Analysis of viral signatures from Marine Microorganisms by Single-cell Amplified Genomes and Metagenomic Assembled Genomes

Submitted by Ashley Grenville Bell to the University of Exeter as a thesis for the degree of Master of Science by Research in Biological Sciences in August 2020.

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

(Signature) …………………………………………………………

Acknowledgements

I would like to express my deepest appreciation to my supervisor, Dr Ben Temperton whose charismatic attitude and sense of humour have been a joy to work for. This project would not be possible without his invaluable advice and guidance, inspiring me to continue my work in research. I am greatly fortunate to be able to tap into his limitless fountain of knowledge, but most of all admire his patience to read all 8 drafts of my thesis without growing too many grey hairs.

I would also express my gratitude to Rebekah Boreham for giving up her time to read and re-read every word of this thesis. Without your proof-reading and motivation, this project would never have come together. I am now far too scared to ever start a sentence with “with” again.

In addition, I would like to thank the Temperton Lab, especially Michelle Michelsen for her guidance during wet-lab work, being a great colleague and even better friend. This project would have been much more difficult without the combined knowledge of the Temperton Lab.

Most of all, thanks to my parents who have supported me financially to complete my Masters and that one day I may make enough money to pay them back.

Abstract

70% of the world's surface is covered by oceans; its impact on the global carbon cycle, climate change, and acid-base biochemistry remain crucial to our understanding of the natural world. The oceans act as important buffers against climate change, absorbing 25% of anthropogenic carbon and over 90% of rising temperatures. 90% of the ocean's biomass is composed of marine microorganisms and their impact on global systems, particularly in the face of anthropogenic climate change, remains an active area of research. Marine microorganisms are critical in the energy cycle and are the foundation for marine life. Warmer waters have led to increasingly stratified and nutrient-depleted water masses at the ocean surface, favouring low-nutrient microbial specialists. One group of these, known as the SAR11 clade, comprise up to 40% of the microbial community and are estimated to convert up to 20% of all global primary production back to atmospheric CO_2 as well as being an important biological source of methane. Increasing SAR11 abundance in warming oceans and concomitant increases in remineralisation of CO_2 and methane may create a positive feedback loop for global warming.

A potential brake on the influence of SAR11 carbon remineralisation is their associated viruses, which are predicted to lyse up to 20% of cellular biomass daily. These viruses also encode an enormous array of genetic diversity and its relationship with both physical and biological factors is key to understanding the marine biome's population dynamics. Predation of cells by viruses is a major driver of carbon export to the deep ocean, but our knowledge of these interactions in the SAR11 clade is limited, in part due to the paucity of host-virus model systems for this clade.

However, studying these microorganisms remains challenging since only a few SAR11 strains have been isolated and cultured for *in vitro* experimentation. Alternative study methods include obtaining genomes via metagenomics studies and Single-cell Amplified Genomes (SAGs). Therefore, the goal of this project is to extract and explore SAR11 host and associated phage genomes from metagenomic and SAG data. Here, I present a study of 451 SAGs collected from the Tara Ocean expeditions and twelve prokaryotic metagenomic samples from the Bermuda Atlantic Time Series (BATS).

Overall, I summarise the difficulty of obtaining contiguous and high-quality SAR11 genomes from metagenomic data. I conclude possible reasons why existing bioinformatics tools are ineffective at recovering such sequences and suggest improvements through long-read technology. Through SAG data, I identified and evaluated genomic regions associated with phage defence to improve our understanding of SAR11-associated viral dynamics in the oceans. Additionally, I characterised two previously undescribed clades of SAR11, both phylogenetically and ecologically. Our 451 SAGS contained fewer phage sequences than SAGs from other taxa, indicating the SAR11 clade does not conform to the expected statement that 20% of all marine microorganisms are infected at any given time. Lastly, I confirmed that a hypervariable region identified as a putative site for host-virus Red Queen dynamics is present within all clades of SAR11, and concluded these regions are enriched in genes related to cell wall biosynthesis. I hypothesise that these genes are related to phage defence, altering the cell wall receptors and preventing recognition of a host by SAR11 phages, therefore resisting infection. These findings together increase our understanding of additional host-phage interactions SAR11 has and impact current models when calculating SAR11 phage carbon-sequestering via the viral shunt.

Table of Contents

List of Figures and Tables

List of Abbreviations

Abbreviation	**Meaning**
ANI	Average Nucleotide Identity
BATS	Bermuda Atlantic Time-series
DCM	Deep Chlorophyll Maximum
DOM	Dissolved Organic Material
GC	Guanine Cytosine
HMM	Hidden Markov Model
KotM	King of the Mountain
KtW	Kill the Winner
MAG	Metagenomically Assembled Genome
MDA	Multiple Displacement Reaction
MSA	Multiple Sequence Alignment
OLC	Overlap Layout Consensus
Pfam	Protein families
PtW	Piggy-back the Winner
RQH	Red Queen Hypothesis
SAG	Single-cell Amplified Genome
SAM	Sequence Alignment Map
SCG	Single Cell Genomics/Genome
UEZ	Upper Euphotic Zone
UMP	Upper Mesopelagic
UMI	Unique Molecular Identifier
WGA	Whole Genome Amplification
WGS	Whole Genome Sequence

List of Software used

Software	Version	Description	Link
anvi'o	5.2.0	Analysis and visualisation platform for 'omics data	http://merenlab.org/software/anvio/
BamM	1.7.3	Parser for BAM files	http://ecogenomics.github.io/BamM/
Barrnap	0.9		
Bbmap suite	38.22	A suite of bioinformatics tools designed for analysis of DNA and RNA sequence data	https://jgi.doe.gov/data-and-tools/bbtools/
BinSanity	0.2.8.2	Unsupervised Clustering of Environmental Microbial Assemblies Using Coverage and Affinity Propagation	https://github.com/edgraham/BinSanity
BioPython	1.72	Tools for biological computation written in Python	https://biopython.org/
BLAST	2.5.0	Find regions of similarity between biological sequences	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Bowtie2	2.3.4.3	Aligning sequencing reads to long reference sequences	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
CheckM	1.0.13	Uses Single Marker Copy Genes to infer species completeness and redundancy	http://ecogenomics.github.io/CheckM/
Diamond	0.9.21	Sequence aligner for protein and translated DNA searches	https://github.com/bbuchfink/diamond
eggNOG	1.0.3	Provides Orthologous Groups of proteins at different taxonomic levels, each with integrated and summarised functional annotations	http://eggnogdb.embl.de/#/app/home
fastANI	1.1		
GTDB-TK	0.2.2	Software toolkit for assigning objective taxonomic classifications to bacterial and archaeal genomes	https://github.com/Ecogenomics/GTDBTk
HDBSCAN	0.8.22	Hierarchical clustering algorithm	https://github.com/scikit-learn-contrib/hdbscan
IQ-Tree	1.6.9	Phylogenetic Analysis	http://www.iqtree.org/
MAFFT	7.407	Multiple sequence alignment program	https://mafft.cbrc.jp/alignment/software/
Matplotlib	3.1.0	Python 2D plotting library	https://matplotlib.org/

Maxbin	2.2.6	Automated binning method to recover individual genomes from metagenomes	(Wu et al. 2014)
metaBAT2	2.12.1	A statistical framework for reconstructing genomes from metagenomic data	https://bitbucket.org/berkeleylab/metabat/src/master/
metaQUAST	5.0.2	Evaluates and compares metagenome assemblies based on alignments to close references	http://bioinf.spbau.ru/metaquast
Minimap2	2.15	Versatile sequence alignment program that aligns DNA or mRNA sequences against a large reference database	https://github.com/lh3/minimap2
MUSCLE	3.8.1551	MUltiple Sequence Comparison by Log-Expectation for multiple sequence alignment of protein and nucleotide sequences	https://www.drive5.com/muscle/
Ocean Data Viewer (ODV)	5.1.7	Software package for the interactive exploration, analysis and visualization of oceanographic and other geo-referenced profile	https://odv.awi.de/
Prodigal	2.6.3	Protein-coding gene prediction for prokaryotic genomes	https://github.com/hyattpd/Prodigal
Prokka	1.12	Rapid prokaryotic genome annotation	https://github.com/tseemann/prokka
Python	3.6.8 / 2.7.16	Programming language	https://www.python.org/
RAxML	8.2.12	Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies	https://github.com/stamatak/standard-RAxML
Samtools	1.9	Manipulate alignments in the BAM format	http://www.htslib.org/doc/samtools.html
Seaborn	0.9.0	Python data visualization library based on matplotlib	https://seaborn.pydata.org/index.html
Snakemake	3.13.3	Workflow management system	https://snakemake.readthedocs.io/en/stable/
Seqtk	1.3	Fast and lightweight tool for processing sequences in the FASTA or FASTQ format	https://github.com/lh3/seqtk
SPAdes	3.13.0	Assembly toolkit containing various assembly pipelines	https://github.com/ablab/spades
T-Coffee	11.0.8	Multiple sequence aligner for Protein, DNA and RNA sequences	http://tcoffee.crg.cat/

trimAl	1.4.1	Automated removal of spurious sequences or poorly aligned regions from a multiple sequence alignment	http://trimal.cgenomics.org/
UMAP	3.3.9	Uniform Manifold Approximation and Projection. Dimension reduction technique that can be used for visualisation	https://github.com/lmcinnes/umap
vConTACT2	0.9.5	A tool to perform guilt-by-contig-association classification of viral genomic sequence data	https://bitbucket.org/MAVERICLab/vcontact2/src
VIRSorter	-	Mines viral signal from microbial genomic data	https://github.com/simroux/VirSorter

1 Literature Review

1.1 The role of the oceans in the global carbon cycle

Oceans have a large impact on the carbon cycle and affect the Earth's climate. They absorb half of anthropogenic carbon (Sabine et al. 2004), store up to 900 Gt of carbon within the ocean surface and 36,400 Gt within the deep ocean, compared to an atmospheric carbon pool of 750 Gt (S. W. Wilhelm and Suttle 1999). CO_2 enters the ocean surface via diffusion and dissolves into the water, particularly in colder waters which sink and take dissolved carbon with them in downwelling events. Upwelling events occur where colder water rises upon approaching shallow coastlines which can warm and release dissolved CO_2 back into the atmosphere (Takahashi et al. 2002). Of the 10 Gt of carbon produced yearly, 2.4 Gt is stored within the ocean (Ciais et al. 2014). Organic carbon storage within the oceans is often microscopic in scale and is divided based on its size. Dissolved organic carbon (DOC) has multiple definitions, historically defined as any organic matter able to pass through a 0.7μm glass fibre filter (Mostofa et al. 2013; Craig A. Carlson et al. 2010). This has more recently been updated to exclude marine microorganisms with a 0.2μm pore-sized plastic filter (Craig A. Carlson and Hansell 2015), but probably should exclude viruses with a new definition as organic carbon passing through a 0.02 filter (Farooq Azam and Malfatti 2007). Regardless, particulate organic carbon (POC) is carbon captured on these filters, being too large to pass through.

