UNIVERSITY OF CALIFORNIA
SANTA CRUZ

BIOINFORMATIC APPROACHES FOR NEW INSIGHTS INTO OLD MARINE METAGENOMIC DATA SETS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOINFORMATICS

by

Jonathan D. Magasin

March 2016

The Dissertation of Jonathan D. Magasin
is approved:

Professor Joshua Stuart, Chair

Doctor Dietlind Gerloff, Advisor

Professor Jonathan Zehr

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Bioinformatic approaches for new insights into old marine metagenomic data sets

by

Jonathan D. Magasin

Sampling from marine environments followed by *en masse* high-throughput nucleic acid sequencing (metagenomics) will transform our understanding of marine microbial communities and their impacts on Earth biogeochemistry. The interpretation of a marine metagenomic (MMG) data set within its environmental context, and extrapolation to planetary scale, require first that we answer simpler questions about a sample, “who is there?” and “what are they doing?” However, often the majority of sequenced genomic fragments (reads) that comprise a MMG data set cannot be assigned taxonomic or functional annotation by searches in reference sequence databases. This stems partly from the bias of reference databases to microbes amenable to study in the lab, the “culturable 1%.” Also, the complexity of MMG samples—thousands of populations per sample, and quadrillions to quintillions of nucleotides—leads to extreme underrepresentation by the reads. Moreover, the shortness of reads makes annotation difficult. There are now tens of thousands of MMG data sets, each with many thousands to billions of mostly unannotated reads.

Learning from this massive resource of raw data will require new bioinformatics approaches. This dissertation explores approaches that pool different MMG data sets to benefit annotation and the discovery of widespread marine microbes. First, we

hypothesized that pooling MMG data sets, and assembling the reads into longer sequences (contigs), would increase species and functional annotation of the reads. This proved true for the forty-two real MMG data sets we investigated. For simulated MMG data sets, pooled contigs were found to rarely mix reads from different species. This supports that pooled contigs, though a consensus of reads from different populations, are biologically interpretable, and that annotation may be transferred to constituent reads.

Second, given the high computational cost of assembly and the huge number of MMG data sets from which to select for pooling, we hypothesized that ranking data sets would make pooled assembly more efficient. This was correct. Ranking data sets by k-mer profile similarity resulted in pooled assembly rates on a par with ranking based on phylogenetic profile similarity. In practical terms, this means one can exploit pooled assembly to increase annotation without first having to annotate sets individually to create phylogenetic profiles.

Third, we found that pooling of MMG data sets can enable the discovery of ubiquitous and abundant marine microbial species and partial characterization of their genomes, without need for culturing them. This was accomplished by "geographic profiling" of public, pooled-assembled MMG data. Three novel species were predicted in multiple MMG sets and substantiated with orthogonal lines of evidence. Experimental work corroborated predicted sequence fragments for each of the species, and analyses of these fragments supported that they likely represent abundant, ubiquitous novel species.

To my research assistant,

Annie.

Acknowledgments

Thank you to my adviser, Dietlind Gerloff, whose unwavering enthusiasm for science and dedication to her students at all levels continue to inspire me. Thank you to Joshua Stuart and Jonathan Zehr for serving on my committee, and to Jonathan for opportunities to interact with the marine microbial ecology community through workshops at the Gordon and Betty Moore Foundation and discussions with Zehr Lab members. Thank you Brian Palenik and Emy Daniels for being such gracious hosts and teachers during my wet lab work at SIO.

I also wish to thank Mark Akeson, Jake Metcalf and Sandra Dreisbach, the instructors of BME/PHIL 80G (Bioethics). My experiences TA'ing 80G were a core part of my graduate education.

Finally, for their support and love, thank you to my parents, Michael and Dennie. Any contribution that I make with the skills learned over these years of study is in your honor.

The text of this dissertation includes a reprint of the following previously published material:

> Magasin, J.D. and Gerloff, D.L. (2015). Pooled assembly of marine metagenomic datasets: enriching annotation through chimerism. *Bioinformatics.* 31(3):311–7. doi:10.1093/bioinformatics/btu546. Epub 2014 Oct. 11.
>
> URL: `http://bioinformatics.oxfordjournals.org/content/31/3/311`

The co-author listed in this publication directed and supervised the research which forms the basis for the dissertation.

Chapter 1

Marine metagenomics background

1.1 Scales of discovery

A milliliter of seawater can contain a million microbial cells that represent thousands of different genomes [52]. The interactions and dependencies among these community members with one another and their specific environments, and the evolutionary mechanisms that created this diversity, motivate marine microbial oceanography, ecology and biology. These fields span vastly different scales—thirty-six orders of magnitude in volume[1] and over 3.8 billion years [83]—and so it is remarkable that marine metagenomics can contribute to each. This section describes how marine metagenomics can help expand our knowledge across these scales, and highlights some recent inter-scale discoveries already to its credit.[2]

[1] Ocean volume, 1.33×10^{18} m^3, to *Prochlorococcus* cell volume (r=0.35 μm).

[2] "Marine metagenomics" in this chapter includes pre-next-generation sequencing approaches that often created clone libraries for random DNA fragments, or that amplified genes of interest. Chapters 3–5 use data sets produced by direct 454 pyrosequencing, with exceptions noted.

1.1.1 Earth biogeochemical cycles

Multiplied over the world ocean, the microorganisms in that drop comprise over 90% of marine biomass and drive element cycles that the biosphere and climate depend upon [21, 76]. Carbon in that biomass turns over every ∼twenty days [83], with possible fates that include respiration back into the atmosphere, assimilation into higher trophic levels, or conversion to particulate or dissolved organic matter some of which is exported to the deep sea for millennia, or to seafloor sediments for millions of years [31, 65].

Simulations can test our understanding of these cycles at global scale, and models that account for open ocean circulation, microbial community structure, and nutrients have shown robust patterns that agree broadly with observations [45, 95].[3] However, such models include only known species, and typically just the abundant ones. Recently the Tara Oceans project reported ∼35,000 uncultured marine microbes, for just the first 243 of over thirty-five thousand marine metagenomic samples [166]. It is a safe bet that important species and novel metabolisms revealed by these data will improve models, as even relatively rare organisms can be critical to a marine community. For example, marine diazotrophs are often rare community members[4] yet supply fixed nitrogen essential to community growth and metabolism [204]. Most diazotrophs are known only by PCR amplification of nitrogenase genes from environmental samples [74,

[3] "Observations" are calculated/inferred, e.g. primary production estimated from chlorophyll, inferred from measured spectra in satellite images. Coastal regions are usually excluded in such simulations.

[4] Maximal Pacific open-ocean diazotroph abundances reported by Church et al. (2009) and Moisander et al. (2010) would place them at <1% of a community, assuming 10^6 cells/ml [26, 116].

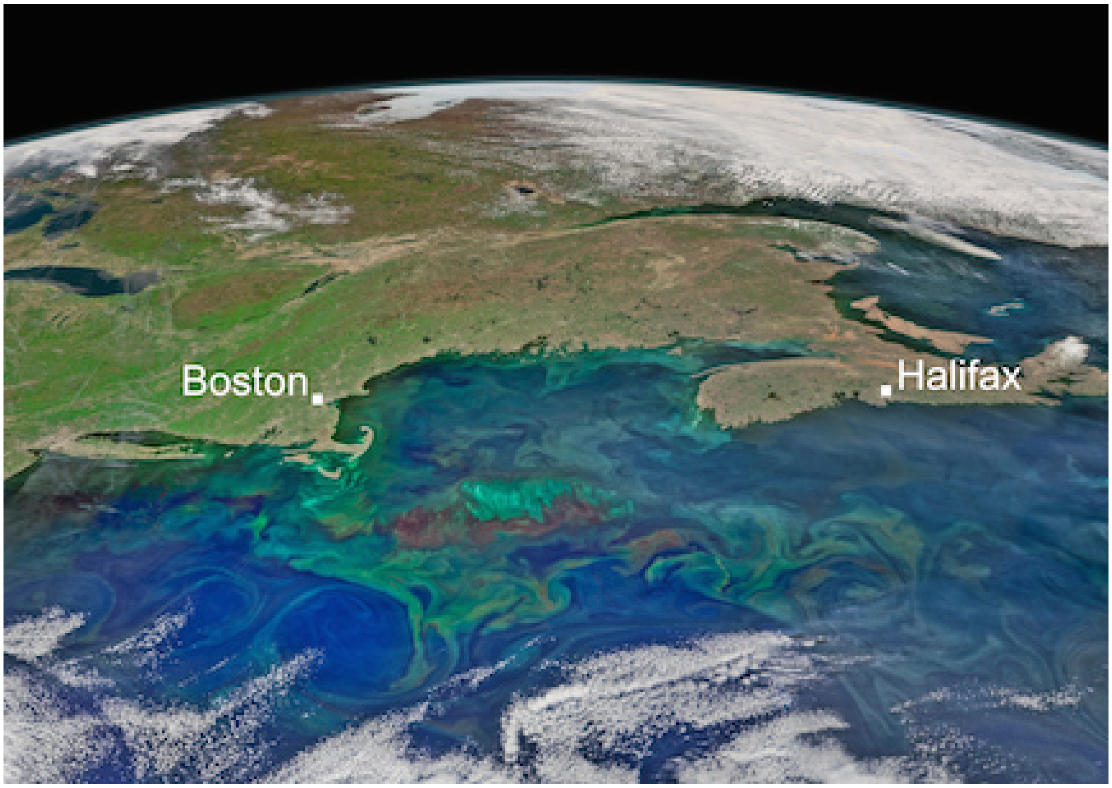


Figure 1.1: North Atlantic spring phytoplankton bloom, May 14, 2015. The brilliant green swirls of phytoplankton appear mixed by the confluence of the Gulf Stream from the south and the Labrador Current from the north. Marine metagenomics focuses on microbes, primarily bacteria and archaea, slightly smaller than the eukaryotic diatoms that likely comprised this bloom, but upon which all ocean and terrestrial life depend. Satellite image composite by Norman Kuring and courtesy of NASA (adapted) [121].

204]. Ironically, their discovery was prompted by discrepancies in models for nitrogen fixation versus denitrification [48].

Field experiments can help us understand biogeochemical cycles at the mesoscale, including how cycles might be influenced to slow climate change. Such experiments fertilize a patch of seawater with nutrients thought to be growth-limiting to trigger a phytoplankton bloom (much smaller than in Fig. 1.1), and then track the fate of atmospheric carbon drawn down into biomass. One such experiment in the Southern

Ocean fertilized a patch of surface water with labeled iron. Although this triggered a massive diatom bloom and significant drawdown of CO_2, carbon export to the deep was not observed [14]. Another fertilization experiment in the phosphorous-limited Eastern Mediterranean Sea unexpectedly stimulated the growth of *non*-photosynthetic microbes and copepod egg production [176]. Our poor understanding of marine biomes at the scales of these experiments make their designs problematic [76].

1.1.2 Community structure

However, some argue that our understanding of global scale impacts will advance through research into marine biomes at the nano- to millimeter scales, i.e. those at which biochemical and physical interactions occur between community members [7]. Many of these interactions are lethal: The typical drop of surface seawater mentioned before also contains 10^7 viruses and 10^4 protists [7]. Rampant viral infection[5] and grazing by heterotrophic protists amount to an average marine microbe lifespan of just 1.5–16 days [191, 155].

Vast numbers notwithstanding, organisms and viruses occupy by volume just one ten-millionth of the drop [136]. In the open ocean a *Prochlorococcus* might be 100–200 diameters (0.5–0.7μm) from its nearest neighbor cell, 40 from its nearest phage [11]. The water is far from homogenous though, peppered with transparent gel colloids (10nm) and mucus sheets and bundles (100μm) [7]. Microbes may attach to structures comprised of these colloids, sheets and bundles. Perhaps some of the genetic diver-

[5] Worldwide phage infections are estimated at 10^{23} per second [64]. This rate is over five orders of magnitude greater than the age of the universe in seconds.

sity of marine microbes, much of it discovered via MMG, stems from adaptations to microniches afforded by such structures [7]. Interactions may also occur in the nutrient-rich microniche surrounding a phytoplankter (phycosphere), or on and in the plume of sinking aggregates of organic detritus (marine snow; ≥0.5 mm) [7].

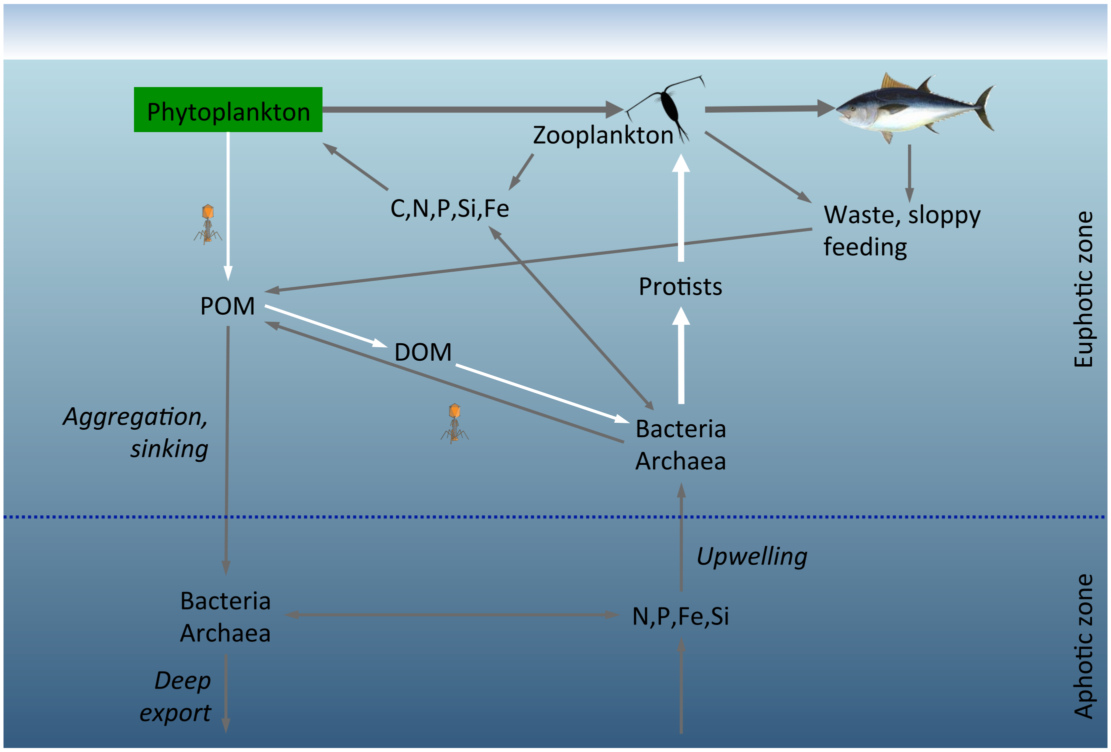


Figure 1.2: The marine microbial loop (white arrows) is the circuit in a marine food web by which dissolved organic matter in the euphotic zone is assimilated and/or respired by heterotrophic bacteria and archaea, and through their predation by heterotrophic unicellular eukaryotes (protists) transferred to higher trophic levels. Phytoplankton growth depends on light and inorganic nutrients (C,N,P,Si,Fe) recycled as shown. Organic matter that sinks below thermocline (dotted line) is mostly respired, driving the CO_2 concentration in the ocean interior well beyond that of the atmosphere (the "biological pump"). A small fraction of the exported organic matter sinks and is buried in the seafloor sediment, for up to millions of years. Thick arrows represent trophic interactions; thin represent mostly physical or chemical processes, some of them heavily influenced by phage-induced lysis. Figure based on that by Azam and Malfatti in [7].

The microbial loop is a conceptual model that places the interactions among microbial community members in the context of dissolved organic matter and connections to higher trophic levels (Fig. 1.2). Bacteria, archaea, and picoeukaryotes that comprise the microbial players are termed autotrophs or heterotrophs according to whether they fix carbon dioxide themselves to produce organic matter, or must obtain it from others. Prefixes may be added to these terms to specify the source of metabolic energy, often sunlight (photo-) or chemical reduction (chemo-). Photoautotrophic microbes such as the cyanobacteria *Prochlorococcus* and *Synechococcus* introduce organic matter at the base of the marine food web (primary production) by photosynthesis. Much larger eukaryotic photoautotrophs, including single-celled diatoms and many dinoflagellates, and also multicellular macroalgae, also contribute substantially to primary production. Sometimes this is via large phytoplankton (diatom or dinoflagellate) blooms, whether seasonal such as the North Atlantic Spring Bloom (Fig. 1.1), or episodic for example the massive diatom bloom that occurred in the Monterey Bay this summer. Heterotrophs obtain particulate (POM) or dissolved (DOM) organic matter from algal and microbial exudates, debris from viral lysis, and in some cases by enzymatically cleaving polymers from the cell surfaces of phytoplankton [7].

Several recent studies highlight the potential of marine ‘omics to reveal mechanisms underlying the microbial loop, specifically coordinated metabolisms between microbial autotrophs and heterotrophs [126, 127, 6]. While diel expression patterns for marine autotrophs have been known for some time (e.g. the upregulation of some *Prochlorococcus* photosynthesis genes before sunrise each day [49]), circadian genes that would drive diel patterns are mostly unknown among heterotrophic bacteria [159]. Thus it was surprise for Ottesen et al. (2013, 2014) to observe bacterial and archaeal heterotroph expression patterns that were sinusoidal and staggered relative to autotroph patterns, both in the Monterey Bay and the North Pacific Subtropical Gyre [126, 127]. Possibly the heterotrophs respond to autotroph metabolites—growth and nutrient acquisition genes were upregulated—but environmental cues or some other mechanism may drive this daily cascade of coordinated expression [126, 127].

1.1.3 Pan-genomes and novel metabolisms

Metagenomics is often introduced by the two questions it asks of an environmental sample, who is there, and what are they doing. However, the "who" that researchers would count and characterize, "species," has proven elusive [85]. The commonly used criterion of ≥97% id 16S rRNA cannot resolve *Synechococcus* from *Prochlorococcus* species, or 'ecotypes' therein, despite differences in niche (nutrient-rich coastal versus nutrient-poor open ocean, usually), physiology, and phage interactions [47]. Striking differences in gene content between type strains and environmental strains of the same species (e.g. in *Prochlorococcus* [82]) have expanded the concept to that of the pan-genome. The pan-genome comprises 'core' genes present in every member of the species (1–2K genes; 80% of the genome) and 'dispensable' (or 'flexible') genes present in one or several strains [109]. Dispensable genes, which can outnumber the core by orders of magnitude, may confer resistance to phage, or may adapt an ecotype to physical and nutrient conditions in a specific environment. Thus one expects members from a population from one site to be more genomically similar compared to members from another population. That said, hundreds of stable subpopulations can exist within a single milliliter of water, as has been shown for *Prochlorococcus* [80]. Models for many species suggest that they have open pan-genomes, i.e. that the curve for discovered genes versus newly sequenced strains (often tens per strain) has not saturated [109], even after eleven years of whole-genome sequencing data [183].

Horizontal gene transfer (HGT, or lateral gene transfer) is the broad term for mechanisms of microbial genetic exchange that include conjugation, viral infection (transduction), and uptake from the environment (transformation). Biller et al. (2014) recently proposed a connection between the latter two for *Prochlorococcus*, namely that small DNA-containing vesicles it produces may facilitate genetic exchange, and/or may act as decoys for viral infection [12].[6]

MMG has also led to some unexpected insights into eukaryogenesis. Metagenomic sequencing and assembly of deep marine sediment samples revealed the first genome (92% complete) of the candidate phylum Lokiarchaeota, which appears to share a common ancestor with Eukarya within the archaeal TACK[7] superphylum, based on marker gene phylogenomics [160]. Moreover, the authors propose that some eukaryotic machinery may have existed in the common archaeal ancestor: They predicted 175 homologs to eukaryotic "signature" proteins, many of them associated with cell membrane remodeling and vesicular trafficking [160].

With respect to "what are they doing," several landmark MMG papers have

[6] Both possibilities had experimental support. About 50% of the genome was represented by DNA fragments recovered from vesicles, and TEM images of some vesicles showed signs of phage infection [12].

[7] TACK comprises the Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota.

shown the versatility of marine microbes in exploiting available energy sources, with consequences at global scale for element cycles. The widespread use of proteorhodopsin by marine bacteria to capture solar energy, first suggested by its abundance in MMG data [9], changed our picture of the carbon cycle, as did the discovery of widespread aerobic anoxygenic photosynthesis [10, 202, 203]. The expansion of ammonia oxidizers to include Crenarchaea did so for the nitrogen cycle [183, 180, 84]. Interestingly, rhodopsin-based phototrophy was already known within Archaea, and ammonia oxidation was already known within Bacteria. Their cross-presences may illustrate how metabolic pathways have evolved as functional modules that are exchanged, co-opted, and sometimes run in reverse, by different community members [41].

1.2 Complementary approaches in marine microbial ecology

Community sampling and sequencing are the common denominators of meta-'omics approaches. Usually these entail whole-genome shotgun sequencing with a next-generation platform. However, several complementary approaches are still used in marine microbial ecology (Tab.1.1), briefly described in this section.

Approach	**Main use**	**How complements metagenomics**

metagenomics	Phylogenetic profiles;
	Functional potential;
	Variation	

metatranscriptomics	Gene expression profiles	vs. functional potential
microarrays	Phylogenetic profiles via proxy genes	Highly sensitive detection of rare members; Inexpensive; Shared designs
PCR, qPCR	Detect genes of specific function or marker (16S)	Highly specific, sensitive; Low cost; Supports follow-up e.g. gene screens for single-cell sequencing; Quantifiable (qPCR)
single-cell sequencing	New reference genomes; In-sample variation	Annotate reads from uncultivable, numerically dominant

Table 1.1: Relationship of other popular approaches in marine microbial ecology to whole-genome shotgun sequencing of community DNA.

1.2.1 Metatranscriptomics

Metatranscriptomic studies typically apply next-generation sequencing to cDNA libraries derived from an environmental sample in order to understand community gene expression. As expression and composition shed light on one another (and for normalization of transcript abundances), metatranscriptomic studies may also produce metagenomic data sets for the same samples, as part of a time series for a specific location (e.g.

[46, 53]) or for sample sites on an expedition (e.g. [166]).

A challenge with metatranscriptomic studies is obtaining sufficient quantities of cDNA, reverse-transcribed from the extracted prokaryotic mRNA. Because as much as 95% of the total RNA extracted from a sample may be ribosomal (rRNA) and tRNA [158], a depletion step is required. Kits such as MICROBExpress deplete rRNA via primers for known 16S and 23S and thus miss novel rRNA or other non-coding RNAs. To avoid such biases, Stewart et al. (2010) developed an rRNA subtraction protocol that uses ribosomal DNA extracted *from the sample* as a PCR template. PCR products are labeled and then hybridized to rRNA in the sample-extracted RNA, enriching mRNA greatly [164].) After mRNA enrichment, the amount of cDNA subsequently reverse-transcribed may only be picograms, in which case amplification is necessary. Multiple displacement amplification with random hexamers has been used [52, 143].

As with metagenomics, annotation is challenging: Typically only one-third of the transcripts can be matched to known genes or functional gene categories [25].

1.2.2 Gene-targeted surveys

With the important exception of the Sargasso Sea study [183], prior to ∼2004 high Sanger sequencing costs limited marine metagenomics to specific genes, e.g. ribosomal RNA for phylogeny-based community profiling, or genes relevant to the cycling of specific elements (e.g. nifH for nitrogen fixation). Although marine metagenomics shifted largely to whole-genome shotgun sequencing with the advent of affordable next-generation platforms, two gene-targeted approaches that complement NGS remain in use.

Marine microarrays can detect with extremely high sensitivity marker genes or transcripts of functional or phylogenetic interest [156], and at a cost well-suited to longitudinal studies over vast ocean areas (e.g. ∼\$15 for an array [146]). By comparison NGS approaches must increase sequencing effort to detect rare taxa or specific genes, at costs easily into the thousands of dollars.[8] Several standardized marine microarray designs could ease sharing and interpretation of data generated by different labs at different sample sites [146, 156]. Low-resolution microarrays have also been used on prototype environmental sample processors in the Monterey Bay, to simultaneously measure environmental variables (temperature, salinity, chlorophyll) and gene expression [137]. Challenges with microarrays include probe limitation to known genes, non-specific binding to probes, and noise that impacts normalization and thus relative abundance estimates.

PCR and qPCR of ribosomal RNA continues to be important for diversity studies (e.g. [207, 30]). In particular, PCR amplification in combination with NGS probe more deeply into the so-called “rare biosphere,” the many species in any sample that remain in the tail of the abundance curve [157, 133]. It also provides a basis for comparison against NGS approaches [190]. Additionally PCR offers a low-cost way to test predictions based on MMG data sets but for which there was insufficient sequence. For example, Palenik et al. (2009) used inverse PCR to verify the circularity of a

[8] For example, the Cornell Genomics Facility offers Illumina paired-end, 2 × 250 bp, single-lane sequencing with HiSeq 2500 for US\$5680 and with MiSeq for US\$1720, excluding DNA sample library preparation.

repA-containing MMG contig (1479 bp), supporting that it is a plasmid [131]; and our approach in chapter 5 used PCR to verify several microbes that were predicted to be in a sample (at non-high abundance). Finally, PCR underlies homology-based screens within metagenomic clone libraries for the discovery of novel enzymes from uncultured organisms ("functional metagenomics") [81].

In short, cheaper and deeper sequencing offered by NGS will probably never represent all the DNA in a single sample, or even milliliter, of seawater [52], and so gene-targeted approaches such as microarrays and PCR will play an important and complementary role in marine metagenomics.

1.2.3 Single-cell genome sequencing

With 99% of species thought to be uncultivable [4], annotation by reference to the tiny fraction of the remaining 1% that comprise most of the reference database sequences severely limits MMG annotation. Single-cell genome sequencing bypasses culturing and thus can produce reference genomes for species/ecotypes that are numerically dominant but unrepresented, or that have genes of interest. The two main and collaborative single cell genomics (SCG) centers, at the JGI and the Bigelow Laboratory for Ocean Sciences, have to date produced many hundreds of single-amplified genome (SAG) sequences.[9] In surface waters single-cell sequencing has revealed hundreds of distinct and stable subpopulations of *Prochlorococcus* within the same milliliter of water [80], and two novel species of *Flavobacteria* that were abundant (up to 1%) in some

[9] Over 1000 SAGs were produced by the Bigelow Laboratory for [80], and the JGI has 2070 in-progress or complete single-cell sequencing projects as of Aug. 28, 2015.

coastal GOS samples [198]. The SAGs for these *Flavobacteria* have helped us begin to identify which taxa account for the widespread use of proteorhodopsin over the ocean surface, by sequencing proteorhodopsin and phylogenetic marker genes within the same DNA fragment. In the ocean interior single-cell sequencing has revealed numerically dominant thaumarchaea that appear capable of fixing their own carbon and oxidizing ammonia [168], within implications for carbon and nitrogen cycles.

Production of SAGs from an environmental sample begins with the separation and capture of individual cells from the population of interest by fluorescence activated cell sorting (FACS) (reviewed in [162]). Cell size and fluorescence of the target population determine the FACS gating parameters. DNA stains can be used to capture abundant community members; intrinsic fluorescence can help capture some types of rare members (e.g. chlorophyll for cyanobacteria). As most of the volume of the rapidly sorted microdroplets is cellular, contamination by cell-free DNA is minimized, but well contamination with foreign cells is an issue [163]. Captured individual cells are next lysed and amplified to create shotgun libraries with sufficient DNA for sequencing. Multiple displacement amplification (MDA) is favored despite uneven coverage, sensitivity to contaminant DNA, and for marine samples a fairly low rate ($\sim$18%) of successful amplification of sorted cells [27]. Typically the next steps are to screen the libraries for (candidate) phyla or genes of functional interest,[10] and to sequence with an NGS platform.

[10] SCG has also been used to estimate an upper bound for specific rare genes among community members, by showing absence/presence in replicate wells with 100 cells each [163].

1.3 From sample collection to sequence data

My work responds to challenges with MMG data sets that stem from inherent sample complexity and properties of the sample collection, sequencing technologies, and annotation workflows that produce the data. These are described briefly in this section. For in-depth background, excellent review papers include DeLong 2009 (marine ecosystems, sample collection and sequencing) [32], Gilbert et al. 2011 (overview of microbial metagenomics within the context of recent marine studies) [52], Kunin et al. 2008 (overview of computational analysis and challenges for metagenomics in general; Sanger sequencing, not 454) [87] and McHardy et al. 2007 (phylogenetic classification) [108].

1.3.1 Sample collection and processing

Open ocean water samples for marine environmental sequencing projects are often collected with a rosette sampler (Fig. 1.3). Each bottle of the rosette samples from a different depth and samples are processed separately as shown. Samples may be taken monthly or seasonally from a geographic location representative of a large region. For example since 1988 the Hawaiian Ocean Time-series has sampled at Station ALOHA, 100 km north of Oahu, and produced many public metagenomic and metatranscriptomic data sets, as well as physical oceanographic measurements that provide crucial context for interpretation [77, 63]. Samples may also be taken at semi-regular distances along a cruise surveying a particular coast or ocean region. The GOS samples, collected at

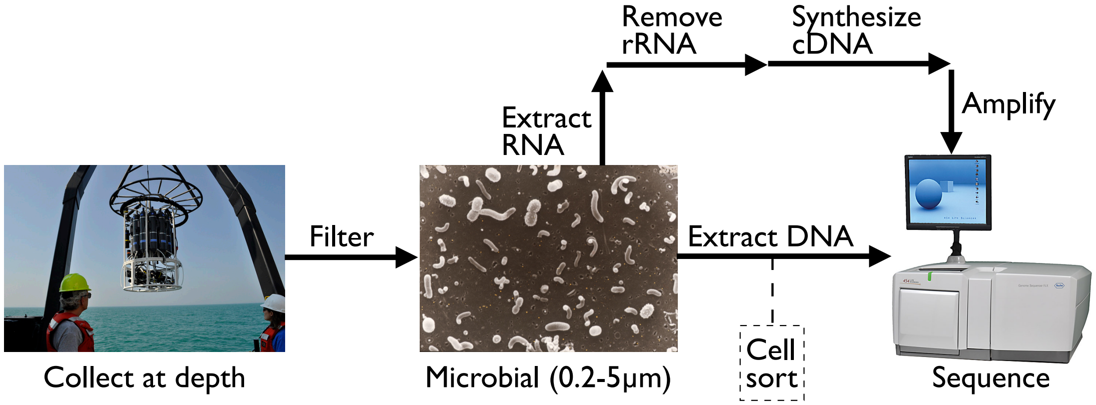


Figure 1.3: Generation of MMG sequence data. Water collected from a target depth is filtered to remove eukaryotes. For metagenomic studies, community DNA is extracted and sequenced without cloning. Single cell genomics sorts and individually sequences cells (see §1.2.3). For metatranscriptomic studies (top path; see §1.2.1) mRNA enrichment and cDNA synthesis are required before sequencing. SEM image by Ed DeLong and used courtesy of Monterey Bay Aquarium Research Institute (c) 2000. Image of 454 sequencer courtesy of Roche. Rosette sampler image from web (Creative Commons License) [29]. Figure based on Fig. 2 in DeLong (2009) [32].

intervals of ∼320 km along an ∼8000 km cruise, are a well-known example [183, 148].

Coastal samples taken from waters over a continental shelf are processed as in Fig. 1.3 except that collection may be far simpler and economical: The sample used in chapter 5 was obtained by lowering a bucket off the Scripps Pier. By comparison research vessels for open ocean sampling run $20–$30K per day and are difficult to schedule [76]. Until sample collection by autonomous vehicles becomes widespread, frequent sampling will only be feasible for coasts.

Next, samples are filtered to remove eukaryotes whose large genomes and repetitive sequence would distort species diversity estimates and confound even partial assembly [87].[11] Often the first filter has a pore size of 5 μm, or smaller if a project

[11] For example, dinoflagellates have some of the largest genomes known (1.5–185 Gbp), not due to extreme polyploidy or repetitive DNA [61].

targets prokaryotes of an expected size range. The flow-through from the first filter is passed through one or more subsequent filters, often a single 0.2 μm filter, to produce size-fractionated samples. These initial steps are largely the same for metagenomic and metatranscriptomic studies and sometimes the same sample will be used for both (§1.2.1). Marine virus studies on the other hand use a tangential flow filtration system to filter at 100 kDa e.g., ideally removing all but viruses and virus-like particles [179].

The volume of water filtered ranges from 1–1000 L [52], with nutrient-rich coastal samples at the low end and nutrient poor (oligotrophic) open ocean samples at the high.[12] Samples for metatranscriptomic analysis use <1 L [32].

At the wet lab, marine metagenomic projects often use off-the-shelf kits to extract DNA (bottom path of Fig. 1.3). 454 sequencing requires only microgram quantities of DNA allowing one to sequence directly, avoiding the labor and biases of a clone library [32].

1.3.2 Sequencing technologies

The 80.61×10^{12} total base pairs that have accumulated on MG-RAST since its start in 2007 are about 4% of the estimated microbial DNA in one liter of seawater (2×10^{15} bp), a typical small volume for a marine metagenomic sample [52].[13] It seems a safe bet that no sequencing technology will ever fully capture the DNA in a sample.

[12] One Antarctic waters researcher at the MicroTools II workshop mentioned that he pumps "until the filter clogs."

[13] MG-RAST accessed on August 19, 2015. The 80.61 Tbp are in 644 billion reads and 199,591 data sets, across marine, soil, host-associated and other biomes. They exclude ~35,000 Tara Oceans data sets which when all sequenced I estimate will total 835–1037 Tbp, based on the raw and cleaned sizes of the first 243 sets.

However, each new generation of the major NGS platforms has enabled a deeper and cheaper look at the complexity of marine communities, with trade-offs between read length and data volume (Tab.1.2).

Platform	**Read length**	**Mega-bases**	**Error rate**	**MG-RAST**		**IMG/M**	
				sets	**proj**	**sets**	**proj**
Sanger (ABI 3730) [87]	~750	$100^{\dagger}$	<1%	78	4	71	6
454/Roche				154	25	86	13
GS20 [69]	~100	33	0.49%				
GS FLX [142]	249	35	-				
FLX Titanium [107, 42]	350	~175–350	0.68%				
Illumina				101	24	448	26
Genome Analyzer [102]	25–35	-	-				
GA_{IIx} [107, 141]	$\leq 150^{\times 2}$	30,000	≤0.34%				
HiSeq 2000 [141]	$\leq 150^{\times 2}$	600,000	0.26%				
MiSeq [141]	$\leq 150^{\times 2}$	1500–2000	0.80%				
Life Technologies							
SOLiD [102]	25–35	2000–4000	-	0	0	1	1
SOLiD 3 [178]	~50	6000	≤3%	0	0	0	0
SOLiD 5500xl [152]	75	155,000	>0.01%	0	0	0	0
Ion Torrent PGM [141]	~200	20–1000	1.71%	69	4	0	0
Totals				402	57	$539^{\ddagger}$	46

Table 1.2: Comparison of selected sequencing platforms used in marine metagenomics. Platforms are ordered within manufacturer from first/early to recent generations. Last columns show public, marine biome, whole-genome shotgun data sets on MG-RAST (searched on Aug. 17, 2015) and IMG/M (searched on Aug. 25, 2015). Reported metrics (columns 2–4) are from indicated studies but representative of the literature. “×2” denotes paired end reads. †: Sanger sequencing effort used by Joint Genome Institute pre-NGS [87]. ‡: IMG/M sets can have multiple sequencing types so total is less than column entries.

The main messages of this section are:

- 454 pyrosequencing was the early and predominant NGS platform used in marine metagenomics because it was the first to produce reads long enough for (limited) annotation [196, 197]. Importantly for this dissertation, they were long enough for assembly (chapters 3,5) and compositional signatures (chapter 4) before those of Illumina. Moreover, during most of my research 454 sets were the ∼only public, MMG, NGS sets available.

- Eventually Illumina reads surpassed in length even those of the first 454 generations. Illumina sets are orders of magnitude larger than 454 sets, an advantage over 454 for detecting rarer organisms in a sample. Innovative assembly of such volumes has recovered complete and nearly complete genomes from non-marine environments (e.g. [1]), but for marine yields many contigs of length ∼1K.[14] Some Illumina assembly challenges in metagenomics are discussed in §1.4.

- Although Illumina platforms now predominate, the hundreds of Sanger- and pyrosequenced marine metagenomic data sets available on public data repositories are invaluable snapshots of microbial communities at one place and time. Resequencing archived samples may be possible in some cases, but it would seem prudent to learn to leverage "old" data for which we have much expertise and tools.

The remainder of this section briefly describes Sanger and NGS sequencing, to highlight why the latter generates vastly more data—why (marine) metagenomics took

[14] The 243 Tara Oceans data sets had N50 contig lengths of 1121±176 nt [171].

off. Some interesting history and particularities of the platforms are described but not essential to understanding later chapters. Mardis [102] and Metzker [111] have written excellent reviews on NGS.

Sanger sequencing

Metagenomics began with Sanger sequencing of environmental 5S and 16S rDNA [161, 129], but the term “metagenomics” was introduced and at first used strictly for untargeted whole-community shotgun sequencing [62]. By the mid-2000’s shotgun studies had applied Sanger sequencing to large-insert clone libraries, most famously the Sargasso Sea pilot study [183] and the Global Ocean Sampling Expedition (GOS) [148]. Despite the enormous amounts of sequence data produced by those studies, at enormous cost, the complexity of those samples restricted representation to abundant community members. Nevertheless, those and other Sanger MMG data sets remain a valuable resource against which to compare modern MMG sets. For example, the Tara Oceans project reported that over 81% of the >40 million non-redundant genes identified so far were not found in GOS, which uncovered 1700 novel protein families [166, 201]. Such sets are also valuable because Sanger sequencing errors are, in comparison to NGS technologies, infrequent and well-understood.

For comparison to NGS sequencing and context for some of the chapter 5 discussion, an overview of Sanger sequencing is now given. From the template DNA molecule, four populations of PCR products are created, one for each base type. The PCR reactions for each include a small proportion of fluorescently labeled ddNTP’s for the specific base only that will prematurely terminate elongation. These ddNTP’s are stochastically incorporated, thus each population comprises PCR products of different lengths but ending in dd{A,C,G,T}NTP respectively. For each population, gel electrophoresis through a capillary tube separates and orders the products by length while a laser strobes for the presence or absence of the ddNTP. Software interleaves the four channels into a “chromatogram” or “trace” (called a “read” in NGS); strobe events become base positions. At each position the software calculates a quality score and calls the base A,C,T or G dependent upon which of the four channels had the largest and highest quality peak.[15] In practice, Sanger reads usually exceed 1000 bp, however the quality often deteriorates noticeably after ∼800 bp: Isolated, low-quality bases manifest first in the chromatogram, followed by peaks that gradually decrease in sharpness, amplitude, and differentiability especially after about 1000–1100 bp.

Next-generation sequencing (NGS) commonalities

Pyrosequencing (454 Life Sciences, then Roche), Illumina, and SOLiD (Applied Biosystems) were the early and predominant NGS technologies that enabled high-throughput genome sequencing (described well in [102]). All three employ variants of “sequencing by synthesis” in which the addition of one nucleotide by DNA polymerase, to a growing DNA oligomer, triggers chemistry that ultimately releases light. High-resolution imaging optics and hardware detect the light, simultaneously for many oligomers (up to tens of millions pre-2010) synthesized in parallel. The cycle repeats many times to produce NGS sequencing reads of technology- and chemistry-dependent

[15] Phred quality scores equal the estimated (based on training) probability of an erroneous base call. Bases from about 40 to 800 usually have qualities ≥ 30, where a score of 30 means the base has 1 chance in 1000 (from $10^{30/10}$) of being wrongly called.

lengths (see Tab.1.2).

The dramatic reduction in cost per base pair made NGS attractive to metagenomics, which requires deep sequencing to represent the diversity and dynamic range (abundant versus scarce) of populations in a sample. Also attractive was the non-requirement of clone libraries for the environmental sequence fragments. Such libraries can be biased against genes that encode proteins toxic to the bacterial host of the library, among other issues.

454 pyrosequencing

Roche 454 pyrosequencing was the first NGS technology to produce reads long enough for gene annotation and homology-based searches and sufficiently many of them that non-abundant species were observed [196, 25, 103]. When applied to amplified variable regions of microbial small subunit rDNA, 454 revealed extraordinarily high diversity, the so-called “rare biosphere” [157]. Some debate arose as to how much of the observed diversity was due to pyrosequencing error [88]. However, subsequent work that clustered conservatively to avoid overestimating diversity [207], as well as Illumina-based work [166], have shown hundreds to thousands of microbial types within single pelagic samples.

