome Function.” *Nucleic Acids Research* 48 (16): 8883–8900.
Sharpe, Garrett C., Scott M. Gifford, and Alecia N. Septer. 2020. “A Model Roseobacter, Ruegeria Pomeroyi DSS-3, Employs a Diffusible Killing Mechanism To Eliminate Competitors.” *mSystems* 5 (4). https://doi.org/10.1128/mSystems.00443-20.
Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.
Sun, Jing, Laura Steindler, J. Cameron Thrash, Kimberly H. Halsey, Daniel P. Smith, Amy E. Carter, Zachary C. Landry, and Stephen J. Giovannoni. 2011. “One Carbon Metabolism in SAR11 Pelagic Marine Bacteria.” *PloS One* 6 (8): e23973.
The pandas development team. 2023. *Pandas-Dev/pandas: Pandas*. https://doi.org/10.5281/zenodo.7794821.
Uchimiya, Mario, William Schroer, Malin Olofsson, Arthur S. Edison, and Mary Ann Moran. 2022. “Diel Investments in Metabolite Production and Consumption in a Model Microbial System.” *The ISME Journal* 16 (5): 1306–17.
Van Trappen, Stefanie, Tjhing-Lok Tan, Jifang Yang, Joris Mergaert, and Jean Swings. 2004. “Alteromonas Stellipolaris Sp. Nov., a Novel, Budding, Prosthecate Bacterium from Antarctic Seas, and Emended Description of the Genus Alteromonas.” *International Journal of Systematic and Evolutionary Microbiology* 54 (Pt 4): 1157–63.
Veseli, Iva, and Zac Cooper. 2022. “Ruegeria Pomeroyi Digital Microbe Databases.” https://doi.org/10.5281/zenodo.7304959.
Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods* 17 (3): 261–72.
Waskom, Michael. 2021. “Seaborn: Statistical Data Visualization.” *Journal of Open Source Software* 6 (60): 3021.
Wetmore, Kelly M., Morgan N. Price, Robert J. Waters, Jacob S. Lamson, Jennifer He, Cindi A. Hoover, Matthew J. Blow, et al. 2015. “Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons.” *mBio* 6 (3): e00306–15.
Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *A Grammar of Data Manipulation [R Package Dplyr Version 1.1.2]* (version 1.1.0). Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=dplyr.
Wickham, Hadley, and Dana Seidel. 2022. *Scale Functions for Visualization [R Package Scales Version 1.2.1]* (version 1.2.1). Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/scales/index.html.
Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2023. *Tidyr: Tidy Messy Data [R Package Tidyr Version 1.3.0]* (version 1.3.0). Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=tidyr.
Xu, Shuangbin, Zehan Dai, Pingfan Guo, Xiaocong Fu, Shanshan Liu, Lang Zhou, Wenli Tang, et al. 2021. “ggtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data.” *Molecular Biology and Evolution* 38 (9): 4039–42.
Xu, Shuangbin, Lin Li, Xiao Luo, Meijun Chen, Wenli Tang, Li Zhan, Zehan Dai, Tommy T. Lam, Yi Guan, and Guangchuang Yu. 2022. “*Ggtree* : A Serialized Data Object for Visualization of a Phylogenetic Tree and Annotation Data.” *iMeta* 1 (4). https://doi.org/10.1002/imt2.56.
Yin, Yanbin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu. 2012. “dbCAN: A

Web Resource for Automated Carbohydrate-Active Enzyme Annotation." *Nucleic Acids Research* 40 (Web Server issue): W445–51.

Yu, Guangchuang. 2020. "Using Ggtree to Visualize Data on Tree-Like Structures." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 69 (1): e96.

———. 2022. *Data Integration, Manipulation and Visualization of Phylogenetic Trees*. CRC Press, Taylor & Francis Group.

———. 2023. *Decorate a "Ggplot" with Associated Information [R Package Aplot Version 0.1.10]*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=aplot.

Yu, Guangchuang, Tommy Tsan-Yuk Lam, Huachen Zhu, and Yi Guan. 2018. "Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree." *Molecular Biology and Evolution* 35 (12): 3041–43.

Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 2017. "Ggtree : An R Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data." *Methods in Ecology and Evolution / British Ecological Society* 8 (1): 28–36.