Marine phytoplankton contributes to the biological carbon cycle, producing up to half of the world's photosynthetic carbon (Field et al. 1998) within the ocean's surface. This is referred to as the biological pump where photosynthetic marine microorganisms fix CO_2 into organic carbon which is transported into the ocean for long term storage (Eppley and Peterson 1979; Hugh W. Ducklow, Steinberg, and Buesseler 2001). Leaky cells, predation and lysis of photosynthetic organisms by eukaryotic grazers and associated viruses release POC and DOC into the marine environment (H. W. Ducklow et al. 1995; Biddanda and Benner 1997) providing nutrients for heterotrophic organisms. The microbial loop is, therefore, created where lysis of marine microorganisms provides DOC for other heterotrophic microorganisms (Jiao et al. 2010). The viral shunt (Curtis A. Suttle 2007a) plays an important role within this loop as viruses on average predate upon 20-40% of all marine microorganisms (Proctor and Fuhrman 1990; Bergh et al. 1989; Curtis A. Suttle, Chan, and Cottrell 1990). The viral shunt is responsible for releasing up to 25% of all DOC into the marine environment (S. W. Wilhelm and Suttle 1999; Wommack and Colwell 2000; Fuhrman 1999). This forms a spectrum of DOC types dependent on molecular weight (F. Azam et al. 1994; Verdugo et al. 2004). Labile DOC refers to lower molecular weight organic carbon molecules like sugars and amino acids. These are highly concentrated within the ocean surface and easily metabolised by other marine organisms (C. A. Carlson and Ducklow 1995; Bauer, Williams, and Druffel 1992). Semi-labile carbon refers to larger organic molecules like protein structures and forms the intermediate size of DOC. The rate of metabolism can also play a part in DOC classification where a structure requiring large amounts of energy to degrade is considered refractory DOC. Simpler organic compounds were rare in analytical carbon sampling of marine waters suggesting reactivity also plays an important role in DOC classification. Lower molecular weight but stable molecules such as benzene are considered refractory DOC due to its double-bonded ring structure (Gruber et al. 2006; Amon and Benner 1996). Refractory DOC and POC are generally higher in molecular weight and consist of more complex compounds such as aggregated cell bodies. These are evenly dispersed throughout the ocean and its sedimentation contributes to long term carbon storage (Hopkinson and Vallino 2005; Bauer, Williams, and Druffel 1992). Aggregation of dead microorganisms sink and contribute to the deep ocean carbon sink, sequestering 10% of all carbon aggregate by marine biota (Ciais et al. 2014). This is only brought back to the surface by thermohaline circulation (Hugh W. Ducklow, Steinberg, and Buesseler 2001).

Image removed due to copyright. See Ciais, P., C. Sabine, G. Bala, L. Bopp, V. Brovkin, J. Canadell, A. Chhabra, R. DeFries, J. Galloway, M. Heimann, C. Jones, C. Le Quéré, R.B. Myneni, S. Piao and P. Thornton, 2013: Carbon and Other Biogeochemical Cycles. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. Figure 6.1

Figure 1.1 Simplified schematic of the global carbon cycle, (Ciais et al. 2014). Units in PgC/Gt per annum.

1.2 The Great Plate Count Anomaly

Examination of seawater under the microscope revealed a vast diversity of microbial life and disproved earlier ideas of sterile oceans (Jannasch and Jones 1959). Microscopy revealed within 1ml of seawater, up to 1 million bacteria and ten times as many viral particles present, (Bergh et al. 1989; Proctor and Fuhrman 1990; Hara, Terauchi, and Koike 1991; Wommack et al. 1992) which contributes to 90% of the oceans total biological carbon in the oceans (S. W. Wilhelm and Suttle 1999).This contrasted previous attempts to evaluate marine microbial abundance and diversity through plated samples where it was concluded that only 1% of the total bacteria could be enumerated (Gregory 1979). This plate count anomaly was attributed either to plating of non-viable cells or the unknown conditions and nutrients required for marine microorganism growth (Staley and Konopka 1985a). Determination of *in situ* activities of these microorganisms was done with microautoradiography using radiolabeled substrates to observe metabolic activity (Iturriaga and Hoppe 1977; Meyer-Reil 1978; Zimmermann, Iturriaga, and Becker-Birck 1978). Results showed microbial cells were metabolically active and therefore the lack of microbial growth on plates was hypothesised to be a result of unknown nutrient conditions required to support growth (Staley and Konopka 1985b). However, measured metabolic activity highlighted the importance of marine microbes in global carbon biogeochemistry.

1.3 The dominance of SAR11

Although microscopy gave indications of marine microbial abundance, it was not until the rise of DNA sequencing that phylogenetic categorisation of these marine microorganisms revealed their extraordinary diversity. In one of the earliest applications of environmental microbial community profiling, microbial communities from the Sargasso Sea were used to produce 16S rRNA amplicon libraries (S. J. Giovannoni, DeLong, et al. 1990), leading to the discovery of a new clade of heterotrophic alphaproteobacteria (Thrash et al. 2011) known as SAR11, or the Pelagibacterales (S. J. Giovannoni, Britschgi, et al. 1990). Later, it was confirmed to be one of the most ubiquitous marine bacteria with an estimated population size of 2.4 x 10^{28} (Stephen J. Giovannoni 2017). SAR11 comprises on average 25% of all surface water cells (Morris et al. 2002; Biers, Sun, and Howard 2009; Herlemann et al. 2014), rising to 40% in stratified waters (Becker et al. 2019). They are the most abundant aquatic microorganisms having also diversified into freshwater habitats (Eiler et al. 2016; Henson et al. 2018).

SAR11 are free-living chemoheterotrophic bacteria capable of oxidising dissolved organic matter (DOM) within the ocean. Comparative genomic analysis identified a core genome across the clade that was the most conserved ever observed (Grote et al. 2012). This high genome conservation facilitates the highest homologous recombination rates of any prokaryote, allowing for high rates of lateral propagation of genes (Vergin et al. 2007; Y. Sun and Luo 2018). All SAR11 genomes are highly ‘streamlined’ (Lynch and Conery 2003), meaning they possess genomes adapted for replication in low-nutrient conditions, selecting against encoding of regulatory mechanisms, pseudogenes and non-coding regions in favour of reduced metabolic costs of replication. Consequently, they have one of the smallest sized free-living cells in both size and genome length (Stephen J. Giovannoni et al. 2005). The removal of unnecessary protein-coding regions has pointed towards its oligotrophic lifestyle and allowed it to have a small genome size. This suggests that it was selected to reduce replication costs in nutrient-poor conditions (Stephen J. Giovannoni, Cameron Thrash, and Temperton 2014). This is in direct contrast to other small streamlined cells where fewer nutrient requirements and a higher surface area to volume ratio allow for faster genome replication rates and undergo a copiotrophic lifestyle. SAR11 undergoes a longer cell division time (A. E. White et al. 2019) in comparison to other prokaryotic species. This explains SAR11’s high abundance in nutrient-poor waters as they are at an advantage when the system is in a steady date. Their decline is heralded by changes within this system, for example, water layer mixing events like storms where upwelling increases the abundance of nutrients (Parsons et al. 2012; Craig A. Carlson et al. 2009).

The SAR11 clade is currently divided into five phylogenetic groups. Each subclade is represented on a phylogenetic branch consisting of similar species of SAR11 and sometimes a unique ecological niche within the ocean. For example, the Ia clade of SAR11 can be split into Ia.1 and Ia.3 subclades of cold-water and warm-water ecological types respectively, based on latitudinal distributions (M. V. Brown et al. 2012). Additionally, ecological types are dependent on ocean currents, as seasonal mixing of DOM from the surface to the mesopelagic allows for blooms of clade IIb to occur (Craig A. Carlson et al. 2009, 2010).

Image removed due to copyright. See Giovannoni SJ. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. Ann Rev Mar Sci. 2017 Jan 3;9:231-255. doi: 10.1146/annurev-marine-010814-015934. Epub 2016 Sep 28. PMID: 27687974. Figure 1

Figure 1.2 Figure 1 from (Stephen J. Giovannoni 2017) detailing ***(a)*** Phylogenetic tree of SAR11 diversity. ***(b)*** the spatiotemporal distribution of SAR11 ecotypes at the Bermuda Atlantic Time Series. DCM = deep chlorophyll maximum; SBL = spring bloom; UEZ = upper euphotic zone; UMP = upper mesopelagic.

1.4 Pelagiphages and their ecological impact

Although the origins of viruses are unclear, it has long been established that viruses have a large impact on ecosystems (Fuhrman 1999; Rohwer, Prangishvili, and Lindell 2009; Middelboe and Brussaard 2017). Viruses are key factors in regulating bacterial and eukaryotic microorganism populations (Curtis A. Suttle 2007b; Wommack and Colwell 2000; Middelboe and Brussaard 2017) and can infect both multicellular and single-cell organisms, often with detrimental effects to the host (Rohwer and Thurber 2009; Weinbauer 2004). Viral particles are hugely abundant, reaching up to 10^8 ml^{-1} in surface waters (Noble and Fuhrman 1998; Hennes and Suttle 1995); ten times greater than cellular microorganisms (Bergh et al. 1989; Hara, Terauchi, and Koike 1991). They act as agents of horizontal gene transfer between cells (McDaniel et al. 2010; Paul 2008) and during infection they can increase both the phenotypic and genotypic diversity of the host (Breitbart et al. 2007; Sharon et al. 2011; L. R. Thompson et al. 2011; Anantharaman et al. 2014).

Generally, viruses undergo a continuous spectrum of lifestyles. This ranges from a lysogenic state where viruses integrate within the genomes of their host and can excise themselves when specific conditions are met. Alternatively, they can exist as free virions that infect their hosts and immediately replicate, followed by host lysis and progeny release (lytic phages). All viruses are able to undergo a lytic form of lifestyle which can be repressed in certain conditions. High abundance and growth rates of hosts can be a condition which shifts the viral population’s lifestyle into lysogeny and therefore increases provirus abundance (Silveira and Rohwer 2016). Temperate phages are of interest due to their ability to integrate themselves within their hosts. Some bacterial genomes have been shown to have up to 20% of their genomes from viral origins (Casjens et al. 2000) through horizontal transfer events. This allows for genetic variation to be introduced into the host genome, allowing for the introduction of genes that may confer advantages or be detrimental to the host (Lindell et al. 2005; Roux et al. 2014; Sharon et al. 2009; Hurwitz, Brum, and Sullivan 2015). It has also been hypothesized that lysogeny acts as a putative mechanism for the transfer of antimicrobial-resistant or pathogenic genes from one bacterium to another (Waldor and Friedman 2005). Prophages - phages that have integrated themselves within a host genome, can become unviable due to the loss of excision/essential genes and therefore confer permanent DNA mutations and additional genes without being pathogenic (Canchaya, Fournous, and Brüssow 2004). Studying these genes may offer insight into the ability of these viruses to add advantages to the cell, increasing their fitness. Additionally, conditions may change the viral lifestyle from a lysogenic to a lytic one where changes in the environment detrimental to the host can cause provirus excision. Moreover, virocells, cells undergoing viral infection, respond metabolically differently based on the type of viral infection (Howard-Varona et al. 2020). This changes the impact of ecological models and carbon sequestering depending on the viral predation.

The dynamics of viral predation of marine microorganisms have been described by several ecological models. Kill the Winner (KtW) states that, in the presence of a high density of prey, there is an increased chance of a prey-predator interaction which eventually results in predator-prey density equilibrium (Avrani, Schwartz, and Lindell 2012; Winter et al. 2010). Thus, prey and predatory numbers are proportionally linked. This is a common interaction in fast-growing and large communities of bacteria. This is further subsetted with Piggyback the Winner (PtW) strategy where viruses integrate within hosts at a high abundance and prevent subsequent additional phage infections (Silveira and Rohwer 2016).

However, SAR11 dominance in the ocean challenges existing ecological theories where its ubiquitous nature contradicts the KtW hypothesis. Hypotheses to explain this contradiction include cryptic escape: the maintenance of abundant but small and slowly replicating cells reduce total biomass turnover and therefore decrease predation (Yooseph et al. 2010). Other explanations include K-strategy defensive specialism where resources are invested in survival over replication (Curtis A. Suttle 2007b; Klappenbach, Dunbar, and Schmidt 2000), where low replication rates prevented SAR11 phages from becoming established. SAR11 can be described as defensive specialists when a large proportion of its resources are spent on maintaining its carry capacity rather than replication (Våge et al. 2014). However, a study by (Y. Zhao et al. 2013) indicated SAR11 phages (pelagiphages) are highly abundant and comprise some of the most ubiquitous phages in the world, in direct contrast to what would be observed under a defensive specialist model. This would indicate that additional ecological models would be needed to explain host-prey interactions. In the presence of a high abundance of both phage and host, the likelihood of contact with each other would logically remain high. Therefore, it is possible SAR11 has evolved a mechanism to escape predation.