One GS FLX Titanium run generates 400–600 megabases in 100–150K high-quality reads. This excludes the 11–35% of reads that are “artificial duplicates,” copies of a single biological molecule appearing on multiple beads due to an error in emulsion PCR [57].

Illumina sequencing

Most Illumina platforms now have read lengths greater than those of early 454 (e.g. 150 bp or 250 bp, paired end) and at staggering volumes (>30 million reads per sequencing run). Although many hundreds of Sanger- and pyrosequenced marine metagenomic data sets now exist on public repositories, Illumina is now the NGS platform of choice.

Other recent sequencing technologies used in marine metagenomics

Other sequencing platforms have enjoyed limited use in marine metagenomics. Iverson and Armbrust [72] produced one closed and one nearly complete genome from SOLiD data sets, an impressive feat given the short read lengths.[16] However, latest generation SOLiD reads are still short and the platform has not seen recent use in MMG (Tab. 1.2). The Ion Torrent Personal Genome Machine (PGM) produces reads in the same length range as 454 depending on the kit (200–400 bp). Although an early evaluation [15] found it less accurate than light-based NGS technologies and revealed biases in the library preparation against high and low G+C% organisms—a concern for meta-'omic studies— it has been used in marine metagenomics, usually in combination with other NGS platforms [96, 94, 184, 206].

1.3.3 Metadata

Data set interpretation depends on the environmental conditions at the time of sample collection, recorded as "metadata" such as temperature, depth, salinity, and

[16] The authors drew some criticism for not distributing their custom assembly software. At the MicroTools II workshop they commented that the raw reads and assembled genomes were available for the evaluation of their work. Releasing their software would have obligated them to maintain it for the community, which would have been very expensive timewise as Iverson completed his Ph.D.

inorganic nutrient availability. Additional metadata related to sample processing is essential: Filter size impacts observed organisms, and sequencing platform impacts annotation and interpretation (e.g. mistaking sequencing errors for genomic variation [69]). Finally, metadata can enable data set reuse beyond the goals of its original study, for example, species discovery guided by sample locations of public sets (chapter 5).

	2007	2008	2009	2010	2011	2012	2013	2014
MMG data sets	58	91	58	5	16	100	65	32
Complete MIMS	86%	96%	78%	80%	94%	100%	100%	97%
Sample coordinates	95%	98%	83%	100%	100%	100%	100%	97%
Feature or material	41%	1%	0%	80%	100%	100%	100%	94%
Depth	23%	100%	57%	100%	100%	64%	57%	75%

Table 1.3: Recording of minimal required metadata for MMG sets on MG-RAST. Metadata is from the MIMS and MIxS standards, first published in 2008 [43] and 2011 [200], respectively. Data sets were created on MG-RAST in the years shown. The second row gives the proportion of sets with most MIMS fields non-blank. (E.g. country is excluded due to open ocean samples.) The last three rows indicate whether specific MIMS fields that would help data set discovery are present and/or useful. Feature and material fields were added to MIMS in the MIxS publication; pre-2011 data sets appear to have had those fields updated automatically by copying the *biome* field. Also note that depth is reported with respect to data sets with environmental package "water."

CAMERA and MG-RAST were early adopters of the MIMS metadata standard for minimal information about a metagenomic data set [43], and both require(d) compliance from uploaded data. MIMS fields include information about the habitat, latitude/longitude, depth, and sampling method. With the creation of standardized formats and ontologies for these fields [19], and software-assisted recording [13], metadata-based searches have become possible at sites that host MMG sets, or at earth sciences metadata/data repositories such as PANGAEA [36]. This is extremely useful for data set discovery, comparison on-site, and reuse (e.g. sets in Aim 1). However, non-compliance is an issue, whether due to errors or upload prior to the MIMS standard. Some errors can be programmed around, e.g. "pyrosequencing" and "454" are the same *seq_method*. Missing data, which is common, cannot. Some projects do not distinguish between their meta-*genomic* and transcriptomic sets, and +/- sign errors in longitude occasionally place ocean samples miles inland. Tab. 1.3 suggests that MMG sets on MG-RAST usually have required MIMS fields, though it has been reported that <10% of all sets on MG-RAST have complete minimal metadata by GSC standards [13].

1.4 DNA sequence assembly in metagenomics

Genome assembly is rarely possible for metagenomic samples, with notable exceptions (pyrosequencing,[181]; SOLiD, [72]; Illumina, [1]). However, given the shortness of NGS reads (Tab. 1.2) assembly of individual data sets can facilitate annotation [196] and lighten the burden on annotation servers. This section overviews the main approaches and challenges with metagenome assembly.

1.4.1 Reference-based and *de novo* assembly

Of the two main *genome* assembly approaches, the computationally cheaper reference-based approach presupposes a close reference genome against which to map newly sequenced reads. Often the goal is to identify variation, e.g. among human populations or pathogens, and in such cases the consensus of the mapped reads—that is, the assembly—is secondary. For microbial sequencing however, reference genomes are rare, and so one often uses a *de novo* assembler to accumulate overlapping reads

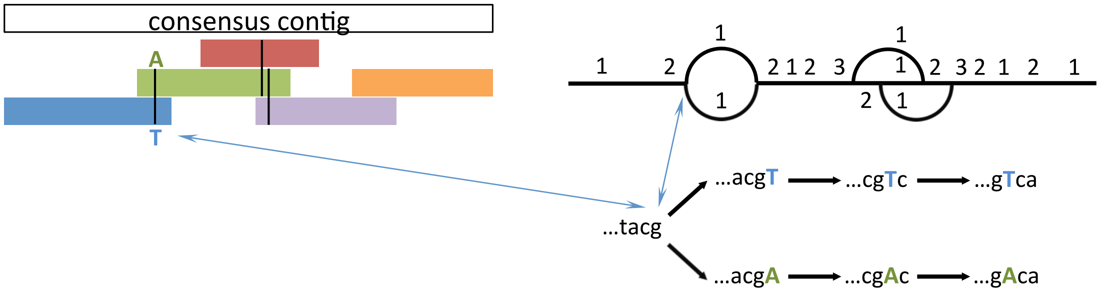


Figure 1.4: OLC and de Bruijn *de novo* assembly: (Left) OLC assembly of five reads (colored bars) with identical overlapping regions except for three single-base differences between indicated read pairs (black lines). (Bottom) Zoom-in of the de Bruijn graph at the blue/green mismatch, where the first bubble (right) begins. For each node, the last four of k total bases are shown. The bubble starts at a shared k-mer between the blue and green reads, and ends k nodes later where the reads are again sequence-identical. (Right) The de Bruijn graph for the same five reads. For clarity nodes are not shown. Edges count k-mers that occur in the reads (numbers), but reads themselves are not tracked. Small bubbles may be interpreted by an assembler as sequencing error and collapsed. Metagenomic assemblers use heuristics to break up more complicated graphs into subgraphs that represent different organisms.

into, ideally, few but very long contigs (hundreds of kilobases) that represent most of the genome. With the true size of the target genome unknown, a common statistic for describing the quality of an assembly is the N50 contig length.[17]

The vast differences in data volume produced by Sanger and 454 versus the latest Illumina and SOLiD platforms require different underlying representations of the reads in computer memory, which in turn motivate two fundamentally different *de novo* assembly strategies (Fig. 1.4). The overlap-layout-consensus (OLC) strategy came first and OLC assemblers have been used for Sanger sequencing projects, notably the Human Genome Project and GOS,[18] and also for 454. Intuitively OLC assemblers

[17] The N50 contig can be found by sorting the contigs by length and then adding, largest to smallest, the lengths until the sum reaches half the total of all contig lengths. The N50 contig is the one whose length was added last (consistent with definition in [115].

[18] The single-genome Celera Assembler [118] was used for the private effort of the Human Genome

pairwise compare all reads, and merge into consensus sequences (the contigs) reads that overlap better than user-specified length and percentage DNA identity thresholds. OLC algorithms assemble using a directed graph data structure: Nodes are reads, edges are overlapping reads, and contigs are strongly supported paths through the graph. Sequencing errors can be averaged out due to high read depth. For Illumina and SOLiD on the other hand, there are far too many reads to pairwise compare, or to represent each as a node in a graph. Instead, a so-called de Bruijn graph is built in which nodes are k-mers observed in the reads, and edges connect nodes whose k-mers overlap by their first/last $k - 1$ positions (Fig. 1.4). For example, if $k = 8$, then a read containing the nine base substring AATCAGCAT will introduce, or add weight to, an edge between nodes AATCAGCA and ATCAGCAT.[19] As for OLC, contigs are heavily-weighted paths through the graph.[20] Unlike for OLC, and importantly for chapters 3 and 5, individual reads are not contained within specific contigs; though tools for post-assembly mapping of reads to contigs exist (e.g. BWA [91]).

1.4.2 Metagenome assembly challenges

The fundamental assumption of single-genome assemblers is that highly similar reads represent the same region in one genome [119]. For an environmental sample that

Project and also for GOS. Today metagenomics assemblers are favored, some of them adapted from single-genome predecessors: Celera Assembler to CABOG [114], Velvet [205] to MetaVelvet [120].

[19] Typically $k \geq 31$, though $k \geq 20$ will produce k-mers that tend to be unique to a species [44]. Fixing k bounds the needed memory proportional to 4^k.

[20] The algorithms first simplify the graphs. Node chains with single in/out edges are merged to represent longer ($> k$) sequences, and bubbles collapsed when possible (Fig. 1.4, right)

contains hundreds to thousands of different operational taxonomic units (OTU's),[21] this is false. Consequently, similar reads that represent different OTU's will sometimes be assembled into "chimeric" contigs, with the frequency of such assembly-induced chimerism dependent upon interplaying factors: how conservatively one sets assembly parameters relative to sequencing error rates and how conserved the region represented by the reads is.[22] Whether chimerism is acceptable depends on one's goals. Limited chimerism can help annotation (chapter 3) and species discovery (chapter 5), but should be avoided in the creation of reference genomes or in studies of variation among closely related organisms.

Another consequence of the complexity of metagenomic samples is that coverage of the genomes will be sparse and non-uniform. This can cause problems for single-genome assemblers with heuristics based on the depth of assembled reads; e.g. low coverage paths through a bubble in de Bruijn graph can be mistaken for sequencing errors rather than a homologous region from a less abundant OTU [134].[23] Until recently single genome *de novo* assemblers were the only option, and generally performed poorly on complex metagenomic sets [134]. Several metagenomic assemblers have been produced, which in essence try to identify subgraphs of the de Bruijn graph that represent distinct species, with each subgraph ultimately reported as a contig. For example, the first metagenome assembler, Meta-IDBA, uses bubble size to discern sub-

[21] OTU's define a threshold percentage DNA identity, with respect to some marker gene, for considering two organisms to be different types. For 16S rRNA this is usually 97%.

[22] Assembly-induced chimerism is distinct from real chimerism due to horizontal gene transfer, and also from false chimerism due to DNA fragment ligation during sequencing.

[23] The number of reads for NGS depends on which version of the platform and reagents are used, while for Sanger sequencing the sequencing effort depends on budget (and clone library effort).

species[24] and multiply aligns the subgraph paths to create a contig [134]. MetaVeltet looks for subgraphs with distinguishable node coverage, reflecting different abundances of community members [120].

For this thesis we assembled 454 data sets with Newbler version 2.7, the OLC assembler from Roche. Reasons for choosing Newbler are given in chapter 3, along with a comparison to Genovo, a non-OLC metagenome assembler [89].

1.5 Annotation pipelines and data repositories

1.5.1 CAMERA, MG-RAST and others

Metagenomics servers are essential resources for annotation and long-term hosting of MMG data. By creating *de facto* standard annotation pipelines they enable comparison of species and functional diversity across data sets. By promoting metadata standards (§1.3.3) they enable reuse beyond initial publication. Servers also provide tremendous computational power and practical bioinformatics support, e.g. updated software and reference sequence databases, that a small lab could not afford/maintain.

CAMERA

The Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) was one of the early main servers and especially popular for MMG [165]. As with other servers, a web interface supported upload, analysis, and post-publication data sharing. Protein annotation by its RAMMCAP pipeline [92] en-

[24] Small bubbles correspond to short differences in mostly identical sequence. Big bubbles correspond to long differences; each path through the bubble likely comprises species-specific k-mers if $k \geq 20$ [44].

tailed ORF calls on the reads, followed by searches for every ORF (unlike MG-RAST) within Pfam, TIGRfam and COG.[25] CAMERA also supported on-site custom analyses of all hosted public data sets, e.g. via BLASTX searches against NCBI/nr (mirrored) or specialty databases (e.g. all 454 reads on CAMERA). It had also begun to support custom analysis via uploadable Kepler workflows. Alternatively, data sets and RAMMCAP results could be downloaded for in-house bioinformatics. While this flexibility (and computational generosity!) distinguished CAMERA from other servers, wait times were frustrating: BLAST-based workflows took a week, and small RAMMCAP analyses took days. CAMERA shut down in 2014 (§1.5.3). Its data sets are now hosted at iMicrobe, which aims to extend the existing iPlant cyberinfrastructure for data, tool, and computational resource sharing to the microbial research community [68].

MG-RAST

Since its start in 2007 the MG-RAST server has been a main workhorse for metagenomics annotation across biomes (host-associated, soil, marine and others), and now processes thousands of jobs per month [112]. Originally annotation was based on SEED subsystem analysis [128]. Uploaded data sets were searched against the SEED database to annotate *reads* with metabolic pathways or complexes, the subsystems,[26]

[25] A separate analysis path in RAMMCAP clustered ORFs to identify novel protein families (at 30% or 60% id as a heuristic). To speed up annotation, optionally just the cluster representatives could be searched in Pfam, TIGRfam and COG, and annotation transferred to cluster members. However Li et al. recommended against this: Sister ORFs might have been misannotated (e.g. if they had fewer protein domains than the cluster representative), and the annotation times for their test sets they thought fast enough (hundreds of CPU hours).

[26] A subsystem is a curated list of components ("functional roles") that constitute a metabolic pathway or complex, with a table that describes how organisms (rows) with that subsystem fulfill its functional roles (columns) with specific genes (cells).

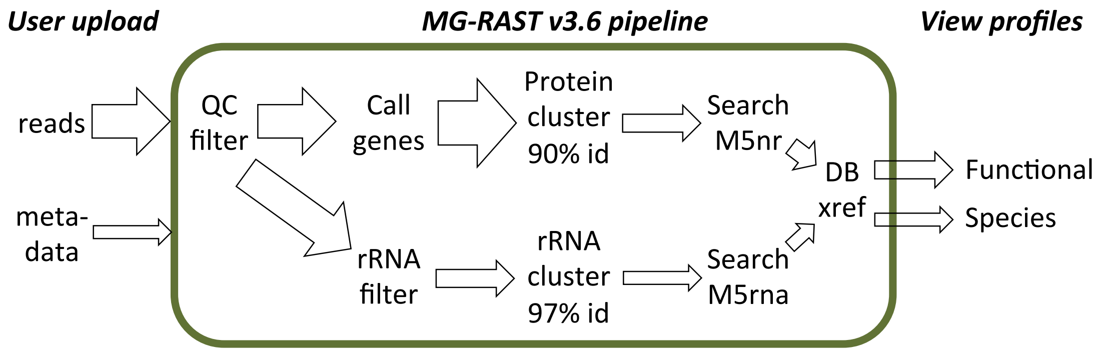


Figure 1.5: MG-RAST annotation pipeline overview: Arrow thicknesses illustrate changes in data volume. Gene calling can identify multiple “features” per sequencing read, thus data volume can grow. However, only cluster representatives are searched against the M5nr/rna databases, greatly reducing volume and increasing MG-RAST job throughput. Unique identifiers for hits in the M5nr/rna allow fast cross-referencing to reference databases (see text). Thus the user can view results with respect to a database of interest, and at specified E-value, percentage identity, and minimum length cutoffs.

thereby producing community composition and functional profiles that could be compared across data sets. At that time annotation typically took 1–2 weeks. In contrast to CAMERA, all annotation and on-site analysis and comparison among public sets was with respect to the pipeline; one could not run custom workflows using their cluster. However, one could download public sets, and results from intermediate stages, for local analysis.

This is still true as of MG-RAST version 3 (Fig. 1.5), except that annotation is now with respect to integrated protein and ribosomal databases, the M5nr and M5rna respectively. The former integrates major protein databases—RefSeq, GenBank, SEED, IMG/M, UniProt, KEGG, and others—into a single non-redundant protein database, such that hits to M5nr sequences can be mapped to the source databases that have them. The M5rna is analogous but integrates the major RNA databases from Silva, GreenGenes, and the Ribosomal Database Project (RDP). Another significant difference from CAMERA is that MG-RAST v3 does not annotate reads[27] directly, as this became too expensive especially with the rise of Illumina. Rather, uploaded sequences receive gene calls—possibly multiple depending on sequence length—and the predicted genes are then clustered (90% id). Just the cluster representatives are forwarded through the annotation stages of the pipeline, greatly reducing computation. Annotations (possibly multiple) are transferred from the representatives to their co-clustered genes. The rRNA steps are similar but predicted RNA is clustered at 97% id. Functional and species composition reported by MG-RAST are therefore estimates based on accumulated annotation counts.

Annotated public data sets can be viewed and compared with respect to their community compositions or functional profiles, using on-site tools to generate dendrograms, heatmaps or downloadable tables. For local analysis one can download public data sets at each step of the MG-RAST pipeline: as submitted, after quality control, at each stage of annotation. Moreover, all of this functionality, and all data, is available programmatically through a web services based API, which was used to create tables and figures in this section.

Other resources

The data sets used in this work were downloaded from MG-RAST, but several other annotation pipelines and/or resources should be mentioned. The Joint Genome

[27] Reads or contigs may be uploaded to MG-RAST, usually the latter for Illumina and SOLiD sets.

Institute's IMG/M[28], like CAMERA and MG-RAST, annotates 'omic data sets by similarity searches for predicted proteins against local copies of major external databases and also custom on-site databases [104]. It also provides web tools for functional and phylogenetic comparisons against metagenomes and genomes hosted at the sister resource IMG, which hosts genomes and metagenomes sequence by the JGI. Functional comparisons are with respect to protein annotations assigned by searching COG, Pfam-A, and TIGRfam. Phylogenetic comparisons are based on searches against a non-redundant database of reference genomes in KEGG and all genomes at IMG.[29] As a major sequencing center, the JGI sometimes has new reference genomes before e.g. NCBI RefSeq (see §1.2.3 and also §5.3.2.1), so the IMG search is a notable advantage of their pipeline. Also remarkable is that all predicted proteins are searched, not just cluster representatives as with MG-RAST.

The Genomes Online Database (GOLD) [130], also run by the JGI, catalogs externally hosted data sets and enables their discovery through metadata-based searches. GOLD follows the metadata standards of the GSC, in fact, IMG/M ensures compliance by starting data set submission with the creation of a GOLD record.

Sequence archives such as the NCBI Sequence Read Archive or the EMBL-EBI's European Nucleotide Archive (ENA) are useful for retrieving sequences from data sets already known, for example from journal article references. However, the EBI recently released a metagenomics pipeline [67] which was used (in addition to custom

[28] Technically IMG/M refers to the public sets and analyses. Registered users upload and analyze their sets using IMG/M ER (expert review).

[29] ORFs are assigned the taxa of their best BLAST hit against. The IMG database had over 13K sequences in late 2013. [104]

analysis) on the first 243 sets released by the Tara Oceans project [166]. Sequence and assay/library related metadata are available for those sets at the ENA, though much more is available at PANGAEA [36].[30]

The megx.net marine ecological genomics portal catalogs marine data sets and interpolates environmental data to help discover sets with shared environmental factors [97].

1.5.2 Biological interpretation challenges

Taxonomic and functional annotation of MMG data sets are the starting point for understanding the biological significance of a sample within its environmental context, and extrapolating to larger scales (§1.1). However, NGS read shortness and the bias of reference databases to the cultivable minority [197] have resulted in low annotation rates for MMG reads (e.g. Fig. 1.6). Illumina reads are still likely too short for reliable annotation but assembly should help.[31] However the second factor will remain a serious limitation (or discovery opportunity) for a long time. Examining <1% of the data sets collected on their ten year, ten expedition study, Tara Oceans estimated over 40 million mostly novel microbial genes [166], which is coincidentally the total number of bacterial and archaeal RefSeq proteins (release 72, July 2015).

[30] Photos of casting records for each collected sample are available, beyond the call of the GSC!

[31] Based on the observations by Wommack et al. (2008) in their Sargasso Sea simulation. Simulated reads with lengths >150 nt had markedly more, and more presumably correct, BLAST hits (Fig. 4C) [196]. The Illumina contig N50's from Tara Oceans were 1121±176 nt [171].

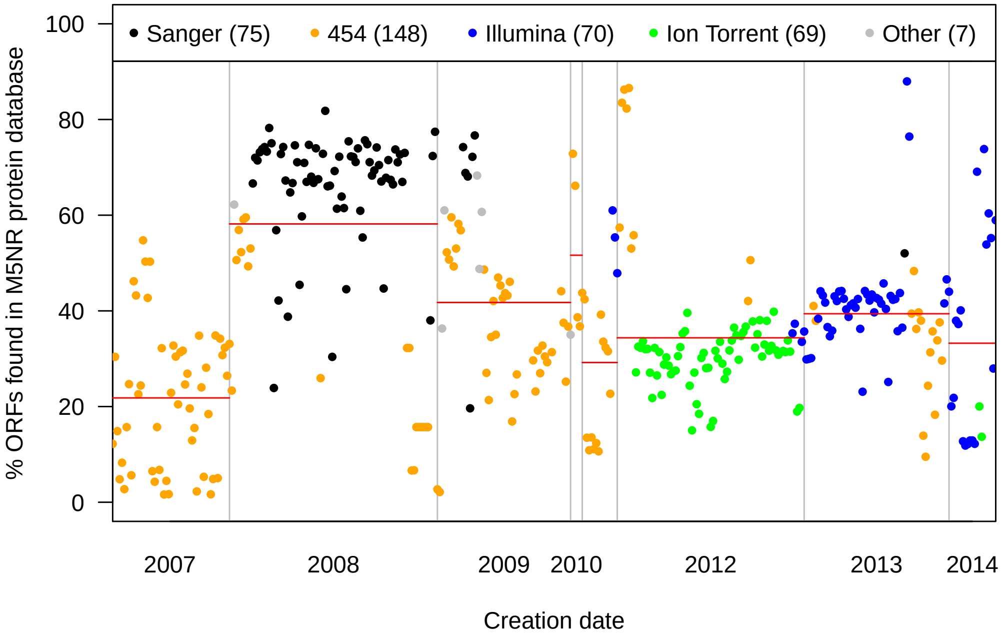


Figure 1.6: Protein annotation for MMG data sets on MG-RAST. Predicted genes (cluster representatives only) for each data set were searched against the M5nr integrated protein database (late 2012; E-value <0.001). Shown are 339 of the 402 sets in Tab. 1.2. GOS comprises 71 of the 75 Sanger sets. Red lines show mean annotation rates.

1.5.3 Practical challenges: computation, storage, funding

Computational challenges

The computational bottleneck for annotating new metagenomic data sets was recognized by 2009 [154] when the NCBI RefSeq database had 9.3 million protein sequences. Today RefSeq has nearly 52 million (40 million microbial), and annotation servers, which spend most of their time performing sequence similarity searches on reference databases [170], may take a week to produce annotation often for less than half the reads in a data set. With uploads of terabase Illumina sets and thousands of jobs submitted per month,[32] the computational demands on annotation servers have become extreme. These factors have already forced MG-RAST into the cloud,[33] with the understanding that cloud-based analysis with current algorithms is financially infeasible for the long-term [193, 194].

Storage challenges

Storage will also challenge data hosts as sequencing costs continue to decrease much faster than storage costs [103, 60, 189]. This is especially problematic in metagenomics where annotation can scale the required storage by five orders of magnitude [125]. Using the samples as a storage medium might work in some cases [167, 143]. However, data sets from resequenced marine samples could vary greatly due to the high species complexity and intra-population variation versus limited DNA capture per sequencing run. There is also the computational cost of reanalysis to consider, and that not storing raw reads would stifle the development of methods that depend on them (e.g. assemblers, as well as the approaches developed in this thesis). Another proposed solution for NGS data sets is to compress them by reference to sequenced genomes [60]. For MMG data sets however, lack of close references for many reads will make this approach problematic.

[32] MG-RAST alone processes thousands of jobs per month. Tang et al. (2014) also note that it processed 1 Tbp in its first five years, and ∼50 Tbp in the next three (10 Tbp from the Human Microbiome Project) [170].

[33] MG-RAST 3.0 uses distributed resources from Argonne National Laboratory, the DOE Magellan cloud, and Amazon's EC2 [193]. Moving all MG-RAST computation to Amazon's EC2 was decided too costly in 2009 [194, 170].

With NSF proposals now requiring data management plans [122], and the uncertainty of funding for centralized repositories (see below), data hosting is being pushed upon the producing laboratories. Indeed, MG-RAST deletes private data sets older than 120 days, and gives lower priority to new jobs from users with unpublished data sets. Conceivably this will result in only a subset of all produced MMG data sets being available at centralized repositories. If so, we will need approaches to discover and share MMG data sets that are hosted at the producing laboratories.

Funding

The NCBI Sequence Read Archive nearly shut down in 2011 due to budget constraints [172]. Ironically, an editorial discussing its planned closure mentioned CAMERA and MG-RAST as alternative resources for metagenomics data [172], but CAMERA shut down in July 2014 after losing support from the Gordon and Betty Moore Foundation [20]. MG-RAST remains, but its August 2015 newsletter warned that the lack of funding opportunities could interfere with system maintenance. One would hope for continued funding for the two U.S.-based annotation pipelines, MG-RAST and IMG/M. They are tremendously valuable resources for the metagenomics community in terms of: computer clusters, bioinformatics expertise, and the assurance of comparable results due to standardized analyses and metadata enforcement.

Chapter 2

Results overview

2.1 Challenges addressed by this work

Two fundamental challenges discussed above are the fragmented, sparse nature of NGS MMG data sets, and the "cultured bias" of the reference sequence databases by which we interpret them. This work responds to these challenges by leveraging a strength of MMG data, the large, growing body of public data sets. By pooling MMG data, the approaches developed in this work have shown promise for gaining new biological insights despite these two challenges.

2.2 Specific aims

Aim 1

I proposed to show that pooled assembly of MMG data sets could be used to increase species and functional annotation.

I hypothesized that pooled assembly of MMG data sets would produce weakly chimeric contigs and thereby enable increases in species and functional annotation. This was correct. In simulated data, chimerism was rarely above the NCBI Taxonomy level of "species," giving justification that species annotations for pooled contigs would be biologically relevant. With real data, both species and functional annotation increased with pooling, sometimes substantially. Moreover, there were indications that unannotated contigs could be a resource for discovering new species.

Aim 2

For Aim 2 I proposed to investigate strategies for ranking MMG data sets to guide selection for efficient pooled assembly.

I hypothesized that ranking MMG data sets based on the similarity of k-mer profiles of the metagenomes, and assembling in rank order, would enable efficient pooled assembly. This was correct. Ranking and assembly of data sets in order of 5-mer profile similarity was comparably efficient (in reads assembled per unit computation) to ranking and assembly in order of phylogenetic similarity, and significantly better than assembly without ranking.

Aim 3

For Aim 3 I proposed to discover at least one uncharacterized marine microbial species represented in public MMG data using an approach I developed, "geographic profiling."

I hypothesized that pooling MMG data sets would enable the discovery of ubiquitous and abundant marine microbial species and partial characterization of their genomes, without need for culturing them. This I showed by applying geographic profiling to public, pooled-assembled MMG data. Three novel species were predicted in multiple MMG sets and substantiated with orthogonal lines of evidence. Experimental work corroborated predicted sequence fragments for each of the species, and analyses of these fragments supported that they likely represent abundant, ubiquitous novel species.

Chapter 3

Improving annotation by pooled, chimeric assembly

3.1 Introduction

This chapter has identical content to the paper 'Pooled assembly of marine metagenomic datasets: enriching annotation through chimerism' [100] mentioned in the Acknowledgements.

High-throughput DNA sequencing of microbial communities (metagenomics) has greatly expanded our knowledge (reviewed broadly in Schloss and Handelsman, 2005 [151]). For example, new marine biology, a better understanding of biogeochemical cycles and more are keenly anticipated from this (currently still main) option for characterizing unculturable species at the molecular level [9, 12, 37, 72, 148, 183]. In marine metagenomics (and for samples derived from other habitats), full realization of the potential of next-generation sequencing (NGS) relies on high-quality species and functional annotation of the now hundreds of datasets available to the scientific community [52]. However, metagenome annotation is confounded by (i) the shortness of NGS reads ([196]; Temperton and Giovannoni (2012) review this challenge and others [174]), which merely characterize small fragments of genes, and (ii) by the small fraction of known microbes represented in sequence databases, often described as "the culturable 1%" [4, 197]. In traditional genomics, assembling reads into contigs is common practice. By contrast, metagenomes are primarily released to the community as reads, predominantly from Sanger or pyrosequencing ("454") to date, and even some "gold standard" annotation pipelines are optimized for unassembled data. Concern over chimeric contigs may explain the reluctance to assemble. For example, only 3 of the 42 datasets in this work were assembled according to the publications in which they first appeared. This may seem justified to an extent by pyrosequencing read lengths, which typically range from around 100–600 nt in publicly available sets, depending on how recently the data were obtained. While some annotation can be derived from individual reads in this range [177, 196], it is plausible that assembly should add value to the data by improving annotation. However, the potential benefits for annotation have not been quantified rigorously.

Previous smaller-scale work with simulated assemblies already suggests that chimeric contigs may not be as common as feared. For example, Mavromatis et al. (2007) [105] reported that 85–95% of contigs from assemblies of the most species rich set of their simulated Sanger-sequencing data did not mix genomes. Neither did the majority of contigs assembled from simulated pyrosequencing reads in other work by the Moya group [94% in [135]; 68–97% in [182]]. Below, we give this question a deeper look, but we also take a step further to ask: Is it advisable for deriving species and function annotation, to cautiously consider not only data from one sample, but also from others? By allowing limited chimerism within and across samples we may strike a balance between information gain and precision (with respect to an individual cells genome) that is beneficial if used strictly for annotation. For example, a contig identified as *Trichodesmium erythraeum* is informative even if the contig mixes strains of this bacterium, whereas refusal to call the species because a contig fails to match precisely a known strain tells us nothing.

3.2 Methods

3.2.1 Marine RefSeq database

Marine RefSeq database: To assign reads and contigs (simulated and real) to marine taxa, we created a searchable, non-redundant database of all known marine bacterial and archaeal protein sequences in NCBI RefSeq (August 2012). The database included 2,585,466 sequences from 754 taxa. See §3.5.2.

3.2.2 Assembly

Assembly: Intra-set assemblies were done with Roche's GS *de novo* Assembler ("Newbler") v2.7 and required read overlaps of ≥40 nt at ≥90% identity. Pooled assemblies used the same parameters. However, because of the large number of reads (8,730,323) in the 42 real sets, we pooled via an iterative approach in which contigs aggregated reads over multiple rounds of assembly. Simulated sets were also assembled with Genovo [89] to investigate algorithm impact on chimerism. See §3.5.2.

The proliferation of new algorithms and tools to assemble (e.g.) millions to billions of (Illumina e.g.) reads has delayed/prevented “selection of standard best practice tools.” [152]

3.2.3 Species annotation and correctness estimation

Species annotation: To obtain conservative species calls for this discussion (§3.3.3), we applied the following protocol, to real and simulated data. Marine RefSeq was searched with each read or contig using translated nucleotide UBLAST [40] with E-value cutoff 1E-1. Only hits exceeding bit score 110 and 90% id were used for consensus-based species calls. For reads, perfect species consensus of the hits was required. For contigs, calls were made only if the species entropy of the hits was 0.242—a 100-read contig with three minority reads, each from a different species, would have this entropy—and only if the hits covered >33% of the contig. (Coincidentally, Charuvaka and Rangwala (2011) [22] measured chimerism the same way.) For simulated sets, correctness of consensus-based species annotation was evaluated by comparison with the true species of the sequence. For reads, the source genome was always known. For contigs, the true species was the majority species of the constituent reads. See §3.5.2.

Estimating the odds of correct species calls: To compare the odds of correct species calls for unassembled versus intraset-assembled versus pooled-assembled reads, we counted correct and incorrect calls (but not non-calls) for reads and contigs from the 10 simulated datasets. We also modeled the impact of species unknown to Marine RefSeq (to estimate the risk of incorrect species calls because of homologous hits in Marine RefSeq) by creating 10 decoy sets in which non-marine NCBI RefSeq genomes served as proxies for unknown marine species. Application of Bayes' Theorem produced the Figure 3.3 odds curves, with probabilities of a correct or incorrect species call estimated from the observed counts. See §3.5.1 and §3.5.2.

3.2.4 Protein annotation

Protein domain annotation: Open reading frames (ORFs) were called on reads and contigs with FragGeneScan v1.16 (model `454_10` was used allowing a 1% error rate [144]) and used to search Pfam A release 26.0 [140] with HMMER v3.0 `hmmscan` [default parameters with trusted cutoff bit scores defined for each hidden Markov model (HMM) [38]]. The per-domain "independent" E-value served as proxy for annotation confidence. Hits to Pfam domains of unknown function were ignored in our evaluation.

3.2.5 Gene family annotation

Gene family annotation: ORFs were also searched against PANTHER 8.1 [113] gene family and subfamily HMMs using the PANTHER script `pantherScore.pl`. Reads were annotated only if they or the contig they belonged to aligned to an HMM with E-value <1E-23, recommended by PANTHER.

3.3 Results and Discussion

3.3.1 Simulated sets approximated species diversity of their real counterparts

To evaluate chimerism and annotation, we created simulated metagenomic datasets from reference sequences (complete genome sequences from known species). Simulated metagenomes in which the provenance of individual reads is known have contributed greatly to method development and discussion of experimental data in this field [22, 105, 110, 135, 187]. To ensure that any impact of chimerism in simulation is likely predictive for real data, we created 10 simulated sets (s prepended to names) using the program MetaSim [147], based on the taxonomic profiles of 10 real datasets in which 559 species (727 taxa) were identified. Thus, we aimed to approximate specific real sets with respect to species composition and total base pairs. In effect the simulated sets included data from 317 species in total (394 taxa; Fig. 3.1). This deficit in complexity, compared with the real sets, was caused by MetaSim reaching the target dataset size before sampling all genomes requested in each input profile. Nonetheless, to our knowledge, our simulated sets are superior in complexity and "realism" to previously published studies, and approximated their real counterparts well, in aggregate and individually. Both in the real and the simulated sets, a small number of identifiable species contributed ∼99.9% of the reads [63 species (real), 56 (simulated), with 51 species in common]. Individually, each real set and its simulated version overlapped by 12 species on average in their respective 20 most abundant species (data not shown).

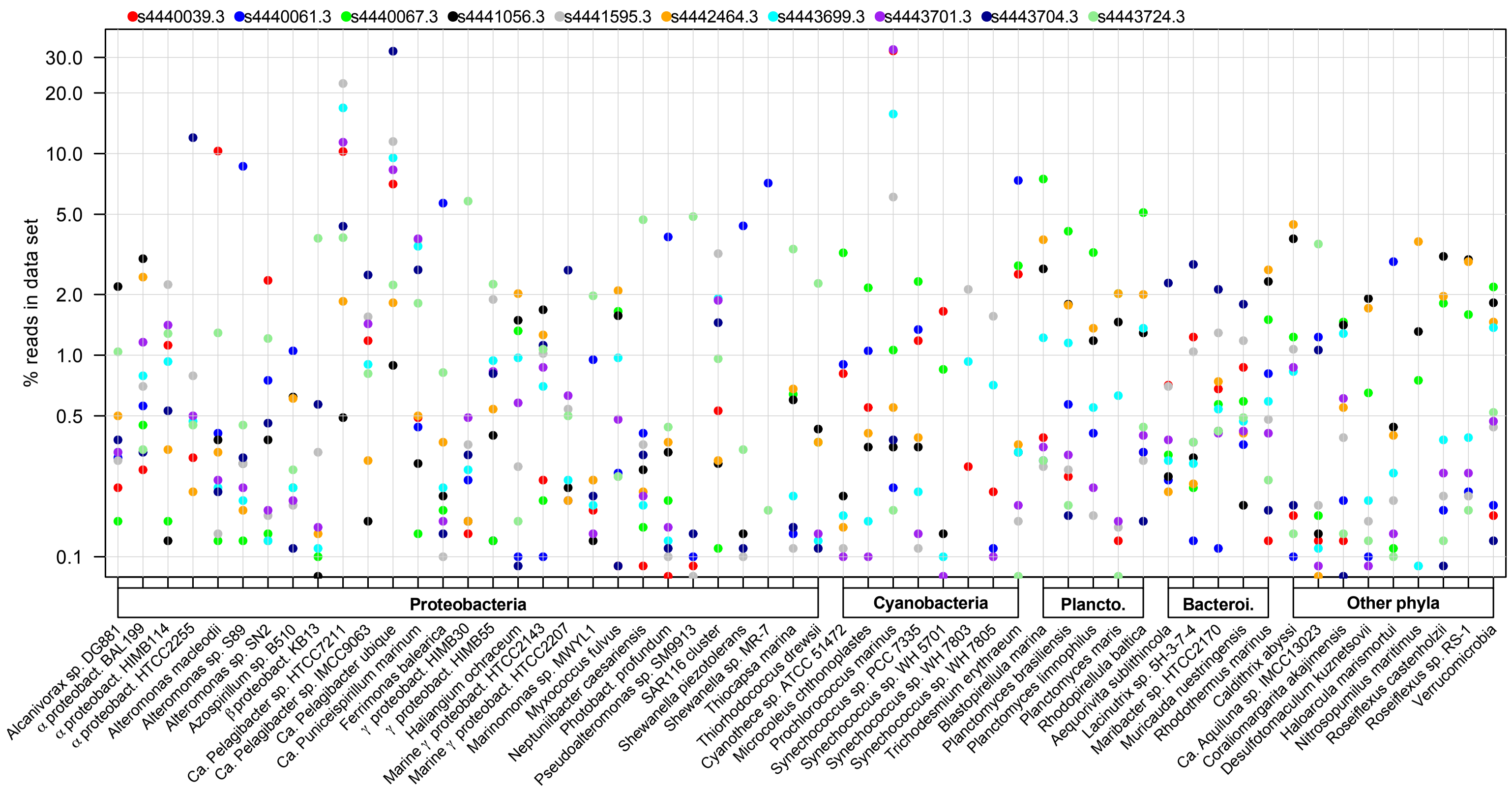


Figure 3.1: Species composition of simulated datasets. The 56 species shown contributed 99.9% of the total reads across all 10 simulated datasets. Sixty-three species (51 of them shown above) contributed 99.9% in the real data counterparts (data not shown). On average, each simulated set and its real counterpart shared 12 of their top 20 species. The commonly used cutoff (UBLASTN top hit >50 bit score) was used for this illustration, which may yield false-positive matches (see text). Abbreviations: Plancto, Planctomycetes; Bacteroi, Bacteroidetes; bact, bacterium. See also §3.5.1

Several patterns are evident in Figure 3.1. First, many species are found in several sets, as is evident from the nearly 10 datasets shown in each species column. Indeed, 73% (230/317) of species appeared in more than eight of the simulated sets. This was also true of the real sets: 76% (425/559) of species appeared in more than eight sets. Nine of the sets derived from subtropical or tropical waters, four of them from the mid-Pacific, but the high proportion of shared species seems noteworthy nonetheless. This may also reflect that marine environmental samples harbor many rare species in the tail of their taxa distributions [132] and that homology-based species annotation at standard stringency risks to (mis)associate their reads with a related species. Based on this observation, we use only a highly conservative species calling protocol in our analyses below (see §3.2).