The Red Queen Hypothesis (RQH) states that an organism must constantly evolve to survive against constantly evolving predators (Van Valen 1973; Brockhurst et al. 2014). This results in a genetic arms race where genes related to predator-prey relationships are under high selection pressure. An example of this is the diversity of bacteria populations maintained by viral predation (Rodriguez-Valera et al. 2009). These ecological models aim to explain predator-prey dynamics where different models are needed to explain different scenarios. This proliferation of successful genes that can be then co-evolved has led to the characterisation of a new ecological model called King of the Mountain (KotM) (S. Giovannoni, Temperton, and Zhao 2013). The KotM hypothesis states that the SAR11 clade is highly abundant due to their ability to share genes that contribute to their success at a rate higher than phages can evolve to predate on them. Their high abundance allows increased interactions and for a large diversity of genes to co-evolve. They have high homologous recombination rates (Vergin et al. 2007) whereby successful genes are passed quickly within a population. A likely candidate to facilitate recombination associated with phage-defence within the SAR11 clade is their Hypervariable Regions (HVRs) (Grote et al. 2012) also known as genomic islands (Avrani et al. 2011): areas within a genome that have higher than normal evolution rates in comparison to the entire genome as a whole. These regions are genetic “playgrounds”, undergoing higher mutation rates with the hopes of developing advantageous genes against phage predation. Recombination events would pass these HVRs to other SAR11 individuals, conferring subsequent immunity to phage infection. This is explored further in **Chapter 5** of this project.

Although the SAR11 clade and its phages are highly abundant, the number of cultured isolates remains low due to its difficulty in culturing. Therefore, relatively new methods including the field of metagenomics and SAGs provide genetic material to study its genome on a population-wide scale.

1.5 Sequencing the difficult to culture

Traditional methods for studying microorganisms include the isolation and axenic culturing of organisms, often on solid agar medium. However, it is estimated that only around 1% of all described organisms can be studied in this way (Rinke et al. 2013; Amann, Ludwig, and Schleifer 1995), although there has been some debate towards this topic (Martiny 2019; Steen et al. 2019). This limitation is hypothesised to result from unknown, but specific nutrient requirements or growth conditions of each species. Additionally, viruses obtained from the environment remain difficult to culture as, without their known host, only virions can be extracted without knowledge of their host they infect. To overcome this, there exist two main methods for obtaining the genomics sequence of these microorganisms. The field of metagenomics (Ghosh, Mehta, and Khan 2019) and single-cell amplified genomes (SAGs) (Kogawa et al. 2018). These methods are regarded as culture-independent methods and rely solely on the ability to capture these organisms with environmental samples or through single-cell sorting.

1.6 Shotgun Metagenomics

One of the first uses of metagenomics was to clone an entire soil metagenome to study its genomic content (Handelsman et al. 1998). It has since evolved into the study of sequenced genetic material or assays of a culture-independent environmental sample (Wooley and Ye 2010). This acts as a “snapshot” of the genetic content obtained at one spatio-temporal point. It involves collecting an environmental sample followed by extraction and sequencing of the DNA using sequencing technology (M. B. Scholz, Lo, and Chain 2012). This allows for a broad range of studies into an environment’s population dynamics, currently impossible with axenic cultured organisms obtained with most traditional microbial culturing techniques, for example, plating. Microbial communities are diverse and metagenomics provides a method to study such interactions en masse. These studies divide into large-scale shotgun studies, aimed at sequencing whole genomes (Pesant et al. 2015), as well as amplicon surveys aimed at specific genes of interest, common in 16S rRNA population community studies (B. J. Campbell et al. 2011). These studies combined with computational pipelines have allowed for the scrutiny of hard to culture microbial populations (Steen et al. 2019). These are important as interactions of these microbes in natural communities prove difficult to reconstruct in traditional axenic cultures. Although extraction and purification of DNA from an environment has been well documented and studied, (Djurhuus et al. 2017; Mygind et al. 2003; Oliveira et al. 2014; De Medici et al. 2003) computational approaches inferring the source of each DNA fragment have been challenging (Sczyrba et al. 2017; Nayfach and Pollard 2016). Current informatics-based software still struggles to determine the identity of DNA sequences from complex communities with similar DNA compositions. However, the field of metagenomics offers promise, allowing us to address important questions relating to the diversity of microbes in natural environments, (Sharon et al. 2011; Sunagawa et al. 2015) inter and intraspecies microbial interactions (Turnbaugh et al. 2006) and evolutionary differences between environmental populations (Hooper et al. 2008).

A typical shotgun metagenomic study comprises of several steps: (i) extraction, isolation and sequencing of DNA, (ii) preprocessing, quality control and assembly of sequenced reads, and (iii) post-assembly analysis of contigs to deduce taxonomic and functional features (Quince et al. 2017). Although simple in theory, in practice a multitude of specialised tools have been created for each different microbial community to allow accurate and efficient recovery of genetic material for better downstream analysis.

1.6.1 DNA Extraction

When DNA is extracted it is important to consider the amount of biomass and therefore DNA required to perform downstream analysis of rare taxa, whilst reducing environmental contamination such as human DNA accidentally introduced due to poor sterilisation techniques. Although DNA amplification allows for the formation of additional starting material (Dean et al. 2001), these run the risks of amplification bias within samples, skewing population abundance data towards taxa that are enriched disproportionally (Probst et al. 2015; Binga, Lasken, and Neufeld 2008; Yilmaz, Allgaier, and Hugenholtz 2010; Direito et al. 2014; Marine et al. 2014). Additionally, physical versus chemical cell lysis DNA extraction methods can skew population abundance data (Wesolowska-Andersen et al. 2014). Therefore, DNA extraction must be equally effective on a diverse range of microbes or risk under or over-representation of certain taxa, resulting in its untargeted approach being referred to as “shotgun” metagenomic sequencing. Regardless, extraction of DNA sequences usually requires fragmentation due to the requirements of short-read sequencing. However, if long unbroken lengths of DNA are required for long-read technologies, chemical over physical extraction is preferred due to its aggressive but less biased cell lysis technique, which shears DNA into shortened fragments lost in fragment selection techniques (Yuan et al. 2012; Kennedy et al. 2014). Contamination of samples is also a risk and disproportionately affects low biomass samples where background contamination is more pronounced (S. J. Salter et al. 2014). These contaminants can either be reduced by adapting existing protocols to include ultraclean steps or performing blank sequencing runs included with background contamination sequenced to be removed from sample data through downstream informatics methods (Schmieder and Edwards 2011).

1.6.2 DNA Sequencing

A variety of methods exist to obtain the genetic sequence of microbial samples in an analysable form for computational analysis. These are broadly categorised into short-read or long-read technologies, dependent on the length of contiguous DNA sequences produced per read. Short reads are normally around and below 500 DNA base pairs in length and are predominantly sequenced on Illumina platforms (Bentley et al. 2008) due to high throughput, fidelity and availability. It is highly accurate, with sequencing error rates between 0.1% to 1%. Alternatives include the Ion Torrent platforms (Rothberg et al. 2011) with similar accuracies. Long read technologies are capable of sequencing longer DNA fragments in excess of 10 kb (Loman, Quick, and Simpson 2015) and include the Nanopore technologies (M. Jain et al. 2015) as well as the Pacific Biosciences platform (Eid et al. 2009). These have higher error rates of up to 15% but have improved substantially since their introduction. Both have methods to increase their accuracy, with PacBio using circular DNA templates and the Nanopore first using “2D” reads, and now relying upon improved machine learning approaches during base-calling (Weirather et al. 2017). The construction of sequencing libraries usually requires fragmentation of the sample into suitable sizes depending on the sequencer of choice, along with barcoding of samples if they are being run in a multiplex. Such processes can reduce the rate of read recovery due to “indexing hopping” in Illumina platforms (Sinha et al. 2017), the failure for barcodes to adhere to DNA fragments or the incorrect binning of barcodes due to read errors. Additionally, PCR bias of GC rich regions may skew read abundances (Laursen, Dalgaard, and Bahl 2017; Y.-C. Chen et al. 2013).

Coverage of a metagenomic sample is also critical downstream as a lack of data can impact consensus read error correction or assembly. Although no true figure exists for the optimum coverage, deeply sequenced metagenomes allow for the detection of more rare taxa (Nayfach and Pollard 2016). To provide a rough estimate of sequencing depth required, the amount of throughput generated is divided by the number of multiplexed samples and the expected abundance of the targeted organisms within a sample. This produces a one-fold coverage depth of the targeted organism and can be further increased to desired levels. Additionally, providing multiple metagenomic samples can help provide a deeper coverage depth but may also increase variability. Furthermore, in repeat regions, additional metagenomic samples are unlikely to provide longer contiguous sequences due to the problems associated with assembling repeat regions (Treangen and Salzberg 2011).

1.6.3 Assembly

After metagenomic sequencing, DNA is translated from its physical form for bioinformatics analysis. Millions of reads are usually generated due to fragmentation by DNA extraction methods and require assembly back into their original contiguous form. De novo assembly of a metagenome is usually performed as metagenomes seldom have reference genomes on which to base their assembly (Simpson and Pop 2015). This is conceptually similar to de novo assembly of axenic cultures where reads are assembled based on their composition and not guided by a reference sequence. The most popular approach involves the usage of a de Bruijn graph (Pevzner, Tang, and Waterman 2001). A de Bruijn graph requires the splitting of reads into short but overlapping lengths of k. These short overlapping sequences are referred to as *k*-mers with k corresponding to the length of the sequence (**Fig 1.3**). For example, a sequence of four base pairs would be considered a 4-mer. These *k*-mers then provide the edges of a de Bruijn graph. Overlapping *k*-mers would then be described as *k*-mers with a similar sequence composition as their overlap would provide the next edge in a de Bruijn graph. The purpose of the assembler would be to find the most suitable or longest path through a de Bruijn graph. This path links other overlapping reads, creating a longer contiguous sequence (contig). This is a complicated process with erroneous reads providing non-biological sequences due to sequencing errors or repeat regions providing no path through to the next set of contigs. This causes fragmentation and misassembly, resulting in small fragments or incorrect compositions of an organism's genetic code (Tørresen et al. 2019). However, erroneous reads can be error corrected via consensus read error corrections. Here sequencing depth is important as a high sequencing depth can help correct erroneous reads where multiple overlapping reads can produce the correct sequence of base pairs by consensus (Schröder et al. 2009). In addition, long reads can help span or “scaffold” across repeat regions, providing areas of the genome before and after repeat regions, reducing fragmentation (Luo et al. 2020).

Sequence:

AGATGCATGCA

k-mer (k = 4)

Split into length of 4 sequences:

```
AGAT
 GATG
  ATGC
   TGCA
    GCAT
     CATG
      ATGC
       TGCA
```

Remove duplicate 4-mer:

ATGC, TGCA

Resulting 4-mers:

AGAT, GATG, ATGC, TGCA, GCAT, CATG

Figure 1.3 An example dataset visualising how *k*-mers are determined.