Second, there is similarity across some sets for species within the same genus. For example, sets s4441595.3 (gray), s4443699.3 (cyan), s4443701.3 (purple), s4440039.3 (red), s4443724.3 (light green) and s4442464.3 (orange) show the same order by read abundance of three *Candidatus Pelagibacter* species: HTCC7211, ubique, IMCC9063. These proteobacterial species are members of the SAR11 cluster, ubiquitous and abundant in marine microbial communities [117], but only 13 complete or near-complete *Candidatus Pelagibacter* genomes have been released (seven of them of *ubique* strains). One daring explanation for the pattern could be that just one uncharacterized *Candidatus Pelagibacter* species is actually present in the real samples, which has three close relatives among the species represented in our database (which is deemed closest may vary depending on which parts of the uncharacterized genome are represented by the reads). This might be a more plausible explanation than three species appearing in the same rank order in six samples, which is also possible, of course. However, two of the six datasets (4443724.3 and 4442464.3) are from distant sample sites from the other four (mid-Pacific). We cannot resolve the cause conclusively but note that we found reminiscent rank order patterns in the corresponding real datasets as well, while for example Cyanobacteria seemed devoid of this phenomenon. Our conservatism in species calling in the analyses described below serves to ensure the validity of their results in either case, although it necessarily leads to higher proportions of "non-calls" than metagenomicists are used to seeing in the literature (50–80% [177]).

3.3.2 Assembly of simulated sets yielded mostly strain- and few species-chimeric contigs

3.3.2.1 Chimerism and impact on species calls

Intuitively, we categorize chimerism at three levels: whether a contig combined reads from different species, or different strains (from the same species), or different individuals (cells). While it is extremely relevant for biodiversity questions and even systems biology, the third level is of less interest here (neither positively nor negatively) than the first two levels. Admittedly species and strain classification criteria can neither be expected to be straightforward, especially for microorganisms, nor consistent. However, this distinction respects the taxonomical classification of species, and pursues our intuitive premise that strain chimerism will unlikely impact on annotation negatively (when annotation protocols are largely homology-based). When Newbler was

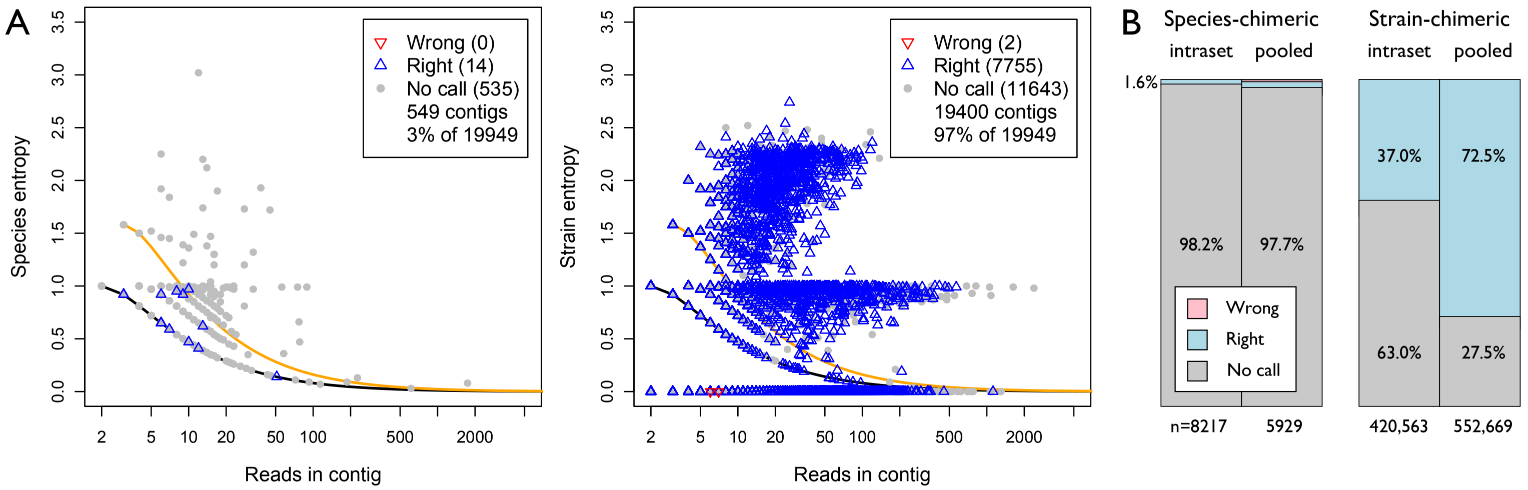


Figure 3.2: Contig chimerism and its impact on species call correctness in simulated data. **(A)** Size and chimerism are shown for contigs from all 10 simulated datasets, separately assembled. There were 549 species-chimeric contigs (A, left) and 19,400 species-homogenous contigs (A, right). For reference, the black curves show the Shannon entropy if exactly one read differs in taxon from all others in the contig. The orange curves show entropy if a contig has two minority reads each from a different taxon. Of the 549 (3% of 19,949) species-chimeric contigs, 513 have entropy ≤1 bit. There are 237 contigs within 0.01 bits of the black curve and 157 on the curve. Of the 19,400 (97%) contigs that are chimeric below the level of species, 14,019 (70%) have entropy 0 bits (not chimeric) and another 3160 have entropy ≤1 bit. Corresponding plots after pooled assembly are in §3.5.3. **(B)** Despite chimerism species calls for reads, based on the calls for contigs into which they assembled, were usually correct. Pooled assembly of the 10 sets increased the number of assembled reads (n) and nearly doubled the proportion of correct calls for reads in strain-chimeric contigs

used to assemble the 10 simulated sets individually ("intraset"), with a minimum required read overlap of 40 nt at 90% identity, i.e. standard parameters, this yielded mostly small contigs (Fig. 3.2A; average N50 contig size of 1734±897 nt), but of high quality, given the hundreds of species present in each sample. Of the 19,949 intra-set contigs 97% (19,400) were species-homogenous and 70% (14,019) were strain-homogenous (Fig. 3.2A). Only 3% (549) were chimeric at the level of species or higher. Moreover, the degree of chimerism, measured as Shannon entropy of the species or strains of the reads within a contig in consideration of its length (plot in Fig. 3.2A), is low. For example, 157 of the 549 species-chimeric contigs had only one read that differed in species from all others in the contig. Accordingly, chimerism did not impact the correctness of species calls made with our conservative protocol (see §3.2). If we defined the true species of a contig to be the majority species of its reads (there was almost always a strong majority, which is evident by the low entropies) few contigs elicited incorrect species calls: 0 species-chimeric contigs and only 2 strain-chimeric contigs (Fig. 3.2A), all short (510 reads).

We noted that 43 reads received species calls before assembly (i.e. raw reads) that were challenged by the calls made for the contigs into which they were later assembled. The contig-based call was correct for 27 of these reads. These are marginal occurrences by reference to the data volume analyzed (∼1.1M total reads in the 10 sets) of which >428K we found in contigs (Fig. 3.2B).

When probing dependency of our findings on the type of assembler and data used (§3.5.2 and §3.5.3; Figs. 3.10–3.12), we noted higher species-chimerism rates (10%)

for intra-set contigs with Genovo as a non-overlap-layout-consensus (non-OLC) algorithm example. Species-misannotation with non-OLC (8.7% as opposed to 0.0% with Newbler) was containable by eliminating short contigs, but we recommend OLC. Generally, assembly seems worthwhile for annotation if such caution is applied, although the extent of improvement will vary.

Next, we assembled all 10 sets in a "pooled assembly" (see §3.2) and observed a modest increase in the number of contigs (19,949 intra-set versus 21,408 pooled; Fig. 3.9). This is not surprising given that intra-set contigs that recruit additional reads when pooled leave the count unchanged and intra-set contigs that merge decrease the count, i.e. only newly assembled reads from different datasets would be expected to increase the contig count in our experiment. More relevant, the read count assembled in contigs increased by 12% through pooling (429K intra-set versus 559K pooled, Fig. 3.2B) and the proportion of in-contig reads with correct species calls with our conservative species caller nearly doubled (37% intra-set versus 73% pooled). While we acknowledge that metagenomic contigs will mask some diversity [34], these are striking improvements due to pooling.

3.3.2.2 Species calls with consideration of novel marine species

Figure 3.3 answers the question: Does assembly improve the odds of a correct species call for a read, given that only a fraction of species are represented in Marine RefSeq? The odds ratio curves weigh correct against incorrect species calls, by counting the reads in three categories: unassembled, intraset-assembled and pooled-assembled.

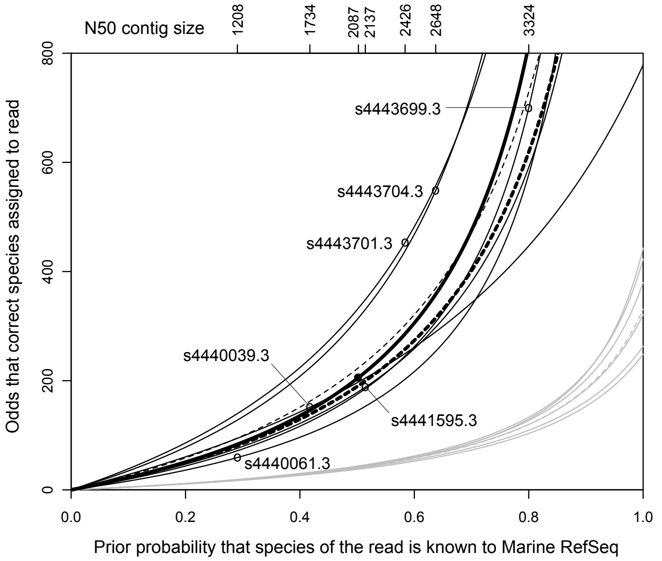


Figure 3.3: Odds of correct species calls for reads that are unassembled, intraset-assembled and pooled-assembled. The odds ratio curves account for whether a read is from a species represented in Marine RefSeq or not. Each gray curve is based on measured rates of correct and incorrect species calls for unassembled reads in a simulated set and 10 decoy sets (§3.5.2). Similarly black curves represent calls for assembled reads. Only datasets for which 41% of reads assembled are shown. The dashed curves show the mean odds ratios with no assembly (gray), intra-set assembly (black), or 10 pooled assemblies with each set withheld (bold). The solid bold curve shows the odds when all 10 sets are pooled. The discs show the N50 contig size for each assembly and suggest that the odds of a correct species call for a read increase with N50

(Reads are not reflected if no species calls were made with our conservative protocol (see also Fig. 3.2B)). The six marine sets for which 41% of the reads assembled show a clear improvement from unassembled (gray) to intra-set (black), in step with the N50 contig size, i.e. the higher the N50, the better the odds. The odds curve for pooled assembly of all 10 sets (bold) is consistent with its N50 contig size. This is consistent with a population of pooled contigs that somewhat resembles the intra-set populations but has contigs with more deeply aligned reads.

3.3.3 Assembly of 42 real data sets improved species annotation

Encouraged by the low species-level chimerism we observed in simulation, we selected 42 real marine microbial metagenomic sets from the MG-RAST public repository [112] and assessed the impact of assembly on species identification. All data were obtained with pyrosequencing ("454") technology. Figure 3.4 (left panel) shows the proportion of reads that elicited calls before assembly (gray) and after (colored). For each dataset, the proportions are illustrated as two bars, for intra-set (upper bar) and pooled (lower) assembly, respectively. Sets are arranged top to bottom by increasing degrees of assembly, measured as intra-set N50 contig size (right panel), and tended to experience greater increases in reads for which species calls were made (colored) the more they assembled. The correlation between intra-set call rate and N50 was strong (Kendall's $\tau = 0.53$), in spite of being impacted by some sets that assembled well but yielded few (or no) species calls with our protocol (see below).

The percentage of new species calls (blue), i.e. calls possible only after assembly, increased with intra-set N50 ($\tau = 0.48$; max 5.9% for 4444077.3). Pooling further increased the new call rate (max 6.6% for 4444077.3) by adding reads to intra-set contigs or creating new ones. In this context, the ratios of new calls to other kinds (green, yellow) depend on the conservativeness of the species caller, of course. A less conservative protocol would more readily assign species to raw reads and thus preempt potential new calls on assembly, though at a cost in misidentifications of homologs for known species as we discussed above (§3.3.2).

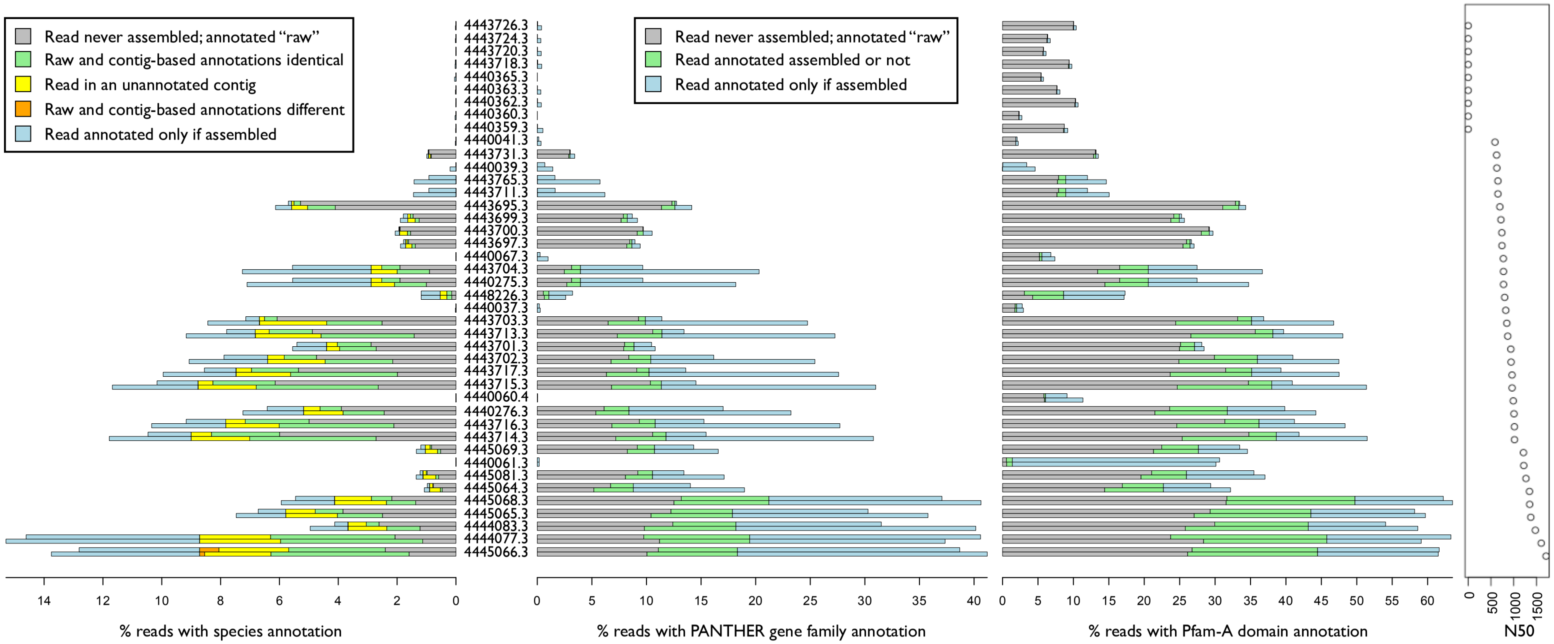


Figure 3.4: Species, PANTHER and Pfam-A annotation changes because of intra-set and pooled assembly of real datasets. **Left:** Species call change (quantity) relative to the total read count in each of the 42 real datasets. Intra-set bars are above the corresponding pooled assembly bars in each pair. Datasets are sorted vertically by increasing intra-set N50 contig size (far right): 589 nt for 4440041.3, and 1711 nt for 4445066.3. Gray: reads that did not assemble i.e. elicited annotation only as individual raw reads. Colored: reads that assembled into contigs and elicited annotation (see legend). Middle: A comparable graphic of gene family annotation increases for intra-set and pooled contigs, based on alignments to PANTHER HMMs, mirrors the species results. Annotation increases with N50 with some of the same exceptions (e.g. 4440061.3). **Right:** Pfam-A domain annotation also increases with assembly. This was especially pronounced for some sets with poor species and gene family annotation despite N50>0 nt (e.g. 4440061.3)

Often, individual reads elicited species calls before assembly (gray) that could be compared with calls based on contig membership. Usually, a contig-based call provided corroboration for the same species as was called beforehand (green), and more so for sets with a higher intra-set N50 ($\tau = 0.57$). Pooling increased the proportion of corroborated calls in each set. For example, the intra-set assembly of 4443715.3 produced contigs that supported calls for 2.1% of reads in this set (green); pooled assembly increased support to 4.1% of these reads. Pooling also tended to increase the proportion of reads for which raw species calls could not be corroborated (yellow), i.e. no call was made for the relevant contig. On inspection, such contigs typically recruited too few translated UBLAST hits from our Marine RefSeq protein database to cover 33% of their length (required for a species call, see §3.2). Rarely, a discrepancy arose between the contig-based and the pre-assembly calls for a read (orange). With these real data, we cannot firmly establish which (if either) of the calls is accurate. However, from the trends observed with simulated data (Fig. 3.3) it seems reasonable to predict that assembly will have helped rectify an incorrect species assignment more often than not.

The species annotation impact of pooling over intra-set assembly for each dataset is also captured by the ratio of pooled to intraset-assembled reads in a set (i.e. the ratio of colored regions in Fig. 3.4). On average, pooling increased species calls by 15% (minimum 0% for 4448226.3, maximum 57% for 4443711.3), in the 27 sets with >0.5% total species calls. Novel species might be responsible for the lack of improvement in the other 15 sets where <0.5% of reads elicited a species call. These 15 sets also yielded comparatively low annotation rates in MG-RAST (0-26.1%). For example, sets 4440060.4, 4440061.3 and 4440067.3 received no calls by our species caller in spite of assembling well (with N50 values of 991 nt, 1221 nt and 773 nt, respectively). These sets derive from a phage study of marine microbialites in which viral metagenomes from the same locations also had low annotation rates (>97% of reads unidentified; [35]).

3.3.4 Assembly of 42 real data sets improved domain and gene family annotation

3.3.4.1 Pfam-A protein domain annotation

We investigated whether assembly would improve annotation of Pfam-A domains (Fig. 3.4, right panel). As with species calls, domain assignments increased after assembly, in positive correlation with N50 ($\tau = 0.75$). Notably though, the number of post-assembly domain assignments far exceeded those of species. For the 33 intra-set assemblies with acceptable N50 values, the proportion of corroborated or new domain assignments was 0.1–40% (mean = 10.6%), compared with the 0.0–12.5% (mean = 2.1%) for species calls. Considering only new annotations, domain assignments ranged from 0.0-29% (mean = 5.9%) and species calls from 0.0–5.9% (mean = 1.0%). The conservation of domains across taxa is illustrated nicely by this increase and domain/family annotation is where we anticipated (and saw) the potential benefits of assembling chimeric contigs from close relatives best reflected. Set 4440061.3 is an extreme example: Neither raw reads nor contigs matched to the data/species in Marine RefSeq, yet 30% of reads assembled into contigs for which domain assignments resulted (using a standard HM-

MER3 protocol, see §3.2). Interestingly, for this set ~40% of the domain assignments were to Bacteriophage Replication Gene A (PF05840) and ~1% to Capsid Protein F (PF02305), possibly reflecting an abundance of prophages, which were not in Marine RefSeq.

Some reads that assembled could not be mapped to pooled contigs for technical reasons (This caused the fractions of reads in pooled-assembled contigs shown in Fig. 3.4 to fall below the corresponding intra-set fractions (e.g. 4445066.3 Pfam domains), even though we define pooling as a superset of intra-set results. We note that the consequence with respect to our evaluation is merely that the annotation gains depicted in Fig. 3.4 are in fact underestimates.

3.3.4.2 PANTHER gene family annotation

To gauge whether assembly would improve identification of multidomain protein architectures (a step toward functional annotation), we used raw and assembled reads in searches against PANTHER 8.1 gene family HMMs. The PANTHER results strongly mirrored the species call results, as is evident in Fig. 3.4. Pearson correlations corroborated this symmetry for total annotations ($r_{intraset}$=0.68, r_{pooled}=0.71) and also for only the new calls enabled by assembly (blue; $r_{intraset}$=0.85, r_{pooled}=0.80). The surprising symmetry between the plots was not reasonably explained by shared reference species between PANTHER HMMs and our database. Only five of the 147 PANTHER species (UniProt Reference Proteomes) were in our database. Therefore, the much higher annotation rates for gene families versus species (Fig. 3.4) suggest that the PAN-

THER models identified contigs representing marine species not in our database. For example, set 4445081.3 has few species calls (even in MG-RAST, only 14.9% of reads have protein or rRNA annotation) but many gene calls, likely because of conserved, homologous genes in uncharacterized species.

3.3.5 Prospects for species discovery

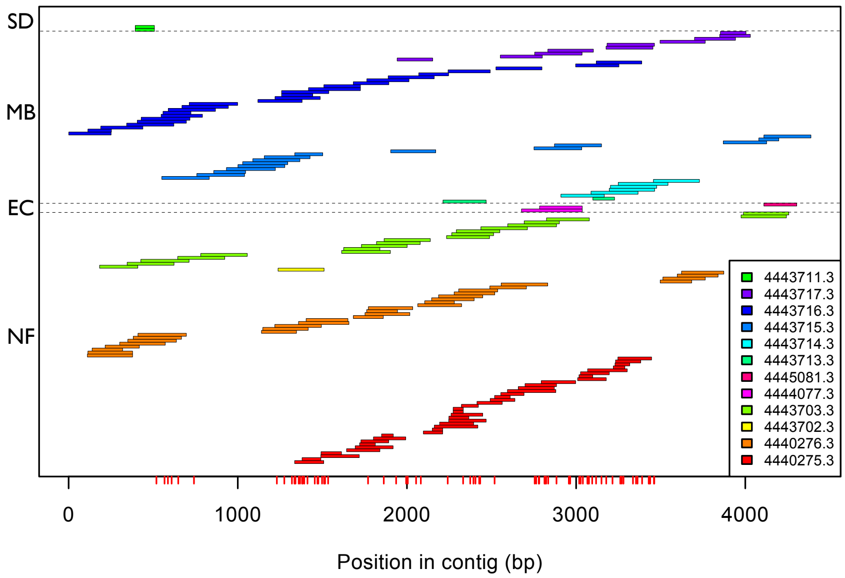


Figure 3.5: Pooled contig03272 was composed of reads from 12 samples from four geographic regions: San Diego (SD), Monterey Bay (MB), western English Channel (EC), near a Norwegian fjord (NF). Although reads lacked significant hits to reference databases, the contig was mappable and may represent a novel, geographically diverse type of γ-proteobacterium. Ticks on the bottom axis show 76 positions where reads from each geographic region were nearly unanimous but differed across regions (as described in the text). These proportions are not likely to occur by chance and may reflect geographically specific differences

Can pooled assembly of metagenomic data help discovery of novel species, especially those that are ubiquitous? To answer this, we first checked whether pooled contigs that lacked species calls by reference to our database Marine RefSeq, which is a protein database and also excluded, e.g. marine viruses (see §3.2), would perhaps elicit hits to NCBI RefSeq genome data. We searched with all 3806 unannotated contigs from pooled assembly with lengths >5kb using a CAMERA BLASTN workflow (E-value <0.1; [165]). The vast majority of the contigs (3688) failed to align at >80% sequence identity over >80% of their lengths to any reference genomes (3779 failed if 90% cutoffs were set). Thus, most of the contigs likely represented novel microbial species.

To illustrate the potential of pooled assembly in this context, we investigated a single pooled contig that lacked a species call, contig03272, in detail. This 4388 bp contig was picked at random from a list of contigs that had not elicited a species call, contained reads from multiple samples, and were >4kb in length. These criteria served to enrich for contigs that might represent novel ubiquitous species and that would overlap multiple genes. Our example contig03272 was built from 148 reads from 12 surface water samples from widely separated oceanic regions (Fig. 3.5). Read depth was relatively uniform (2 reads minimum; mean = 8.9 reads) and all reads mapped to the contig (BLASTN) over nearly their full lengths (96.5% minimum; mean = 99.7%) and with high sequence identity (92.7% minimum; mean = 97.5%). Two computationally predicted protein sequences aligned to over 78% of contig03272 (BLASTX search of NCBI/nr, E-value <10), both of them proteins from SAR86 clade members (B and A): a Dehydroquinate Synthase (71% and 65% id) and a Membrane Carboxypeptidase/Penicillin-binding protein (Mrca; 59% and 62% id). These predicted proteins are encoded by neighboring genes in both SAR86 genomes. At the time of analysis, these observations in combination suggested to us that contig03272 might represent a geographically widespread, novel marine microbe.

Indeed, during the preparation of this manuscript, a draft genome assembly for the γ-proteobacterium SCGC AAA076-P13 (NCBI BioProject accession PRJNA195664) was submitted to the NCBI as results of a Joint Genome Institute (JGI) project applying single-cell sequencing technology to 30 ubiquitous, uncultured marine microbes. The timing is fortuitous, as the new data corroborated both the ubiquity of the microbe represented by contig03272 and that our contig had been assembled correctly. BLASTN of contig03272 against the draft SCGC AAA076-P13 genome assembly yielded coverage of 97% of the contig, in three match alignments with 97%, 97% and 95% sequence identity, respectively.

Interestingly, the draft genome assembly was based on a sample collected in the Gulf of Maine, far from the four oceanic regions that contributed to contig03272 (Fig. 3.5). As the latter were widely separated, population-level differences between the isolates of this new species (SCGC AAA076-P13) should be apparent in our data. A cursory inspection of non-unanimous positions in the contig corroborated this. Of the 106 non-unanimous positions covered by 8 reads, 76 were near-unanimous (we allowed one read not to match) within oceanic regions. Random reshuffling of the geographic labels in the same contig (same read positions and label proportions), as a null model, failed to attain this degree of segregation coincidentally (P $\approx$ 0; 0 instances in 10,000 trials).

3.4 Conclusions

While the notion that annotation (quality and quantity) should improve with assembly is intuitive, this is to our knowledge the first deep effort at quantifying the benefits of strain-chimeric contigs from a practical perspective for metagenomics. In simulation and for real datasets, we showed that intra-set and pooled assembly of marine metagenomic data (i) produce chimeric contigs that rarely violate marine species boundaries; (ii) lead to quality and quantity improvements in species, protein domain and gene family annotation; (iii) may help identify geographically diverse but novel species and population differences within those species. We focused exclusively on data from pyrosequencing ("454") efforts as the currently predominant type in marine metagenomic repositories. However, most key concepts and the strategy of our analyses are transferrable to other sequencing technologies. With their read lengths now surpassing those from the first "454" generation, Illumina-sequenced sets may be the next data to investigate.

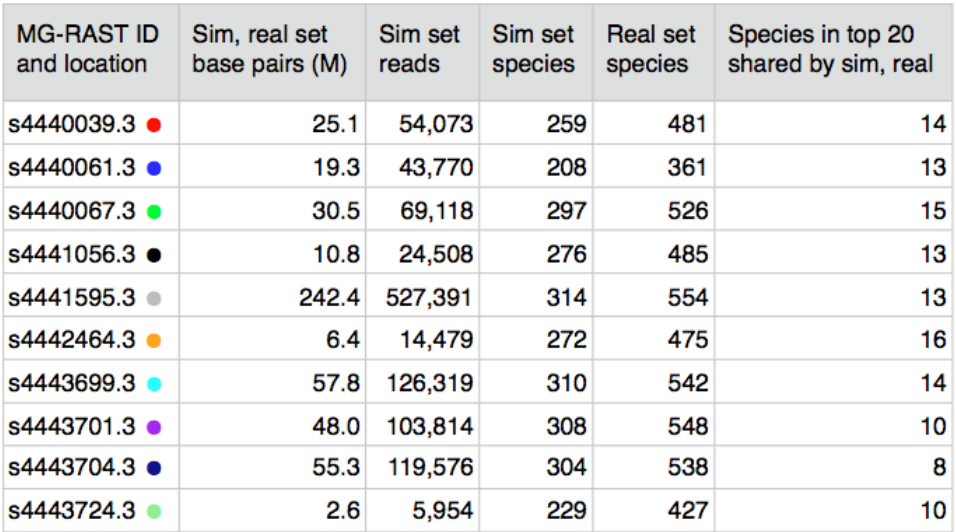
3.5 Supplementary Materials

3.5.1 Data

Note: As of June 2014, MG-RAST hosts over 160 public pyrosequenced sets representing 104 sample locations and 34 investigators. In contrast, today there are few public Illumina-sequenced marine metagenomic sets on MG-RAST (12, all from the same study and sample location). We anticipate similar benefits to annotation by assembling sets derived from Illumina technologies, which now produce reads longer than those from early pyrosequencing (including many in this study) and at far greater depths.

3.5.1.1 Simulated metagenomic data sets

It stands to reason that an estimate of chimerism ought to be based on simulated samples that are faithful to the community compositions (species and their relative abundances) of real samples from the habitat of interest. Moreover, organisms not known to live in that habitat should be excluded. For this reason we selected ten real metagenomic data sets to model simulated counterparts. Of the 559 species identified in ten real data sets (Fig. 3.6), 317 were represented in their simulated counterparts (Tab. 3.1). This required that we supplement the MetaSim genome database, which contains only complete, prokaryotic RefSeq genomes, with contigs from unfinished marine genomes that were abundant in the real sets. For the thirty-three such unfinished genomes, we added their RefSeq contigs >100kbp to the MetaSim database. (We excluded shorter contigs to avoid a bias toward assembly of non-chimeric contigs.) The following unfinished genomes occurred in two or more of the simulated data set profiles: *Candidatus Pelagibacter* spp. HTCC7211 and *ubique* HTCC1002; *Prochlorococcus marinus* str. MIT 9202; SAR116 cluster α-proteobacterium HIMB100; α-proteobacteria spp. BAL199 and HIMB114; and γ-proteobacteria spp. HIMB30, HIMB55 and HTCC2207.



MG-RAST ID and location	Sim, real set base pairs (M)	Sim set reads	Sim set species	Real set species	Species in top 20 shared by sim, real
s4440039.3 ●	25.1	54,073	259	481	14
s4440061.3 ●	19.3	43,770	208	361	13
s4440067.3 ●	30.5	69,118	297	526	15
s4441056.3 ●	10.8	24,508	276	485	13
s4441595.3 ●	242.4	527,391	314	554	13
s4442464.3 ●	6.4	14,479	272	475	16
s4443699.3 ●	57.8	126,319	310	542	14
s4443701.3 ●	48.0	103,814	308	548	10
s4443704.3 ●	55.3	119,576	304	538	8
s4443724.3 ●	2.6	5,954	229	427	10

Table 3.1: The ten simulated data sets with their real counterparts from which the MetaSim species profiles were made. Colors correspond to Figs. 3.1 and 3.6.

Even with a loose cutoff (best translated nucleotide UBLAST hit >50 bits in Marine RefSeq), the majority of reads in some real sets could not be assigned to a genome (Fig. 3.6). One might expect a simulated metagenomic set based on a low percentage of called reads in its real counterpart, yet scaled to the same number of base pairs (Tab. 3.1), to show a bias toward assembly. We did not observe this to be a problem. For example simulated set s4440061.3 was based on 4% of reads called in the real set but its intra-set N50 fell below those of five other simulated sets based on rates from 32–94% (Fig. 3.6, see Fig. 3.3 for N50s). Set s4440061.3 had fewer reads than six sets, but this did not likely mask any strong bias toward assembly. Even the largest set by far, s4441595.3, was based on a low call rate (36%) but had only the fourth largest N50.

The N50 contig sizes for the intraset-assembled simulated marine sets averaged 1700±897 nt, or 1.6–5.8 times the mean length of a simulated read (450 nt). The high standard deviation is due to the wide range of data set sizes, some of them too small for much assembly (Tab. 3.1), and the unequal proportions of species. Pooling the simulated sets increased the N50 contig size to 2087 nt, unsurprising given the high level of shared species (73% of species were shared by eight or more sets; Fig. 3.1).

3.5.1.2 Decoy metagenomic data sets

To estimate the impact of unknown species (i.e. those not in Marine RefSeq) to the annotation of unassembled, intraset-assembled, and pooled-assembled sequences (Fig. 3.3), we created ten decoy sets from non-marine, microbial, NCBI RefSeq genomes. Four community composition profiles, each with 86–100 equally proportioned species, were input to MetaSim to create the decoy sets. This species complexity was allowed to be lower than those of the simulated marine sets in order to encourage assembly, so that the denominator in Equation 3.3 (in §3.5.2.4) would be based on many assembled decoy reads (107K in fact, for intra-set). This slightly lower complexity made the odds curves in Fig. 3.3 conservative. Each decoy set had 100K reads, with the same mean read size (450 nt) and error rate (1%) as the simulated marine sets. The mean N50

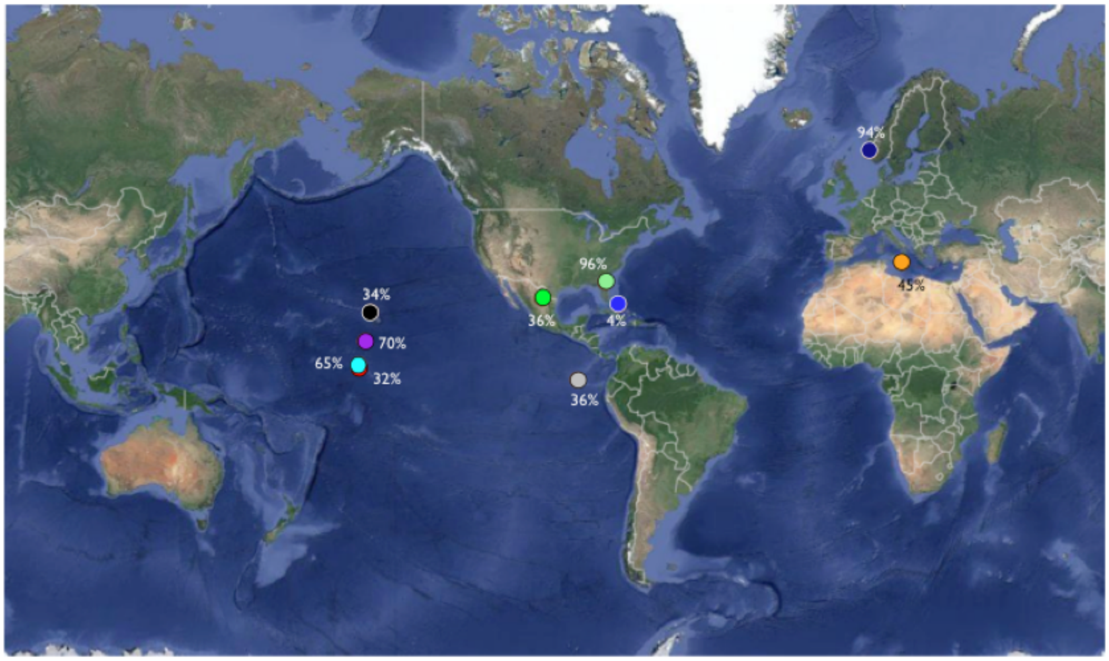


Figure 3.6: Sample locations of the real data sets on which the ten simulated sets were based. Beside each set is the percentage of reads with species calls (best hit >50 bits). Profiles based on these species calls were input to MetaSim to create the simulated data sets. A copy of this map with each sample location linked to its MG-RAST summary page is at `https://mapsengine.google.com/map/edit?mid=zdoMT32zZMKM.kR4y-S6rMR_M` and also provided in `SF2.SimulatedSamples.kml` (importable by GoogleEarth).

contig size for intra-set assembly of the decoy sets was 1306±18 nt.

3.5.1.3 Real metagenomic data sets

Selection from MG-RAST

The forty-two marine microbial metagenomic samples were selected, from published sets at MG-RAST, to span a broad geographic range and set of principal investigators (Fig. 3.7). They are summarized in Tab. 3.2.

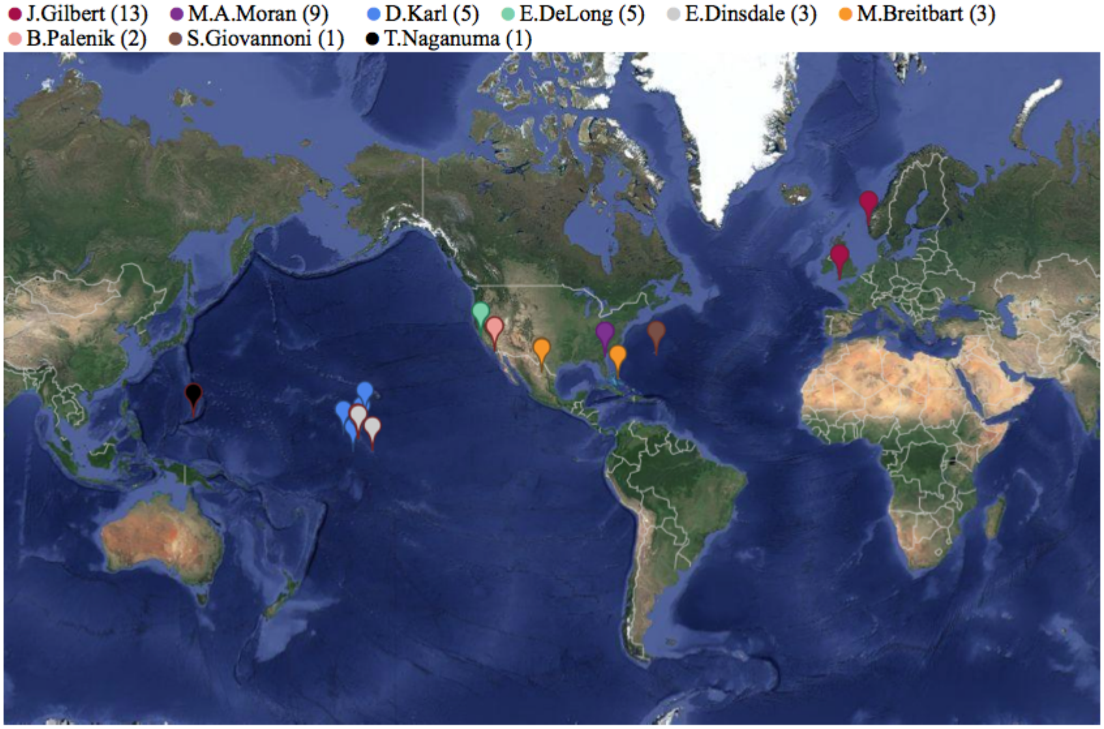
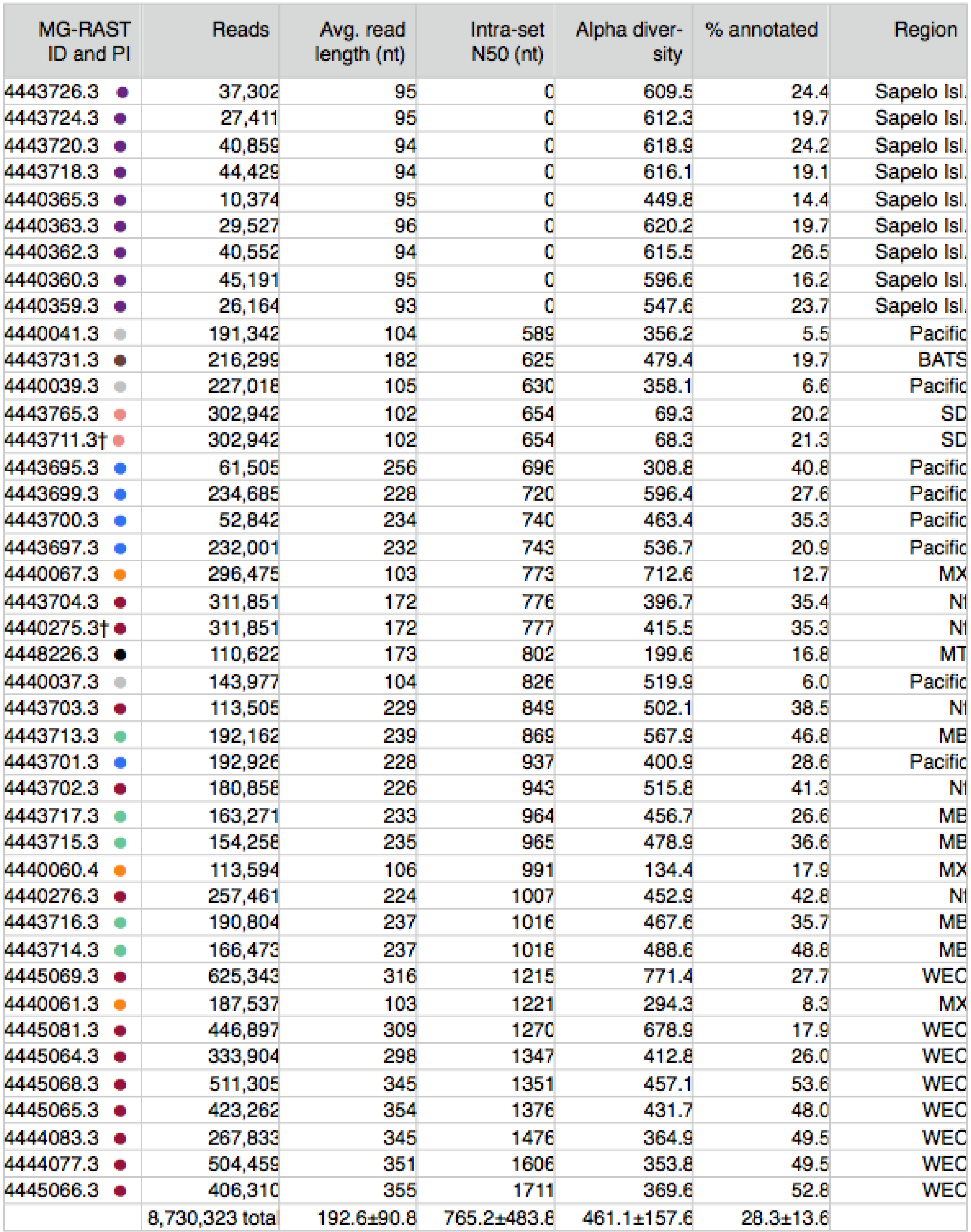


Figure 3.7: Sample locations of the forty-two real data sets colored by principal investigator, with number of sets in parentheses. Original map at `https://mapsengine.google.com/map/edit?mid=zdoMT32zZMKM.k0exTQ4KG1OE`, or in the KML file `SF3.RealSamples.kml` (importable by GoogleEarth). The original and KML maps include links to the MG-RAST summary pages for each sample, which include metadata (project information, sample collection coordinates, links to publications, etc.) and results from MG-RAST analysis.