A unique challenge of metagenomics is the lack of uniform coverage across all its member genomes. In single genome assemblies, coverage of reads across the genome is deemed uniform and therefore repeat regions and sequencing errors can be deduced by abnormal coverage (Simpson 2014). However, within metagenomics, coverage is dependent on species abundance, with higher coverages of common taxa. This results in rare taxa having lower sequencing coverage and, in extreme cases, no coverage across parts of their genomes (Gagic et al. 2015). This results in fragmented assemblies as a de Bruijn graph is unable to span across these regions. Additionally, with multiple closely related species, nucleotide compositions only contain slight variations and therefore create branching points within de Bruijn graphs (Iqbal et al. 2012). Consequently, being unable to resolve these paths, a graph may terminate and lead to fragmentation to avoid misassemblies. *K*-mer sizes may be shortened to assist in providing additional branches in de Bruijn graph formation, but this can result in larger numbers of edges due to repetitive *k*-mers obscuring the correct path direction. Longer *k*-mer lengths are better at providing correct path formation with fewer branching options, but risk poor representation in low coverage genomes and therefore rare taxa (Wick et al. 2015).

There are a wide variety of metagenomic assemblers which use the benefits of both long and short *k*-mer lengths to assist in contig reconstruction. Metagenomic assemblers like metaSPAdes (Nurk et al. 2017), MEGAHIT (D. Li et al. 2015), MetaVelvet (Namiki et al. 2012) and IDBA-UD (Peng et al. 2012) use multi *k*-mer approaches to avoid choosing a *k*-mer length that may enrich for high or low abundance organisms. Additionally, with the increasing usage of long-read data, some hybrid assemblers like metaSPAdes and Unicycler (Wick et al. 2017) now contain algorithms to work with multiple different sequencing platforms to improve assemblies. Subsequently, with the rise of multiple metagenomic assemblers, large numbers of comparative studies have compared algorithms with little consensus to the most effective assembler in all metagenomic dataset (Bradnam et al. 2013; Forouzan et al. 2018; Simpson and Pop 2015; Vollmers, Wiegand, and Kaster 2017; Ayling, Clark, and Leggett 2019; Bertrand et al. 2019; W. Zhang et al. 2011; Sczyrba et al. 2017). It is more likely the microbial community and sequencing platform play a more important role in the “best” assembler for each dataset (Sutton et al. 2019). Moreover, due to these limitations, assemblers are unlikely to produce contigs representing entire individual genomes (Tully, Graham, and Heidelberg 2018). Until current bioinformatic algorithms improve or new sequencing methods are developed, it is likely individual populations will be represented by multiple contigs within an assembled metagenome representing one individual. Although long-read metagenomics has resulted in some assembly free viral genomes (Beaulaurier et al. 2019), it is likely that sorting contigs from the same individual into “bins” will still be required. Therefore, another set of bioinformatics algorithms is required that seeks to classify contigs into bins that represent the genome they derived from.

1.6.4 Binning

Metagenomic assemblies are usually highly fragmented, with many contigs representing small proportions of an organism's DNA. There is now a need to group contigs that belong to the same species together to create a Metagenome Assembled Genome (MAG). However, without knowing a priori which contig belongs to which species' genome, assigning contigs into "bins" based on its species proves difficult. Instead, there are two main methods for binning genetic material to form MAGs. The supervised method relies on external data to classify contigs, whereas unsupervised methods rely on the content of the contigs themselves. Both methods rely on comparing themselves to either existing databases or to other contigs within the same sample and assigning matches into bins.

Supervised methods provide a reference genome against which to compare contigs from metagenomes. A high-quality match against a reference genome may indicate this is the same species and is the basis for taxonomic classification (Altschul et al. 1990). However, with many microbial species being unknown and unsequenced (Steen et al. 2019), supervised methods are limited by the content of the database. If large proportions of a metagenomic assembly are unable to map to existing databases, unsupervised methods would be more suitable for binning contigs.

Species genomes can be differentiated by their genome composition, either in whole or in part. Examples of this include ribosomal subunit 16 rRNA, present in all bacteria. The small difference within this rRNA allows for the construction of 16S phylogenetic trees, which is based on the knowledge that sequence compositions are more similar within species that are closely related (Lane et al. 1985). This is due to their more recent divergent event from their last common ancestor, with mutations and variations having less time to develop. Therefore, sequence composition of contigs can be used for binning via the composition of *k*-mer frequencies (Karlin, Mrázek, and Campbell 1997) where regions are compared based on their *k*-mer frequency to indicate how similar sequences are to each other. *K*-mer frequencies can be described as the number of different *k*-mers a sequence may have (**Fig 1.4**). Tetramers, a *k*-mer frequency of four, is the most commonly used *k*-mer (Dick et al. 2009) with 4^4 = 256 number of possible 4-mers. The frequency of each 4-mers is recorded for each sequence and it is expected that if another contig has the same frequencies of these 4-mers, they are closely related or the same. This is because a similar sequence would have a similar composition of nucleotides. This provides a mathematical representation tabulating a nucleotide composition into numeric categories for data analysis (Marçais and Kingsford 2011). Many algorithms use *k*-mer frequency to differentiate genetic sequences (Rosen et al. 2008; Beaulaurier et al. 2019; Andersen 2018) but often fails to differentiate between species where *k*-mer frequencies are closely related (Alneberg et al. 2014; Strous et al. 2012; Dick et al. 2009). Increasing the *k*-mer size can help with differentiating between more similarly related species due to the increasing numbers of *k*-mers providing a higher resolution. However, a higher number of k exponentially increases the number of *k*-mer values (5-mer = 4^5 *k*-mers = 1024), making increasing this metric exponentially computationally intensive. Additionally, with at least 4 dimensions (k=1), visualisation and plotting of contigs are impossible on 2D plots, especially with higher numbers of k. Therefore, dimension reduction algorithms are critical in the visualisation of *k*-mer counting techniques. Dimension reduction algorithms allow for clustering of similar sequences, as with unlimited dimensions each *k*-mer frequency distribution would be plotted on its own plane, providing no clustering of data points. Dimension reduction techniques allow for the plotting of a representation of high-resolution data onto a lower-dimension for visualisation of a dataset's structure.

Sequence:

AGATGCATGCA = 11

Resulting 4-mers:

AGAT, GATG, ATGC, TGCA, GCAT, CATG

***K*-mer frequency:**

AGAT: 1

GATG: 1

ATGC: 2

TGCA: 2

GCAT: 1

CATG: 1

Figure 1.4 Example dataset visualising how *k*-mer frequencies are determined

Dimension reduction techniques like the T-distributed Stochastic Neighbor Embedding (t-SNE) algorithm (Maaten and Hinton 2008) can be used to reduce high dimensional data into a lower-dimensional space. It uses a probability distribution of high dimensional data to construct a similar distribution on a lower-dimensional map. UMAP is an algorithm similar to the t-SNE and is effective at visualising high dimensional data (McInnes, Healy, and Melville 2018). Contigs with a similar *k*-mer frequency group together, forming clusters when plotted on a scatter plot. Clusters would represent genetic sequences with similar *k*-mer frequencies and therefore come from the same or similarly related organisms. Identifying statistically supported boundaries between these clusters is then required for ‘binning’ contigs to obtain MAGs. This can be done manually with human input (Andersen 2018) by extraction of all contigs within a human-determined cluster. However, clusters may be located close to each other and, with an unknown amount of organisms within a metagenome, automation of cluster determination becomes critical to reduce human error and to ensure reproducibility. Algorithms like HDBSCAN (McInnes, Healy, and Astels 2017) cluster data points on a scatter plot based on density. Dense points are grouped as clusters and sparser background noise are ignored from the algorithm. Single-linkage clustering of clusters is used to produce a dendrogram of data points and cut at points that minimise cluster fragmentation through user-defined parameters. This allows for the automation of cluster determination to aid extraction of MAGs.

Bins produced by *k*-mer frequency often do not have the resolution to exclude species of similar genetic sequences (Beaulaurier et al. 2019). Therefore, calculating the Average Nucleotide/Amino-acid Identity (ANI/AAI) (Konstantinidis and Tiedje 2005) of coding regions can help further discriminate multiple species in a bin. Rapid evaluation of AAI used by tools such as CompareM (D. Parks 2018) works by pairwise comparison of *k*-mer counting nucleotides or amino acids in protein-coding regions of all genetic material within a bin. Here, higher values of k can be used as reduced amounts of data reduces computational resources. All contigs are provided with a value of percentage similarity to each other, creating clusters of similar contigs within a bin. These groupings represent a high resolution of different genetic material within a bin.

Decreasing sequencing costs has resulted in an increase in time series and multisample metagenomics. This has popularised the use of contig coverage as another mechanism for MAG creation within a metagenomic sample. If multiple samples are taken from the same spatio-temporal point, species composition and abundance would be expected to be the same. Although increased read depth at bacterial origins of replication can skew these assumptions (C. T. Brown et al. 2016; Korem et al. 2015), generally, a disproportionate number of reads is generated for each species within a population. After assembly, reads are mapped to the contigs providing a contig coverage depth value. Upon the knowledge that species are present at differing abundances, the expectation is that genetic material derived from the same genome of any given species should have similar abundance and therefore similar coverage depth within a sample. Therefore, contigs with a similar coverage depth would then be predicted to be from the same species. To achieve this, bioinformatics programs are used to create a Sequence Alignment Map (SAM) file. This indicates the mapped location and quality of a query against a reference genome. Examples of mapping algorithms include Bowtie2 (Ben Langmead and Salzberg 2012), BBtools (Brian Bushnell, Rood, and Singer 2017) and Minimap2 (H. Li 2018). This has proved effective at MAG creation (Sharon et al. 2013; Albertsen et al. 2013). Combined with *k*-mer frequencies, GC-content and taxonomy, automated tools are used to assist in contig binning with MetaBAT2, (Kang et al. 2019) BinSanity (Graham, Heidelberg, and Tully 2017) and MaxBin (Wu et al. 2014) among others have now automated this process. Combined with human input, this may be the most effective method in MAG creation (Delmont et al. 2018).

There exists a multitude of different binning software, taxonomic classifications and other methods of binning, a combination of these algorithms along with human input may assist in MAG creation (Delmont et al. 2018). Therefore, the consensus of which taxonomic classifier and binning software group which contigs would increase the accuracy of determining the contigs genome identity. Anvi'o (Eren et al. 2015), a bioinformatics software, assists in the visualisation of metagenomes. Anvi'o uses a dendrogram to display the metagenome with hierarchical clustering were the closest related contigs are combined and grouped together. Layers are used to represent different statistics about the metagenome. Statistics like Guanine-Cytosine (GC) content are used as species generally have the same GC content. This provides additional support for binning confirmatory methods. Tracks represent different metagenomic coverages from coverage files. Additional layers on the anvi'o plot can be any additional label provided for each contig, for example, taxonomy or different binning software.

However, even with the advances in bioinformatics protocols, the clustering of high quality (Bowers et al. 2017) MAGs remain elusive (Tully, Graham, and Heidelberg 2018). Therefore, advances in automation of MAG creation, completeness metrics become important in judging how effective these tools are at producing whole genomes. Single-copy marker genes, defined as genes present only once within a genome, have been shown to be an effective metric and determine genome completeness and therefore quality (D. H. Parks et al. 2015; Rinke et al. 2013; B. J. Campbell et al. 2011). Should these genes be missing from a MAG, it can be said that the genome is incomplete as these are missing a gene that would be expected to be there. Alternatively, should duplicate copies of single-marker copy genes be present, this would indicate the presence of contamination. This is an important metric and defines the efficacy of MAG creation. However, these percentages are only estimations of genome completion as they rely on existing genome databases to determine single-copy marker genes. If a majority of organisms are unsequenced, determination of these sequences are only partially effective (**Table 2.3**).

Taxonomic profiling can be used as a proxy for estimating presence-absence data as well as microbial abundances within metagenomes. This process can be performed without metagenomic assembly, whereby external sequences consisting of the genetic information of microbes of interest are recruited against metagenomic reads. This process of mapping reference genomes to metagenomic reads is carried out without the need for assembly of metagenomic data, avoiding problematic metagenomic assembly problems and reducing computational resources. Additionally, this allows for the capture of rare taxa or organism without complete genomic coverage throughout a metagenome. However, this profiling method is difficult without external reference genomes.

The increasing popularity of single-cell amplified genome techniques (Lasken 2007; Rinke et al. 2013) and new understandings in the cultivation of hard to grow microbes (Carini et al. 2014; Stewart 2012), reference genomes diversity are now increasing rapidly. Success has been shown in low diversity microbiomes including the human gut microbiome (Human Microbiome Jumpstart Reference Strains Consortium et al. 2010). However, for diverse environments including soil and aquatic microbiomes, representative reference genomes are still lacking especially for rare taxa. This reduces the success of MAG creation by taxonomic profiling. This is further compounded where the mapping of reads to reference genomes can result in false positives, especially with many closely related species with similar genetic sequences.

1.6.5 Limitations and opportunities

There are still multiple limitations that metagenomics has to overcome in order to be used as an effective method for obtaining high quality and complete genomes from any environmental sample. MAG recovery has been shown to be more favourable with additional similar samples (Alneberg et al. 2014), which can become prohibitively expensive when combined with the need to increase sequencing depth to obtain rare taxa (Gagic et al. 2015). In addition, binning of assembled contigs is limited by the lack of reference genomes on which to base sequence differentiation (Sangwan, Xia, and Gilbert 2016). De novo binning algorithms rely on differences of genomic sequences which are often similar in complex highly diverse metagenomes, making binning computationally difficult (Ayling, Clark, and Leggett 2019). Metagenomic assemblers are also often computationally demanding, requiring specialist equipment for assembling high depth metagenomes (Nurk et al. 2017; Sangwan, Xia, and Gilbert 2016). More importantly, computational algorithms are currently unable to differentiate between closely related organisms, therefore MAGs are more likely to represent a population of closely related strains over an individual organism (Sangwan, Xia, and Gilbert 2016). It still has a relatively high cost as specialised sequence equipment is often prohibitively expensive and assembly requires high-end computational facilities. These costs are expected to reduce over time with decreasing sequencing costs, new cheaper sequencing platforms (Loman, Quick, and Simpson 2015) and the rise of cloud server computing. The lack of representative reference genomes, often consisting of model or easily cultivable organisms, limit the effectiveness of assigning taxonomic identity. This is widespread in diverse environments and shifts recovery of genomes towards these well-documented species (D. H. Parks et al. 2017). This also extends towards genetic-based studies as only well-characterised genes from model organisms have been experimentally deduced. Gene functional studies have low throughput in comparison to gene discovery through bioinformatic methods resulting in many uncharacterised genes. This makes the study of the transcriptome of a MAG difficult to determine, and hence its relationship with other organisms or roles within a population unknown. This is further exacerbated by the unknown viability of a cell as DNA from dead cells or environmental DNA still persist within an environment (Collins et al. 2018). If the goal is to study the active metatranscriptome of a community, it is difficult to pair metatranscriptomic data from living cells with metagenomics. This is because DNA from dead cells can be detected days to weeks after death within a marine system (Collins et al. 2018; Harrison, Sunday, and Rogers 2019; I. Salter 2018) making it difficult to draw conclusions on the origin and expression levels of each active gene due to the unknown mortality of metagenomic data.

The potential for metagenomes to study a population remains significant where multiple ‘omics studies can be combined together. Metagenomics can be complemented with RNA sequencing, metabolomics and metaproteomics assays to allow for detailed characterisation of a microbial population, without having to recreate such populations ex-situ. This extends to viromes which depend on not only their hosts to replicate in order to provide sufficient biomass for sequencing but also depend on their host’s nutrient and environmental requirements. Viral metagenomics sidesteps the culturing phase and allows for studies into the virome. Metagenomics is also well suited to time series studies where changes in microbial populations are critical in understanding questions from fields ranging from human health (Turnbaugh et al. 2006) to environmental conditions (Pieterse, de Jonge, and Berendsen 2016; Gupta, Rovira, and Roget 2011). These studies can provide multiple insights into population dynamics or biomarkers describing a particular condition which can be experimentally validated by *in vitro* assays. Lastly, such metagenomic data is publicly available (Biller et al. 2018; Pesant et al. 2015) providing a resource for other studies to compare against. Currently, new bioinformatic tools are being developed monthly with noticeable improvements. As a result, recovery of unknown taxa will be improved, enriching current databases with reference genomes to improve MAG recovery (Tully, Graham, and Heidelberg 2018). This has increased the recovery of microbial features allowing for strain-level comparisons of MAGs (Quince et al. 2016; M. Scholz et al. 2016; Truong et al. 2017).

Long read technology provides transformative DNA recovery improvements for metagenomics. Current metagenomic assemblies suffer from the fragmentation of genomes making MAG recovery more difficult. Long read technology with recorded read lengths of over 100kb will help reduce fragmentation due to repeat regions and scaffold across unresolvable de Bruijn graph paths. Although higher coverages are needed to provide read error correction due to high error rates, some success has been had with hybrid assemblies using highly accurate short reads to error correct error-prone long reads (Vaser et al. 2017; Walker et al. 2014). Lastly, with portability and reduced costs the Oxford Nanopore MinION and Flongle, costs are less of an obstacle and raise the tantalising possibility of in-field metagenomic sequencing.

Metagenomics has become an increasingly popular field in the study of environmental communities since its first applications (Tyson et al. 2004; Venter et al. 2004). Its increasing popularity has allowed for the analysis of more complex datasets leading to insights into human health and environmental communities. Although limitations still exist to the general scientist with bioinformatics tools requiring training to use along with high costs to sequencing diverse metagenomes at high depths to obtain rare taxa, its potential coupled with other ‘omics fields would provide unparalleled and unique insights into the function of a microbial population.

1.7 Single-cell amplified genomes

The introduction of metagenomic studies in the mid-2000s provided extensive amounts of genetic information, expanding databases and our understanding of microbial populations (Venter et al. 2004; Tringe et al. 2005; Rusch et al. 2007). However, issues with reconstructing complete genomes from metagenomes proved problematic. Although some success in MAG recovery has been shown in human gut microbiomes (Sharon et al. 2013), technical challenges still exist in the recovery of complete genomes at a strain level from diverse and complex metagenomes. MAG recovery from marine (Tully, Graham, and Heidelberg 2018) and soil microbiomes (Wilkins et al. 2020) still remain inefficient. This complicates studies into the individual dynamics of microbes such as metabolomics, transcriptomics and evolutionary events within a population. Here, Single Cell Genomics (SCG) provides a culture-independent alternative to obtain genetic information of individual cells from an environment (Marcy, Ouverney, et al. 2007). This is an advantage in the analysis of hard to culture organisms. This allows for the generation of high-resolution genomic data of an individual from an environment or for the targeting of specific taxa, which would be challenging with metagenomic studies. However, SCG relies on the technically difficult process of isolating and sequence a DNA molecule from a single isolated cell. The process of SCG can be divided into three steps: (i) isolation of individual cells from its environment without contamination; (ii) amplification of the isolated cells DNA into quantities required for DNA sequencing; and lastly (iii) assembly of Single-cell Amplified Genomes (SAGs) (Eberwine et al. 2014).

Individual cells first need to be isolated from their environment to allow for pure sequencing of the target organism. This allows for the generation of sequencing data free of contamination and, in principle, similar to the sequencing of axenic cultured organisms (Woyke et al. 2010; K. Zhang et al. 2006). In the case of cells attached to physical structures like tissue or filters, mechanical or chemical processes are required for dissociation. Once target cells are in suspension, a variety of isolation methods are available.

Manual methods for isolation of single cells include various dilution methods including serial dilution (Staszewski 1984; Fuller, Takahashi, and Hurrell 2001), micro pipetting or optical tweezers (Landry et al. 2013). However, most popular approaches usually require some sort of automation for sorting high quantities of single cells. These include the well documented fluorescence-activated cell sorting (FACS) (Basu et al. 2010; Rinke et al. 2014; Shapiro 2005; Navin et al. 2011) based on fluorescent cell labelling and distribution of samples into individual containment vessels. Microfluidics is another cell sorting mechanism (A. K. White et al. 2011; Leung et al. 2012; Macosko et al. 2015; Marcy, Ouverney, et al. 2007) involving manipulation of samples within microdroplets for containment. Although microfluidics lack tools for cell differentiation in comparison to FACS, opportunities such as individual cell experimentation and lab-on-a-chip remain possible. Lastly, microscopy is often performed for confirmation of the physical isolation of single cells within capture devices along with 16S/18S rRNA sequencing for taxonomic identity. Subsequently, cell lysis is often required to access genomic DNA within isolated single cells. Lysis needs to remain consistently effective against a wide range of taxa without damaging DNA or leaving chemicals that may impede downstream analysis. Therefore, chemical cell lysis is usually employed by alkaline solutions (Lasken 2007) or hydrolytic enzymes (Swan et al. 2011; Marcy, Ouverney, et al. 2007).

An individual bacterial cell only contains femtograms worth of DNA and is generally regarded as too little for current DNA sequencing processes. Therefore, Whole Genome Amplification (WGA) of genomic DNA is required while minimising artefact introduction, amplification bias and chimaeras. Initially, attempts at WGA used PCR-based amplification of recurring sequences within a genome (Lichter et al. 1990) or random priming (Telenius et al. 1992; L. Zhang et al. 1992) by degenerate oligonucleotide-primed PCR (DOP-PCR). However, this resulted in the random amplification of fragments of DNA and the loss of large proportions of unamplified DNA sequences producing low coverage. Later, isothermal methods of WGA were introduced, the most common being the Multiple Displacement Amplification (MDA) method. MDA uses random hexamer primers and DNA polymerase Φ29 through rolling-circle amplification (Dean et al. 2001; D. Y. Zhang et al. 2001). This had a higher coverage and lower error rate than previous polymerases due to Φ29 polymerase’s high fidelity (de Bourcy et al. 2014). After cell lysis, random hexamer primers consisting of sequences of six random nucleotides bind to their complementary location on the template DNA. These primers allow for Φ29 polymerase to bind and begin DNA synthesis with the template strand as a reference. When Φ29 polymerase reaches another starting site, it displaces the newly sequenced DNA strand and continues synthesis, creating branched structures normally 12 to 100 kb in length. Additional primers can anneal to this branched structure and DNA synthesis of these branched structures is repeated. This results in exponential amplification of the loci that are amplified first, resulting in uneven coverage of the template DNA (de Bourcy et al. 2014). Additionally, chimaeras can be created through mispriming events (Lasken and Stockwell 2007; Marcy, Ishoey, et al. 2007) creating artefact sequences. Chimaera formation can be reduced by endonuclease cleaving of branching structures (K. Zhang et al. 2006).

To counter low and uneven coverage of MDA reactions, hybrid methods seek to limit isothermal amplification. Displacement DOP-PCR or PicoPLEX relies on the formation of hairpin loops of amplified structures to prevent ensuing primer binding and amplification. This prevents exponential amplification and therefore amplification bias (Langmore 2002). MALBAC uses a similar approach using complementary primers on amplified strands that results in a loop, preventing further amplification cycles (Chapman et al. 2015; Gawad, Koh, and Quake 2016). Regardless, both hybrid and MDA methods are both effective at WGA and, when compared MDA, had a higher genome coverage vs hybrid methods (MDA 84%, MALBAC 72% DOP-PCR 39%) (Huang et al. 2015) but produced less uniform genome coverage and fewer chimaeras as expected (Hou et al. 2015). Background contamination is also cited as an issue (de Bourcy et al. 2014) as it is amplified along with target samples. Here, microfluidic methods were shown to reduce contamination (Blainey and Quake 2011) in MDA reactions as well as increase uniformity of coverage due to the usage of nanolitre over microlitre reactions (Marcy, Ishoey, et al. 2007). This indicates the potential for high throughput and automation of single-cell sorting and WGA (Fu et al. 2015).