MG-RAST ID and PI	Reads	Avg. read length (nt)	Intra-set N50 (nt)	Alpha diversity	% annotated	Region
4443726.3 ●	37,302	95	0	609.5	24.4	Sapelo Isl.
4443724.3 ●	27,411	95	0	612.3	19.7	Sapelo Isl.
4443720.3 ●	40,859	94	0	618.9	24.2	Sapelo Isl.
4443718.3 ●	44,429	94	0	616.1	19.1	Sapelo Isl.
4440365.3 ●	10,374	95	0	449.8	14.4	Sapelo Isl.
4440363.3 ●	29,527	96	0	620.2	19.7	Sapelo Isl.
4440362.3 ●	40,552	94	0	615.5	26.5	Sapelo Isl.
4440360.3 ●	45,191	95	0	596.6	16.2	Sapelo Isl.
4440359.3 ●	26,164	93	0	547.6	23.7	Sapelo Isl.
4440041.3 ●	191,342	104	589	356.2	5.5	Pacific
4443731.3 ●	216,299	182	625	479.4	19.7	BATS
4440039.3 ●	227,018	105	630	358.1	6.6	Pacific
4443765.3 ●	302,942	102	654	69.3	20.2	SD
4443711.3† ●	302,942	102	654	68.3	21.3	SD
4443695.3 ●	61,505	256	696	308.8	40.8	Pacific
4443699.3 ●	234,685	228	720	596.4	27.6	Pacific
4443700.3 ●	52,842	234	740	463.4	35.3	Pacific
4443697.3 ●	232,001	232	743	536.7	20.9	Pacific
4440067.3 ●	296,475	103	773	712.6	12.7	MX
4443704.3 ●	311,851	172	776	396.7	35.4	NH
4440275.3† ●	311,851	172	777	415.5	35.3	NH
4448226.3 ●	110,622	173	802	199.6	16.8	MT
4440037.3 ●	143,977	104	826	519.9	6.0	Pacific
4443703.3 ●	113,505	229	849	502.1	38.5	NH
4443713.3 ●	192,162	239	869	567.9	46.8	MB
4443701.3 ●	192,926	228	937	400.9	28.6	Pacific
4443702.3 ●	180,858	226	943	515.8	41.3	NH
4443717.3 ●	163,271	233	964	456.7	26.6	MB
4443715.3 ●	154,258	235	965	478.9	36.6	MB
4440060.4 ●	113,594	106	991	134.4	17.9	MX
4440276.3 ●	257,461	224	1007	452.9	42.8	NH
4443716.3 ●	190,804	237	1016	467.6	35.7	MB
4443714.3 ●	166,473	237	1018	488.6	48.8	MB
4445069.3 ●	625,343	316	1215	771.4	27.7	WEC
4440061.3 ●	187,537	103	1221	294.3	8.3	MX
4445081.3 ●	446,897	309	1270	678.9	17.9	WEC
4445064.3 ●	333,904	298	1347	412.8	26.0	WEC
4445068.3 ●	511,305	345	1351	457.1	53.6	WEC
4445065.3 ●	423,262	354	1376	431.7	48.0	WEC
4444083.3 ●	267,833	345	1476	364.9	49.5	WEC
4444077.3 ●	504,459	351	1606	353.8	49.5	WEC
4445066.3 ●	406,310	355	1711	369.6	52.8	WEC
	8,730,323 total	192.6±90.8	765.2±483.8	461.1±157.6	28.3±13.6	

Table 3.2: The forty-two real data sets listed by increasing intra-set N50 contig size. The number of reads we downloaded are given; these were quality-filtered by MG-RAST. The average read length, alpha species diversity, and percent annotated are as reported by MG-RAST. The geographic regions we assigned: Bermuda Atlantic Time Series (BATS), San Diego (SD), Mexico (MX), Norwegian fjord (Nf), Mariana Trough (MT), Monterey Bay (MB), western English Channel (WEC). Colored dots designate the principal investigator as given in Fig. 3.7. Additional information on each data set is available at MG-RAST via the Google map link in the Fig. 3.7 caption, or by visiting `http://metagenomics.anl.gov/metagenomics.cgi?page=MetagenomeOverview&metagenome=MGID` where MGID is an MG-RAST identity. † indicates the set is a duplicate of the set in the previous row. Excluding the Sapelo Isl. sets which did not assemble, the average N50 contig size was 973.8±300.6 nt.

Similarity of the forty-two data sets

The numbers of reads contributed by data sets to pooled contigs should depend on the similarities of the underlying communities sampled. E.g. sets from the same geographic regions would be expected to overlap in their most abundant organisms and thus to show much cross-set assembly. Retrospectively, that is what we found when we measured for the pooled contigs all pairwise contributions by the forty-two sets (Fig. 3.8). However, while within-region cross-set assembly was most common, pooled contigs also often recruited reads from different regions, for example from San Diego, the Monterey Bay and the western English Channel, all coastal surface water samples.

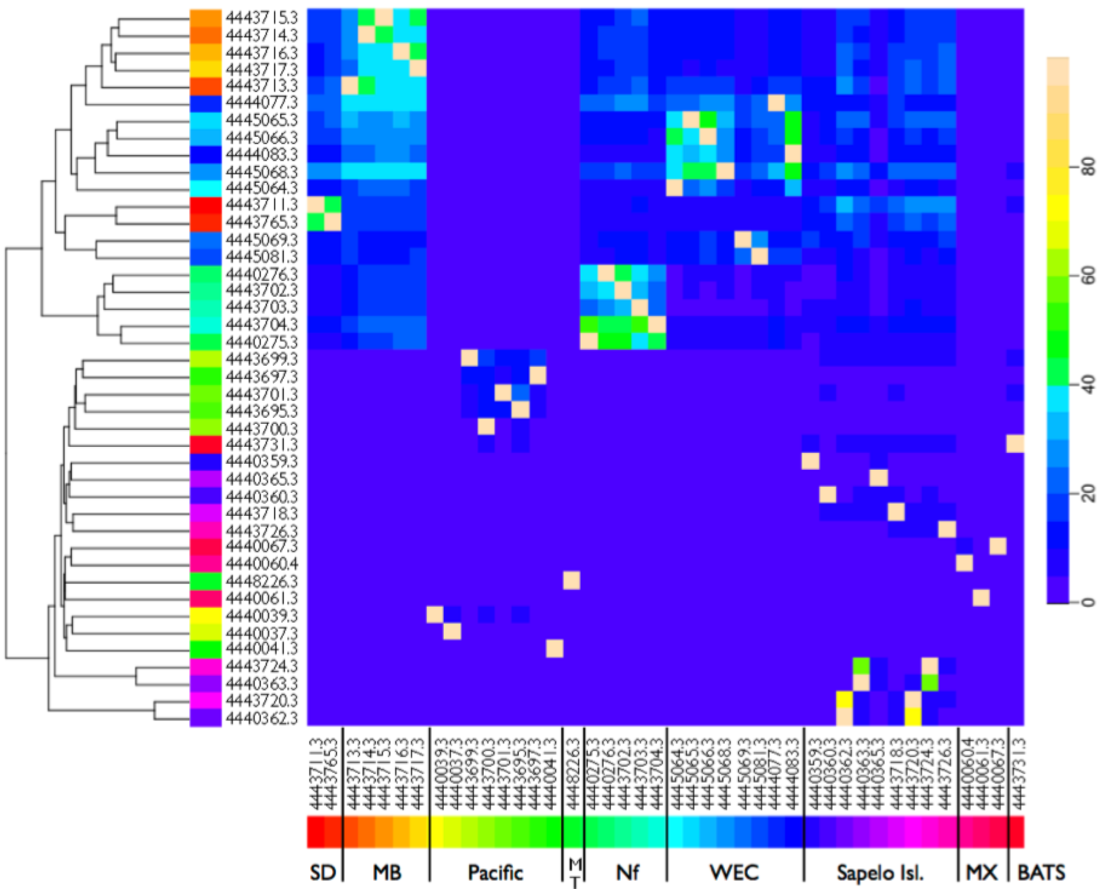


Figure 3.8: Pairwise intersection of data sets within the pooled contigs. Heatmap columns organize sets by their geographic regions (defined in Tab. 3.2). Each cell in column j and row i represents the proportion of all the pooled contigs that had reads from data set j that also had reads from set i. Thus column j can be viewed as a profile for how data set j contributed reads to the same contigs as each other set. (The lightest cells represent 100% of contigs shared between data set j with itself.) Rows cluster the data sets by column similarity (Euclidean distance). To aid visualization, data sets are colored by their geographic regions, so one can see where clusters respected those regions. For example contigs that had reads from several Monterey Bay sets (orange) were common while contigs that had reads from the Mexico sets and any other geographic region were rare. Overall, for 35.4% of the cross-set cells the proportion of pairwise shared contigs is at least 5%.

3.5.2 Methods

3.5.2.1 Marine RefSeq database

To create a searchable, non-redundant database of all known marine bacterial and archaeal species, NCBI RefSeq [139] was searched with taxids for all marine microbes in the Gordon and Betty Moore Foundation's Microbial Genome Sequencing Project (`http://camera.calit2.net/microgenome`), and those in the Genomes Online Database [130] with habitat "marine" or "sea water." The supplementary file `SF1.MarineRefSeq.txt` lists the 754 marine taxa represented among the ∼2.6 million protein sequences in Marine RefSeq. For each taxon are given the NCBI taxid and the number of protein sequences contributed to Marine RefSeq.

3.5.2.2 Assembly with Newbler

Rationale

Newbler version 2.7 was selected because it is a mature assembler specifically created for pyrosequencing reads. It provides copious text output on assembly results, including the assembly details for every input read, e.g. whether the read assembled and into which contig. Only reads with status "Assembled" in `454ReadStatus.txt` output files were counted toward chimerism and annotation measurements. "PartiallyAssembled" reads we excluded because the ends of such reads map poorly to contigs.

The following command line parameters were used for both intra-set and pooled assemblies: `-ml 40` and `-mi 90` for reads to overlap at least 40 nt and 90%

id; `-rip` to ensure reads appeared in at most one contig; `-urt` to use the entire read at contig ends; `-notrim` to disable primer trimming, as reads were already quality filtered.

Tracking of individual reads, a feature of overlap-layout-consensus (OLC) assemblers such as Newbler, was essential for quantifying chimerism and species call accuracy. Another advantage of using an OLC assembler for our marine metagenomic data was the ability to construct contigs from reads representing low-abundance organisms in the presence of sequencing errors and strain variation. By comparison a de Bruijn assembler would have required us to check post-assembly that reads mapped at nearly their full lengths to contigs. More significantly, sequencing errors and strain variation would have precluded exact matching of k-mers from the rare overlapping reads contributed by low-abundance organisms [119]. (Although pre-assembly correction of sequencing errors is a often done for Illumina-sequenced single-genome samples, this seems a poor strategy to deal with strain variation, and was impractical in our case because the reads spanned so many different samples and pyrosequencing generations.) Indeed, MetaVelvet generally assembled an order of magnitude fewer reads from our simulated data sets (counting the five sets for which both Newbler and MetaVelvet assembled >1000 reads) and into shorter contigs (MetaVelvet N50s were 2–10 times smaller; data not shown) [120].

Round robin pooled assembly of the real metagenomic data sets

While Newbler had no problems assembling individual data sets, we could not get it to complete the assembly of all forty-two data sets at once (i.e. as a single assembly project). (Despite weeks of infrequent updates to the status file `454NewblerProgress.txt`, Newbler never finished.) As a workaround, we used an iterative, round robin approach where for each data set, an initial set of intra-set contigs aggregated the intra-set unassembled reads from all the other data sets. This entailed 42×41 assemblies. Reads could only assemble into a single contig, ensured by synchronized updates to the FASTA files of unassembled reads by forty-two parallelized assembly jobs. The contigs from the 42×41 assemblies were then assembled into the final set of pooled contigs. Note that this approach did not exhaust pairwise encounters of unassembled reads.

The round robin approach required a workaround to Newblers 2kbp maximum input sequence length. To input previously assembled contigs >2kbp to Newbler, we tiled them at length 1kbp with 800 bp overlap. This is similar to the approach used by Wurm et al. (2011) [199]; however our tiling parameters more often, though not always, caused Newbler to reassemble the original contigs. When Newbler did reassemble tiles, it was possible to unambiguously trace reads in the tiles to a final pooled contig. When Newbler did not reassemble tiles—not even at a much later iteration—then the final pooled contigs of the reads in those tiles were ambiguous. This happened for 1,159,998 of the 2,969,128 reads that pooled-assembled. For 395,077 of these ambiguous-contig reads, their final pooled contig could be resolved by searching for them against the pooled contigs (UBLAST nucleotide search of the reads at 95% sequence identity, 100% of read aligned). The reads that could not be resolved caused the pooled percentages in Fig. 3.4 to sometimes fall below the intra-set percentages for PANTHER and Pfam-A results. This did not happen for species annotation. Even if the final pooled contig of a read was ambiguous, it was often possible to assign it to a species when all of its candidate final pooled contigs had been called as the same species. Note that the round robin approach was not necessary to pooled-assemble the ∼1.1E6 reads in the ten simulated sets.

Newbler assembly of the simHC metagenomic data from Mavromatis et al.

For comparison we assembled simHC, the simulated high species complexity FAMeS set [105], with Newbler with parameters used throughout this work, except we included quality scores for the 116,771 real, Sanger-sequenced reads. (Quality scores were not available for the pyrosequenced data sets.) For measuring chimerism we used BLASTN [3] to identify reads that mapped to contigs at >90% id and >95% coverage. PartiallyAssembled reads we did not automatically exclude, as done for our marine set assemblies, because such reads often mapped well to contigs. They may have been marked as PartiallyAssembled due to low-quality end sequence. The 793 reads that mapped at our cutoffs belonged to 435 contigs, shown in Fig. 3.12.

3.5.2.3 Chimerism rates with a non-overlap-layout-consensus assembler

We wanted to know whether one might expect substantially different chimerism rates with a non-OLC assembler. To this end the Genovo *de novo* assembler was ideal because it is based on a fundamentally different approach—a probabilistic model is used to infer a set of contigs that are likely to have produced the observed reads—and is specifically designed for pyrosequenced metagenomes [89]. We performed only intra-set assemblies with Genovo, all with forty iterations. This was sufficient for convergence based on the read status files, which showed that no reads had changed contigs by the fortieth iteration. One of our data sets was significantly larger (s4441595.3 in Tab. 3.1) than the largest set for which Genovo was developed and tested (311K reads). Nevertheless, all but one (s4443704.3) could be assembled with Genovo given sufficient time on our Linux cluster (e.g. CPU times of 10 min for s44437.4.3, 11 hrs for s4440039.3, and 309 hrs for s4441595.3). We could not however complete a pooled assembly of all ∼1.1E6 simulated reads. (Only eighteen iterations completed after twenty-six days on our cluster.)

3.5.2.4 Species annotation for simulated data sets

True species assignment

Correctness of consensus-based species annotation was evaluated by comparison to the true species of the sequence. For reads the source genome was always known by MetaSim annotation in the FASTA defline. For contigs the true species was defined as the majority species of the constituent reads, but only if the species entropy of the reads was ≤0.242. (Contigs can grow to hundreds of reads, therefore it seemed reasonable to tolerate some mis-assembly, here about three reads in a 100-read contig.) This entropy was exceeded for 539 intra-set and 695 pooled, simulated contigs. During species annotation (by our consensus-based caller) these contigs were only permitted to be non-calls or wrong-calls.

Species assignment for decoy sets

Species calls made on reads in decoy sets, or contigs assembled from those sets, were considered incorrect because they mistakenly identified novel species for homologs in Marine RefSeq. The proportion of such calls was used in the application of Bayes' Theorem to produce Fig. 3.3.

Application of Bayes' Theorem for Fig. 3.3

For each category of simulated read (unassembled, intraset-assembled, pooled-assembled), probabilities of making a correct or incorrect species call were estimated from observed counts of correct and incorrect calls. Note that contigs could contain both correctly and incorrectly called reads; the true species of each read was compared to the species called for the contig. The observed counts were used to estimate the probabilities for correct (c) and incorrect ($\bar{c}$) species calls in each category, for reads that were from species represented in Marine RefSeq (m) or not ($\bar{m}$). By application of Bayes Theorem the probability of a correct species call for a read, given it is from one of the three categories, is:

$$P(c) = P(c, m) + P(c, \bar{m}) = P(c|m)P(m) + P(c|\bar{m})P(\bar{m}) \tag{3.1}$$

where the second term is 0 because calls on non-marine (decoy) reads must be incorrect. Similarly for the denominator of the odds ratio, the probability of incorrectly calling a species for a read is given by:

$$P(\bar{c}) = P(\bar{c}, m) + P(\bar{c}, \bar{m}) = P(\bar{c}|m)P(m) + P(\bar{c}|\bar{m})P(\bar{m}) \tag{3.2}$$

The final odds ratio is then:

$$\frac{P(c)}{P(\bar{c})} = \frac{P(c|m)P(m)}{P(\bar{c}|m)P(m) + P(\bar{c}|\bar{m})P(\bar{m})} \tag{3.3}$$

which is shown in Fig. 3.3 for each read category (raw, intra-set, pooled).

3.5.2.5 Pooled contig follow-up analyses

A subset of pooled contigs that failed to elicit species calls by reference to Marine RefSeq were investigated further by a CAMERA [165] BLASTN search against NCBI RefSeq genomes (E-value <0.1). This served to detect if they could be from known marine viruses, contaminants or other genomes excluded from our database; or rather derived from novel (unknown) species. Moreover, positions in some of these contigs at which the constituent reads lacked unanimity were analyzed to ascertain if they might segregate by geographic regions from which the reads were sampled (Fig. 3.5).

3.5.3 Results and Discussion

3.5.3.1 Chimerism in pooled contigs

Pooled assembly of simulated metagenomic sets produced contigs of similar entropies and sizes to those from intra-set assembly (compare Fig. 3.2A and Fig. 3.9), though of higher coverage, evident by a near doubling of the number of assembled reads with only a small increase in the number of contigs (1459 or 7%; Fig. 3.2B). Pooling also increased the proportion of strain-homogenous contigs from 70% for intra-set to 85%.

3.5.3.2 Chimerism observed when using different assemblers

Genovo assembly of the simulated metagenomic data sets

Although the Genovo *de novo* metagenome assembler uses a fundamentally different algorithm than Newbler (§3.5.2.3), both assembled comparable proportions of

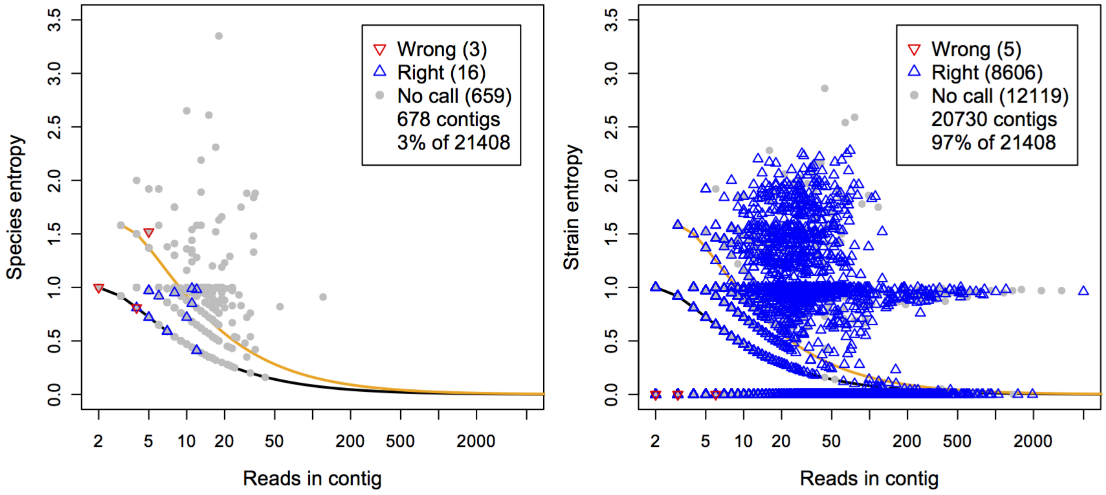


Figure 3.9: Pooled assembly of the simulated data sets produced few species-chimeric contigs (left) and mostly species- or strain-homogenous contigs (right). Species-chimeric contigs made up 3% of all pooled contigs and usually had just one read from the non-majority species (459 contigs are within 0.01 bits of the black curve). Species-homogenous contigs accounted for 97% of pooled contigs and were more likely to get a correct species call. Most (85%) of the species-homogenous contigs were also strain-homogenous (entropy 0 bits).

reads (35.8% and 42.0%, respectively). However, Genovo more frequently produced species-chimeric contigs than Newbler in our hands (10% vs. 3%; Figs. 3.10 and 3.2A). Genovo contigs were smaller (average N50 contig sizes of 835±186 nt vs. 1700±897 nt for Newbler), most (65,734, or 89.0%) assembling with five or fewer reads. We observed a higher proportion of reads with incorrectly called species in Genovo contigs (8.7% vs. 0.0% in Newbler contigs), with most of the affected reads in contigs with five or fewer reads (Figs. 3.10 and 3.11). Based on this, and the observation that the species caller rarely (0.3%) made incorrect calls on unassembled reads (which are shorter still), we would advise a minimum of twenty reads mapped to a Genovo contig for the purpose assigning species to its reads.

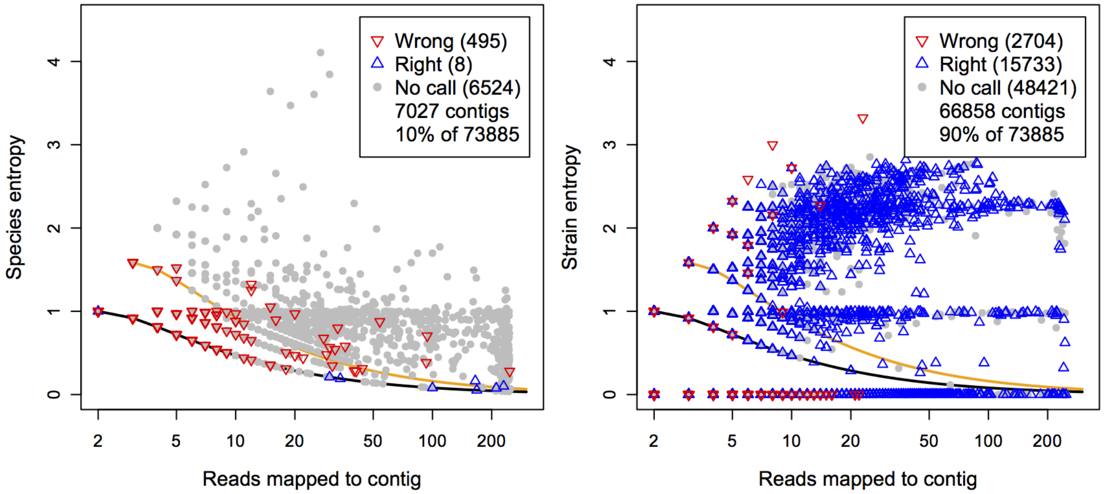


Figure 3.10: Intra-set assembly by Genovo of the simulated data sets, except s4443704.3, produced a higher proportion of species-chimeric contigs (left) than Newbler (Fig. 3.2A), yet also a higher proportion of strain-homogenous contigs (entropy 0 bits for 84.0% of contigs, right plot). Most (65,734) of the contigs in both plots have five or fewer reads, and those contigs account for 77.1% of the reads with incorrect species calls.

Newbler assembly of the simHC metagenomic data from Mavromatis et al.

For comparison, we used Newbler to assemble simHC, the FAMeS gold standard set of highest species complexity (113 genomes, 105 species) [105]. Newbler assembled 2584 Sanger reads (2.21% of all 117K) into 677 contigs, a result similar to that obtained by Mavromatis et al. [105] with ARACHNE [8]: 2744 reads (2.32%) assembled into 558 contigs. We observed that 6.2% of the Newbler contigs were species-chimeric (Fig. 3.12, left), a proportion only slightly higher than that obtained for our simulated marine set assemblies (3% species-chimeric in Figs. 3.2A and 3.9). We further observed that 93.8% of contigs were species-homogenous and 90.1% were strain-homogenous, the latter close to the proportions Mavromatis et al. reported for Phrap [58], ARACHNE and JAZZ [5]: 85.5%, 85.5% and 94.9%, respectively (Tab. 1 in [105]). Such similar

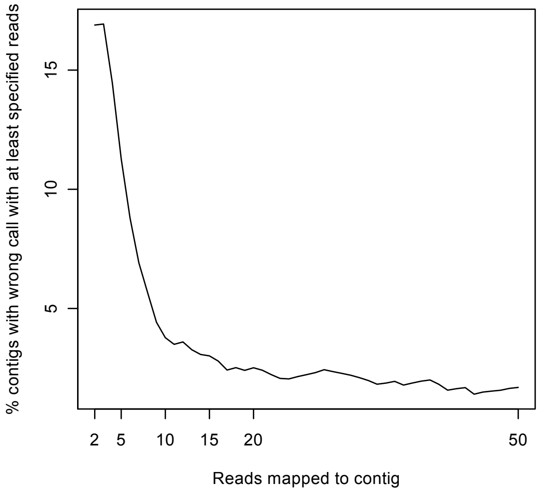


Figure 3.11: Wrong species calls as a function of Genovo contig size. For a given number of reads, the plot shows the proportion of incorrectly called contigs with at least that many reads. This right side accumulation ensures that every point on the plot is based on hundreds of contigs. E.g. there were 237 contigs with 50 or more reads. Only 2.5% of contigs with 20 or more reads (n=836) had incorrect species calls, versus 8.8% of contigs with 6 or more reads (n=3956).

rates of chimerism for four assemblers suggest that algorithms based on read overlap (OLC in the cases of Newbler, ARACHNE and JAZZ, greedy in the case of Phrap) are well-suited to metagenomic assembly, at least for Sanger-length reads. The results for our simulated marine sets make us optimistic for pooled assembly of pyrosequenced samples, as well as those sequenced with recent generations of Illumina platforms.

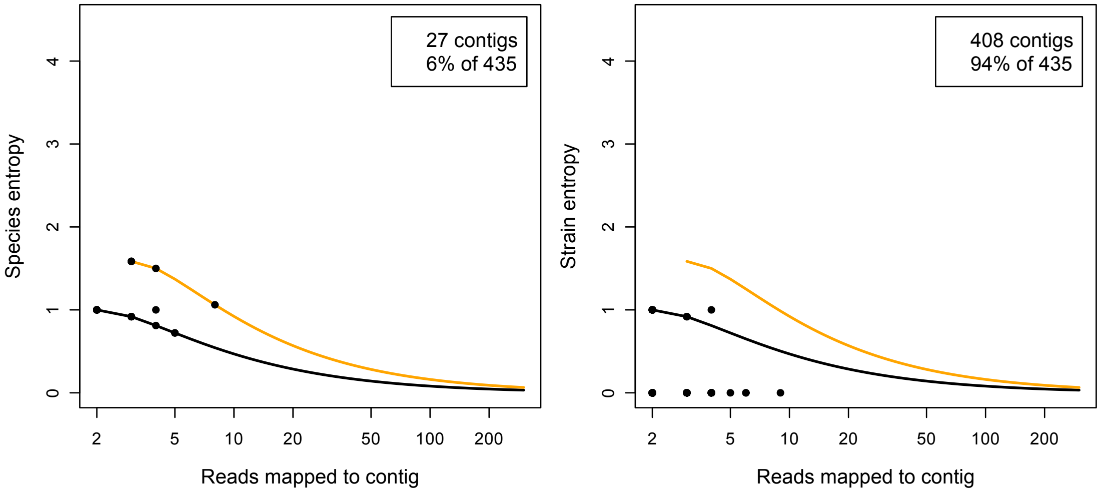


Figure 3.12: Newbler assembly of the FAMeS simHC data set. Most of the 435 contigs used to assess chimerism were strain-homogenous (392, or 90.1%, with entropy 0 bits, right). All but four of the 435 contigs had five or fewer reads.

Chapter 4

Data set selection for efficient pooled assembly

4.1 Introduction

Given that pooled assembly can enable increased annotation but is computationally expensive, how can we pool efficiently? That is, how should MMG data sets be ordered to rapidly increase assembled reads but not computation? Intuitively, sets with similar community compositions (microbial types and relative abundances) should be pooled first, but similarity is difficult to measure. Taxonomy-based approaches (e.g. BLAST hits) use the often small fraction of annotated reads. Alternatively, *ab initio* approaches based on k-mers use all the reads from a metagenome. Much has been published on the use of k-mers to distinguish microbial genomes [78, 79] or to find genes [98, 124]. Far less exists on k-mer signatures of whole metagenomes, but several works have demonstrated the distinguishability of metagenomes from very different biomes (soil, host-associated, marine, etc.) [195, 73, 101].

This chapter, however, determines whether *marine* metagenomes can be distinguished and thus ranked, for the practical purpose of efficient pooled assembly. The two main results are: (1) MMG data sets have distinguishable k-mer profiles that can proxy for community member phylogeny; (2) The rate of pooled-assembly when MMG data sets are ranked by, and assembled in order of, k-mer profile similarity is competitive with that when sets are ranked by phylogenetic profile similarity, and consistently better than random. These support that k-mer-based ranking of MMG data sets is a practical strategy for exploiting pooled assembly to improve annotation (chapter 3) or to discover novel marine microbes among the unannotated contigs (chapter 5). The large increases in assembled reads that we observed for some data sets, despite low annotation rates, make the second point especially exciting.

k-mer-based ranking, here evaluated on a single computer cluster, would also work in a network as illustrated in Fig. 6.1. The figure also shows the relationship among the three approaches developed in this thesis and may be useful to review now.

4.2 Data and Methods

This section is organized into two parts. First §4.2.1 describes data and methods used in preparatory work to determine whether MMG data sets could be distinguished by 5-mer profiles (with results in §4.3.1). Given that they could, §4.2.2 then describes data and methods used in more quantitative work to compare the performance of pooled assemblies when data sets were ranked by 5-mer versus phylogenetic profile similarity (with results in §4.3.2). Note that each part used different data sets (Tabs. 4.1 and 4.2). All analyses (correlations, PCA, clustering) and plots were done in R.

4.2.1 Characterizing data set 5-mer profiles

Data sets: In preliminary work twenty-two data sets were used to test whether marine metagenomes have distinct and robust 5-mer profiles (Tab. 4.1). The MMG sets were selected to capture a variety of projects (with consistent sample preparations within a project), geographic locations (tropical and cold waters of various depths), data types (metagenomic and metatranscriptomic), and sequencing technologies (454 and Sanger). Two non-marine sets were included as outgroups with lower expected similarity to MMG sets and each other. All sets were downloaded from MG-RAST and included only those reads that passed quality screening in the MG-RAST pipeline.

Data set profiles: For each data set, a custom Perl script created 5-mer profiles by sliding a five-nucleotide window across the quality-filtered reads (both strands). A separate script created phylogenetic profiles by counting the phyla of reads whose ORFs had hits to the MG-RAST M5nr database.

Clustering data set profiles: The same projection and clustering procedure was used for 5-mer and phylum matrices: Matrix entries were $\log_2$-transformed and each column centered and scaled (variance 1). Singular value decomposition was then applied, followed by projection to a lower dimensional space that captured 98% of the

Data set	**Reads**	**% annot.**	**Location**	**Depth**	**MG id**
xa41^{c,e}	191,342	4%	S. Pacific (Xmas Atoll)	10m	4440041.3
wec64^{c}	333,904	28%	W. English Channel	surface	**4445064.3**
nf02^{c}	180,858	48%	Norwegian fjord	surface	**4443702.3**
nf07^{c}	91,817	33%	Norwegian fjord	surface	4443707.3
nf63^{c}	89,247	33%	Norwegian fjord	surface	4440163.3
ha51^{s}	6323	63%	N. Pacific (Stn. ALOHA)	10 m	4441051.3
ha57^{s}	8400	56%	N. Pacific (Stn. ALOHA)	70 m	4441057.4
ha55^{s}	3690	62%	N. Pacific (Stn. ALOHA)	130 m	4441055.3
ha41^{s}	6082	59%	N. Pacific (Stn. ALOHA)	200 m	4441041.3
ha62^{s}	10,763	64%	N. Pacific (Stn. ALOHA)	770 m	4441062.3
ha56^{s}	9779	67%	N. Pacific (Stn. ALOHA)	4000 m	4441056.3
spsg95	61,505	44%	S. Pacific	5 m	**4443695.3**
pe97	232,001	33%	Equatorial Pacific	5 m	**4443697.3**
pne98	57,553	15%	Equatorial N. Pacific	5 m	4443698.3
pne99	234,685	32%	Equatorial N. Pacific	5 m	**4443699.3**
hew35^{n,e}	63,742	23%	Subtropical S. Pacific	5 m	4443735.3
hew36^{e}	65,685	26%	Subtropical S. Pacific	5 m	4443736.3
hew38^{n,e}	71,736	23%	Subtropical S. Pacific	5 m	4443738.3
hew40^{e}	69,267	18%	Subtropical S. Pacific	5 m	4443740.3
hew43^{n,e}	64,230	15%	Subtropical S. Pacific	5 m	4443743.3
am37	91,645	59%	acid mine drainage	-	4441137.3
tg01	44,844	62%	termite gut	-	4442701.3

Table 4.1: Data sets used in 5-mer profile characterization. Data sets are named by location or investigator and the last two digits of their MG-RAST identifier (MG id). ^{s} denotes a Sanger-sequenced data set; all other sets are 454. ^{n} denotes a night sample; all others are daytime. ^{c} denotes a coastal sample; all other marine samples are open ocean. ^{e} denotes a metatranscriptomic data set; all others are metagenomic. Sets with emboldened MG-RAST identifiers were also used in the performance analysis.

variance (six dimensions for phylum, ten for 5-mer). Then, the data sets were clustered in the reduced spaces using k-means (Euclidean distance and $k = 6$). Random subsets of reads for 5-mer profile robustness tests were selected via a custom Perl script, and clustered as just described.

4.2.2 Measuring rates of pooled assembly

Data sets: From the data sets that assembled well when pooled (i.e. that had contigs comprised of reads from multiple data sets and N50 not near zero; see Fig. 3.8), twenty were selected (Tab. 4.2).

Monterey Bay	4443713.3 4443714.3 4443715.3 4443716.3 4443717.3
Norwegian fjord	4443702.3 4443703.3 4443704.3 4440276.3
W. English Channel	4444083.3 4445064.3 4445065.3 4445066.3
Pacific	4443695.3 4443697.3 4443699.3 4443700.3 4443701.3
San Diego	4443765.3
BATS	4443731.3

Table 4.2: MMG data sets used to measure pooled assembly rates. Details for each set are in Tab. 3.2. Most data sets had 114–423K reads, except for 4443695.3 (62K reads) and 4443700.3 (53K reads).

Data set profiles: As in §4.2.1, the 5-mer profiles were simple counts generated by sliding a five nucleotide window over reads.[1] Phylogenetic profiles were counts of MG-RAST v3 phylum, family, or species assignments based on open reading frame searches against RefSeq ($E \leq 10^{-10}$; ≥60% DNA identity; ≥30 alignment length).[2]

Terminology: A "flight" is a succession of assemblies that begins with a specific data set and adds each of the remaining nineteen data sets (Tab. 4.2) one at a time and in rank order. Data set are ranked 1–19 in decreasing order of similarity to the first set in a flight, by an approach in Tab. 4.3. Step 1 in a flight is an intraset assembly,

[1] Though profiles comprised of z-scores for over- or underrepresented k-mers have been used to measure *fragment* similarity (Schbath (1995) [149] which influenced TETRA [173]), we observed that they would have made no difference for whole metagenomes: Pairwise Pearson correlations of data set profiles comprised of z-scores, or simple counts, were equal.

[2] A custom shell script searched MG-RAST via its RESTful API, and so results were compiled by MG-RAST from pre-computed runs of the pipeline, by filtering to the specified cutoffs (main text). The returned phylogenetic abundances were not read counts, rather they were the number of hits to RefSeq proteins elicited by ORFs predicted for the reads (described in §1.5.1). Abundances counted all hits above the cutoffs, thus reads often had multiple hits due to multiple ORFs and multiple hits per ORF.

and the step 1 data set is the "intra-set." Each of steps 2–20 are pooled assemblies. The "gain" at each step is the number (or proportion) of newly assembled reads from the intra-set, i.e. gain does not count reads from the pooled sets.

Category	Specific approaches for selecting data sets
5-mer profile	Highest Pearson or Kendall correlation
Phylogenetic	Phylum, family or species Bray-Curtis dissimilarity
Unordered	Randomly selected
Greedy	Next set maximizes gain in assembled reads

Table 4.3: Approaches to ranking and ordering data sets that were compared. "Greedy" is a mock approach in which the gains in assembled reads for each set are known.

Data set similarity metrics: Tab. 4.3 shows the seven similarity metrics ("approaches") for ranking data sets that were compared. The 5-mer approaches ranked data sets, and thus ordered them for assembly, in decreasing order of their 5-mer profile Pearson or Kendall correlations. The phylogenetic approaches ranked data sets in increasing order of Bray-Curtis [16] dissimilarity between their phylum, family or species profiles as assigned my MG-RAST. Bray-Curtis dissimilarity ranges from 0 (identical counts for all taxa) to 1 (no shared taxa). Random rankings served as low bars for comparison. Two Random rankings were created for each data set. For clarity, plots show only one.

The Greedy approach defined a near-optimal ordering of data sets for pooled assembly. For a given intra-set, the Greedy approach ranked the remaining nineteen sets in order of maximal gain in newly assembled reads from the intra-set. Because different data sets can induce some of the same intra-set reads to assemble, gain at each step in a flight depends on which sets have already been pooled. Thus, to define Greedy rankings, all data sets were assembled pairwise and a custom Perl script was used to calculate the Greedy flight. The script accounted for already gained intra-set reads at each flight step. When run on our computer cluster, for the most part Greedy flights proceeded in order of maximum gain.[3]

For every data set and approach, a combination of Perl, R and Unix shell scripts were used to create 5-mer and phylogenetic profiles, and to create flight description files to guide the assembly.

Assembling the flights: Custom Perl, R and Unix shell scripts ran the Newbler assemblies, iteratively adding one data set at a time based on the flight description. Additional scripts recorded information needed to calculate performance metrics after each step (below). All assemblies ran on the UCSC campusrocks Linux cluster.[4]

Measuring assembly efficiency: As in chapter 3, we took the perspective of a researcher who uses pooled assembly to aid interpretation of the reads in a single data set. Thus the number of newly assembled reads at each step in each flight was recorded, i.e. the successive gains with each pooled assembly after the initial intraset assembly.[5] Also at each step, Newbler recorded the number of searches (*numberSearches* in `454NewblerMetrics.txt`). This was a stable measure of computational cost because

[3] Occasionally, the predicted Greedy ordering was not maximal: A gain at step i exceeded a gain at step $i - 1$ or $i - 2$. Only for two data sets (4443700.3 and 4443704.3) did this happen within the first five steps. Fig. 4.4 shows that these out-of-order steps were insignificant (see slight inflections in the black curves).