Multiple sequencing technologies are available for DNA sequencing after WGA, applicable for both long and short-read technologies (Loman et al. 2012). DNA prepared though WGA is able to undergo traditional sequencing methods with the most common approach being pair-end Illumina reads (Chitsaz et al. 2011).

Subsequent assembly is similar to normal de Bruijn graph assembly with the exception of being performed by specialised single-cell assemblers designed to deal with uneven coverage (Nurk et al. 2013; Peng et al. 2012; Chitsaz et al. 2011). The percentage completeness of genomes from SAGs derived from single-copy marker genes varies widely from zero to complete due to the random nature of hexamer binding. Within this study, I find from MDA reactions of 451 SAR11 SAGs performed by Bigelow Laboratory's Single Cell Genomics Center to be on average complete 66% with completeness values ranging from 91 to 4% with little to no contamination (0 - 4%) (**Table 3.2**).

The advantage of SCGs are their ability to obtain genetic information without culturing at high completeness and low contamination rates. This allows for the study of hard to culture organisms *in situ*, which had previously been intractable. Along with high throughput methods, SAGs allow for the study of large numbers of organisms at a high resolution, providing information on the structure and dynamics of microbial populations. However, SAGs require cell dispersion within a solution for single-cell sorting when attached to a physical structure (Clingenpeel, Clum, and Schwientek 2015). In addition, both single-cell sorting and WGA requires training on extensive specialist equipment (Stepanauskas 2012). With femtograms of starting DNA, amplification is required which is far from perfect, resulting in highly variable genomic coverage (de Bourcy et al. 2014) and chimaeras (Lasken and Stockwell 2007; Marcy, Ishoey, et al. 2007). More advancement for uniform, accurate and consistent genome amplification is required to provide high-quality genomes. The emerging fields of individual viral sequencing (Allen et al. 2011) that are difficult to assemble from short-read metagenomic data (Tadmor et al. 2011) has led to the discovery of new phage taxa through the unculturable SUP05 bacteria (Roux et al. 2014). In addition, SCGs coupled with single-cell transcriptomics (Shintaku et al. 2014; Macaulay et al. 2015; Dey et al. 2015) and metabolomics (Rubakhin, Lanni, and Sweedler 2013; Heinemann and Zenobi 2011; Ståhlberg et al. 2012) provides a promising and complete outlook of microorganisms metabolism.

1.7 SAGs and MAGs compared

Both SAGs and MAGs have greatly contributed to our understanding of microbial evolution, phenotype and physiology (Hugerth et al. 2015; Swan et al. 2011; Rinke et al. 2013; Zaremba-Niedzwiedzka et al. 2017; Spang et al. 2015), however, both are limited in their ability to produce complete genomes for downstream analysis. This is exacerbated further with both SAGs and MAGs needing expensive sequencing platforms and reagents to convert biological sequences to an analysable form. Although SAGs and MAGs seek to produce the same result - high quality assembled genomes, their biological methods of isolation and sampling are fundamentally different and should be treated differently. This is summarised by their main differences being that SAGs are derived from DNA within a cell and are independent of the complexity of a microbial community allowing for strain-level comparisons. Metagenomes are the collection of the genomic content of a population and are effective at describing population dynamics over individual stain level divergences. Both methods can be explored to obtain genomic data (Alneberg et al. 2018), but studies have shown how each compliments each other to produce a more in-depth understanding of microbial populations. Multiple studies have been effective at combining both approaches to directly aid both SAG (Mende et al. 2016) and MAG (Becraft et al. 2016) genome recovery rates, but the main advantage of SCG is the ability to assist in reference sequences binning of rare or undocumented taxa within metagenomic samples (Tringe et al. 2005). This allows for only previously discovered phyla through 16S/18S amplicon sequencing to gain full genomic assemblies furthering understanding of microbial ecosystems.

Additional benefits include the study of organism biogeography, distribution and abundance within metagenomic data. SCG provides reference genomes which can be recruited against metagenomic reads. Successful recruitment provides assembly free presence-absence data of SCG in metagenomic data, allowing rare or uncultured taxa biogeography to be studied from environmental data. This has been used to discern the biogeographic distribution of an uncultured

Flavobacteria in marine environments (Woyke et al. 2009). This was combined with other informatics tools to discover a novel chemotrophic pathway in deep ocean bacteria (Swan et al. 2011) and expanded along with metaproteomics to identify marine bacterioplankton responsible for the degradation of hydrocarbons after the Deepwater Horizon oil spill (Mason et al. 2012). These strengths allow for both methods to complement each other, with metagenomics providing environmental data and SAGs used to study and quantify rare and uncultured microorganisms within the context of an environment.

1.8 Project Outline

Although SAR11 and its phages are now well studied with multiple genomic assemblies in public databases, only a handful of assemblies are fully complete. This is mainly due to the difficulty and time needed to culture SAR11 and therefore its phages, resulting in most assemblies existing as MAGs (NCBI, 2019; JGI, 2019). As a result, it is possible multiple ecotypes may remain undiscovered. In addition, clade-wide studies are difficult to perform without quality assemblies. Prediction of the host would be reliant on finding phages infecting their host at the time of capture or integrated prophages. Some bioinformatic pipelines exist to predict hosts based on viral sequences (Ahlgren et al. 2016; Galiez et al. 2017) but are still far from perfect accuracy (Galiez et al. 2017). SAR11 phages are one of the most ubiquitous in the marine habitat (Y. Zhao et al. 2013), but complete full-length genomes from isolates remain uncommon compared to their ubiquity.

Currently, there are difficulties in obtaining SAR11 genomes and SAR11 bacteriophages, the usage of metagenomes and SAGs provides a unique opportunity to obtain genetic sequences related to these two ubiquitous organisms. Therefore the goals of this project are to: (1) Identify novel SAR11 ecotypes, particularly at depth and in the Arctic, where sampling has traditionally been poor; (2) To evaluate the use of SAGs to identify novel pelagiphages in order to determine infection rates and host-specificity within the SAR11 clade; (3) To evaluate the function encoded within HVRs function and their degree of conservation within and between SAR11 clades. Such findings would improve our knowledge of the biography of the SAR11 clade, highlight novel genetic features and suggest mechanisms which allow for SAR11's ubiquity alongside its predatory phages.

2 Marine Metagenomics

2.1 Introduction

2.1.1 Abstract

Analysis of prokaryotic genome sequences can provide valuable insights into the phenotypic and functional characteristics of a microbial population, furthering our understanding of microbial life. However, hard to culture organisms remain elusive to culture ex-situ due to unknown nutrient and environmental requirements. Metagenomics provides an alternative source of genetic material; studying the sequenced DNA of a culture-independent environmental sample. This allows for the analysis of hard to culture microorganisms such as the SAR11 clade, one of the most ubiquitous marine bacteria. Although highly abundant, relatively few SAR11 genomes are available within public databases. Therefore, extraction of Metagenome Assembled Genomes (MAGs) provides an alternative source of genetic material for the study of SAR11 and its phages. Microbial populations are known to exhibit "bloom and bust" relationships with its decline explained by negative density-dependent selection by viruses. Alternatively, host microorganisms may be undergoing viral infections, where viruses are unable to complete their lifecycle and replicate due to the lack of required nutrients. Diel cycle metagenomic datasets are ideal for answering these biological questions in combination with providing additional information on microorganism diversity, geographical ranges and its ecological relationship with its viral predators over time. However, bioinformatic methods still face challenges in obtaining complete genomes from fragmented metagenomic data. Since the inception of metagenomics, multiple bioinformatic methods have been applied to improve MAG recovery. Within this study, a variety of state of the art algorithms were explored to recover prokaryotic MAGs with a particular focus towards SAR11. I show that current bioinformatic MAG binning algorithms are ineffective at extracting high-quality draft genomes from marine metagenomic samples, particularly those associated with SAR11. I suggest improvements in the

sequencing method to aid MAG recovery as well as critically analyse existing binning algorithms. I hope that this study can be used to inform future studies on the minimum sequencing depth required to obtain MAGs from highly diverse metagenomes as well as advocate for the usage of long reads metagenomics and its potential to increase MAG recovery.

2.1.2 Where do the marine metagenomes come from?

The Bermuda Atlantic Time-series Study (BATS) site in the Sargasso Sea has been collecting data about physical, biological and chemical ocean properties at monthly intervals since 1988 (Bates, Michaels, and Knap 1996). BATS is a long-term site situated in a subtropical gyre resulting in it being an ultra-low nutrient zone in a seasonal oligotrophic system. The maximum depth of this region is 4000m, with the water column divided into two main layers of study. During calmer summer months, the upper euphotic zone (UEZ) is between 0 and 120m and the upper mesopelagic (UMP) between 120 to 300m (Stephen J. Giovannoni and Vergin 2012). Spring blooms occur after winter mixing events where the average mixed layer depth is at its deepest of 260m. Mixing allows the upper euphotic and upper mesopelagic zones to homogenise, exporting refractory Dissolved Organic Matter (DOM) to deeper waters, increasing nutrient concentrations in surface water and promoting eukaryotic picophytoplankton blooms (Stephen J. Giovannoni and Vergin 2012; Vergin et al. 2013). As surface water is heated throughout the summer months, the water column stratifies (Michaels et al. 1994; Doney, Glover, and Najjar 1996) and is dominated by cyanobacteria (Vergin et al. 2013).

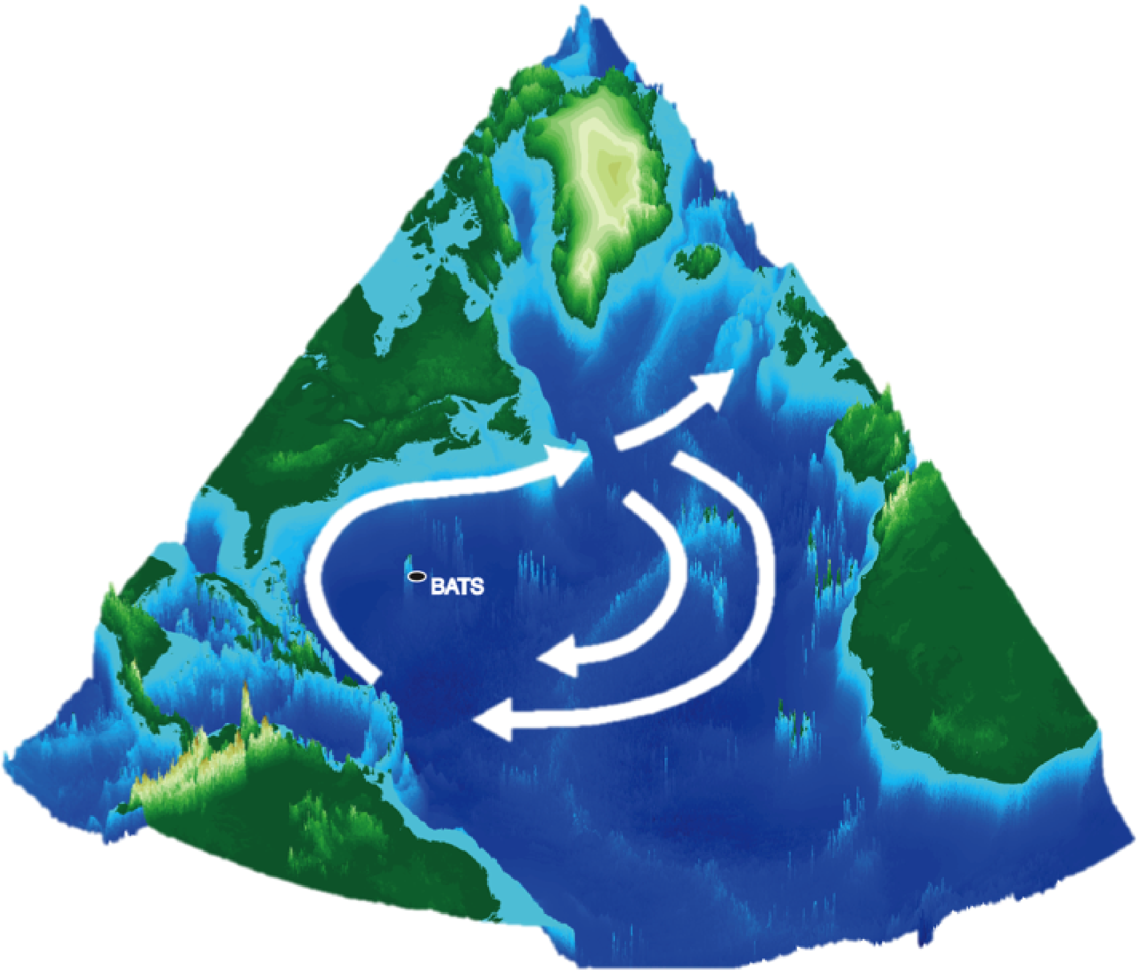


Figure 2.1 Location of BATS on the globe with subtropical gyre displayed in white arrows (Bermuda Institute of Ocean Sciences | About)

2.1.3 Metagenomic time series

Microbial communities are key players in marine biogeochemistry, with culture-independent methods from environmental samples allowing us to study its complex biological interactions and diversity (Pesant et al. 2015; Biller et al. 2018). Despite the availability of hundreds of marine metagenomic datasets and Whole Genome Sequences (WGS), gaps still remain in the understanding of microorganism diversity, distribution and biological factors that structure a community over time and space.

SAR11 are the dominant heterotrophs within oligotrophic marine systems and therefore key to understanding the cycling of nutrients within these types of systems. Time series data from 2003-05 (Craig A. Carlson et al. 2009) were successful in describing annual population fluctuations of the SAR11 clade within euphotic (0-120m) and upper mesopelagic zones (160-300m). The cause of these shifts in population abundances is unknown but is probably related to genetic adaptations to changing nutrient gradients. However, viruses are also important regulators in microorganism mortality, lysing 20-40% of marine microorganisms (Curtis A. Suttle 2007a; Proctor and Fuhrman 1990) and metabolically reprogramming them through horizontal transfer of genetic material. Therefore, the study of any marine microbial community would be incomplete without a description of the impact of its viral population. Biological samples taken from varying depths in a metagenomic time series like BATS can help us understand how viral abundances correlate to abiotic and/or biotic elements. Previous studies have been successful in identifying biological abundance and classification. For example, during summer and autumn stratification, a virioplankton maximum develops between 60 and 100m and is lost during winter mixing events (Parsons et al. 2012), resulting in a study by (Goldsmith et al. 2015) to suggest the viral community composition exhibits a winter and summer state. However, such studies may only capture a subset of the community responsive to the chosen species. Instead, virome studies would be more impactful to study a wide range of populations (Sullivan 2015). This is important in understanding the impact of viruses on existing ecological niches and its temporal and spatial variability in context with other physical and biological parameters.

2.1.4 Diel Cycles

Diel cycles are an important component of short-term dynamics in viral populations (Winter et al. 2004). (Aylward et al. 2017) suggested an increase in cyanophage transcription coinciding with host *Prochlorococcus* replication during afternoon hours indicated coupling of host and viral reproduction. (Yoshida et al. 2018) also observed increased cyanophage gene expression in afternoon and evening times. This was followed by increases in viral infections of heterotrophs. This may suggest cyanophages drive the diel release of organic matter into the environment. I hypothesised that these are then taken up by heterotrophic microorganisms, providing nutrients to complete viral replication of already-infected cells within the heterotrophic community. Increases in viral abundance would reflect increases in nutrient uptake and host replication rates rather than increased contact rates. An identical sustained and constant ratio of phage and host abundance would indicate this coupling of viral and host replication. This would further be reinforced by large numbers of phages within hosts awaiting desired nutritional conditions. Additionally, analysis of differences in the abundance of lytic and lysogenic phages may indicate ecological strategies of phages at different time points of a diel cycle. Time series data would allow for the identification of phages populations within ecological niches, indicating if different populations of phages occupy different locations based on physical and temporal factors.

2.2 Materials and Methods

2.2.1 Metagenomic sampling

Samples were obtained from BATS during a research cruise in July 2017. Cellular and viral metagenomes were prepared from 40L of seawater collected by Niskin at ~06:00 and ~19:00 each day over three days from two depths: 80m and 200m. 80m was chosen to obtain samples from the virioplankton maximum zone: between 60m and 100m during Summer months (Parsons et al. 2012). 200m was picked as this was outside of the euphotic zone. If a relationship exists between heterotrophs and phototrophs due to interactions within the euphotic zone, this trend should not be replicated in 200m samples outside of the phototroph ecological niche. It is acknowledged that vertical migratory zooplankton may be present due to water currents at both depths and act as active transport for carbon export to depth (J. Sun et al. 2011).

Preparation of marine metagenomic samples and sequencing were not performed in this study but briefly, for each sample, 40L of seawater was obtained at 80m and 200m by Niskin and filtered using a 0.22 µm Sterivex filter to separate the cellular fraction from the viral fraction. The cellular fraction was released from the filters and DNA isolated using a phenol-chloroform DNA purification protocol. DNA was prepared and sequenced via short-read Illumina technology by Nextera library preparation kit, creating 2 x 250 bp paired-end read lengths.

2.2.2 Read preprocessing

Resulting reads were checked for quality using a pipeline developed by JGI using BBTools (B. Bushnell 2019) - a suite of bioinformatics tools- for preprocessing of Illumina reads before assembly. Error correction and read normalisation were skipped, as both steps are included within the SPAdes (Bankevich et al. 2012) algorithm. The SPAdes assembly pipeline is designed to take error-corrected reads using the BayesHammer algorithm (Nikolenko, Korobeynikov, and Alekseyev 2013). Additionally, read normalisation in more recent versions of SPAdes (3.10 and up) use differential coverage to resolve ambiguities (Nurk et al. 2017) and would impact the assembly adversely.

```
# Sort reads by k-mer frequency to decrease computational
resources and remove duplicates
clumpify.sh in=<reads> out=<reads_clumped> dedupe optical

# Filters reads based on k-mer frequency and quality score
based on location within an Illumina flow cell
filterbytile.sh                                in=<reads_clumped>
out=<reads_filtered_by_tile>

# Trim adaptors based on matching k-mer frequency
bbduk.sh in=<reads_filtered_by_tile> out=<reads_trimmed> \
ktrim=r k=23 mink=11 hdist=1 tbo tpe minlen=70 ref=adapters
ftm=5 ordered

# Trim synthetic spike-ins based on matching k-mers frequency
bbduk.sh in=<reads_trimmed> out=<reads_filtered> k=31 \
ref=artifacts,phix ordered cardinality
```

2.2.4 Metagenomic assembly and quality assessment

After quality control, reads were assembled using metaSPAdes v3.13.0 (Bankevich et al. 2012) with the parameters outlined below. SPAdes was used as it has been shown to be an effective assembler for short read metagenomics (Vollmers, Wiegand, and Kaster 2017; Forouzan et al. 2018; Sutton et al. 2019). Basic statistics (Table 2.1) were performed on each assembly to ascertain the quality using BBTools.

```
#Assemble metagenome with SPAdes
spades.py  --meta  --phred-offset  33  -k  25,55,95,125  --threads
16 \
--pe1-1 <metagenome_fwd_reads> \
--pe1-2 <metagenome_rev_reads> \
-o <metaSPAdes_output>
```

2.2.3 Taxonomic relative abundance

To confirm that our metagenomic samples contained marine organisms including SAR11, reads were compared against the non-redundant database from the NCBI containing unique protein-coding regions of genomes. Kaiju (Menzel, Ng, and Krogh 2015) provided a taxonomic classification of reads and relative abundance. SAR11 is an order level clade, therefore relative abundance was produced at an order taxonomic level to confirm its presence within metagenomic data.

```
# Run kaiju
kaiju -t <nodes.dmp> -f <kaiju_db_nr.fmi> -i \
<QC_metagenomic_reads> -o <output_file> -z 16

# Get relative abundance
kaiju2table -t <nodes.dmp> -n <names.dmp> -o <output> \
-r order <kaiju_output_file>
```

2.2.5 Binning by Coverage

To create coverage files for MAG creation, bowtie2 (Ben Langmead and Salzberg 2012) was used to recruit reads to metagenomic assemblies from all samples taken from the same depth (80m and 200m). This would allow for the extraction of MAGs from each metagenome. This resulted in six SAM files for each sample. SAM files were then filtered for reads with a 95% identity minimum cutoff using BamM (Imelfort and Lamberton 2015). A 95% identity was chosen as this is the general boundary used to delineate species (Konstantinidis and Tiedje 2005; Richter and Rosselló-Móra 2009).

```
# Create index
bowtie2-build <metagenome.fna> <metagenome_name> --threads 16
\
2>&1 | tee <log_file>

# Map reads to metagenomic assembly
for i in <all_reads>; do
bowtie2 -x <metagenome_name> --no-unal  --threads 16 \
-1 <QC_forward_reads> \
-2<QC_reverse_reads> \
-S <mapping_file> \
 2>&1 | tee -a <log_file>;
done

# Sort and index the bam files to order them
for i in <mapping_files>; do
samtools view -F 4 -@ 16 -buSh <mapping_file> | samtools sort
-@ 16 - -o <sorted_mapping_files> 2>&1 | tee -a <log_file>;
samtools index -@ 16 <sorted_mapping_files> 2>&1 | tee -a
<log_file>;
done

# Filter and remove mappings below >95% identity
for i in <sorted_mapping_files>; do
bamm filter -b <sorted_mapping_files> --percentage_id 0.95 \
2>&1 | tee -a <log_file>;
done
```

2.2.6 Binning contigs into MAGs

Relative abundance of contigs for each sample and *k*-mer composition was used to 'bin' contigs into putative MAGs. Multiple binning algorithms were used to create a consensus. Here MetaBAT2 (Kang et al. 2019), BinSanity (Graham, Heidelberg, and Tully 2017) and MaxBin (Wu et al. 2014) were used.

```
# BinSanity
get-ids -f <working_dir> -l <metagenome.fna> -o <ids.txt> -x 1

Binsanity-profile -i <metagenome.fna> -s <working_dir> \
 --ids <ids.txt> -c <coverage> -T 16

Binsanity-wf -f <working_dir> -l <metagenome.fna> \
 -c <coverage.cov.x100.lognorm> --threads 16 -o <output_dir>

#MetaBAT2 using recommended settings from manual
jgi_summarize_bam_contig_depths \
--outputDepth <depth.txt> <sorted_mapping_files>

metabat -i <metagenome.fna> -a <depth.txt> -o <output_dir> \
--maxP 90 --maxEdges 500 --minS 75 --noAdd --numThreads 16

#Maxbin
for i in <sorted_mapping_files>; do
pileup.sh in=<sorted_mapping_files>
out=<sorted_mapping_files_pileup>;
cut -f1,2 <sorted_mapping_files_pileup> >
<maxbin_coverage_files>;
done

run_MaxBin.pl -thread 16 -contig <metagenome.fna> -out
<output_dir> \
-abund <maxbin_coverage_files_1> \
-abund2 <maxbin_coverage_files_2> \
-abund3 <maxbin_coverage_files_3> \
-abund4<maxbin_coverage_files_4> \
-abund5 <maxbin_coverage_files_5> \
-abund6 <maxbin_coverage_files_6>
```