[4] The UCSC campusrocks computer cluster ran CentOS Rocks release 6.2. It had one head node and seven compute nodes with 304 cores in total: three 64-core nodes with 256GB of RAM; two 48-core nodes with 192GB of RAM; and two 8-core nodes with 16GB of RAM. Campusrocks used the Sun Grid Engine 6.0 queuing system.

[5] In light of our goal (chapter introduction), only rates of assembly—not annotation—were measured. Annotation rates would have offered little guidance on how to rank data sets for pooled assembly due to sensitivity to search parameters and the state of reference databases. Several well-assembled but poorly annotated sets in Fig. 3.4 make the point.

numberSearches is unaffected by other jobs running on a host, file system I/O, and performance differences among hosts in the cluster. In figures the x axes accumulate *numberSearches* with each step in a flight. In total Fig. 4.4 shows 2780 assemblies: 20 data sets, each with 7 flights and 20 assemblies per flight, except for 4443765.3 for which Bray-Curtis species dissimilarity could not be calculated.

Computation time (Unix system + user time, not wall clock time) was also recorded for all flights but this was only used for comparison to numbers of searches in Fig. 4.3.

For all data sets and approaches, areas under the gain versus searches curves (AUC's) were calculated, both over entire curves and also at each 5% increase in searches completed. For each data set, maximum possible AUC was normalized to one so that approaches could be compared; i.e. approaches for a data set were plotted on the same "canvas" of area one, and then ranked by AUC. AUC rankings were then averaged across all data sets and are reported in the right plots in Figs. 4.5 and 4.6. Bootstrapping of AUC at percentages of gains in assembled reads entailed 10× resampling of the data sets, each time plotting the AUC versus percentage gained as in Figs. 4.5 and 4.6, and examining the approach rankings for consistency.

4.3 Results

This section is organized into two parts, preparatory work (§4.3.1) and performance comparisons (§4.3.2), with different data and methods as explained in §4.2.

4.3.1 5-mer profiles were distinct, robust and consistent with phylum profiles.

4.3.1.1 Observation 1: Data sets were distinguishable by 5-mer profile.

Five-mer profile correlations were higher within data sets than across data sets, suggesting that data sets have distinguishable 5-mer profiles. Two tests supported this:

- Data sets were partitioned at 90%/10%, 70%/30%, and 50%/50% of their reads, and 5-mer profiles created for each side of a partition. The Pearson correlations between 5-mer profiles were higher for intra-set comparisons (mean of correlations = 1.00) than for inter-set comparisons (mean = 0.63).

- Data sets were also partitioned into reads with and without phylum annotation, and the 5-mer profiles for each compared. Profiles for reads from the same data set, but with versus without annotation, had strong Pearson correlations (mean = 0.85). This was not the case for cross-set profile comparisons for annotated versus unannotated reads (mean = 0.58). This supported Observation 1 and also suggested that unannotated reads provide complementary information for distinguishing data sets.

4.3.1.2 Observation 2: Similar clusters whether profiles based on phyla or 5-mers.

Next I compared phylum-based and 5-mer-based clusters of the data sets. The clusters looked similar suggesting that for the purpose of ranking data sets, 5-mer-based profiles could be a computationally cheap approximation of community composition (Fig. 4.1).

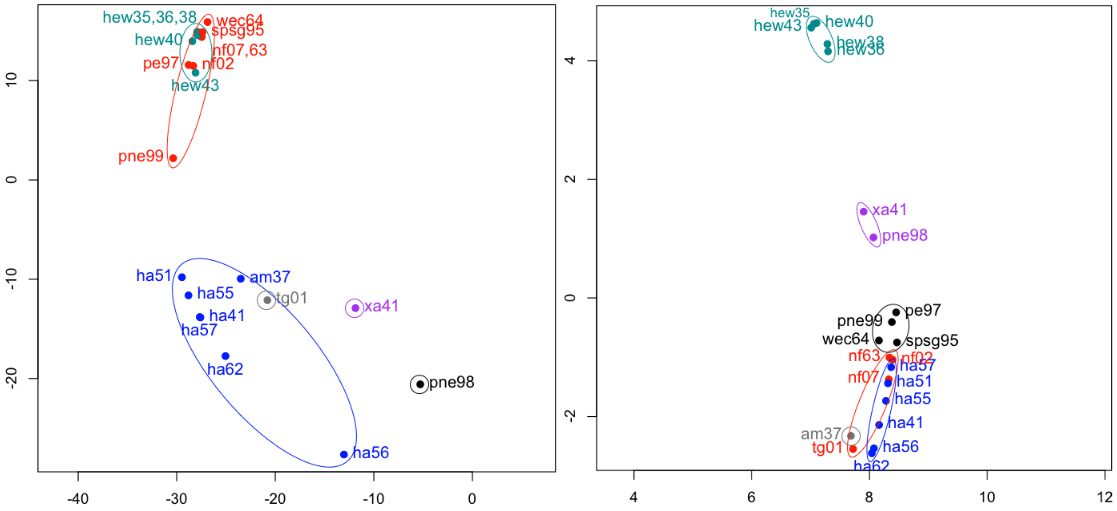


Figure 4.1: Data sets clusters appeared similar whether profiles were based on 5-mers or phyla. Left: Five-mer profiles for annotated reads from each data set were projected to a reduced dimensional space that captured 98% of the variance, and then clustered with *k*-means (colors). Right: By the same procedure, phylum profiles were projected and clustered. Just the first two principal axes are shown, which capture 80% (5-mer) and 95% (phylum) of the variances.

As shown in Fig. 4.1, data sets formed similar clusters whether represented by phylum or 5-mer profiles. The Hewson sets (cyan) are abundant in *C. watsonii* (samples taken during a phytoplankton bloom). The HOT/ALOHA sets (blue), all sampled from the same location, also cluster under both representations, though am37 joins the cluster in the 5-mer case. The Norwegian fjord sets (nf) cluster in both cases (red). So do most of the Pacific samples (pne99, pe97, spsg95) and western English Channel (wec64), however those four sets form a separate cluster (black) when represented by phylum profiles but join the Norwegian fjord cluster (red) under 5-mer profiles. Lastly, for both representations xa41 and pne98 are remote, clustering only with each other or not at all.

One inconsistency between the two projections is that only the two non-marine data sets, am37 and tg01, cluster with MMG sets in one projection but do not cluster in the other.

4.3.1.3 Observation 3: Unannotated reads had little impact on clusters.

Including reads that lacked phylum annotation in the 5-mer profiles yielded nearly identical clusters to those based on just the annotated reads (Fig. 4.2 vs. Fig. 4.1, left). The exceptions were pe97 and pne99 which switched to the HOT/ALOHA (blue) cluster. So for some data sets, ranking would be influenced by the inclusion of all reads.

4.3.1.4 Observation 4: 5-mer profiles were robust.

To assess whether 5-mer profiles are robust, random samples of 10% of the reads in each data set were profiled and projected. In Fig. 4.2 these samples appear as spots scattered near the data sets. For the 454 data sets, the random samples are always bound tightly. For the Sanger-sequenced sets (ha), the variation is greater likely because those sets have fewer reads. Overall, the 5-mer profiles appear robust. This

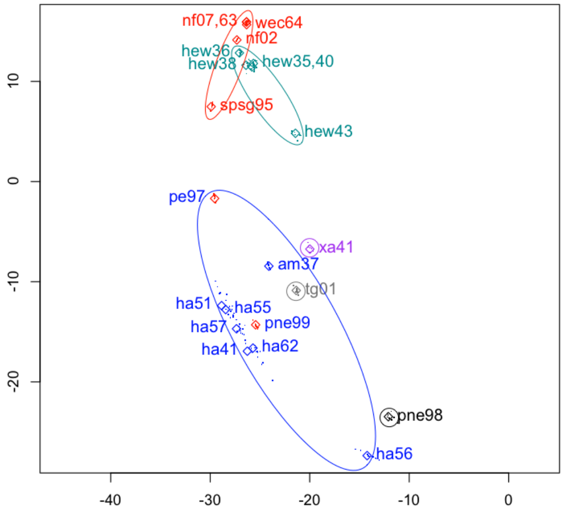


Figure 4.2: Clusters when 5-mer profiles include both annotated and unannotated reads. Compared to Fig. 4.1, clusters are identical except for pne99 and pe97. For each data set (diamonds), small spots that represent random 10% samples of its reads are shown (mostly within the diamonds) and support that data sets have robust 5-mer profiles.

was also true for profiles for random samples of the annotated or unannotated reads.

4.3.2 Ranking data sets increased pooled assembly efficiency.

Please see §4.2.2 which describes the terminology, data and methods in this section.

4.3.2.1 Finding 1: Computational cost was comparable across assemblies.

Assemblies occurred on different host machines of the campusrocks cluster. To check that Newbler performance, in terms of number of reads assembled and *numberSearches*, did not depend on which host assembled the data sets, six flights were assembled in replicate (n=120 pairs of assembly steps to compare). For five flights the replicate and original assemblies occurred on different hosts. The mean differences in number of reads assembled, at corresponding steps in replicate and original flights, were negligible ($0.08 \pm 0.22\%$ of reads assembled in original), as were the mean differences in computational cost ($0.00{\pm}0.01\%$ of *numberSearches* in original). Therefore the assembly performances reported below, though measured on different hosts of the campusrocks cluster, can be compared.

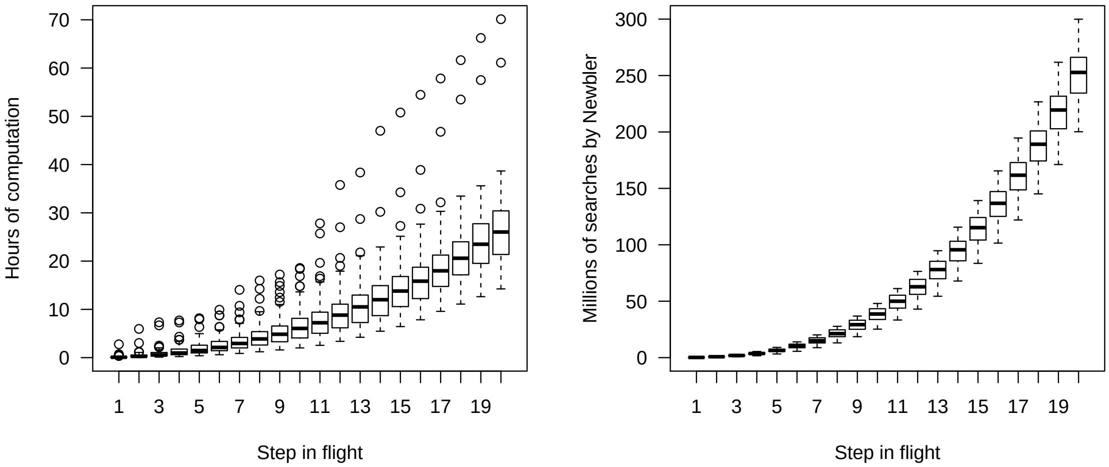


Figure 4.3: Cumulative hours of computation (Unix system + user time; left) and cumulative searches (right) at every step for all flights. Step i is a pooled assembly of i data sets. The distribution at each step represents 133 flights (seven approaches to ranking (Tab. 4.3) for all twenty data sets except 4443765.3).

CPU hours (not wall clock time) provide an intuitive measure of computational cost and are compared to the number of searches in Fig. 4.3. In this figure, each step on the x axes bins the assembly costs across all approaches (Tab. 4.3) and all data sets. Successive steps show an exponential trend, consistent with the pairwise read comparisons that characterize OLC assembly (§1.4.1). The low variation in search costs (right), even at step ten when the number of ways to choose from twenty sets for pooling is maximal,[6] supports that differences in data set sizes did not greatly impact search costs. Thus performance differences discussed below do not depend on when larger versus smaller data sets were pooled. Lastly, outliers in the left plot are not mirrored in the right by increased Newbler searches, thus they likely stemmed from cluster or software issues. They indicate that search costs described in this chapter underestimate time costs, with financial consequences if one pays for compute time in a commercial cloud (e.g. MG-RAST uses Amazon's EC2 [194]).

4.3.2.2 Finding 2: 5-mer and phylogenetic flights ranked sets differently.

We verified that 5-mer and phylogenetic approaches ranked data sets differently, so it was worthwhile to compare their assembly performances. The lower triangle in Tab. 4.4 shows low Kendall correlations between data set rankings by 5-mer and phylogenetic approaches ($\tau \leq 0.39$). In contrast, intra-approach rankings correlated strongly (5-mer, $\tau = 0.86$; phylogenetic, $0.86 \leq \tau \leq 0.91$). A similar observation was made when only the top five ranked data sets were considered (upper triangle): 5-mer and phylogenetic approaches overlapped in only 2–3 of their top five ranked data sets, compared to ∼4 sets for approaches from the same category. Finally, the 5-mer and phylogenetic rankings differed from the Random rankings by both statistics, but they were weakly similar to Greedy rankings (up to $\tau = 0.57$ at the maximum σ_τ).

We checked the sensitivity of data set rankings to k-mer length, for $k \in \{4, 5, 7\}$. Pairwise Kendall correlations between rankings based on different k's were fairly high

[6] Different data sets were in fact chosen for pooling at each step by 5-mer and phylogenetic approaches (described in the next section).

top5 τ	5-mer P	5-mer K	Species	Family	Phylum	Random	Greedy
5-mer P		**4.2**	2.7	2.8	2.5	1.2	2.4
5-mer K	**0.86**		2.6	2.7	2.4	1.2	2.5
Species	0.38	0.37		**4.5**	**4.1**	1.4	2.7
Family	0.39	0.39	**0.91**		**4.3**	1.3	2.9
Phylum	0.34	0.34	**0.86**	**0.90**		1.4	2.6
Random	-0.02	-0.01	0.01	0.02	0.01		1.4
Greedy	0.32	0.33	0.32	0.34	0.26	0.02	

Table 4.4: Similarity of approaches in how they ranked data sets. All entries average a statistic (τ, *top5*) for a pair of approaches over all data sets. Lower triangle entries give Kendall correlations (τ) between flights (steps 2–20). Upper triangle entries count the number of data sets that approaches shared among their top five ranked (steps 2–6). Intra-category comparisons are emboldened. Standard deviations within 5-mer and phylogenetic approaches were small: $0.03 \leq \sigma_\tau \leq 0.14$ and $0.51 \leq \sigma_{top5} \leq 0.95$. Greedy standard deviations: $0.17 \leq \sigma_\tau \leq 0.24$ and $0.68 \leq \sigma_{top5} \leq 1.12$.

(means: $\tau_{4,5} = 0.64$, $\tau_{7,5} = 0.77$ and $\tau_{4,7} = 0.64$). Rankings also strongly overlapped among the top five sets (means: $top5_{4,5} = 4.7$, $top5_{7,5} = 4.5$ and $top5_{4,7} = 4.2$). Therefore we can assume that flights with data sets ranked by 4-mer or 7-mer profile similarity would have assembled with similar performance curves as were observed for 5-mer flights.

4.3.2.3 Raw results – gain versus computational cost

Other findings depend on the raw results illustrated in Fig. 4.4 and introduced here. For each of the twenty data sets, all seven ranking approaches (Tab. 4.3) were measured, thus each plot has seven colored curves.[7] Each x axis shows the $\log_{10}$ of the cumulative number of searches as data sets are individually added to the assembly pool in rank order, from the “intra-set” (step 1) to the least similar data set (step 20). Each

[7] Except 4443765.3 which has only six.

y axis shows the cumulative percentage of reads from *only* the intra-set. For example, for the Monterey Bay (MB) data set 4443714.3, plotted last on the next page, 10% of its reads assembled without pooling, but 46% had assembled by the time the last data set was added to the pool. The one-letter codes on the optimal Greedy (black) curve show that the first set pooled by that approach was from the Monterey Bay (M), the second from the western English Channel (W), and so on for the top five. Note that y axes differ across data sets, and plots are ordered by increasing intra-set assembly, i.e. by the minimum values on their y axes.

Using the terminology defined in §4.2.2, each plot shows seven flights, one for each approach in Tab. 4.3, as a relation between gained reads and computational cost. Gains and costs were measured at the end of each flight step, so each curve interpolates between steps.

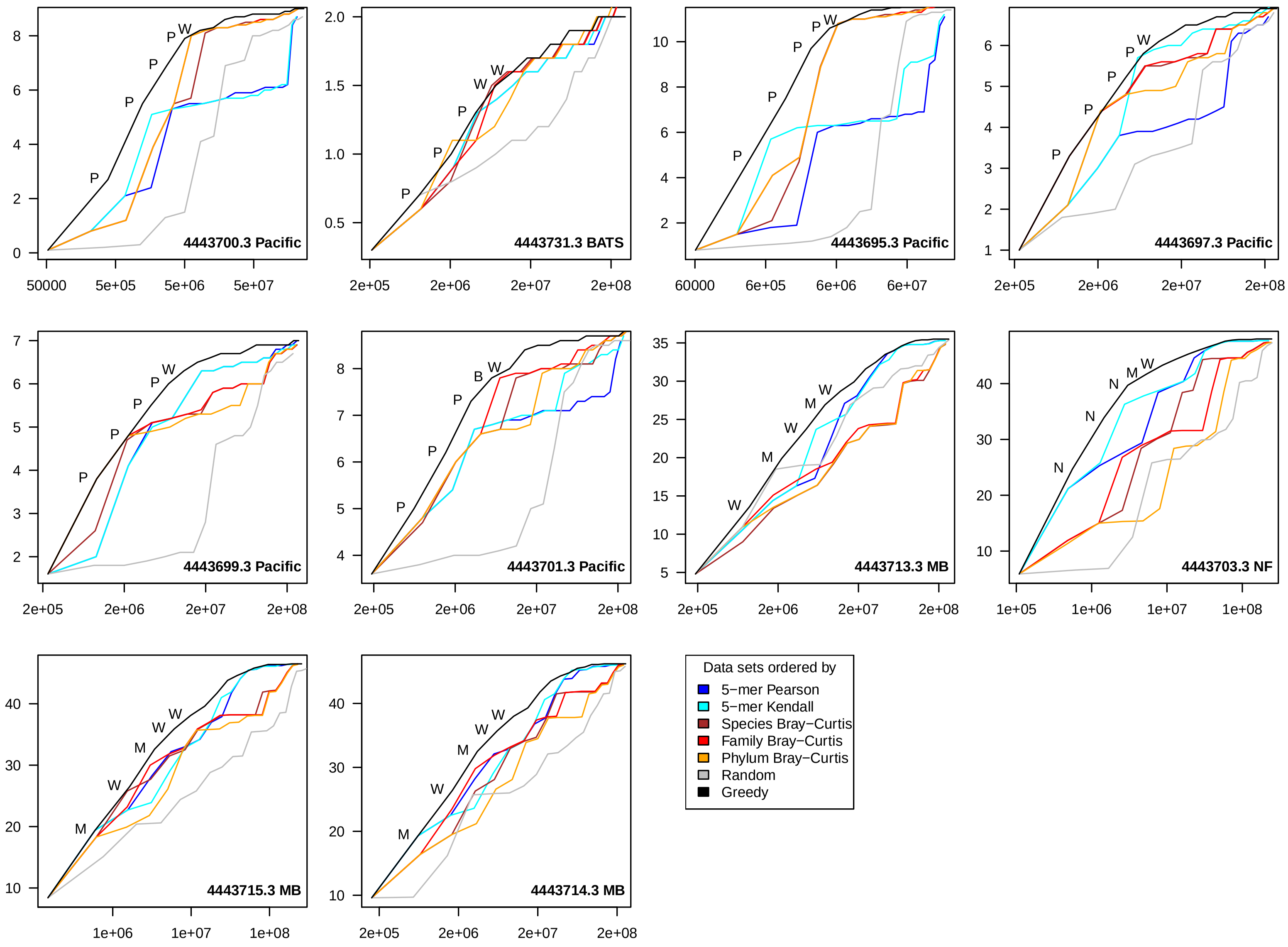


Figure 4.4: Assembly rates when data sets are pooled in rank order as assigned by the seven approaches. X axes are the cumulative number of Newbler searches. Y axes are the percentage of gained reads from the intra-set. Each plot shows seven flights (legend, bottom right) for the indicated data set. For the Greedy flight, one-letter codes denote (continues)

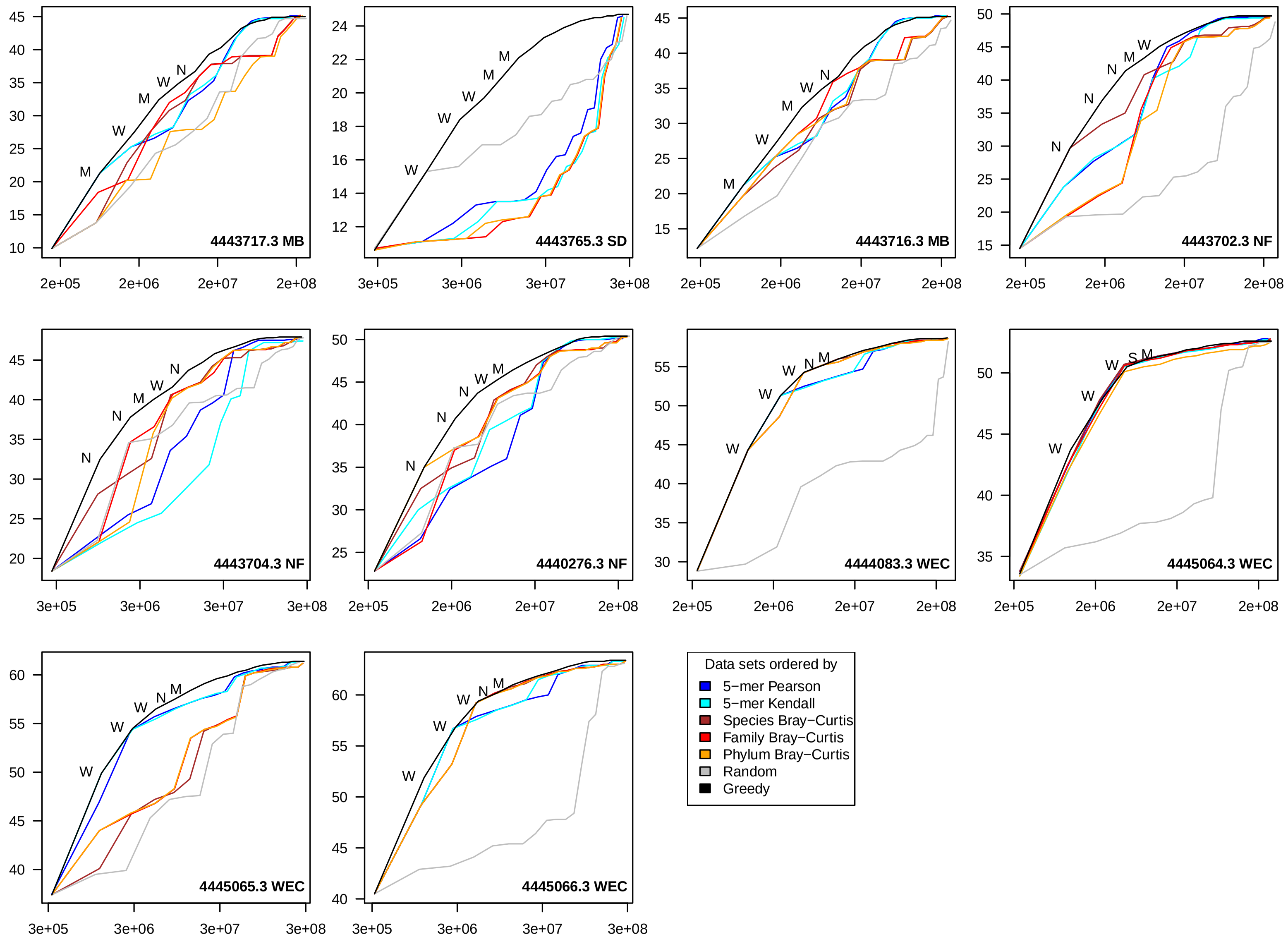


Figure 4.4: (continued) the geographic regions of the top five ranked sets: P=Pacific, M=Monterey Bay, B=BATS, N=Norwegian fjord, S=San Diego, and W=western English Channel.

4.3.2.4 Finding 3: Data sets from the same geography often ranked highly.

Given that samples collected from the same or nearby locations should share microbial types, would ranking by geographic proximity maximize assembled reads (e.g. see Fig. 3.8)? To answer this, the geographies of the top-ranked sets by the Greedy approach (based on maximizing gain in assembled reads, not geography; §4.2.2) were examined: Often, but not always, the top sets derived from the same geographic region as that of the intra-set, summarized in Tab. 4.5.

Pacific	W. English Channel	Monterey Bay	Norwegian fjord
3.8 of 5	3.00 of 4	2.00 of 5	3.00 of 4

Table 4.5: Average number of sets from the same geographic region as the intra-set among the top five ranked by Greedy. For example, looking over all Monterey Bay flights in Fig. 4.4, on average Greedy ranked 2.00 of the other Monterey Bay sets among the top five, and thus three from other geographic regions. The maximum for each region excludes the intra-set and thus is one less than the total sets from the region (shown). Lone sets from San Diego and BATS omitted.

Consistent with the expectation, for western English Channel intra-sets, Greedy pooling always ranked the remaining WEC sets among the top five (top three in fact). This was also true of the Norwegian fjord sets and all but one of the Pacific sets. Within each of these regions, at least one sample had been collected days and/or miles apart from the others,[8] so the table counts are not due to sets being near-replicates. Alternatively, open ocean versus coastal sampling may partly explain why Pacific sets (open ocean) ranked highly among each others' top five. Consistent with this, the only other

[8] Three of the WEC sets represent samples that were collected only 6–12 hours apart (on August 27, 2008 for 4444083.3, 4445065.3, 4445066.3), which may partly explain why Greedy pooling favored them. However, the fourth set, was collected months earlier (on April 4, 2008 for 4445064.3). Pacific sets (Tab. 4.2) represented samples hundreds of miles apart (gyres north and south of the equator, and equatorial samples). Norwegian fjord sets represent one location but samples were collected about one week apart (4443702.3 and 4443703.3 on May 13, 2006; 4440276.3 and 4443704.3 on May 19, 2006).

open ocean set (BATS, 4443731.3) had its top three sets from the Pacific.

In contrast to the expectation that one should pool first by geography, for the Monterey Bay the top five ranked sets by Greedy included two to three from the western English Channel (mostly ranked second and fourth). Four of the Monterey Bay samples were collected during or shortly after a spring 2001 phytoplankton bloom (4443714.3–4443717.3). So it interesting that among the Monterey Bay top five, the only WEC set that *never* occurred was collected during the North Atlantic spring phytoplankton bloom in 2008 (4445064.3). That is, all the other WEC sets, from summer 2008 non-bloom conditions, occurred among the top five for Monterey Bay sets, which were mostly from a bloom. This is taken up in the §4.4.

The 5-mer and phylogenetic approaches, like Greedy, tended to first select sets from the same geographic region as the intra-set, with on average 2.2 same-region sets among the top five ($0.65 \leq \sigma \leq 0.94$; compare to Greedy in Tab. 4.5).[9] With this in mind, we next compare the assembly rates of the approaches.

4.3.2.5 Finding 4: Ranking data sets by 5-mer or phylogenetic similarity increased pooled assembly efficiency.

The 5-mer and phylogenetic ranking approaches performed comparably (Fig. 4.5), and significantly better than Random (left, area under the curve (AUC) comparisons; Bonferroni corrected $P < 0.001$; one-sided Mann-Whitney-Wilcoxon tests).

[9] Mean calculated from the five averages for each of the 5-mer and phylogenetic approaches. For each approach, its average was over all (n=18) data sets; e.g. 5-mer Pearson flights on average had 2.2 of their top five sets from the same geographic region as the intra-set, while Phylum flights had 2.1. SD and BATS excluded because they were the only sets from their regions.

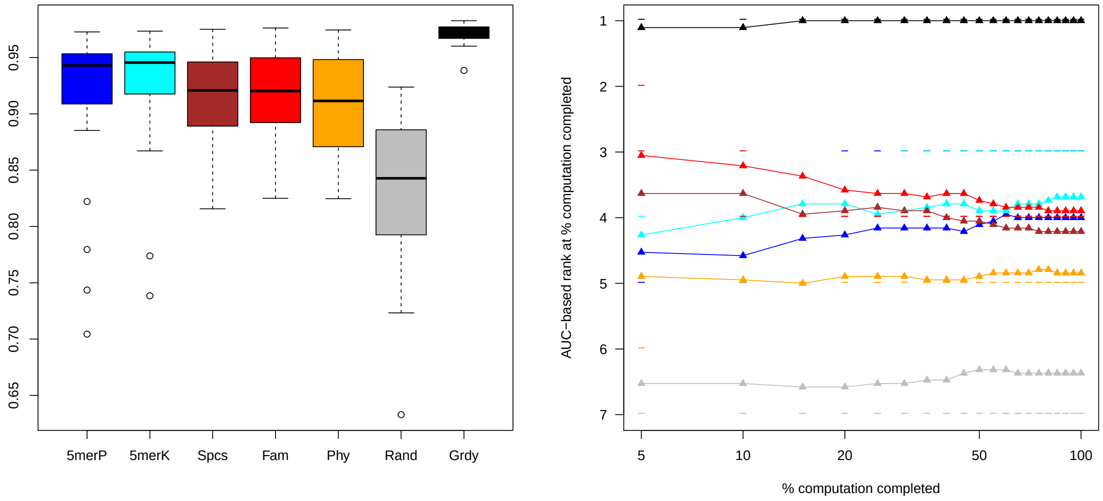


Figure 4.5: Approaches for ordering pooled assemblies compared by area under the curve (AUC). The two plots summarize all those in Fig. 4.4 and use the same color scheme. **Left** AUC of each approach, averaged over all data sets and flight steps 1–20. **Right** *Approaches* (not data sets) ranked 1–7 by AUC at percentage of searches completed (§4.2.2). The x axis starts at 5% to ensure rankings begin after some assembly completed. Connected colored triangles show mean ranks; dashes show medians. Ranks are averaged over data sets.

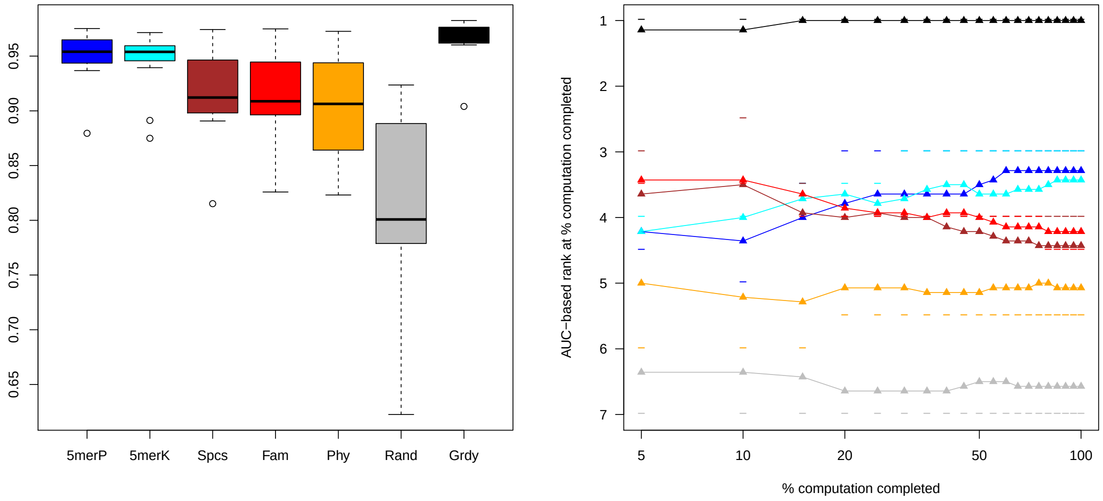


Figure 4.6: Approaches ranked by AUC as in Fig. 4.5 except that Pacific data sets were excluded. Remaining sets are all coastal.

The right plot in Fig. 4.5 helps one see which approaches performed better or worse as more data sets were pooled. Notwithstanding limitations of ranking the approaches,[10] some trends are evident, all of which appeared robust in bootstrapping.

- Family and Species outperformed 5-mer based approaches in early steps—that is, pools with fewer data sets—though the four approaches were close in rank by the twentieth step (100% computation completed). Probing this further, when Pacific data sets were excluded[11] the 5-mer approaches outperformed Family and Species by the time ∼20% of searches completed (Fig. 4.6).
- Family and Species performed comparably well and better than Phylum, as shown in the plots and supported by bootstraps. Because all three approaches were based on the same MG-RAST assignments to RefSeq genomes—just described at three taxonomic levels (Methods)—it seems that Phylum offered comparatively poor resolution.

4.3.2.6 Finding 5: Pooling greatly increased the percentage of assembled reads.

It is noteworthy that pooling produced appreciable gains in assembled reads, usually several thousand to 100K new reads for each of the first few sets added (Fig. 4.4). By step 20 the median assembly level was 32% of the intra-set reads. Excluding the six

[10] Rank is discrete and does not reflect small differences in AUC.

[11] Pacific sets seemed problematic for 5-mer approaches in Fig. 4.4. This might suggest a problem with those approaches for open ocean samples. As a counterexample, however, 5-mer approaches performed well for the open ocean BATS set (4443731.3), and like Greedy they ranked Pacific sets among the top three to pool with BATS.

data sets with poor intraset assembly (<5%) and 4443765.3,[12] by step 20 the *minimum* was 35% of reads assembled. In contrast the six data sets with the lowest intraset assembly percentages rose to at most 11.1% (4443695.3 in Fig. 4.4). However, even for these data sets most of the gains came from pooling, and usually with one or two sets.

At step 20, pooled contigs had on average N50 = 1480 nt, that is, they were much longer than a typical Sanger read.

4.4 Discussion

It came as a surprise that 5-mer and phylogenetic approaches ranked data sets differently but performed comparably overall (compare Tab. 4.4 to Figs. 4.5 and 4.6). Some of the ranking differences may stem from low MG-RAST annotation rates for some sets (median 58.8%, range 45.3–79.5%) leading to profiles based on a fraction of the reads, versus all reads for the 5-mer profiles. However, it may also be that 5-mer profiles capture not just taxonomic but also functional information. Dinsdale et al. (2008) showed that the biomes of forty-five microbial 454 metagenomic data sets could be predicted based on their metabolic profiles, and hypothesized that the environments were selecting for genes, not taxa [37].[13] With the same data sets now represented by dinucleotide profiles, Willner et al. (2009) showed that they clustered by biome and with similar levels of observed variance (80.8% versus 79.8% for Dinsdale et al.)

[12] Set 4443765.3 was cell-sorted for *Synechococcus*, which might explain why it intraset-assembled well. Curiously the only other single-site set, the BATS set 4443731.3, also did not assemble much when pooled.

[13] Their hypothesis was motivated by the high variation in gene content they observed across biomes, despite low taxonomic variation reported in the literature.

[195, 37].[14] Though these two papers distinguished biomes, later work by Gianoulis et al. (2009) showed that marine metagenomes could be resolved by functional profiles [51].[15] Moreover, 5-mers and even smaller k-mers have been used to train HMM-based gene finders [98].

That geographic metadata alone will not suffice for ranking data sets is the main message of §4.3.2.4. In that section we saw that the ∼optimal approach, Greedy, sometimes ranked geographically distant sets among the top five for pooling. In particular, the Monterey Bay (MB) sets gained the most assembled reads when pooled with western English Channel (WEC) sets. However, there was one curious exception: WEC set 4445064.3, the only one derived from a sample collected during the 2008 North Atlantic Spring Bloom, was also the only WEC set that Greedy excluded from the MB top five pools. Why? The most striking difference between 4445064.3 and the other WEC sets was a reduction in the abundance of an unknown *Rhodobacteraceae* member from ∼45% to ∼10% [55]. This member had been identified with 16S rDNA v6 tags and was presumably the same population across the other WEC sets, which were also all sampled from the L4 site, but in *summer* 2008 [55]. Remarkably the MB sets were also dominated by *Rhodobacteraceae* members (28% NAC11-7 based on 16S microarray probes) [146]. It is plausible that these *Rhodobacteraceae* differed from the population identified in 4445064.3, and caused that set to assemble poorly with MB sets.[16] I.e. the

[14] The 80.8% is the variance captured by the first three principal components (Table 1 in [195]). The 79.8% is for the first two canonical discriminant analysis axes (Figure 1a in [37]).

[15] Essentially both the Dinsdale et al. (2008) and Gianoulis et al. (2009) papers performed multivariate analyses of 454 metagenomic data sets for which pathway abundances (by reference to KEGG or SEED) had been measured.

[16] Differences in the second most abundant community members between the WEC and MB sets were not likely a factor because they always differed yet only 4445064.3 was excluded. Specifically, all WEC

exceptional WEC set may have been excluded because it represented phytoplankton bloom conditions.

Unlike Greedy however, the 5-mer and phylogenetic approaches ranked *all* the WEC sets poorly ($rank > 10$) for pooling with MB sets. Instead they favored pooling with the Norwegian fjord (NF) data sets, possibly due to strong overlaps between two of the dominant members in both: *Rhodobacteraceae* (20.3–25.4% in MB; 15.6–27.2% in NF) and *Flavobacteraceae* (10.8–17.4% in MB; 8.1–14.2% in NF).

For a given data set, the three phylogeny-based approaches used the same reads, but counted them at phylum, family or species level. It is then noteworthy that Family, which had an intermediate number of features between Phylum and Species, performed on average slightly better than those approaches (Figs. 4.5 and 4.6).

The data set for which 5-mer and phylogenetic approaches fared the worst, the San Diego set (4443765.3), sheds light on when the approaches will fail and succeed. That set was cell-sorted for *Synechococcus*, which comprised perhaps 70% of the reads [131]. Consequently it differed so greatly from the other sets that they were all poor choices for pooling. For example, profile correlations between the San Diego set and all others fell within narrow ranges (0.71–0.90 Pearson, 0.49–0.69 Kendall), and Bray-Curtis dissimilarities to all other sets were ∼1, the maximum. In comparison, set 4443700.3 had slightly wider correlation ranges (0.70–1.00 Pearson, 0.42–0.93 Kendall), that enabled better discrimination among sets (see Fig. 4.4). Its family Bray-Curtis dissimilarities to

sets had SAR11(α-proteobacteria) as their second most abundant member (18–20%), versus SAR86 (γ-proteobacteria) for the MB sets. The summer 2008 WEC sets assembled well with the MB sets despite this difference. Because only WEC set 4445064.3 assembled poorly, it seems likely that the change in *Rhodobacteraceae* from spring to summer was the major factor.

other sets fell within a wide range, 0.23–1.00, and that approach performed best.

Chapter 5

Discovery of ubiquitous marine bacteria by geographic profiling

5.1 Introduction

Chapter 3 hinted at the potential of pooled contigs for species discovery (§3.3.5 and Fig. 3.5). To this end a strategy to identify contigs that collectively may represent the same species would be essential for guiding follow-up research. For example, the decisions to sequence a type strain or to do a targeted gene study might be based on abundance, geographic range, and metabolic capacities inferred from the contigs thought to represent the same species.

In this chapter I describe the application of a new strategy for species discovery that works in concert with pooled assembly. Geographic profiling is an analytical aid to identify pooled contigs that might represent a single species—that is, closely related populations from different samples.[1] It does not use reference sequences and therefore can be used when close reference genomes are not known. On the other hand, if homologous sequences are available it can provide orthogonal evidence for species discovery.

With our new strategy we predicted three uncharacterized α-proteobacteria. To validate our predictions experimentally, we recovered DNA fragments for all three α-proteobacteria in a sample that was used to predict them. This entailed PCR with primers derived from pooled contigs, followed by cloning and sequencing of several PCR products. Clones were nearly sequence-identical to their associated pooled contigs. This corroborated the assembly. It also supported the ubiquity of the predicted α-proteobacteria, given the worldwide distribution of data sets that contributed reads to their pooled contigs. Finally, phylogenetic analysis placed each clone among the α-proteobacteria, and at <90% DNA sequence identity to its closest RefSeq species. This is consistent with our belief that the contigs may represent uncharacterized species.

5.2 Data and methods

5.2.1 Computational methods used with MG-RAST and clone data

Geographic profiles: Geographic profiling (described below) depends on read assignments to pooled contigs. Because the round robin approach to pooled assembly (§3.5.2.2) assigned reads to the first pooled contig that assembled them, it

[1] As in chapter 3, “species” is here used as an operational term, not as a precise biological kind. Limitations with the term were discussed in §1.1.3.

risked to introduce bias into the profiles. To avoid this, profiles were instead based on clustering. CD-HIT-454 [93] was used to cluster reads at $\geq$90% id to the pooled contigs, which served as the set of cluster seeds. Twenty-seven of the 42 data sets were clustered. The MX, MT and Sapelo Isl. sets were excluded due to poor intra-set assembly (Tab. 3.2) or poor mixing with other sets when pooled (Fig. 3.8).

Geographic regions were defined to group data sets derived from sample sites within a few hundred kilometers of each other (see Fig. 3.7 and Fig. 3.5). Pooled contigs were represented as vectors whose components were read counts from the geographic regions. However, the vectors were scaled to length one so that contig lengths would not impact clusters during principal component analysis. The scaled vectors were the geographic profiles of the pooled contigs.

Principal component analysis (PCA): PCA on pooled contigs was performed using R. Data was centered but not scaled. Although not scaling meant that large data sets could attract some contigs to the corners of the plot, we focused on off-corner clusters in which widespread geography outweighed this bias. Additionally, we only included contigs with reads from at least three geographic regions.

Percentage GC simulation: The percentage GC distribution for thirty cluster E contigs (166Kbp) that were thought to represent *Rhodobacteraceae* was calculated with the EMBOSS program `geecee` [145], and summary statistics calculated with R (Figs. 5.1 and 5.2). In order to see if a similar percentage GC distribution could be produced from a single *Rhodobacteraceae* genome, the thirty-contig cluster E distribution was compared to simulated distributions produced for each of 47 genomes downloaded from www.roseobase.org. Each simulated distribution included thirty randomly drawn sequences from a genome that, like the cluster E subset, totaled 166Kbp. Additionally, the thirty-contig cluster E distribution was compared to simulated distributions that included multiple *Rhodobacteraceae* genomes. For these simulations thirty contigs were drawn randomly from 2–47 genomes, also selected randomly. Genome mixtures at each level (2–47) were repeated ten times.

Codon usage: FragGeneScan [144] was used as described in chapter 3 to call ORFs for all contigs that were >5Kbp and assembled reads from at least three geographic regions. Then the EMBOSS program `cusp` [145] was used to calculate the codon frequencies per 1Kbp. Next a custom R script was used to (1) cull the contigs by selecting fifty at random from each of the colored groups assigned by geographic profiling; (2) perform PCA on the codon usage profiles for the selected contigs; (3) cluster the contigs by k-means, with k=10 (because the squared error leveled off at k=10 in a curve for $k \in \{1 \ldots 20\}$); (4) count the geographic profile cluster assignments within each codon usage cluster. The R script was run ten times.

Multiple sequence alignments and phylogenetic trees: Protein and DNA multiple alignments for the protein-encoding regions of cloned sequences (§5.2.2) were made with MUSCLE (default parameters) [39]. Only sequences likely to represent the same gene as the cloned sequence were included in multiple alignments: RefSeq genes had to span the entire ORF predicted for the clone by FragGeneScan, and Global Ocean Sampling (GOS; [148]) traces had to extend over protein domain boundaries. Also, GOS traces were rejected if their ends failed to match the cloned sequence, excluding low-quality initial and final bases typical of Sanger sequencing. (That is, GOS traces with truncated alignments were rejected.) Sequences meeting these coverage criteria were selected from the top BLAST [3] hits to NCBI RefSeq [139] proteins and the GOS reads in the Trace Archive [90] (database WGS 13694) for inclusion in multiple alignments.

Some cloned sequences were based on pooled contigs that assembled many reads from a single data set (e.g. contigs 06386 and 01453). In such cases the reads were presumed to represent a gene in an abundant species in a different (non-SIO Pier) population. For comparison to the cloned sequence, those reads were reassembled into a sub-contig that was included in the multiple alignment.

Multiple alignments were edited in Jalview version 2.9 [188], then uploaded to the phylogeny.fr web site [33]. There, PhyML version 3.0 [59] was used to create protein and DNA phylogenetic trees using LG and HKY85 substitution models, respectively, with 100 bootstraps. Trees were drawn with TreeDyn version 198.3 [23].

DNA multiple alignments and trees for intergenic regions used the same RefSeq and GOS sequences as above unless they diverged so strongly as to be unalignable.

5.2.2 Experimental validation with an SIO Pier sample

SIO validation sample: The sample we used was collected from the end of the SIO Pier in October 2006 and filtered to retain cells in the 0.45–2.0 μm range, as described in Palenik et al. (2009) [131]. Palenik et al. subsequently FACS-sorted for *Synechococcus*, and the derived metagenomic data set from which all our predictions were made is on MG-RAST as set 4443765.3. Ideally we would have performed PCR on the post-FACS-sorted sample on which we based our predictions. However, as only the whole-genome, sub-2.0 μm sample was available, we used that, herein referred to as the SIO validation sample.

Primer design: Contigs selected for PCR were reassembled from the reads they attracted during CD-HIT-454 clustering (see above). Using the EagleView [66] assembly viewer, the reads comprising the reassembled contigs were checked for areas of poor consensus over their whole lengths. If there were no clear signs of mis-assembly, candidate sites for primer design were selected, almost exclusively from sites spanned by reads from the SIO sample in order to target that specific population. Subsequences of an SIO read that exactly matched reads from other geographic regions were considered first. If no such subsequence could be found among any SIO reads within PCR range (<2Kbp), then SIO reads with few mismatches to reads from other geographic regions were used. As a last choice, non-SIO reads that exactly matched the consensus contig were used. Web-based primer design tools were used to promote high annealing temperatures because of the sample complexity [138], and to check for self-dimer and hetero-dimer interactions [71, 175]. Only primers for which none of the design tools found issues were then checked against NCBI/nt for specificity. The primers are described the Tab. 5.1. All were synthesized by Integrated DNA Technologies, Inc. (San Diego office).

PCR and imaging: PCR reactions were performed on a 10× diluted, whole-genome, sub-2-micron DNA lysate from the 2006 SIO Pier sample [131], and with the Promega GoTaq Green Master Mix kit according to manufacturer directions (Promega

Product	Primers		Melt
contig01965 to	green	5'-AGCTTGGGCAGCACCATA-3'	65 °C
contig02395	blue	5'-CAGCCGTGGTCACGTATACAA-3'	66 °C
	black	5'-GCGCTCGGTTGATATTCTTTGG-3'	65 °C
	red	5'-GTCCATGTGCCATCCTGTGA-3'	65 °C
contig06386	plum	5'-TCGGAGAATAAATGGCTCAAGT-3'	63 °C
	moss2	5'-gTAACAAATCATCATCAAGACGACCTG-3'	65 °C
contig01453	tangelo	5'-GTGAAGATTCTGATATGAGCAATGTC-3'	63 °C
	pink	5'-ATCTGCCCTATATGCCATTCC-3'	63 °C

Table 5.1: Primers for PCR products. All primers were designed from reads in data set 4443765.3 from the SIO Pier, except for moss2 which was designed from a Monterey Bay read. The one lower-case nucleotide in moss2 was a mismatch introduced to reduce self-dimerization. High melting temperatures were desirable to encourage specific amplification within the species-rich sample.

in Madison, WI, USA). The products discussed in this thesis were produced with the following thermocycling program: (1) Initial denaturation at 94 °C for 2 min.; (2) Denature at 94 °C for 30 sec.; (3) Anneal at 57 °C for 30 sec.; (4) Elongate at 72 °C for 30 min.; (5) Repeat steps 2–4 twenty-nine times; (6) Final elongation at 72 °C for 10 min.; (7) Hold at 4 °C. All PCR products were imaged by DNA gel electrophoresis, with 1.5% agarose gels stained with SYBR® Safe DNA Gel Stain and 1 Kb Plus DNA Ladders, both from Life Technologies (Carlsbad, CA, USA).

Cloning: Cluster E (yellow) and cluster F (cyan) clones were prepared in the same way. Following manufacturer instructions, we used the pCR® 2.1-TOPO® TA Cloning Kit with TOP10F' Chemically Competent *E. coli* cells (Invitrogen in Carlsbad, CA, USA). Ten clones and a negative control were grown for cluster E, and five clones and a negative control for cluster F.

Sequencing: Clones were Sanger-sequenced in both directions by Eton Bioscience, Inc. (San Diego, CA, USA) using universal primers M13F(-21) and M13R

Reverse that were complementary to the cloning vector. Clones were short enough that Sanger reads from opposite ends of the clones overlapped. However, internal primers were additionally used for two of the blue–green clones that bridged contig 01965 and contig 02395, to ensure high-quality base calls over the full length. (See red and black internal primers in Fig. 5.3.) This facilitated end-to-end DNA sequence for each clone by manually overlapping the Sanger reads and then trimming off the vector sequence. The end-to-end, trimmed DNA sequences will be referred to as "cloned sequence" in the remainder of this chapter.

5.3 Prediction of three uncharacterized, ubiquitous α-proteobacteria

By exploiting a feature specific to pooled contigs, we produced a "geographic profile" plot that when overlaid with other evidence suggested multiple uncharacterized marine species. The depth of reads assembled in the pooled contigs, and the geographically widespread sample sites of the data sets, could indicate that the species are both ubiquitous and abundant. Contigs representing three of the suspected species were selected for experimental follow-up.

5.3.1 Geographic profiling

Geographic profiling aims to cluster pooled contigs by species based on the proportions in which they assembled reads from different oceanic regions. For example, pooled contig03272 in Fig. 3.5 assembled reads from San Diego, the Monterey Bay, the western English Channel, and a Norwegian fjord in the proportions 2:58:3:85. Allowing several assumptions about how reads are sampled from a genome (below), contigs with similar proportions and high total read counts might represent the same species. Such contigs will form clusters when projected by principal component analysis of these proportions—their geographic profiles—and orthogonal evidence can be overlaid onto the clusters to further test whether the contigs represent the same species. In this way geographic profiling leverages the fragmented, partial representation of genomes characteristic of MMG data to help discover new species.

The approach rests on this claim: For a given species, the pooled assembly of two 454 data sets will produce contigs in which the ratio of reads contributed by each set tends to follow the ratio of the species' abundance in each set, that is S_A/S_B. This assumes that contigs usually do not mix species, shown in simulation in chapter 2, and that each 454 read is drawn at random from the genome, generally accepted.[2]

Such a ratio can be extended beyond two data sets by representing the contig as a vector whose components equal the reads contributed by each set. For example,

[2] It is usually assumed that each 454 read derives from a single cell [87], and from a random location in that cell's genome [102, 50]. The latter is sufficient for the ratio but overly strict. As long as the data sets distribute their reads across the genome using the same model, then model terms will cancel in the expected value calculation for the ratio of read counts to leave S_A/S_B.

Proof: Let $X_A(i)$ and $X_B(i)$ be the read counts for two data sets at position i of the (perhaps unknown) genome. They can be modeled as $X_A(i){\sim}S_Ap(i)$ and $X_B(i){\sim}S_Bp(i)$. S_A and S_B are constants that represent the total cells of the target species in each sample. From each cell one DNA fragment is drawn, independently, that ultimately manifests as a pyrosequenced read. The density $p(i)$, for drawing and pyrosequencing a fragment from the genome at position i, and assembling it correctly, is presumed the same for the two data sets because they are both 454 sets and assembled with the same software. The expected value of the ratio of read counts is then $\mathbb{E}[X_A(i)/X_B(i)] = \mathbb{E}[\frac{S_Ap(i)}{S_Bp(i)}] = S_A/S_B$.

contig03272 in Fig. 3.5 would have twelve components. However, since similar waters share species (e.g. see Fig. 3.8), grouping data sets by geographic region will concentrate reads in cases where species representation was low, and thus help with species discovery.[3,4] In the example, contig03272 drops to four dimensions.

With thousands of contigs to visualize, techniques for dimension reduction and clustering, here principal component analysis and k-means clustering, are essential for human recognition of contig clusters that might represent the same species. Although contigs representing different species will sometimes project near one another, resolution can be helped by overlaying orthogonal evidence (discussed below). Moreover, as a discovery approach we can choose to pursue the high-confidence clusters. As we shall see, they can harbor species that are both abundant and ubiquitous.

5.3.1.1 Relationship to differential coverage binning

In parallel with the development of geographic profiling, Albertsen et al. [1] explored a similar approach as a step to improve *de novo* metagenomic assembly of closely related data sets. The Illumina data sets they used represented two samples, before and after addition of hot phenol, and thus the same community but with shifted abundances. By separately assembling the data sets, cross-mapping the reads to the contigs, and clustering based on coverage by the sets, they assembled 31 genomes. That success spurred efforts by several other groups, and there are now several variants of

[3] Grouping recasts $X_A(i)$ and $X_B(i)$ in the proof above as sums across groups of data sets, i.e., $X_A(i) \sim \sum_{k \in A} S_k p(i)$ and similarly for B. The densities cancel and a ratio of read counts for the groups remains. Note that differences in sequencing effort among the data sets have no impact on the ratio.

[4] Without grouping, each data set would add a dimension and make the contigs occupy a smaller volume within which to recognize species clusters, the so-called "curse of dimensionality."

differential coverage binning [70, 75, 2]. All of them exploit coverage differences among pooled-assembled Illumina contigs, but differ in which other contig features they use (e.g. tetramer profiles) and clustering details. None of them group data sets by geography likely because their Illumina sets had high read counts. For our 454 data sets on the other hand, grouping by geography was necessary: Without it we found that contigs that were likely from the same organism, based on strong alignments to the same GOS reads, often were not in the same cluster in a geographic profile plot. In contrast, with geography-based grouping, GOS links rarely joined contigs from different clusters.

5.3.2 Illustrative rediscovery of two recently sequenced SAGs

Before geographically profiling them, we filtered the 242K pooled contigs that lacked species assignments by reference to Marine RefSeq (§3.2.1) to remove those with few reads, using length >5Kbp as a proxy. Our goal was to discover ubiquitous species, so we next filtered out contigs that assembled reads from fewer than three oceanic regions. Fig. 5.1 shows the geographic profile plot for the remaining 1729 unassigned pooled contigs. In what follows we focus on contigs off the corners of the triangle; corner contigs assembled most of their reads from a single geographic region.[5]

[5] The triangularity of the plot stems from the vector components summing to one, and three oceanic regions contributing most reads (WEC, NF, MB). For example, if there were only three geographic regions $x+y+z=1$, then contigs with most reads from x would be near the corner $(1,0,0)$, and contigs with reads from just regions x and y would be on the line from $(0,1,0)$ to $(1,0,0)$.

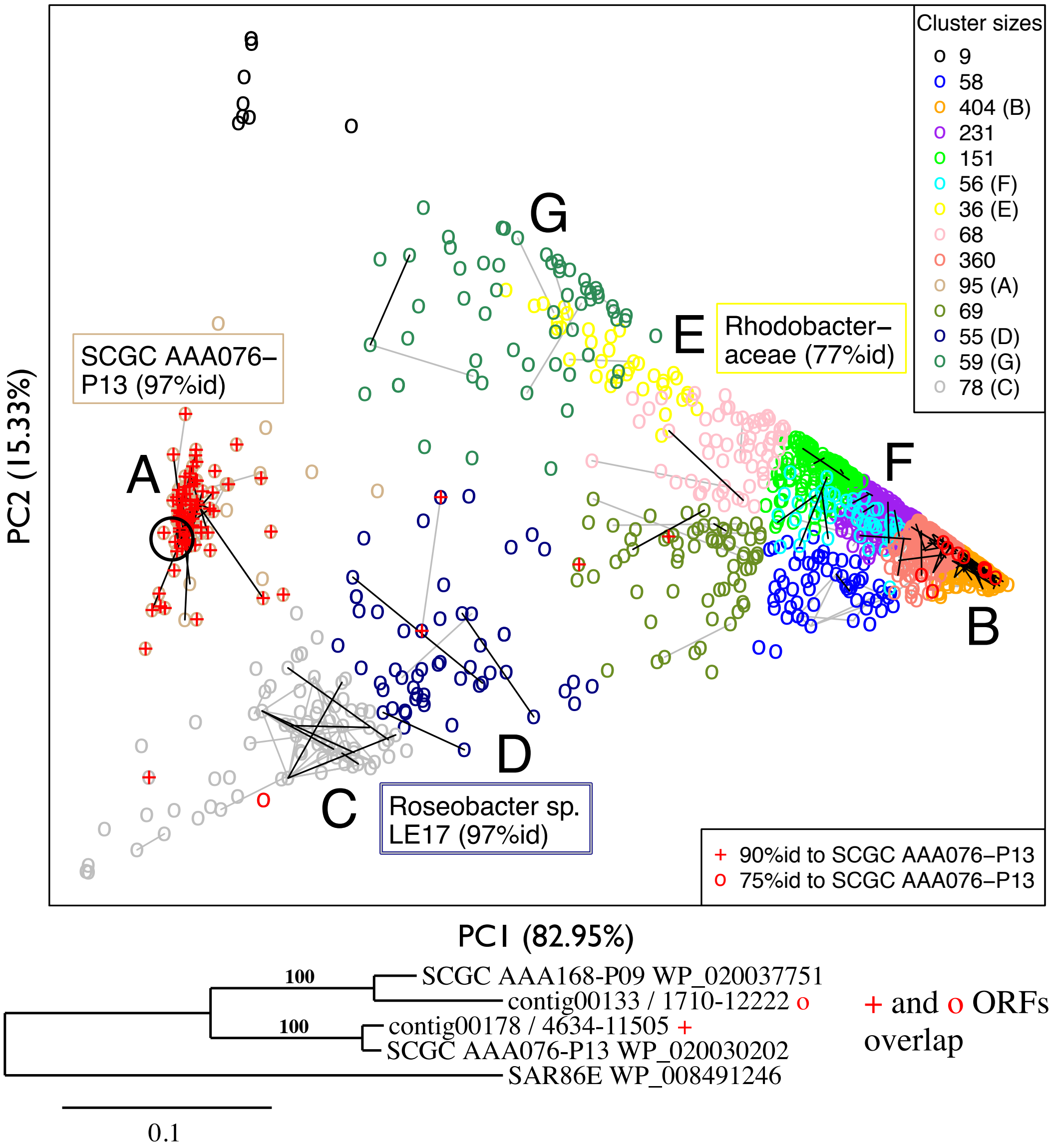


Figure 5.1: Geographic profile plot of 1729 pooled contigs that lacked species assignments by comparison to Marine RefSeq proteins (August 2012). Some clusters comprised mostly one species. For example, cluster A (tan) has 88 contigs that map at ≥90% id to the RefSeq (May 2013) genome for γ-proteobacterium SCGC AAA076-P13 (+). Contigs that map at only ≥75% id to this genome (o) concentrate in cluster B (orange) suggesting a closely related but distinct species. (continues)

Figure 5.1: (continued) Clusters C (gray) and D (navy blue) are enriched in *Roseobacter* sp. LE17. Cluster E (yellow) has hits to various members of family *Rhodobacteraceae*. Gray lines join contigs that align to ≥10 of the same GOS traces at ≥90% id and ≥100 nt. Black lines join contigs that align separately to each trace in a GOS mate pair (same cutoffs). PC1 and PC2 component loadings are, respectively: WEC=0.81, NF=-0.52, MB=-0.28, and MB=0.75, NF=-0.64, WEC=-0.16 (other loadings ∼0). *Bottom:* One of five phylogenetic trees corroborating that the + and o contigs represent different species of γ-proteobacteria, the former more closely related to SCGC AAA076-P13 and the latter to SCGC AAA168-P09.

5.3.2.1 γ-proteobacterium SCGC AAA076-P13

Late during preparation of our manuscript (chapter 3), we determined that contig03272 (Fig. 3.5) represented the recently sequenced γ-proteobacterium SCGC AAA076-P13, a single-cell isolate from the Gulf of Maine, sequenced as part of the Joint Genome Institute effort to generate marine reference genomes.[6] Based on its read depth and broad geographic distribution of data set sample sites, we anticipated that other contigs might represent AAA076-P13. With the reference genome now in hand, we tested our approach for identifying them.

All 1729 unassigned contigs were aligned to the SCGC AAA076-P13 genome at cutoff 90% DNA sequence identity (+). As it turned out, eighty-three of ninety-five contigs in cluster A (tan color) mapped on average at 97% id to the P13 genome (Fig. 5.1, +). Moreover, using a looser cutoff of 75% id (o) we identified fifty-five contigs mostly within cluster B (orange), suggesting that geographic profiling can sometimes separate closely related genomes.

[6] JGI proposal 387: Generating reference genomes for marine ecosystem research: Single cell sequencing of ubiquitous, uncultured bacterioplankton clades

To more carefully check whether the + and o contigs represented different species, six contig pairs (+,o) that represented homologous protein-encoding regions were analyzed phylogenetically. Five of six phylogenetic trees supported that the + and o contigs represented different species, the former more closely related to SCGC AAA076-P13 and the latter to γ-proteobacterium SCGC AAA168-P09. An example tree for carbamoyl phosphate synthase is shown; γ-proteobacterium SAR86E is an outgroup (Fig. 5.1, bottom).

5.3.2.2 *Roseobacter* sp. LE17

Another positive example for the approach was found in clusters C (gray) and D (navy blue) whose contigs represented a close relative of the recently single-cell-sequenced genome of *Roseobacter* sp. LE17. Sixty of 78 contigs in cluster C and 42 of 55 in cluster D aligned on average at 97% id to LE17. Note that k-means clustering is only a visual aid, not a classifier; and so the presence of LE17 in adjacent clusters C and D is not a concern. Overlaid taxonomic information guides cluster interpretation. Additionally, Sanger sequences from the Global Ocean Sampling Expedition (GOS) can be overlaid. For example, the black lines between contigs denote GOS mate pairs where each mate aligned at >90% id to a different contig, suggesting that the contigs likely represent the same species.

Together these two examples demonstrated that geographic profiling can sometimes aggregate contigs from a single species into a clear cluster and such that close relatives can be excluded. They also demonstrated that homology-based evidence and

GOS data can be overlaid to aid human recognition of species clusters.

Cluster	**Closest RefSeq genome to cluster contigs**			**Cluster vs. RefSeq**		
		Size (MB)	%GC	%GC	%id	%cvg
A	SCGC AAA076-P13	1.3	33.6	34.3±2.8	97	37
C	*Roseobacter* sp. LE17†	∼3	54.4	55.7±2.0	98	10
D	*Roseobacter* sp. LE17†	∼3	54.4	48.4±9.3	97	4.7
E	various *Rhodobacteraceae*	4.2	60.0	50.4±3.6	77	<5.4
F	various α-proteobacteria	1.3–2.2	-	31.8±4.8	-	-

Table 5.2: Contig clusters compared to their closest RefSeq genomes. Coverage estimates used only contigs that aligned at ≥90% id and ≥800 nt to the RefSeq genome(s). Percentage identities are contig averages. For E, the median size and %GC for the *Rhodobacteraceae* in Tab. 1 in Newton et al. (2010) [123] are shown. † draft genome.

5.3.3 Cluster selection for species discovery

The remaining clusters in which we set out to discover new species were less dense, had fewer GOS links, and (necessarily) lacked close reference genomes. Nevertheless several lines of evidence led us to select clusters E (yellow) and F (cyan) for discovery and experimental follow-up.

5.3.3.1 Cluster E

As of March 2014 the closest RefSeq genomes to contigs in cluster E are from various genera primarily within family *Rhodobacteraceae*. Sixteen of the contigs aligned to as many species in that family, but at only 72–88% id (mean = 77% id) and with 2–29% contig coverage (Fig. 5.2). The lack of GOS mate pair links between the contigs could suggest that the cluster E contigs represent multiple species.[7] However, given

[7] The single gray line within cluster E represents 24 GOS traces that aligned to the two contigs at the cutoffs described in Fig. 5.1. Thirteen other contigs in cluster E would have been similarly linked except that none of their pairs had ≥10 traces shared.

the low sequence similarity another explanation is that genes that were nearby in GOS sequences (∼4Kbp insert libraries) were not so in whatever genome(s) cluster E represented.

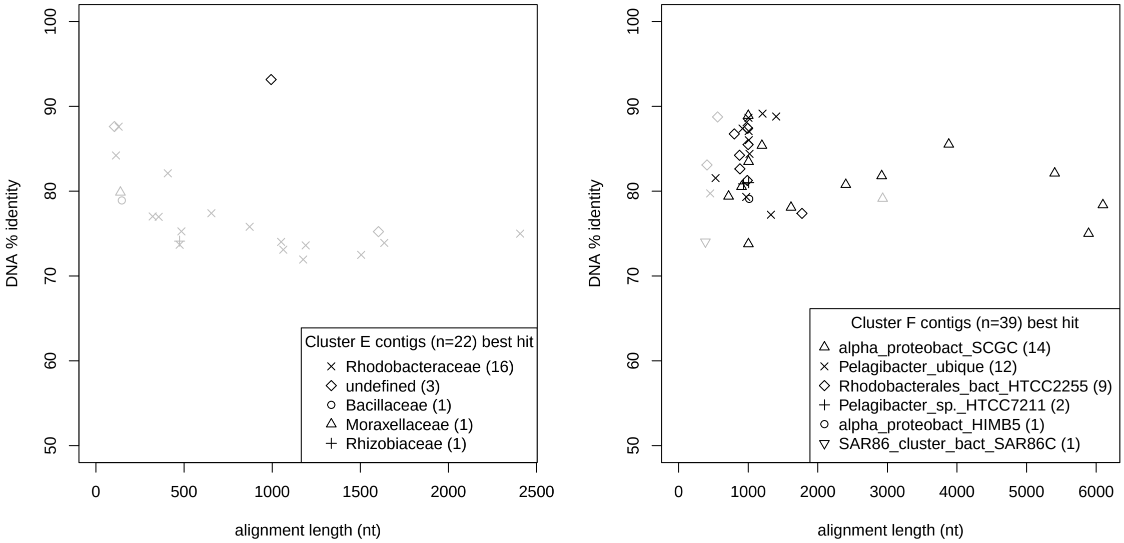


Figure 5.2: **Left** Families of the closest RefSeq bacterial genome hits for cluster E contigs (July 2014 NCBI megablast, filtered at ≥50% id). The sixteen hits to *Rhodobacteraceae* include as many species and twelve genera; *Leisingera*, *Phaeobacter*, *Ruegeria*, and *Sulfitobacter* each have two species. Most alignments covered <50% of the contig (gray); the sixteen hits to *Rhodobacteraceae* covered only 2–29% (median 12%). **Right** Species of the closest RefSeq bacterial genome hits for cluster F contigs (searched as before but also filtered at ≥30% contig coverage). Thirty-four contigs aligned with coverage ≥50% (black), five with coverage <50% (gray). *Pelagibacter ubique* (×) includes seven strains (HTCC8051 and HTCC7217 have four and three contigs). The SCGC α-proteobacteria (△) include five types (AAA076-C03 has six contigs, all others have two).

Nevertheless, we did make several observations that added to evidence from geographic profiles and homology that the contigs "go together:" First, the %GC was 50.4±3.6% for thirty cluster E contigs, the sixteen with best hits to *Rhodobacteraceae* and the fourteen not assigned to other genomes (Fig. 5.2, left). This falls within the %GC range reported by Newton et al. (2010) [123] for thirty-two representative

Roseobacters[8] (38.0–70.0%, median 60.0%) and is nearest that of Rhodobacterales bacterium HTCC2150 at 49%, not among the hits. More surprising was the consistency of %GC among the contigs (s.d.=3.6%), considering that they spanned only 166Kbp, or about 3.1–5.4% of the thirty-two representative genomes. We compared the observed %GC distribution to simulated distributions at the same "sequencing effort" (thirty contigs, 166Kbp) for every available Roseobacter genome at Roseobase (§5.2.1). To our surprise, for each genome the simulated contigs (i) had a mean %GC that differed by just 0.4% on average (max 1.8%) from that of the whole genome, and (ii) had standard deviations of just 2.5% on average (IQR 2.2–2.8%). This suggested that the thirty contigs, if from a single Rhodobacterium, could provide a rough estimate of %GC. In contrast, when simulated contigs from 2–47 different genomes were measured together, the standard deviations were higher, 5.8% on average (IQR 5.5–6.2%). Thus the 3.6% standard deviation observed for the cluster E subset was nearer that expected for a single genome.

Second, codon usage profiles for ORFs in cluster E contigs were more similar to each other than to those of other clusters. Specifically, in each of ten experiments, fifty ORFs were drawn at random from each of the fourteen geographic profile clusters in Fig. 5.2, and clustered based on their codon usage profiles (§5.2.1). On average thirty-nine of the ORFs drawn from cluster E were assigned to the same codon usage cluster (range, 27–46). I.e. usually, for cluster E, *codon usage* clusters for ORFs corroborated *geographic profile* clusters for contigs.

[8] "Roseobacter" refers to a lineage that includes over forty-five members mostly from family *Rhodobacteraceae*. It does not refer to the (italicized) genus *Roseobacter*.

Because cluster E contigs had on average 20% of their reads from a 2006 Scripps Pier data set published by Palenik et al. (2009) [131], we approached Prof. Palenik to collaborate in validating several of the contigs experimentally. We also identified cluster F contigs as having an appreciable proportion of reads from that data set.

5.3.3.2 Cluster F

Cluster F appeared to be comprised largely of three groups of α-proteobacteria (Fig. 5.2, right): One was most similar to several *Pelagibacter ubique* strains; another was most similar to Rhodobacterales bacterium HTCC2255 or SCGC AAA076-C03 which seem closely related;[9] a third group was most similar to four other SCGC isolates, AAA[015-N04, 160-J14, 536-G10 and 536-K22]. Unlike for cluster E, most contigs aligned over nearly their full lengths (median coverage, 98%; median length, 1004 nt), compared to median bacterial protein-encoding gene lengths of 801 nt and 1419 nt, reported in 2005 and 2011, respectively [17, 186].[10] It therefore seems likely that the alignments captured large portions of genes, and were not truncated at protein domain boundaries.

The percentage GC for cluster F also supported that it represented a different species than cluster E. At 31.8±4.8%, it was much lower and more variable than that measured for cluster E, and consistent with the ranges for closest RefSeq species (28.8–29.9% for eleven *Pelagibacter ubique* genomes, and 30.6–36.7% for the SCGC genomes; NCBI, April 8, 2015). The higher variability might also be partially explained by the

[9] They have identical 16S rDNA and are >97% id overall [99].

[10] The 2005 study was based on sixty-seven genomes and the 2011 study on 478.

smaller amount of sequence (74Kbp). However, a mix of species seems the main factor. We also note that unlike for cluster E, cluster F ORFs did not have a similar pattern of codon usage (consistent with a mix of species).

Three GOS mate pairs linked three distinct pairs of cluster F contigs, perhaps supporting that those contigs at least represented the same species (black lines in Fig. 5.1). For one pair the best hits to the two contigs were both to *Pelagibacter ubique* (≥99% id). However a second pair of GOS-linked contigs had best hits to different species, an SCGC genome (90% id) and *Pelagibacter* sp. HTCC7211 (92% id). The third pair of linked contigs did not have hits that met the genome alignment cutoffs (50% id, 100 nt, 30% coverage).

5.4 Experimental validation and analysis of predicted α-proteobacteria sequences

Our biological hypotheses are that clusters E and F represent three novel species of uncharacterized, abundant, ubiquitous α-proteobacteria. The PCR, cloning and sequencing results presented in this section show that several of the pooled contigs from these clusters faithfully represent real DNA sequences from organisms in the SIO sample. We offer them, and other observations summarized in Tab. 5.3, as preliminary evidence that our hypotheses are correct.

Clone (cluster)	**Clone % id to its contig**	**Closest RefSeq genome to clone**	**Evidence clone is chromosomal**	**Where seen**
01965 to 02395[†] (E)	99.1–99.9%	*Rhodobacteraceae* (various, 67–72% id)	SMC gene; RefSeq synteny	SIO, MB, WEC, Med
06386 (F)	98.5%	SCGC AAA076-C03 (81% id)	RefSeq synteny	SIO, MB, WEC
01453 (F)	97.8%	SCGC AAA536-G10 (87% id)	none for or against	SIO, MB, WEC, NS

Table 5.3: Summary of evidence supporting that the contigs/clones represent uncharacterized, ubiquitous species. For the last column, the SIO clones were required to align fully and at $\geq$90% id to sequences from the indicated sample sites: MB = Monterey Bay; WEC = western English Channel; NS = Nova Scotia; Med = Mediterranean.
[†] Five clones for the cluster E bridge were sequenced.

We also note that the successful PCR and sequencing support two methodological hypotheses: First, chimeric, pooled-assembly contigs can be used to interpret uncultured, environmental sequences. This had only been shown in simulation, and only for RefSeq genomes in chapter 3. Second, geographic profiling can help bin contigs that represent the same DNA molecule.

5.4.1 Roseobacter in cluster E

5.4.1.1 Contig selection

Pooled contigs 01965 and 02395 in cluster E were selected for experimental validation because they appeared to represent a novel species within family *Rhodobacteraceae*. Although both contigs contained ORFs with predicted homologs to RefSeq proteins in the family (Fig. 5.3, top axes), at the nucleotide level both contigs differed substantially from all *Rhodobacteraceae* RefSeq genomes: Contig 01965 had no RefSeq genomic hits, and the top hits for contig 02395 spanned five genera (74–79% id, ≤655 nt).

That contigs 01965 and 02395 represented a single genome was suggested by four GOS traces that bridged them, spanning 342 nt of missing sequence between the contigs.[11] That the contigs were not extra-chromosomal was suspected from a predicted chromosome segregation (SMC) protein in contig 02395 (discussed below).

[11] The four GOS traces aligned to the contigs at 94–99% id. The alignments only, and fully, included the ends of the traces and the contigs, up to 444 nt of contig 01965 (the 3' end in Fig. 5.3) and 616 nt of contig 02395 (5' end). We verified that contigs 01965 and 02395 were not separate due to an assembly artifact: No 454 reads that could bridge the two contigs were found, using the GOS traces to search for candidate reads within the forty-two MMG data sets.

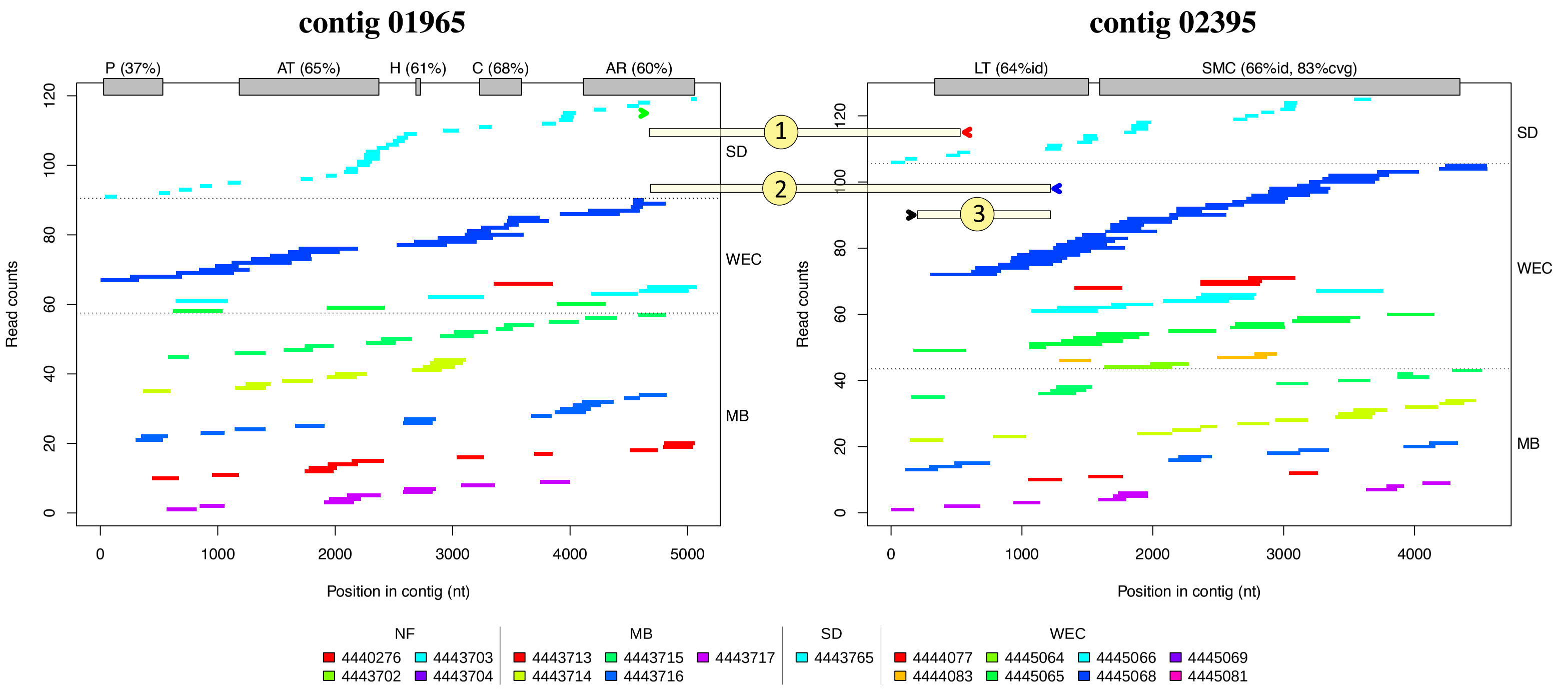


Figure 5.3: Pooled contigs 01965 and 02395 from cluster E. Reads are colored by data set (legend) and grouped vertically by the geographic region of the sample site. Best RefSeq protein hits of the contig ORFs are: P = peptide deformylase; AT = aminotransferase; H = hypothetical; C = CheY; AR = alpha ribazole 5-P phosphatase; LT = lytic transglycosylase; SMC = chromosome segregation protein. ORFs fully cover each protein hit at the indicated DNA percentages identity, except for SMC (83% coverage). PCR primers (colored arrows) were designed from the indicated SD read to target that population. PCR products ① and ② bridged the two contigs to test whether they represent the same species (main text).

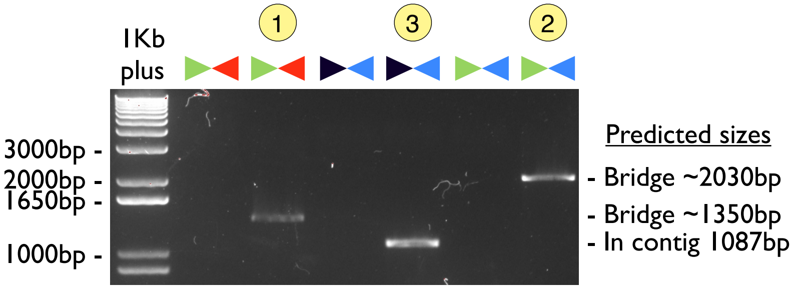


Figure 5.4: PCR products with primers depicted in Fig. 5.3. Products that bridge the contigs (lanes 3, 7) suggest that they represent nearby sequences from the same species. Negative controls (lanes 2, 4, 6) had primers but no template DNA.

5.4.1.2 PCR and sequencing

All PCR targets defined in Fig. 5.3 were successfully amplified as shown in Fig. 5.4,[12] and with sizes close to those predicted. For the bridge products ① and ②, predicted sizes presume 342 missing nucleotides, based on the four GOS traces. Note that had the GOS traces been unknown, we would still suspect that the bridge products represent the same DNA region: Both start at the green primer, but they end at primers blue and red that contig 02395 predicted to be 663 bp apart, quite close to the 500–700 bp observed in the gel.[13]

[12] The red–black product is in another gel (not shown).

[13] If funding were available to try many combinations of primers and with long-range PCR kits, then observations like this could be used to gather evidence for which contigs are connected, i.e. without GOS support.

To test whether the PCR products captured one DNA region or multiple, ten clonal colonies that each contained a ② bridge PCR product were restriction-digested (Fig. 5.5). All ten inserts appeared to be the same size and thus unlikely to represent different DNA regions. Sanger sequencing of five inserts confirmed this more or less: the inserts were $\geq$98.3% id (mean 99.2% id).[14] Whether the cloned sequences represent a single *strain* is less clear, with on average eight differing base pairs per kilobase. However this may reflect Sanger sequencing errors given sequences that derived from the very same clone were not always identical (99.2–100.0% id for end-primed versus internally-primed sequences).

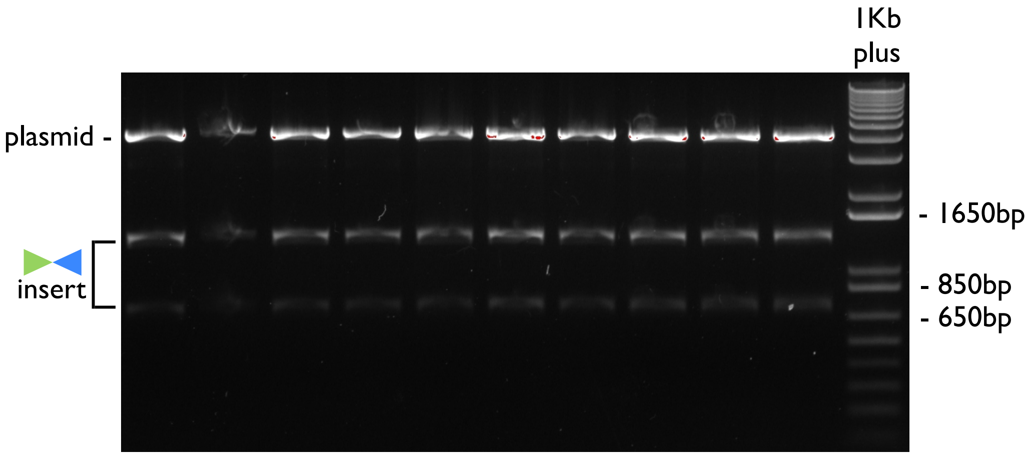


Figure 5.5: Ten clones of the ② bridge PCR products. Lanes show the restriction digests of plasmids harvested from *E. coli* colonies. Each colony was grown from a cell transfected with plasmids into which ② PCR bridge products had been inserted (see Figs. 5.3 and 5.4).

[14] This is based on all-by-all alignments of the higher quality range of the sequences from positions 40 to 800. The full sequences ranged from 1135–1235 nt.

5.4.1.3 Analysis of cloned product

A search among RefSeq proteins suggested that the cloned sequences captured two genes, a cobalamin biosynthesis gene, cobD,[15] and most of a lytic transglycosylase, possibly membrane-bound. Nearly all hits were from *Rhodobacteraceae*. A search among GOS data (>98% id) identified traces that expanded the range of the represented organism to the western shores of the North Atlantic.[16] ORFs predicted for these traces, and the RefSeq proteins, were included in our analysis.

Phylogenetic trees for the clone cobD and lytic transglycosylase protein and DNA sequences placed them consistently in a clade with just the environmental sequences (Figs. 5.6 and 5.7). Both trees showed a more recent Roseobacter ancestor to the clones than *Rhodobacteraceae* bacterium HTCC2255, the most deeply branched Roseobacter in the lineage [123]. A separate analysis with nearly the full-length RefSeq protein sequences also suggested a Roseobacter ancestor (trees not shown).[17] Although clearly among the Roseobacters, with only two genes we cannot more precisely assign our clones, e.g. to a specific Roseobacter clade. Additional sequencing might enable this. Robust Roseobacter phylogenetic trees have been created using universal single-copy genes [123] (though not for 16S rRNA alone [18]).

[15] CobD is absent in Fig. 5.3 because it was only partially covered by contigs 01965 and 02395 and thus not a strong RefSeq hit.

[16] A more recent search for the cloned sequences in selected Tara Oceans data sets further expanded the range to the Mediterranean Sea: Two HSPs, each >94% id, covered 77% of the clones.

[17] *Rhodobacteraceae* bacterium HTCC2150, the second most deeply branched Roseobacter, was used as the outgroup for the full-length protein trees. GOS and WEC sequences were excluded from these trees because they only partially captured the genes and so would have introduced gaps into the multiple alignments. Without GOS and WEC, the cobD and lytic transglycosylase multiple alignments had 285 and 297 positions, respectively.

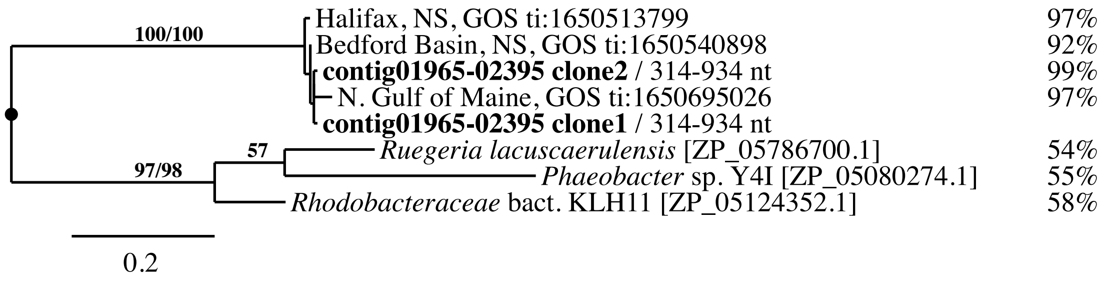

100/100
Halifax, NS, GOS ti:1650513799 97%
Bedford Basin, NS, GOS ti:1650540898 92%
contig01965-02395 clone2 / 314-934 nt 99%
N. Gulf of Maine, GOS ti:1650695026 97%
contig01965-02395 clone1 / 314-934 nt
57
Ruegeria lacuscaerulensis [ZP_05786700.1] 54%
97/98
Phaeobacter sp. Y4I [ZP_05080274.1] 55%
Rhodobacteraceae bact. KLH11 [ZP_05124352.1] 58%
0.2

Figure 5.6: Protein phylogenetic tree for a cobalamin biosynthesis protein (cobD) predicted for 5' nucleotides of clones 1 and 2 (bold) and homologs that also derive from individual cells. The corresponding DNA tree supported nearly the same phylogeny; bootstrap values for shared branches are given (protein/DNA). The trees were rooted (black disc) using as outgroups the Roseobacter cobD sequences from Rhodobacterales bacteria HTCC2150 (ZP_01741958.1) and HTCC2255 (WP_008035845.1). The BLASTX percentages identity of clone 1 to each sequence are shown at the right. Reference sequences were selected because they were among the top hits when clone sequences (DNA and ORFs) were searched for in RefSeq genomes and proteins. NS = Nova Scotia

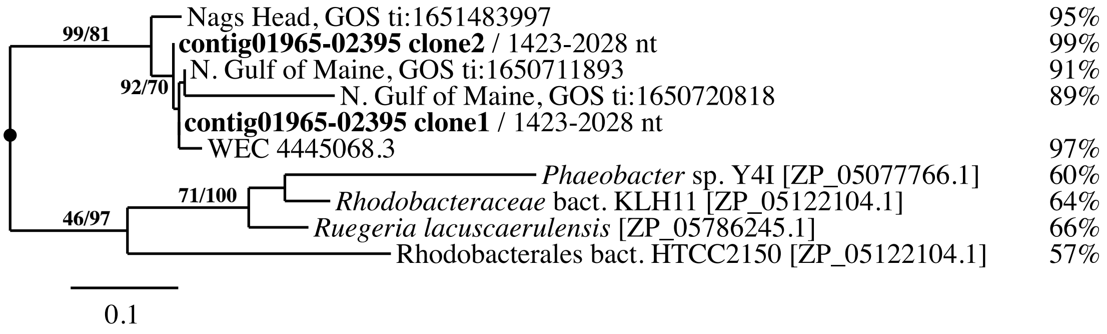

99/81
Nags Head, GOS ti:1651483997 95%
contig01965-02395 clone2 / 1423-2028 nt 99%
92/70
N. Gulf of Maine, GOS ti:1650711893 91%
N. Gulf of Maine, GOS ti:1650720818 89%
contig01965-02395 clone1 / 1423-2028 nt
WEC 4445068.3 97%
Phaeobacter sp. Y4I [ZP_05077766.1] 60%
71/100
Rhodobacteraceae bact. KLH11 [ZP_05122104.1] 64%
46/97
Ruegeria lacuscaerulensis [ZP_05786245.1] 66%
Rhodobacterales bact. HTCC2150 [ZP_05122104.1] 57%
0.1

Figure 5.7: Protein phylogenetic tree for a lytic transglycosylase, possibly murein, predicted for 3' nucleotides of clones 1 and 2 (bold). All sequences derive from individual cells except WEC 4445068.3. Branches that were also supported by the corresponding DNA tree have bootstrap values given (protein/DNA). The trees were rooted (black disc) using as an outgroup a homolog from a Roseobacter clade member, Rhodobacterales bacterium HTCC2255 (WP_008035268.1). BLASTX percentages identity of clone 1 to each sequence are shown at the right. Note that different GOS traces overlapped this gene and cobD in Fig. 5.6. The RefSeq sequences were selected as explained in Fig. 5.6.

Synteny also supported that the contigs and clones represent chromosomal DNA from the Roseobacter lineage: The Roseobacter RefSeq species in Figs. 5.6 and 5.7 all have nearly every gene as shown in Fig. 5.3 on chromosomal DNA. All have an alpha ribazole-5-P phosphatase, cobD, lytic transglycosylase, and SMC. In contrast two bacteria that were reported to be in abundance in the SIO sample, *Pelagibacter ubique* str. HTCC1062 and the coastal *Synechococcus* sp. CC9311, do not [131]. Moreover, SMC is exclusively chromosomal among the *Rhodobacteraceae*, with two doubtful exceptions.[18]

[18] The exceptions are two hypothetical proteins in plasmids from *R.sphaeorides* ATCC 17025 (ABP72178.1) and *D.shibae* DFL 12 (ABV95595.1). The SMC-like regions in these genes are much shorter ($\leq$1542 nt) than the more typical SMC length reflected in Fig. 5.3, and in neither plasmid are the neighboring genes similar to those shown. Both organisms have full-length SMC genes on their chromosomes, with the usual synteny for *D.shibae* DFL 12.

5.4.2 Roseobacter in cluster F

5.4.2.1 Contig selection

Contig 06386 was selected because it appeared to represent one of the three abundant α-proteobacteria groups identified in §5.3.3.2, and a different type of *Rhodobacteraceae* than in cluster E. It was also a practical choice because it had enough SIO reads to design primers for both short (i.e. easy) and long PCR products. Finally, the long product would enable analysis of coding and intergenic base pairs.

5.4.2.2 PCR and sequencing

Both the short (403 bp) and long (1227 bp) PCR products defined in Fig. 5.8 were amplified and found to be near the predicted sizes (Fig. 5.9). Five of five clones for the long product were ~1200 bp, consistent with amplification of only the targeted region (restriction digest not shown). Sanger sequences assembled from both ends of one clone corroborated the length (1227 bp) and were identical over the 225 bp they overlapped.

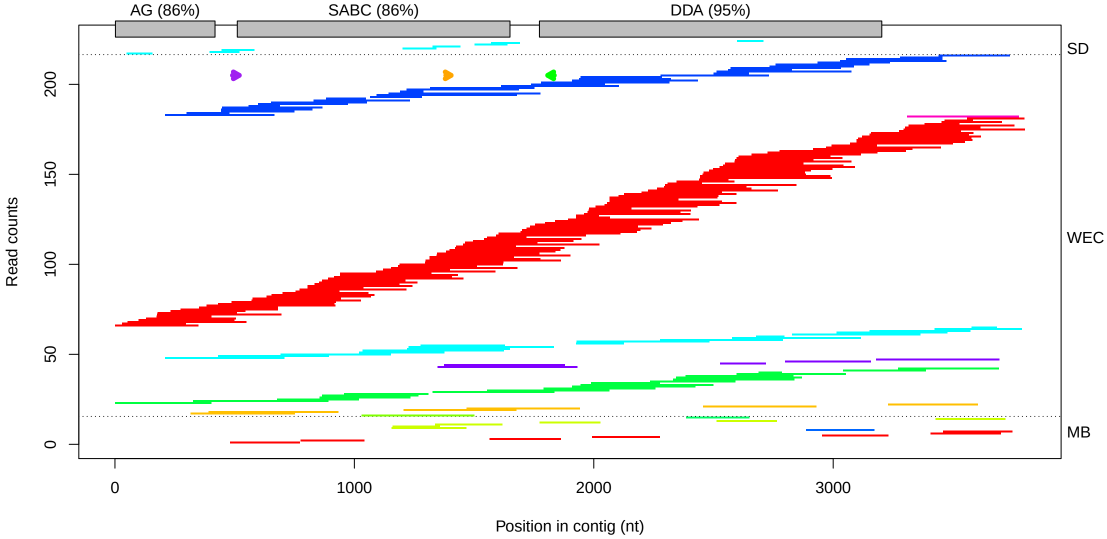


Figure 5.8: Pooled contig 06386. Reads are colored by data set and grouped by geographic region as in Fig. 5.3. Best RefSeq protein hits of the contig ORFs (top) are: AG = alpha-glucosidase; SABC = sugar ABC transporter ABC-binding protein; DDA = dihydropyrimidine dehydrogenase subunit A. ORFs fully cover each protein hit at the indicated DNA percentages identity, except for the alpha-glucosidase (26% coverage). The PCR targets shown here were validated in Fig. 5.9.

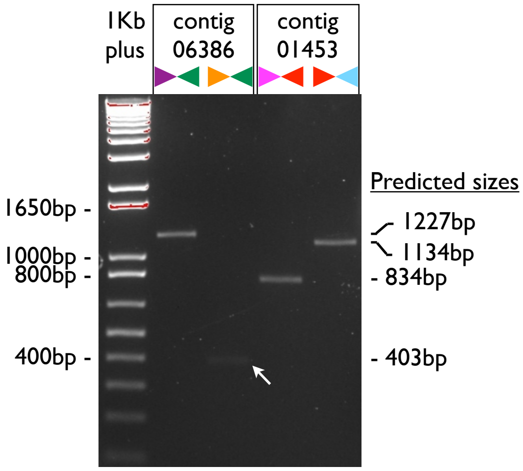


Figure 5.9: Gel image for PCR products for contigs 06386 and 01453 using primers indicated in Figs. 5.8 and Fig. 5.11, respectively. All PCR products had sizes close to those predicted. The plum–moss2 and tangelo–pink products were subsequently cloned and sequenced. Negative control lanes, for the same primers but no SIO template, were run in a gel one month prior and produced no bands (not shown).

5.4.2.3 Analysis of cloned product

The entire cloned sequence aligned to contig 06386 at 98.5% id; therefore the contig, though chimeric, closely approximated real DNA sequence. The seventeen mismatch positions between the clone and contig 06386 did not likely stem from pyrosequencing errors given the depth of the western English Channel reads and their high sequence similarity (Fig. 5.8).[19] Neither were the mismatches explicable only by population differences: At ten of the mismatch positions variation was observed among the western English Channel reads. Only at three positions did variation segregate by geography, with the California reads and the clone unanimous, but different from the western English Channel reads, also unanimous.[20]

The closest RefSeq protein to that represented by the clone was the maltose ABC-type transporter MalK (E-value$\sim$0, 88% amino acid id to WP_020056776), from a single-amplified Roseobacter genome, SCGC AAA076-C03 ("SAG C03") [99]. Serendipitously, when we searched for the clone MalK among the pooled contigs we found a single hit, to a cluster G contig, which aligned to the SAG C03 genome at 99% id and 100% coverage. This is reminiscent of the example in Fig. 5.1 with SCGC AAA076-P13 (+) and its close cousin in cluster B (o). As in that example, here phylogenetic trees showed that the MalK clone was distinct from its RefSeq cousin, SAG C03 (Fig. 5.10). Moreover, nearly the same topology and strong bootstrap support in Fig. 5.10 were observed

[19] Pyrosequencing error rates are <1% (Tab. 1.2). In contrast, 81 of 84 western English Channel reads that fell within the cloned range clustered (single-linkage by blastclust) at 98% id and 30% coverage.

[20] Though not quantitative, these patterns were clear in the assembly viewer. Read depth was always >20 and the differences at the positions referred to were striking compared to the near unanimity in the viewed region.

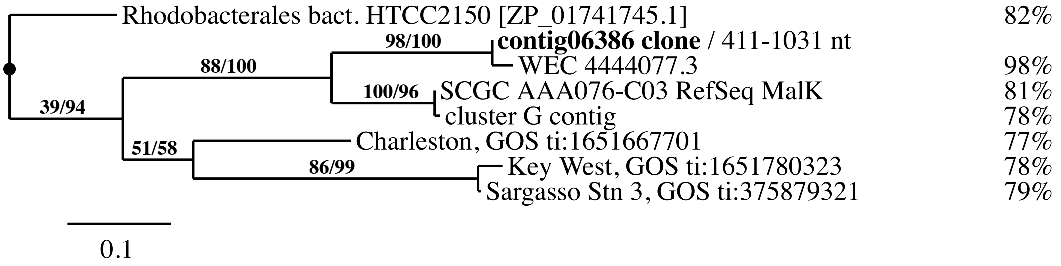

Rhodobacterales bact. HTCC2150 [ZP_01741745.1]
82%
98/100
contig06386 clone / 411-1031 nt
88/100
WEC 4444077.3
98%
100/96
SCGC AAA076-C03 RefSeq MalK
81%
39/94
cluster G contig
78%
51/58
Charleston, GOS ti:1651667701
77%
86/99
Key West, GOS ti:1651780323
78%
Sargasso Stn 3, GOS ti:375879321
79%
0.1

Figure 5.10: Phylogeny of the contig 06386 cloned PCR product and selected homologous sequences also from single cells (except WEC 4444077.3). The closest RefSeq protein is MalK in the Roseobacter SCGC AAA076-C03. The GOS traces are presumed homologous because they partially cover both protein domains in MalK (PFAM00005, PFAM08402). The WEC 4444077.3 contig was assembled only from reads in that data set because it seemed enriched in the target species (Fig. 5.8). The tree is for the 207 C-terminal amino acids of MalK (193 alignment positions). A DNA tree for the same region (559 alignment positions) had identical topology and so bootstrap values for both trees are given (protein/DNA). The trees were rooted (black disc) using as an outgroup a homolog from a Roseobacter clade member, Rhodobacterales bacterium HTCC2255 A05 from the Monterey Bay. On the right are the BLAST percentages identity between the clone and each sequence, not restricted to the multiple alignment positions.

in a DNA tree for the first 141 intergenic bases after MalK (not shown).[21] Altogether the trees and pairwise alignments (Fig. 5.10, right) indicate that the cloned sequence represents a MalK not in RefSeq nor observed by GOS, yet present in an organism(s) off the California coast and the western English Channel. Close cousins, with respect to this gene, have been observed as far south as Bermuda (Sargasso Stn. 3, the only open-ocean site in Fig. 5.10) and along the North American east coast as far north as the Gulf Maine, where SAG C03 was isolated [99].

We assume that contig 06386 was correctly assembled—given that it was 98.5%

[21] All sequences appeared to have homologous intergenic sequence after the MalK stop codon except for Charleston which was excluded from the multiple alignment. The resulting intergenic tree for the other six sequences had bootstraps of 100 and the same topology as the one shown in Fig. 5.10, with one exception: A common ancestor was shared between the contig 06386 clone and the clade with SAG C03 and the cluster G contig (bootstrap = 39), followed by a common ancestor among those three and the WEC contig.

id to the clone and deeply covered by reads (Fig. 5.8)—and note that the gene order in the contig agrees with that in the draft SAG C03 genome: 3496 nt of the 3515 nt contig aligned at 85.6% id to the genome. Because this alignment was to a large scaffold (412Kbp) it seems likely that the DNA fragment we cloned is chromosomal.

5.4.3 α-proteobacterium in cluster F

5.4.3.1 Contig selection

Contig 01453 was selected for similar reasons as contig 06386 except that it represented the third group, a mix of SCGC genomes.

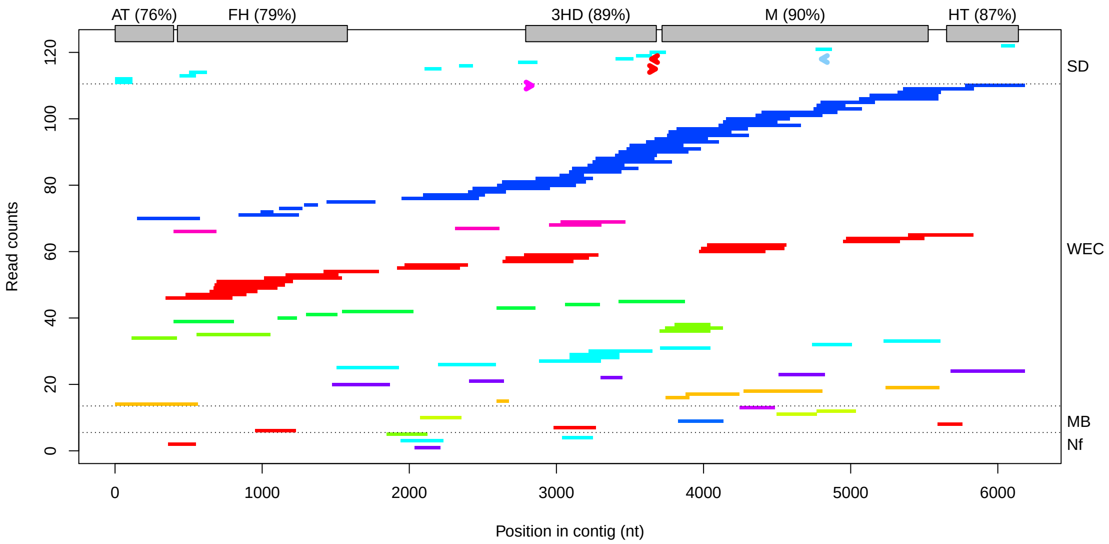


Figure 5.11: Pooled contig 01453. Reads are colored by data set and grouped by geographic region as in Fig. 5.3. Best RefSeq protein hits of the contig ORFs are (top): FH = fumarylacetoacetate hydrolase; 3HD = 3-hydroxyisobutyrate dehydrogenase; M = monooxygenase; AT = acyl-CoA transferase; HT = hydratase. ORFs fully cover each protein hit at the indicated DNA percentages identity, except for AT (34% coverage) and HT (63% coverage). The PCR targets shown here were validated in Fig. 5.9.

5.4.3.2 PCR and sequencing

The two PCR targets defined in Fig. 5.11 were amplified yielding the products shown in Fig. 5.9. The magenta–red product was 50–100 bp smaller than predicted (834 bp), perhaps due to amplification of a homologous sequence such as one found in a follow-up GOS search.[22] Nevertheless, the sharp band in the gel supports that a single DNA region was amplified.

The red–cyan product, on the other hand, had a size closer to the predicted 1134 bp. Five of five clones for this product were ~1200 bp (restriction digest not shown). Sanger sequences from both ends of one clone aligned identically over 320 nt of high-quality base calls, and were assembled into the 1134 nt cloned sequence analyzed below.

5.4.3.3 Analysis of cloned product

The cloned sequence aligned fully to pooled contig 01453 at 97.8% id; this substantiates that the chimeric contig approximated real DNA. Inspection of all twenty-five clone-to-contig mismatch positions with an assembly viewer showed that variation was not specific to geography: At seventeen positions, assembled western English Channel reads differed from each other, even when from the same sample, and the clone usually matched one of two main variants. At seven positions the clone differed from all reads (WEC and SIO).

The two closest RefSeq proteins to the cloned sequence were from the recent single-cell amplified genomes SCGC AAA536-G10 and AAA015-N04. Both were hypothetical flavoproteins involved in K^+ transport (COG2072) based on their 352 C-terminal residues, which aligned to the cloned sequence ORF at 86% id and 78% id,

[22] The existence of the intended PCR target was corroborated by a Gulf of Maine GOS trace (the top hit) that aligned to the contig, between the primers, at 834 nt and 99% id. One Sargasso Sea trace was found that aligned at 726 nt, close to the length observed in the gel. However, it aligned in two HSPs (71% id and 83% id). Each HSP captured a primer and 101 nt between them were unaligned.

respectively.[23]

Protein and DNA phylogenetic trees for single-cell homologs to the cloned sequence ORF were consistent (Fig. 5.12). Remarkably, sequences separated by whether they were open-ocean (SCGC AAA015-N04 and Sargasso Stn. 13) or coastal (all others), but not by geography: For example, two of the North Atlantic sequences (WEC and Halifax) are more similar to the SIO clone and the Mediterranean SCGC AAA536-G10 than to the remaining North Atlantic sequence (Nags Head). The same pattern was observed in an intergenic DNA tree for forty-five bases after the ORF (not shown).[24]

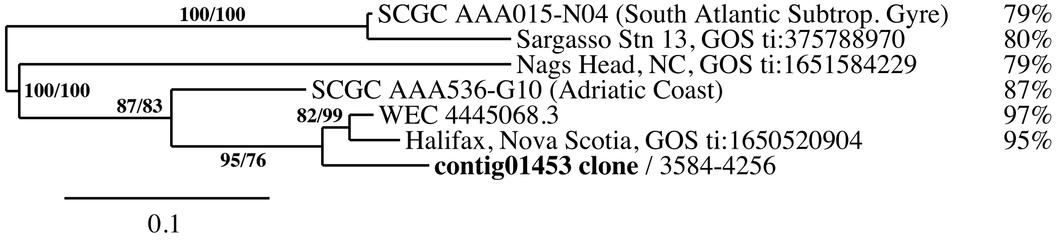

100/100
SCGC AAA015-N04 (South Atlantic Subtrop. Gyre)
79%
Sargasso Stn 13, GOS ti:375788970
80%
Nags Head, NC, GOS ti:1651584229
79%
100/100
87/83
SCGC AAA536-G10 (Adriatic Coast)
87%
82/99
WEC 4445068.3
97%
Halifax, Nova Scotia, GOS ti:1650520904
95%
95/76
contig01453 clone / 3584-4256
0.1

Figure 5.12: Phylogeny for the contig 01453 cloned sequence and homologous genes. Except for the WEC 4445086 contig which represents a population, each sequence derives from a single genome from the sample site. Protein (shown) and DNA trees had the same topology so bootstrap values for both are given (protein/DNA). The outgroup SCGC AAA015-N04 is a member of the SAR116 clade within the α-proteobacteria. Trees for ranges of the cloned sequence outside of 3584–4256 nt were consistent but lacked one or more GOS traces due to incomplete coverage of the range (not shown). On the right are the DNA percentages identity between the clone and each sequence, not restricted to the multiple alignment positions.

In summary, the cloned sequence substantiates that contig 01453 represents real DNA, likely from an uncharacterized coastal α-proteobacterium found in the Pacific

[23] COG2072 includes the flavin-containing monooxygenase domain (FMO; PFAM00743), as in Fig. 5.11.

[24] Bootstraps were all $\geq$94 and the topology was the same as in Fig. 5.12, except the Nags Head sequence was excluded because it terminated prior to the stop codon. The coastal sequences all shared a stop codon at 1 nt. The two open ocean sequences shared a stop codon at 34 nt and identical amino acids prior.

(La Jolla to the Monterey Bay) and both sides of the Atlantic (Nova Scotia and the western English Channel).

5.5 Discussion

5.5.1 Inferred Roseobacter group for cluster E

Rhodobacteraceae 16S rRNA was not found among the SIO reads (set 4443765.3) to help assign the cluster E contigs/clones to one of the major Roseobacter subgroups. However, *Rhodobacteraceae* 16S rRNA was identified in the western English Channel set 4445068.3. Because that set deeply covered cluster E contigs/clones (e.g. Fig. 5.3), some of the identified 16S reads in WEC 4445068.3 likely represented a close relative of the SIO population. Therefore we used them as proxies to search NCBI/nt to suggest a Roseobacter subgroup.[25] The search recovered hundreds of 16S environmental sequences, remarkably often from San Pedro and Monterey Bay time series (99–100% id). Taxonomic information for some of the top hits suggested three Roseobacter subgroups: DC-5-80 (also known as the Roseobacter clade-affiliated (RCA) cluster; 98 and 99% id); NAC11-7 (99% id); or CHAB-I-5 (99% id).

[25] The threshold that characterizes Roseobacter membership, 89% id [18], was used to identify reads in 4445068.3 that likely represented Roseobacter 16S rRNA. These reads were then strictly assembled (99% id, $\geq$100 nt overlap) into fifteen contigs between 576–1060 nt. Searching for the contigs within NCBI/nt, only eight of them had hits to 16S sequences from known lineages (in assembled genomes) or suspected lineages (based on targeted environmental sequences): Two contigs represented uncultured γ-proteobacteria (98–100% id), and one represented *Pelagibacter ubique* (99% id). The former was ruled out because the organism of interest is likely an α-proteobacteria (as Roseobacters are), and the latter because *Pelagibacter ubique* lacks cobD and has a dissimilar gene order around its SMC from that observed in the clones. The remaining five contigs aligned over their full lengths (634–1060 nt) to Roseobacter 16S sequences at $\geq$98% id and were used to infer the likely Roseobacter subgroup (main text).

Sixteen-S rRNA from all three Roseobacter subgroups have often been observed in association with phytoplankton blooms [18, 106]. Consistent with this, the surface water samples for the three data sets that most deeply covered the cloned sequences (Fig. 5.3) were collected shortly after phytoplankton blooms in the North Atlantic (4445068.3) and Monterey Bay (4443714.3, 4443716.3): The western English Channel sample (4445068.3) was collected in the summer of 2008 when chlorophyll A concentration, indicative of eukaryotic phytoplankton abundance, was the highest observed in a six-year time series (2003–2008; [56]). The Monterey Bay samples were enriched in NAC11-7 Roseobacters—not RCA nor CHAB-I-5—likely following a spring phytoplankton bloom in 2001 [146].[26,27] Altogether these observations suggest that the cluster E contigs/clones represent a bloom-associated Roseobacter from the NAC11-7 subgroup. Gene predictions for other contigs in cluster E did not support or refute this hypothesis. Additional sequencing would help verify the subgroup (§5.5.4).

5.5.2 Inferred Roseobacter group for cluster F contig 06386

Deep coverage of contig 06386 by western English Channel set 4444077.3 (Fig. 5.8) enabled us, with the approach just described, to infer that contig 06386 rep-

[26] CHAB-I-5 and RCA were ruled out as follows: In the Monterey Bay, CHAB-I-5 was abundant in 80 m samples, while NAC11-7 was abundant in 0 m samples [146]. This makes CHAB-I-5 unlikely to be the cluster E Roseobacter. Voget et al. (2015) [185] identified the bloom-associated RCA member *Planktomarina temperata* RCA23 in metagenomic data sets used in this dissertation, with samples taken from the Monterey Bay [146], the western English Channel [54], and a Norwegian fjord [53]. They reported RCA23 to be highly abundant in the Norwegian fjord samples, with 84% of the genome covered. In contrast, contigs in cluster E assembled few or no reads from the Norwegian fjord data sets (unlike clusters C/D). This makes RCA23 unlikely to be the cluster E Roseobacter.

[27] Interestingly, set 4443713.3, the only Monterey Bay set from a *fall 2001* phytoplankton bloom, had the lowest representation in Fig. 5.3 despite a similar sequencing effort to other MB sets.

resented either the RCA or NAC11-7 subgroup.[28] Two 16S proxy contigs represented RCA, and one represented NAC11-7. The depth of reads from 4444077.3 over the RCA 16S proxies was about one tenth that of the NAC11-7 proxy, whose read depth was similar to that in Fig. 5.8. It therefore seems more likely that contig 06386, like the cluster E Roseobacter, represented a member of the NAC11-7 subgroup.

That contig 06386 and cluster E represent different NAC11-7 members is suggested by their very different abundance levels in the April 22, 2008 western English Channel sample (Tab. 5.4; not correcting for contig lengths which are close). Interestingly, geographic profiling separated cluster E and contig 06386 despite *not* seeing this striking read depth difference (since WEC sets were grouped).

WEC data set	**Sample date**	**Sample time**	**E: bridge reads**	**F: 06386 reads**	**F: 01453 reads**
4445081.3	Jan. 28	3pm	0	1	3
4444077.3	Apr. 22	4pm	5	109	18
4445068.3	Aug. 27	4pm	57	34	37
4445065.3	Aug. 27	10pm	14	20	6
4445066.3	Aug. 28	4am	11	17	6

Table 5.4: Western English Channel data sets as represented in the contigs/clones. All sets are from samples taken in 2008 at the L4 site [54]. Gilbert et al. (2012) reported a large increase in chlorophyll A from spring to summer, 2008, as well as a decrease in Rhodobacterales [56].

[28] See footnote 25. Here, four 16S proxy contigs, assembled from 4444077.3 reads, elicited hits in NCBI/nt to 16S rRNA. One hit to a γ-proteobacterium was rejected. Two contigs aligned fully (1273 nt and 1526 nt) and at 99% id to 16S from *Planktomarina temperata* RCA23. The third contig, 2253 nt, end-aligned to 1911 nt of a NAC11-7 environmental clone at 99% id.

5.5.3 Phytoplankton bloom associations

Shifts in read depth by western English Channel data sets over the cloned contigs (see Figs. 5.3, 5.8, and 5.11) present another interesting pattern: All three species seemed abundant during the 2008 North Atlantic phytoplankton bloom (April and August samples in Tab. 5.4), but rare outside the bloom (January).[29] This pattern might extend to the SIO Pier sample which also contributed few reads to the three contigs, and also was collected when there was not a bloom.[30] However, because the SIO sample was cell-sorted, this is speculative.

[29] The counts in Tab. 5.4 are uncorrected because contig sizes were similar, as were the sequencing efforts represented by the data sets (617K, 638K, 621K, 526K, 500K in the table order), compared to the large differences in bloom versus non-bloom counts.

[30] Non-bloom: October 10, 2006, when the sample was collected, appears to have been unremarkable with respect to nutrient concentrations, based on Fig. 1b of Tai and Palenik (2012) [169], and inspection of chlorophyll concentrations for that day versus all of 2006, downloaded from the Southern California Coastal Ocean Observing System web site: `http://www.sccoos.org`

5.5.4 Follow-up research into cluster E and F Roseobacters

Members from the RCA cluster alone can constitute 10% of bacteria in temperate to polar coastal waters [153], yet to date there is just one RCA RefSeq genome, *Planktomarina temperata* RCA23 [185]. The other abundant Roseobacter subgroup, NAC11-7, has just one sequenced representative, Rhodobacterales bacterium HTCC2255 (draft), and most of the other major subgroups have none [123, 18]. Given this major gap in our knowledge of the Roseobacter lineage, follow-up on the putative NAC11-7 species in clusters E and F could be especially valuable. Next steps are outlined in Tab. 5.5.

Track	**Description**	**Informs**
16S rRNA	PCR, sequencing with *Rhodobacteraceae* 16S primers.	Roseobacter group placement; distribution by comparison to the rich set of environmental 16S.
Tara Oceans data	Search for clone/cluster E sequences. Bridge cluster E contigs with Tara Oceans contigs.	Distribution, population differences.
Marker genes	Based on 16S, perform PCR, sequencing of genes typical for clade (Fig. 1 in Newton et al., 2010)	More robust phylogeny; comparison to metabolisms of RefSeq Roseobacters.

Table 5.5: Possible follow-up research into the cluster E Roseobacter, from least to most effort. PCR and sequencing would first be done with the 2006 SIO sample (Palenik Lab), but additional samples/collaborations would be valuable.

5.6 Supplement: Geographic profiling of simulated data sets

This chapter began with several examples of mostly species-homogenous clusters produced by geographic profiling (A, C, D in Fig. 5.1). However we also saw that cluster F contained at least two α-proteobacteria. To further check whether mixed or homogenous clusters might be expected in practice, we applied geographic profiling to the simulation pooled contigs (chapter 3), for which the true species were known (Fig. 5.13). Overall, species represented by many contigs clustered more tightly than species with fewer (e.g. γ-proteobacterium HTCC2207 had 68 contigs; left). The exception to this, *Prochlorococcus marinus*, is likely due to thirteen strains of *P.marinus* in the simulation data being lumped together by the NCBI taxonomy into a single species, despite the recognition that *P.marinus* comprises multiple distinct clades and hundreds of subpopulations [80]. Pairs of contigs from the same species tended to be closer than pairs selected at random (right), consistent with the shortness of GOS links in Fig. 5.1.

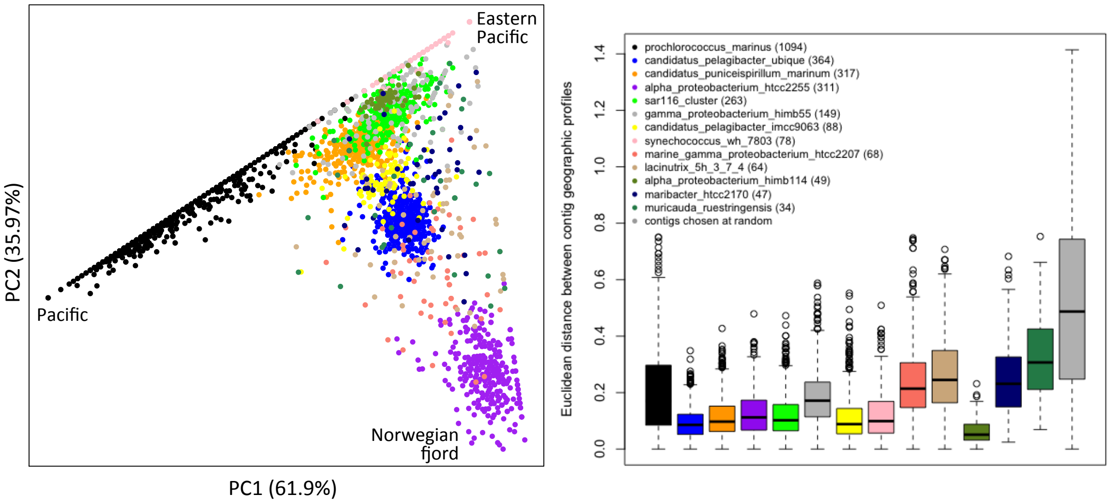


Figure 5.13: Pooled contigs >2Kbp, assembled from the ten simulated data sets, tend to cluster by species when geographically profiled. **Left** Species with at least thirty contigs are shown. Corners of the plot are labeled with the geographic region that most heavily loaded the components of the PCA axes. **Right** Distances between contigs from the same species (non-gray) are smaller than distances without respect to species ($P \approx 0$ in Wilcoxon rank sum test). Between 130 and 1000 pairs of distinct contigs were chosen at random for intra-species distance distributions, and 10,000 pairs at random for the null distribution.

Chapter 6

Future of approaches developed in this work

The past decade has witnessed extraordinary growth in the number and size of metagenomic data sets, but also uncertain funding for annotation servers/repositories that the community depends on (§1.5.3). These trends and challenges will spur new approaches that can exploit the data for its scale while managing costs. For example, MG-RAST's introduction of the M5nr/rna databases in effect parallelized searches of RefSeq, SEED, KEGG, etc., and thereby increased annotation but not computation (§1.5.1; [192]). Other optimizations to the v3.0 pipeline circumvented annotation costs estimated at \$300K per data set [194].[1] With respect to exploiting metagenomic big data, the discovery potential seems especially promising for marine microbial oceanography and ecology: "Old" data from multiple projects can help the detection of long-term

[1] Optimizations included switching to annotate clusters (not reads) with BLAT (not BLAST). The estimate presumed a 100Gbp data set annotated on Amazon's EC2 cloud. Today, HiSeq and MiSeq platforms exceed 100Gbp (Tab. 1.2).

and large-scale patterns and changes, similar to single-project time series and worldwide sampling efforts (e.g. [86]).

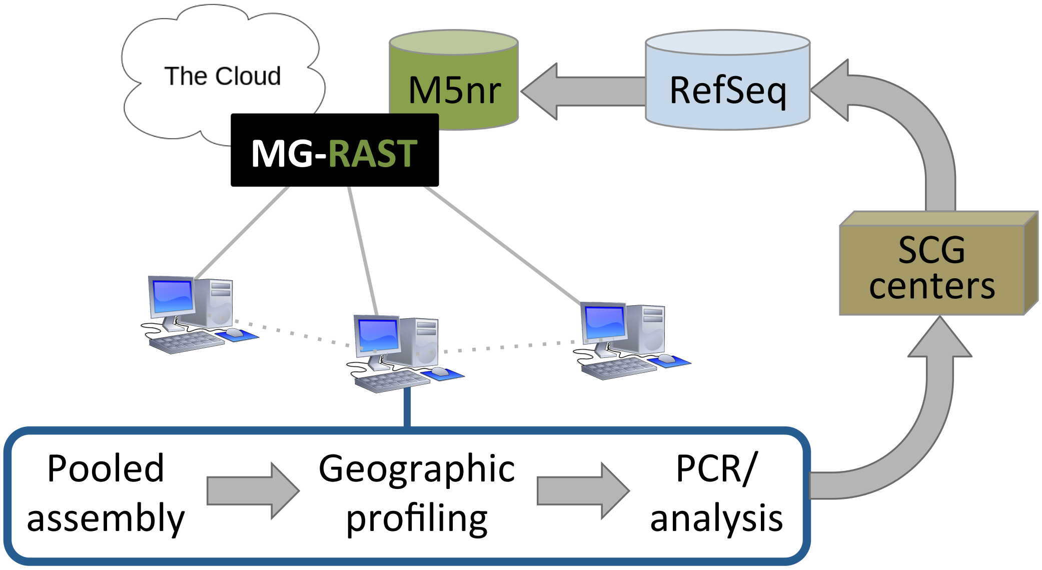


Figure 6.1: Future prospects for approaches in context. Individual labs (computers) use pooled assembly to enrich annotation for their data sets. Labs then use geographic profiling to discover ubiquitous species among the unannotated contigs, followed by inexpensive PCR validation in the original samples. Labs would pursue species of interest by single-cell genome sequencing at the JGI or Bigelow Laboratory SCG centers. New genomes would improve RefSeq and other reference databases (cylinders) used by annotation servers such as MG-RAST (which is partially cloud-based). Planned features for MG-RAST (main text) could enable data sharing among client labs generally (dotted lines), and in particular for pooled assembly of unpublished metagenomic data sets. k-mer profiles would help discovery and ranking of lab-hosted data sets.

The approaches developed in this dissertation complement future plans for MG-RAST as illustrated in Fig. 6.1. Version 4.0 of the pipeline will include multi-metagenome recruitment plots which will map reads from different data sets to reference genomes or genes. This new feature could support geographic profiling if user-supplied sequences (e.g. contigs from a new data set) could serve as recruitment targets. In fact, a second planned feature could enable just that: MG-RAST v4.0 will also include a remote compute client so that users can compare data sets at the lab site (i.e. not uploaded) with data sets hosted at MG-RAST. With this client one could map reads from hosted data sets to a local query sequence, and then pool and assemble the data sets that had reads mapped. The contigs produced would then be geographically profiled to identify those that might represent the same species as the query. This is similar to the chapter 5 example with SCGC AAA076-P13 in which a single contig drove the search for same-species contigs (5.3.2.1), but in this new, networked scenario the query contig must first be used to identify data sets.

Alternatively, the k-mer profile of a data set could be used to identify other sets. These sets could be hosted at MG-RAST, or at lab sites with MG-RAST clients as shown in Fig. 6.1. In fact, we originally envisioned that k-mer profiles would be a low-bandwidth way to discover similar data sets that were distributed across a network of microbiology labs. In this scenario lab data sets would likely be unpublished, so one would not expect metadata or annotation to exist in standardized forms. k-mer profiles would thus be an extremely low-effort way for labs to share these data sets.

The benefits of pooled assembly for species annotation should increase as RefSeq grows (Fig. 3.3), encouraged by single-cell genomics projects to generate reference genomes (e.g. §5.3.2.1). We also suspect that recent Illumina data sets, with read lengths on a par with those in our 454 study, would enjoy similar annotation benefits. However, a new strategy for the pooled assembly itself would be required for massive Illumina sets, and to take advantage of their paired-end reads. Binning reads by approximate phylogeny within data sets (often a step in the assembly of Illumina data), and then across data sets by their *k*-mer profiles, might help.

Pooled assembly also could be used to study variation within and across populations, despite the lack of close reference genomes. Fig. 3.5 and the cloned product analyses in §5.4 are illustrative; higher read depths of post-454 technologies would support more quantitative methods. In contrast, population variation across human microbiomes has been studied with respect to *reference* microbial genomes (e.g. [150]).

In 1770 Captain Cook observed "a kind of brown scum" over the surface of the Coral Sea [28], likely filamentous mats from a *Trichodesmium* bloom. Two hundred and eighteen years later, Chisholm et al. (1988) used shipboard flow cytometry to discover *Prochlorococcus*, possibly the most abundant cellular organism on Earth [24]. Today, high-throughput sequencing is an exciting new lens for marine microbiology, made more powerful by tools that can integrate and compare many data sets. We offer the approaches developed in this work as a catalyst to others that will use raw sequencing data to explore the world within a drop of seawater.

Bibliography

[1] M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, 31(6):533–538, 2013.

[2] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nat. Methods*, 11(11):1144–1146, 2014.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, 1990.

[4] R. I. Amann, B. J. Binder, R. J. Olson, S. W. Chisholm, R. Devereux, and D. A. Stahl. Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl. Environ. Microb.*, 56(6):1919–25, Jun 1990.

[5] S. Aparicio, J. Chapman, E. Stupka, N. Putnam, J. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, 297(5585):1301–1310, 2002.

[6] F. O. Aylward, J. M. Eppley, J. M. Smith, F. P. Chavez, C. A. Scholin, and E. F. DeLong. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *P. Natl. Acad. Sci. USA*, 112(17):5443–5448, 2015.

[7] F. Azam and F. Malfatti. Microbial structuring of marine ecosystems. *Nat. Rev. Microbiol.*, 5(10):782–791, 2007.

[8] S. Batzoglou, D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov, and E. S. Lander. ARACHNE: a whole-genome shotgun assembler. *Genome Res.*, 12(1):177–189, 2002.

[9] O. Béja, L. Aravind, E. V. Koonin, M. T. Suzuki, A. Hadd, L. P. Nguyen, S. B. Jovanovich, C. M. Gates, R. A. Feldman, J. L. Spudich, et al. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, 289(5486):1902–1906, 2000.

[10] O. Béjà, M. T. Suzuki, J. F. Heidelberg, W. C. Nelson, C. M. Preston, T. Hamada, J. A. Eisen, C. M. Fraser, and E. F. DeLong. Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature*, 415(6872):630–633, 2002.

[11] S. J. Biller, P. M. Berube, D. Lindell, and S. W. Chisholm. Prochlorococcus: the structure and function of collective diversity. *Nat. Rev. Microbiol.*, 13(1):13–27, 2015.

[12] S. J. Biller, F. Schubotz, S. E. Roggensack, A. W. Thompson, R. E. Summons, and S. W. Chisholm. Bacterial vesicles in marine ecosystems. *Science*, 343(6167):183–186, 2014.

[13] J. Bischof, T. Harrison, T. Paczian, E. Glass, A. Wilke, and F. Meyer. Metazen–metadata capture for metagenomes. *Standards in Genomic Sciences*, 9(1):18, 2014.

[14] P. W. Boyd, A. J. Watson, C. S. Law, E. R. Abraham, T. Trull, R. Murdoch, D. C. Bakker, A. R. Bowie, K. Buesseler, H. Chang, et al. A mesoscale phytoplankton bloom in the polar southern ocean stimulated by iron fertilization. *Nature*, 407(6805):695–702, 2000.

[15] L. M. Bragg, G. Stone, M. K. Butler, P. Hugenholtz, and G. W. Tyson. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput Biol*, 9(4):e1003031, 2013.

[16] J. R. Bray and J. T. Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4):325–349, 1957.

[17] L. Brocchieri and S. Karlin. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.*, 33(10):3390–3400, 2005.

[18] A. Buchan, J. M. González, and M. A. Moran. Overview of the marine Roseobacter lineage. *Appl. Environ. Microb.*, 71(10):5665–5677, 2005.

[19] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, S. E. Lewis, E. C., et al. The environment ontology: contextualising biological and biomedical entities. *J. Biomedical Semantics*, 4:43, 2013.

[20] `http://phylogenomics.blogspot.com/2014/05/camera-metagenomics-resource-is.html`. Letter from CAMERA to its users, posted on Jonathan Eisen's blog Tree of Life on May 28, 2014.

[21] `http://www.coml.org/projects/international-census-marine-microbes-icomm`. Census of marine Life web site accessed on October 13, 2015.

[22] A. Charuvaka and H. Rangwala. Evaluation of short read metagenomic assembly. *BMC Genomics*, 12(Suppl 2):S8, 2011.

[23] F. Chevenet, C. Brun, A. Bañuls, B. Jacq, and R. Christen. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics*, 7(1):439, 2006.

[24] S. W. Chisholm, R. J. Olson, E. R. Zettler, R. Goericke, J. B. Waterbury, and N. A. Welschmeyer. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. 1988.

[25] L. Chistoserdova. Recent progress and new challenges in metagenomics for biotechnology. *Biotechnol. Lett.*, 32(10):1351–1359, 2010.

[26] M. J. Church, C. Mahaffey, R. M. Letelier, R. Lukas, J. P. Zehr, and D. M. Karl. Physical forcing of nitrogen fixation and diazotroph community structure in the North Pacific Subtropical Gyre. *Global Biogeochem. Cy.*, 23(2), 2009.

[27] S. Clingenpeel, A. Clum, P. Schwientek, C. Rinke, and T. Woyke. Reconstructing each cell's genome within complex microbial communitiesdream or reality? *Frontiers in Microbiology*, 5, 2014.

[28] J. Cook. *Captain Cook's Journal During His First Voyage Round the World Made in HM Bark Endeavour 1768-71*, volume cihm_14800. E. Stock, 1893. Filmed from a copy of the original publication held by the Library Division, Provincial Archives of British Columbia.

[29] `https://www.flickr.com/photos/nysdec/12370374544`. Rosette sampler image is from the Cooperative Science and Monitoring Initiative Lake Ontario, and shared under a Creative Commons License (`https://creativecommons.org/licenses/by-nc-nd/2.0/`). Image accessed on web October 19, 2015.

[30] P. D. Countway and D. A. Caron. Abundance and distribution of Ostreococcus sp. in the San Pedro Channel, California, as revealed by quantitative PCR. *Appl. Environ. Microb.*, 72(4):2496–2506, 2006.

[31] P. A. D. Giorgio and C. M. Duarte. Respiration in the open ocean. *Nature*, 420(6914):379–384, 2002.

[32] E. F. DeLong. The microbial ocean from genomes to biomes. *Nature*, 459(7244):200–206, 2009.

[33] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J. Dufayard, S. Guindon, V. Lefort, M. Lescot, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, 36(suppl 2):W465–W469, 2008.

[34] N. Desai, D. Antonopoulos, J. A. Gilbert, E. M. Glass, and F. Meyer. From genomics to metagenomics. *Curr. Opin. Biotech*, 23(1):72–76, 2012.

[35] C. Desnues, B. Rodriguez-Brito, S. Rayhawk, S. Kelley, T. Tran, M. Haynes, H. Liu, M. Furlan, L. Wegley, B. Chau, et al. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature*, 452(7185):340–343, 2008.

[36] M. Diepenbroek, H. Grobe, M. Reinke, U. Schindler, R. Schlitzer, R. Sieger, and G. Wefer. PANGAEA — an information system for environmental sciences. *Computers & Geosciences*, 28(10):1201–1210, 2002.

[37] E. A. Dinsdale, R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, et al. Functional metagenomic profiling of nine biomes. *Nature*, 452(7187):629–632, 2008.

[38] S. R. Eddy et al. A new generation of homology search tools based on probabilistic inference. In *Genome Inform.*, volume 23, pages 205–211. World Scientific, 2009.

[39] R. C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, 2004.

[40] R. C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[41] P. G. Falkowski, T. Fenchel, and E. F. Delong. The microbial engines that drive Earth's biogeochemical cycles. *Science*, 320(5879):1034–1039, 2008.

[42] L. Fan, K. McElroy, and T. Thomas. Reconstruction of ribosomal RNA genes from metagenomic data. *PLoS One*, 7(6):e39948, 2012.

[43] D. Field, G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, et al. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol*, 26(5):541–7, May 2008.

[44] Y. Fofanov, Y. Luo, C. Katili, J. Wang, Y. Belosludtsev, T. Powdrill, C. Belapurkar, V. Fofanov, T. Li, S. Chumakov, et al. How independent are the appearances of n-mers in different genomes? *Bioinformatics*, 20(15):2421–2428, 2004.

[45] M. J. Follows, S. Dutkiewicz, S. Grant, and S. W. Chisholm. Emergent biogeography of microbial communities in a model ocean. *Science*, 315(5820):1843–1846, 2007.

[46] J. Frias-Lopez, Y. Shi, G. W. Tyson, M. L. Coleman, S. C. Schuster, S. W. Chisholm, and E. F. DeLong. Microbial community gene expression in ocean surface waters. *P. Natl. Acad. Sci. USA*, 105(10):3805, 2008.

[47] J. A. Fuhrman, J. A. Cram, and D. M. Needham. Marine microbial community dynamics and their ecological interpretation. *Nat. Rev. Microbiol.*, 2015.

[48] J. N. Galloway, F. J. Dentener, D. G. Capone, E. W. Boyer, R. W. Howarth, S. P. Seitzinger, G. P. Asner, C. Cleveland, P. Green, E. Holland, et al. Nitrogen cycles: past, present, and future. *Biogeochemistry*, 70(2):153–226, 2004.

[49] L. Garczarek, F. Partensky, H. Irlbacher, J. Holtzendorff, M. Babin, I. Mary, J. C. Thomas, and W. R. Hess. Differential expression of antenna and core genes in Prochlorococcus PCC 9511 (Oxyphotobacteria) grown under a modulated light–dark cycle. *Environ. Microbiol.*, 3(3):168–175, 2001.

[50] D. R. Garza and B. E. Dutilh. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. *Cell. Mol. Life Sci.*, 72(22):4287–4308, 2015.

[51] T. A. Gianoulis, J. Raes, P. V. Patel, R. Bjornson, J. O. Korbel, I. Letunic, T. Yamada, A. Paccanaro, L. J. Jensen, M. Snyder, et al. Quantifying environmental adaptation of metabolic pathways in metagenomics. *P. Natl. Acad. Sci. USA*, 106(5):1374–1379, 2009.

[52] J. A. Gilbert and C. L. Dupont. Microbial metagenomics: Beyond the genome. *Annu. Rev. Mar. Sci.*, 3(1):347–371, 2011.

[53] J. A. Gilbert, D. Field, Y. Huang, R. Edwards, W. Li, P. Gilna, and I. Joint. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One*, 3(8):e3042, 2008.

[54] J. A. Gilbert, D. Field, P. Swift, L. Newbold, A. Oliver, T. Smyth, P. J. Somerfield, S. Huse, and I. Joint. The seasonal structure of microbial communities in the Western English Channel. *Environ. Microbiol.*, 11(12):3132–3139, 2009.

[55] J. A. Gilbert, D. Field, P. Swift, S. Thomas, D. Cummings, B. Temperton, K. Weynberg, S. Huse, M. Hughes, I. Joint, et al. The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS One*, 5(11):e15545, 2010.

[56] J. A. Gilbert, J. A. Steele, J. G. Caporaso, L. Steinbrück, J. Reeder, B. Temperton, S. Huse, A. C. McHardy, R. Knight, I. Joint, et al. Defining seasonal marine microbial community dynamics. *ISME J.*, 6(2):298–308, 2012.

[57] V. Gomez-Alvarez, T. K. Teal, and T. M. Schmidt. Systematic artifacts in metagenomes from complex microbial communities. *ISME J.*, 2009.

[58] P. Green. Documentation for phrap. 1996.

[59] S. Guindon, J. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321, 2010.

[60] M. H. Fritz, R. Leinonen, G. Cochrane, and E. Birney. Efficient storage of high throughput dna sequencing data using reference-based compression. *Genome Res.*, 21(5):734, 2011.

[61] J. D. Hackett, T. E. Scheetz, H. S. Yoon, M. B. Soares, M. F. Bonaldo, T. L. Casavant, and D. Bhattacharya. Insights into a dinoflagellate genome through expressed sequence tag analysis. *BMC Genomics*, 6(1):80, 2005.

[62] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5(10):R245–R249, 1998.

[63] `http://hahana.soest.hawaii.edu/hot/`. HOT web site accessed on October 7, 2015.

[64] R. W. Hendrix. Bacteriophage genomics. *Curr. Opin. Microbiol.*, 6(5):506–511, 2003.

[65] K. Holmén. The global carbon cycle. *Global Biogeochem. Cy.*, 50:239–262, 1992.

[66] W. Huang and G. Marth. EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Res.*, 18(9):1538–1543, 2008.

[67] S. Hunter, M. Corbett, H. Denise, M. Fraser, A. Gonzalez-Beltran, C. Hunter, P. Jones, R. Leinonen, C. McAnulla, E. Maguire, et al. EBI metagenomics — a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.*, 42(D1):D600–D606, 2014.

[68] B. Hurwitz. imicrobe: Advancing clinical and environmental microbial research using the iPlant cyberinfrastructure. In *Plant and Animal Genome XXII Conference*. Plant and Animal Genome, 2014.

[69] S. M. Huse, J. A. Huber, H. G. Morrison, M. L. Sogin, D. M. Welch, et al. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, 8(7):R143, 2007.

[70] M. Imelfort, D. Parks, B. J. Woodcroft, P. Dennis, P. Hugenholtz, and G. W. Tyson. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, 2:e603, 2014.

[71] `https://www.idtdna.com/calc/analyzer`. Integrated DNA Technologies Oligo-Analyzer 3.1.

[72] V. Iverson, R. M. Morris, C. D. Frazar, C. T. Berthiaume, R. L. Morales, and E. V. Armbrust. Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. *Science*, 335(6068):587–590, 2012.

[73] B. Jiang, K. Song, J. Ren, M. Deng, F. Sun, and X. Zhang. Comparison of metagenomic samples using sequence signatures. *BMC Genomics*, 13(1):730, 2012.

[74] A. W. Johnston, Y. Li, and L. Ogilvie. Metagenomic marine nitrogen fixation–feast or famine? *Trends Microbiol.*, 13(9):416–420, 2005.

[75] D. D. Kang, J. Froula, R. Egan, and Z. Wang. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165, 2015.

[76] D. M. Karl. Microbial oceanography: paradigms, processes and promise. *Nat. Rev. Microbiol.*, 5(10):759–769, 2007.

[77] D. M. Karl and R. Lukas. The Hawaii Ocean Time-series (HOT) program: background, rationale and field implementation. *Deep Sea Research Part II: Topical Studies in Oceanography*, 43(2-3):129–156, 1996.

[78] S. Karlin and C. Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, 11(7):283–290, 1995.

[79] S. Karlin, J. Mrazek, and A. M. Campbell. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.*, 179(12):3899–3913, 1997.

[80] N. Kashtan, S. E. Roggensack, S. Rodrigue, J. W. Thompson, S. J. Biller, A. Coe, H. Ding, P. Marttinen, R. R. Malmstrom, R. Stocker, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science*, 344(6182):416–420, 2014.

[81] J. Kennedy, J. R. Marchesi, and A. D. Dobson. Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. *Microbial Cell Factories*, 7(1):27, 2008.

[82] G. C. Kettler, A. C. Martiny, K. Huang, J. Zucker, M. L. Coleman, S. Rodrigue, F. Chen, A. Lapidus, S. Ferriera, J. Johnson, et al. Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *PLoS Genet.*, 3(12):e231, 2007.

[83] Z. Kolber. Energy cycle in the ocean: powering the microbial world. *Oceanography*, 20:82–91, 2007.

[84] M. Könneke, A. E. Bernhard, R. José, C. B. Walker, J. B. Waterbury, and D. A. Stahl. Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*, 437(7058):543–546, 2005.

[85] K. T. Konstantinidis, A. Ramette, and J. M. Tiedje. The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 361(1475):1929–40, Nov 2006.

[86] A. Kopf, M. Bicak, R. Kottmann, J. Schnetzer, I. Kostadinov, K. Lehmann, A. Fernandez-Guerra, C. Jeanthon, E. Rahav, M. Ullrich, et al. The Ocean Sampling Day Consortium. *GigaScience*, 4(1):1–5, 2015.

[87] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz. A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.*, 72(4):557, 2008.

[88] V. Kunin, A. Engelbrektson, H. Ochman, and P. Hugenholtz. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.*, 12(1):118–123, 2010.

[89] J. Laserson, V. Jojic, and D. Koller. Genovo: de novo assembly for metagenomes. *J. Comput. Biol.*, 18(3):429–443, 2011.

[90] R. Leinonen, H. Sugawara, and M. Shumway. The Sequence Read Archive. *Nucleic Acids Res.*, page gkq1019, 2010.

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

[92] W. Li. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics*, 10(1):359, 2009.

[93] W. Li, L. Fu, B. Niu, S. Wu, and J. Wooley. Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief. Bioinform.*, page bbs035, 2012.

[94] Y. W. Lim, D. A. Cuevas, G. G. Z. Silva, K. Aguinaldo, E. A. Dinsdale, A. F. Haas, M. Hatay, S. E. Sanchez, L. Wegley-Kelly, B. E. Dutilh, et al. Sequencing at sea: challenges and experiences in Ion Torrent PGM sequencing during the 2013 Southern Line Islands Research Expedition. *PeerJ*, 2:e520, 2014.

[95] E. Litchman, C. Klausmeier, J. Miller, O. Schofield, and P. Falkowski. Multinutrient, multi-group model of present and future oceanic phytoplankton communities. *Biogeosciences Discussions*, 3(3):607–663, 2006.

[96] K. G. Lloyd, L. Schreiber, D. G. Petersen, K. U. Kjeldsen, M. A. Lever, A. D. Steen, R. Stepanauskas, M. Richter, S. Kleindienst, S. Lenk, et al. Predominant archaea in marine sediments degrade detrital proteins. *Nature*, 496(7444):215–218, 2013.

[97] T. Lombardot, R. Kottmann, H. Pfeffer, M. Richter, H. Teeling, C. Quast, and F. O. Glöckner. Megx. net — database resources for marine ecological genomics. *Nucleic Acids Res.*, 34(suppl 1):D390, 2006.

[98] A. V. Lukashin and M. Borodovsky. Genemark. hmm: new solutions for gene finding. *Nucleic Acids Res.*, 26(4):1107–1115, 1998.

[99] H. Luo, B. K. Swan, R. Stepanauskas, A. L. Hughes, and M. A. Moran. Evolutionary analysis of a streamlined lineage of surface ocean Roseobacters. *ISME J.*, 8(7):1428–1439, 2014.

[100] J. D. Magasin and D. L. Gerloff. Pooled assembly of marine metagenomic datasets: enriching annotation through chimerism. *Bioinformatics*, 31(3):311–317, 2015.

[101] N. Maillet, G. Collet, T. Vannier, D. Lavenier, and P. Peterlongo. Commet: comparing and combining multiple metagenomic datasets. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 94–98. IEEE, 2014.

[102] E. R. Mardis. Next-generation DNA sequencing methods. *Annu. Rev. Genom. Hum. G.*, 9:387–402, 2008.

[103] E. R. Mardis. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203, 2011.

[104] V. M. Markowitz, N. N. Ivanova, E. Szeto, K. Palaniappan, K. Chu, D. Dalevi, I. Chen, A. Min, Y. Grechkin, I. Dubchak, et al. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.*, 36(suppl 1):D534, 2008.

[105] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods*, 4(6):495–500, 2007.

[106] X. Mayali, P. J. Franks, and F. Azam. Cultivation and ecosystem role of a marine Roseobacter Clade-Affiliated cluster bacterium. *Appl. Environ. Microb.*, 74(9):2595–2603, 2008.

[107] K. E. McElroy, F. Luciani, and T. Thomas. GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*, 13(1):74, 2012.

[108] A. C. McHardy and I. Rigoutsos. What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr. Opin. Microbiol.*, 10(5):499–503, 2007.

[109] D. Medini, C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli. The microbial pan-genome. *Curr. Opin. Genet. Dev*, 15(6):589–594, 2005.

[110] D. R. Mende, A. S. Waller, S. Sunagawa, A. I. Järvelin, M. M. Chan, M. Arumugam, J. Raes, and P. Bork. Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS One*, 7(2):e31386, 2012.

[111] M. L. Metzker. Sequencing technologies — the next generation. *Nat. Rev. Genet.*, 11(1):31–46, 2009.

[112] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386, 2008.

[113] H. Mi, A. Muruganujan, and P. D. Thomas. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, 41(D1):D377–D386, 2013.

[114] J. R. Miller, A. L. Delcher, S. Koren, E. Venter, B. P. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry, and G. Sutton. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, 24(24):2818–2824, 2008.

[115] J. R. Miller, S. Koren, and G. Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

[116] P. H. Moisander, R. A. Beinart, I. Hewson, A. E. White, K. S. Johnson, C. A. Carlson, J. P. Montoya, and J. P. Zehr. Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. *Science*, 327(5972):1512–1514, Mar 2010.

[117] R. M. Morris, M. S. Rappé, S. A. Connon, K. L. Vergin, W. A. Siebold, C. A. Carlson, S. J. Giovannoni, et al. SAR 11 clade dominates ocean surface bacterioplankton communities. *Nature*, 420(6917):806–810, 2002.

[118] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, et al. A whole-genome assembly of Drosophila. *Science*, 287(5461):2196–2204, 2000.

[119] N. Nagarajan and M. Pop. Sequence assembly demystified. *Nat. Rev. Genet.*, 14(3):157–167, 2013.

[120] T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.*, 40(20):e155–e155, 2012.

[121] `http://earthobservatory.nasa.gov/IOTD/view.php?id=85921`. NASA Earth Observatory public domain image of 2015 North Atlantic spring bloom accessed on web October 14, 2015.

[122] NSF data mangement plan requirements. `http://www.nsf.gov/eng/general/-dmp.jsp`. NSF Directorate for Engineering web site accessed on October 7, 2015.

[123] R. J. Newton, L. E. Griffin, K. M. Bowles, C. Meile, S. Gifford, C. E. Givens, E. C. Howard, E. King, C. A. Oakley, C. R. Reisch, et al. Genome characteristics of a generalist marine bacterial lineage. *ISME J.*, 4(6):784–798, 2010.

[124] H. Noguchi, J. Park, and T. Takagi. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.*, 34(19):5623–5630, 2006.

[125] National Research Council (US). Committee on Metagenomics, Functional Applications, and National Academies Press (US). *The New Science of Metagenomics: Revealing the Secrets of our Microbial Planet.* Natl Academy Pr, 2007.

[126] E. A. Ottesen, C. R. Young, J. M. Eppley, J. P. Ryan, F. P. Chavez, C. A. Scholin, and E. F. DeLong. Pattern and synchrony of gene expression among sympatric marine microbial populations. *P. Natl. Acad. Sci. USA*, 110(6):E488–E497, 2013.

[127] E. A. Ottesen, C. R. Young, S. M. Gifford, J. M. Eppley, R. Marin, S. C. Schuster, C. A. Scholin, and E. F. DeLong. Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science*, 345(6193):207–212, 2014.

[128] R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, 33(17):5691, 2005.

[129] N. R. Pace, D. A. Stahl, D. J. Lane, and G. J. Olsen. The analysis of natural microbial populations by ribosomal RNA sequences. In *Advances in Microbial Ecology*, pages 1–55. Springer, 1986.

[130] I. Pagani, K. Liolios, J. Jansson, I. M. A. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides. The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, 40(D1):D571–D579, 2012.

[131] B. Palenik, Q. Ren, V. Tai, and I. Paulsen. Coastal Synechococcus metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity. *Environ. Microbiol.*, 11(2):349–359, 2009.

[132] C. Pedrós-Alió. Marine microbial diversity: can it be determined? *Trends Microbiol.*, 14(6):257–263, 2006.

[133] C. Pedrós-Alió. The rare bacterial biosphere. *Annu. Rev. Mar. Sci.*, 4:449–466, 2012.

[134] Y. Peng, H. C. Leung, S. Yiu, and F. Y. Chin. Meta-IDBA: a de novo assembler for metagenomic data. *Bioinformatics*, 27(13):i94–i101, 2011.

[135] M. Pignatelli and A. Moya. Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PLoS One*, 6(5):e19984, 2011.

[136] L. R. Pomeroy, P. J. Williams, F. Azam, and E. A. Hobbie. The microbial loop. *Oceanography*, 20:28–33, 2007.

[137] C. M. Preston, R. Marin, S. D. Jensen, J. Feldman, J. M. Birch, E. I. Massion, E. F. DeLong, M. Suzuki, K. Wheeler, and C. A. Scholin. Near real-time, autonomous detection of marine bacterioplankton on a coastal mooring in Monterey Bay, California, using rRNA-targeted DNA probes. *Environ. Microbiol.*, 11(5):1168–1180, 2009.

[138] ProMega BioMath Calculators.

[139] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott. Ncbi reference sequences (refseq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, 40(D1):D130–D135, 2012.

[140] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, et al. The Pfam protein families database. *Nucleic Acids Res.*, 40(D1):D290–D301, 2012.

[141] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13(1):341, 2012.

[142] N. L. Quinn, N. Levenkova, W. Chow, P. Bouffard, K. A. Boroevich, J. R. Knight, T. P. Jarvie, K. P. Lubieniecki, B. A. Desany, B. F. Koop, et al. Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics*, 9(1):404, 2008.

[143] J. Raes, K. U. Foerstner, and P. Bork. Get the most out of your metagenome: Computational analysis of environmental sequence data. *Curr. Opin. Microbiol.*, 10(5):490–498, 2007.

[144] M. Rho, H. Tang, and Y. Ye. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.*, 38(20):e191–e191, 2010.

[145] P. Rice, I. Longden, A. Bleasby, et al. EMBOSS: the European molecular biology open software suite. *Trends Genet.*, 16(6):276–277, 2000.

[146] V. I. Rich, V. D. Pham, J. Eppley, Y. Shi, and E. F. DeLong. Time-series analyses of Monterey Bay coastal microbial picoplankton using a 'genome proxy' microarray. *Environ. Microbiol.*, 13(1):116–134, 2011.

[147] D. C. Richter, F. Ott, A. F. Auch, R. Schmid, and D. H. Huson. MetaSim — a sequencing simulator for genomics and metagenomics. *PLoS One*, 3(10):e3373, 2008.

[148] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, et al. The Sorcerer II Global Ocean Sampling Expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, 5(3):e77, 2007.

[149] S. Schbath, B. Prum, and E. de Turckheim. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comput. Biol.*, 2(3):417–437, 1995.

[150] S. Schloissnig, M. Arumugam, S. Sunagawa, M. Mitreva, J. Tap, A. Zhu, A. Waller, D. R. Mende, J. R. Kultima, J. Martin, et al. Genomic variation landscape of the human gut microbiome. *Nature*, 493(7430):45–50, 2013.

[151] P. D. Schloss and J. Handelsman. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.*, 6(8):229, 2005.

[152] M. B. Scholz, C. Lo, and P. S. Chain. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotech*, 23(1):9–15, 2012.

[153] N. Selje, M. Simon, and T. Brinkhoff. A newly discovered Roseobacter cluster in temperate and polar oceans. *Nature*, 427(6973):445–448, 2004.

[154] M. Settings. Metagenomics versus Moore's law. *Nat. Methods*, 6(9):623–623, 2009.

[155] E. B. Sherr and B. F. Sherr. Significance of predation by protists in aquatic microbial food webs. *Antonie van Leeuwenhoek*, 81(1-4):293–308, 2002.

[156] I. N. Shilova, J. C. Robidart, H. J. Tripp, K. Turk-Kubo, B. Wawrik, A. F. Post, A. W. Thompson, B. Ward, J. T. Hollibaugh, A. Millard, et al. A microarray for assessing transcription from pelagic marine microbial taxa. *ISME J.*, 8(7):1476–1491, 2014.

[157] M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. Microbial diversity in the deep sea and the underexplored rare biosphere. *P. Natl. Acad. Sci. USA*, 103(32):12115–12120, 2006.

[158] R. Sorek and P. Cossart. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat. Rev. Genet.*, 2009.

[159] M. I. Soriano, B. Roibás, A. B. García, and M. Espinosa-Urgel. Evidence of circadian rhythms in non-photosynthetic bacteria? *Journal of Circadian Rhythms*, 8(1):8, 2010.

[160] A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, and T. J. Ettema. Complex archaea that

bridge the gap between prokaryotes and eukaryotes. *Nature*, 521(7551):173–179, 2015.

[161] D. A. Stahl, D. Lane, G. Olsen, and N. Pace. Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences. *Appl. Environ. Microb.*, 49(6):1379–1384, 1985.

[162] R. Stepanauskas. Single cell genomics: an individual look at microbes. *Curr. Opin. Microbiol.*, 15(5):613–620, 2012.

[163] R. Stepanauskas and M. E. Sieracki. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *P. Natl. Acad. Sci. USA*, 104(21):9052–9057, 2007.

[164] F. J. Stewart, E. A. Ottesen, and E. F. DeLong. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J.*, 2010.

[165] S. Sun, J. Chen, W. Li, I. Altintas, A. Lin, S. Peltier, K. Stocks, E. E. Allen, M. Ellisman, J. Grethe, et al. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res.*, 39(suppl 1):D546, 2011.

[166] S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, et al. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359, 2015.

[167] A. Sundquist. Data management for the first million human genomes. UCSC BME280B talk on March 3, 2011 by DNAnexus CEO Dr. Andreas Sundquist.

[168] B. K. Swan, M. D. Chaffin, M. Martinez-Garcia, H. G. Morrison, E. K. Field, N. J. Poulton, E. D. P. Masland, C. C. Harris, A. Sczyrba, P. S. Chain, et al. Genomic and metabolic diversity of Marine Group I Thaumarchaeota in the mesopelagic of two subtropical gyres. *PLoS One*, 9(4):e95380, 2014.

[169] V. Tai and B. Palenik. Temporal variation of Synechococcus clades at a coastal Pacific Ocean monitoring site. *ISME J.*, 3(8):903–915, 2009.

[170] W. Tang, J. Bischof, N. Desai, K. Mahadik, W. Gerlach, T. Harrison, A. Wilke, and F. Meyer. Workload characterization for mg-rast metagenomic data analytics service in the cloud. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 56–63. IEEE, 2014.

[171] `http://ocean-microbiome.embl.de/statstable.html`. Tara Oceans Microbiome Data Resource accessed on web October 15, 2015.

[172] G. E. Team. Closure of the NCBI SRA and implications for the long-term future of genomics data storage. *Genome Biol.*, 12(3):402, 2011.

[173] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, 5(1):163, 2004.

[174] B. Temperton and S. J. Giovannoni. Metagenomics: microbial diversity through a scratched lens. *Curr. Opin. Microbiol.*, 2012.

[175] ThermoFisher Scientific Multiple Primer Analyzer.

[176] T. Thingstad, M. Krom, R. Mantoura, G. F. Flaten, S. Groom, B. Herut, N. Kress, C. Law, A. Pasternak, P. Pitta, et al. Nature of phosphorus limitation in the ultraoligotrophic eastern Mediterranean. *Science*, 309(5737):1068–1071, 2005.

[177] T. Thomas, J. Gilbert, and F. Meyer. Metagenomics — a guide from sampling to data analysis. *Microb. Inform. Exp.*, 2(3):1–12, 2012.

[178] T. Thomas, J. Gilbert, and F. Meyer. A 123 of metagenomics. In *Encyclopedia of Metagenomics*, pages 1–9. Springer, 2015.

[179] R. V. Thurber, M. Haynes, M. Breitbart, L. Wegley, and F. Rohwer. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.*, 4(4):470–83, 2009.

[180] A. H. Treusch, S. Leininger, A. Kletzin, S. C. Schuster, H. Klenk, and C. Schleper. Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environ. Microbiol.*, 7(12):1985–1995, 2005.

[181] G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, 2004.

[182] J. F. Vázquez-Castellanos, R. García-López, V. Pérez-Brocal, M. Pignatelli, and A. Moya. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics*, 15(1):37, 2014.

[183] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66, 2004.

[184] L. Villanueva, D. R. Speth, T. van Alen, A. Hoischen, and M. S. Jetten. Shotgun metagenomic data reveals significant abundance but low diversity of ”Candidatus Scalindua” marine anammox bacteria in the Arabian Sea oxygen minimum zone. *Frontiers in Microbiology*, 5, 2014.

[185] S. Voget, B. Wemheuer, T. Brinkhoff, J. Vollmers, S. Dietrich, H. Giebel, C. Beardsley, C. Sardemann, I. Bakenhus, S. Billerbeck, et al. Adaptation of an abundant Roseobacter RCA organism to pelagic systems revealed by genomic and transcriptomic analyses. *ISME J.*, 2014.

[186] M. Wang, C. G. Kurland, and G. Caetano-Anollés. Reductive evolution of proteomes and protein structures. *P. Natl. Acad. Sci. USA*, 108(29):11954–11958, 2011.

[187] Y. Wang, H. C. Leung, S. M. Yiu, and F. Y. Chin. MetaCluster-TA: taxonomic annotation for metagenomic data based on assembly-assisted binning. *BMC Genomics*, 15(Suppl 1):S12, 2014.

[188] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton. Jalview version 2a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.

[189] K.A. Wetterstrand. DNA sequencing costs: data from the NHGRI Large-Scale Genome Sequencing Program Available at `http://www.genome.gov/sequencingcosts/`. NHGRI web site accessed on October 7, 2015.

[190] R. A. White, P. C. Blainey, H. C. Fan, and S. R. Quake. Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics*, 10(1):116, 2009.

[191] W. B. Whitman, D. C. Coleman, and W. J. Wiebe. Prokaryotes: the unseen majority. *P. Natl. Acad. Sci. USA*, 95(12):6578–6583, 1998.

[192] A. Wilke, T. Harrison, J. Wilkening, D. Field, E. M. Glass, N. Kyrpides, K. Mavrommatis, and F. Meyer. The m5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics*, 13(1):141, 2012.

[193] A. Wilke, J. Wilkening, E. M. Glass, N. L. Desai, and F. Meyer. Porting the MG-RAST metagenomic data analysis pipeline to the cloud. 2011.

[194] J. Wilkening, A. Wilke, N. Desai, and F. Meyer. Using clouds for metagenomics: a case study. In *Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference*, pages 1–6. IEEE, 2009.

[195] D. Willner, R. V. Thurber, and F. Rohwer. Metagenomic signatures of 86 microbial and viral metagenomes. *Environ. Microbiol.*, 11(7):1752–1766, 2009.

[196] K. E. Wommack, J. Bhavsar, and J. Ravel. Metagenomics: read length matters. *Appl. Environ. Microb.*, 74(5):1453–1463, 2008.

[197] J. C. Wooley, A. Godzik, and I. Friedberg. A primer on metagenomics. *PLoS Comput Biol*, 6(2):e1000667, 2010.

[198] T. Woyke, G. Xie, A. Copeland, J. M. Gonzalez, C. Han, H. Kiss, J. H. Saw, P. Senin, C. Yang, S. Chatterji, et al. Assembling the marine metagenome, one cell at a time. *PLoS One*, 4(4):e5299, 2009.

[199] Y. Wurm, J. Wang, O. Riba-Grognuz, M. Corona, S. Nygaard, B. G. Hunt, K. K. Ingram, L. Falquet, M. Nipitwattanaphon, D. Gotzek, et al. The genome of the fire ant Solenopsis invicta. *P. Natl. Acad. Sci. USA*, 108(14):5679–5684, 2011.

[200] P. Yilmaz, R. Kottmann, D. Field, R. Knight, J. R. Cole, L. Amaral-Zettler, J. A. Gilbert, I. Karsch-Mizrachi, A. Johnston, G. Cochrane, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.*, 29(5):415–420, 2011.

[201] S. Yooseph, G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, et al. The Sorcerer II Global Ocean Sampling Expedition: expanding the universe of protein families. *PLoS Biol.*, 5(3):e16, 2007.

[202] N. Yutin and O. Béjà. Putative novel photosynthetic reaction centre organizations in marine aerobic anoxygenic photosynthetic bacteria: insights from metagenomics and environmental genomics. *Environ. Microbiol.*, 7(12):2027–2033, 2005.

[203] N. Yutin, M. T. Suzuki, H. Teeling, M. Weber, J. C. Venter, D. B. Rusch, and O. Béjà. Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. *Environ. Microbiol.*, 9(6):1464–1475, 2007.

[204] J. P. Zehr. Nitrogen fixation by marine cyanobacteria. *Trends Microbiol.*, 19(4):162–173, 2011.

[205] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 18(5):821–829, 2008.

[206] B. Zhang, C. R. Penton, C. Xue, Q. Wang, T. Zheng, and J. M. Tiedje. Evaluation of the Ion Torrent PGM for gene-targeted studies using amplicons of the nitrogenase gene, nifH. *Appl. Environ. Microb.*, pages AEM–00111, 2015.

[207] L. Zinger, L. A. Amaral-Zettler, J. A. Fuhrman, M. C. Horner-Devine, S. M. Huse, D. M. Welch, J. B. Martiny, M. Sogin, A. Boetius, and A. Ramette. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS One*, 6(9):e24570, 2011.