Kaiju (Menzel, Ng, and Krogh 2016), a taxonomic classifier, was used to assign identities to contigs. Other taxonomic classifiers Centrifuge (Kim et al. 2016) and CAT/BAT (von Meijenfeldt et al. 2019) were evaluated but not used due to the anvi'o's limitation of only one taxonomic classifier usage. Kaiju was used as it provided the most taxonomic hits. Script for this process can be found below.

```
# Setup and run Kaiju
KAIJUDB=$(kaiju_db_dir)

kaiju -t $KAIJUDB/nodes.dmp \
-f $KAIJUDB/kaiju_db_nr.fmi \
-i <metagenome.fna> \
-o <kaiju_metagenome.out> \
-z 16 \
-v 2>&1 | tee -a <kaiju_metagenome.db>

addTaxonNames \
-t $KAIJUDB/nodes.dmp \
-n $KAIJUDB/names.dmp \
-i <kaiju_metagenome.out> \
-o <kaiju_metagenome.names> \
-r superkingdom,phylum,class,order,family,genus,species

anvi-import-taxonomy-for-genes
-i <kaiju_metagenome.names> -c contigs.db -p kaiju
--just-do-it
```

Following *k*-mer, coverage and taxonomic binning, contigs and their associated bin labels were visualised in anvi'o (Eren et al. 2015), a genomic visualisation tool was used to visualise the metagenome. Metagenomes are represented as a Circos plot with a dendrogram of contigs in the middle. Script from assembly to visualisation can be found here. MAGs were evaluated for completeness using CheckM (D. H. Parks et al. 2015), a software program that calculates genomes completeness by single-copy marker genes using this script.

```
#Run CheckM
checkm lineage_wf <MAG_dir> <output_dir> -x <ext_type> \
--tab_table --file <output_file> -t 16 --pplacer_threads 16
```

2.2.7 Binning by *k*-mer counting

Following lack of consensus binning and any SAR11 genomes produced from specialised binning software (**Fig 2.4-6**), a different approach using strictly *k*-mer counting was implemented to improve genome recovery. Samples were k-mer counted with a k-value of 5 (Default in *k*-mer counting script by (Beaulaurier et al. 2019)).

2.2.8 Exploration of UMAP parameters

UMAP (McInnes, Healy, and Melville 2018) is a dimension reduction algorithm that uses BH-tSNE plots to reduce the 512 *k*-mer dimensions to two dimensions visualised on a scatter plot. UMAP parameter space was explored for optimised clustering. The number of neighbours (n_neighbours) was explored to optimise MAG recovery. As n_neighbours determined data plotting, this directly impacted cluster formation. This is the number of neighbouring points UMAP compares a datapoint to, starting from the closest Euclidean distance. A lower n_neighbours directed UMAP to examine more localised differences within the dataset, comparing how similar two points are. This is beneficial for small local changes but performs poorly when looking at the global picture in how a point was related to a wider variety of data points. In summary, a larger n_neighbours revealed the global structure and a smaller n_neighbours quantified how different two points are from each other.

2.2.9 Visualisation and binning of *k*-mer counted contigs

Visualisation of UMAP plots from *k*-mer counted reads are displayed using scripts developed by (Beaulaurier et al. 2019), here repurposed for contigs. *K*-mer counts were normalised against contig size as differing contig sizes would influence the number of *k*-mers, as there was a large distribution of contig sizes from 2500 base-pairs to almost 200 kb. This was done by dividing the total number of *k*-mers over the contig size. Normalised *k*-mer counts were then visualised with UMAP. Lastly, HDBSCAN (McInnes, Healy, and Astels 2017) was used to automate cluster determination. Clusters determined by HBDSCAN were exported as MAGs using seqtk (H. Li 2012). MAG quality was assessed with CheckM.

```
# Assign variables
MG=<metagenome.fna>
PREFIX=<file_prefix>
CUTOFF=<min_contig_length>
THREADS=<num_threads>
KMERS=<k-mer_size>
NEIGHBORS=<n_neighbours>
CLSTR_SIZE=<min_cluster_size>

# Clustering
#make dir to organise saved results
mkdir ${PREFIX}_bin; cd ${PREFIX}_bin

# Run k-mer frequencies counting
kmer_freq.py $MG -t $THREADS -k $KMERS >
${PREFIX}_kmer_freq.out
run_umap.py ${PREFIX}_kmer_freq.out -p $PREFIX -l $CUTOFF -n
$NEIGHBORS
run_hdbscan.py ${PREFIX}.umap.tsv -p $PREFIX -c $CLSTR_SIZE
plot_kmer_bins.py -p $PREFIX ${PREFIX}.hdbscan.tsv

# Get contig names assigned to bins
for i in $(seq $(cut -f5 ${PREFIX}.hdbscan.tsv | grep [0-9] |
sort -nur | tail -n 1) \
$(cut -f5 ${PREFIX}.hdbscan.tsv | grep [0-9] | sort -nur |
head -n 1)); do \
cut -f1,5 ${PREFIX}.hdbscan.tsv | grep -P "\t${i}$" | cut -f1
> bin_${i}; \
done

# Create MAGs from contig bin identities
for i in bin_*; do seqtk subseq $MG $i >> ${i}.fna; done

# QC MAG completeness
checkm lineage_wf ./ checkm_output -x fna --tab_table --file
checkm_output.tsv -t $THREADS --pplacer_threads $THREADS
```

2.2.10 Binning by reference genomes

Binning by reference is an alternative method for potentially recovering high-quality SAR11 genomes. This has been the basis of MAG recovery in low complexity microbiomes (Sharon et al. 2013) and the algorithm for determining metagenome quality in metaQUAST (Mikheenko, Saveliev, and Gurevich 2016). Using access to 451 SAR11 SAGs and nearly 300 publically available SAR11 genomes covering the three main SAR11 clades (I-III), it becomes possible to create a clade size pangenome for recovery of SAR11 like sequences. Here, over 730 SAR11 genomes consisting of SAGs from this study along with publicly available SAR11 genomes consisting of SAGs, MAGs and isolates were concatenated together to form the known SAR11 pangenome. Reads from a metagenome can be mapped against it, with successfully recruited reads described as genetically “SAR11 like”. This process was performed using bbmap.sh.

```
# Recruit SAR11-like reads from metagenomes with bbmap.sh
bbmap.sh ref=<all_SAR11.fasta> \
in=<metagenome_fwd_reads> in2=<meteagenome_rev_reads> \
outm=<SAR11_mapped_fwd_reads> outm2=<SAR11_mapped_rev_reads> \
threads=16 2>&1 | tee <log_file>
```

Reads were assembled using a single-cell assembler described as a “mini metagenome” assembler (Nurk et al. 2013), where assembled contigs would hopefully result in a SAR11 metagenome. A single-cell assembler was chosen over a normal assembler to compensate for a highly uneven coverage expected within metagenomes in general. As this “mini-metagenome” can be assumed to contain only one type of taxa, it has similar characteristics to that of a single cell assembly and therefore a single cell assembler was used instead of a metagenomic assembler.

```
# Assemble SAR11-like reads into contigs
spades.py --sc --careful --threads 16 \
--phred-offset 33 -k 25,55,95,125 \
--pe1-1<SAR11_mapped_fwd_reads> \
--pe1-2 <SAR11_mapped_rev_reads> \
-o <SAR11_SPAdes_output_dir> 2>&1 | tee -a <log_file>
```

It is still likely that contamination from non-SAR11 species is included within the metagenome due to highly similar and conserved sequences (for example 16S sequences). Therefore, binning into MAGs is required. Binning by binning algorithm was performed by BinSanity, as this produced the most high-quality MAGs (**Fig 2.4**).

```
# Mapping metagenome back to reads for coverage
bowtie2-build <metagenome.fna> <metagenome_name> --threads 16
for i in <metagenome_reads>; do \
bowtie2 -x <metagenome_name> --no-unal  --threads 16 \
-1 <metagenome_forward_reads> \
-2<metagenome_reverse_reads> \
-S <mapping_file> \
 2>&1 | tee -a <log_file>;
done

#sort and index the bam files to order them
for i in <mapping_file>; do \
samtools view -F 4 -@ 16 -buSh <mapping_file> | \
samtools sort -@ 16 - -o <sorted_mapping_files> 2>&1 | tee -a
<log_file>;
samtools index -@ 16 <sorted_mapping_files> 2>&1 | tee -a
<log_file>;
done

# Run BinSanity
Binsanity-profile -i <metagenome.fna> -s <coverage_files_dir>\
-c <output_file> -T 16

Binsanity-wf -f <working_dir> -l <metagenome.fna> \
-c  <coverage_file.x100.lognorm> -o <output_dir> --threads 16
```

K-mer counting methods established previously were also performed to compare binning methods. To evaluate if the creation of SAR11 MAGs was successful and to establish completion and redundancy values, scatter plot of bin completion and redundancy was plotted in python using matplotlib (Hunter

2007). Datapoint sizes are in relation to MAG genome size and colour representative of taxonomic identification. The graphical script can be seen here.

2.3 Results

2.3.1 Assembly quality

Metagenomic SPAdes assemblies were assessed for quality using several metrics: Number of Reads after QC, Number of Scaffolds, Scaffold N50, Metagenome Size, percentage read recruitment to assembly and average scaffold coverage. All metagenomes assembled were highly fragmented, with several thousand scaffolds produced. Generally, <40% of reads mapped back to assemblies indicating a high amount of genetic material is not present in assemblies. This may suggest that a majority of individuals are lost resulting in a possible underrepresentation of genetically unique organisms. A low median scaffold coverage also indicates rare taxa were unlikely to be represented. Overall, 200m assemblies were of better quality compared to 80m assemblies based on percentage read recruitment and scaffold coverage, most likely as a result of increased sequencing depth. This may be due to reduced diversity at lower depths due to reduced variation of environmental conditions (Costello and Chaudhary 2017).

S/N	dd-mm-yy hh:mm	Depth (m)	Number of Reads after QC	Number of scaffolds	Scaffold N50	Metagenome Size (bp)	% reads recruitment to assembly	Median scaffold coverage
1	08-07-17 18:30	80	2,700,949	21,815	3,555	32,138,697	29.3	16.1
3	09-07-17 06:00	80	10,578,979	92,869	13,928	116,337,238	33.0	17.5
5	09-07-17 19:00	80	1,016,660	7,686	1,257	12,098,713	21.8	18.5
7	10-07-17 06:00	80	92,906	324	9	372,949	10.2	19.6
9	10-07-17 19:00	80	5,452,977	44,866	7,090	57,824,806	34.0	16.7
11	11-07-17 06:00	80	2,583,040	7,336	1,914	8,834,212	15.3	11.3
2	08-07-17 18:30	200	40,344,052	282,666	44,223	351,706,424	38.5	7.1
4	09-07-17 06:00	200	38,589,724	248,807	40,109	272,855,562	30.6	7.0
6	09-07-17 19:00	200	34,164,731	229,447	38,708	261,121,439	32.7	7.0
8	10-07-17 06:00	200	47,326,956	280,359	43,341	310,198,198	32.3	7.1
10	10-07-17 19:00	200	26,146,107	188,357	32,544	225,677,717	35.4	6.8
12	11-07-17 06:00	200	19,821,910	603,636	165,053	384,671,866	44.6	7.3

Table 2.1 Assembly statistics and metadata of all metagenomic assemblies for the cellular fraction using the metaSPAdes (Nurk et al. 2017).

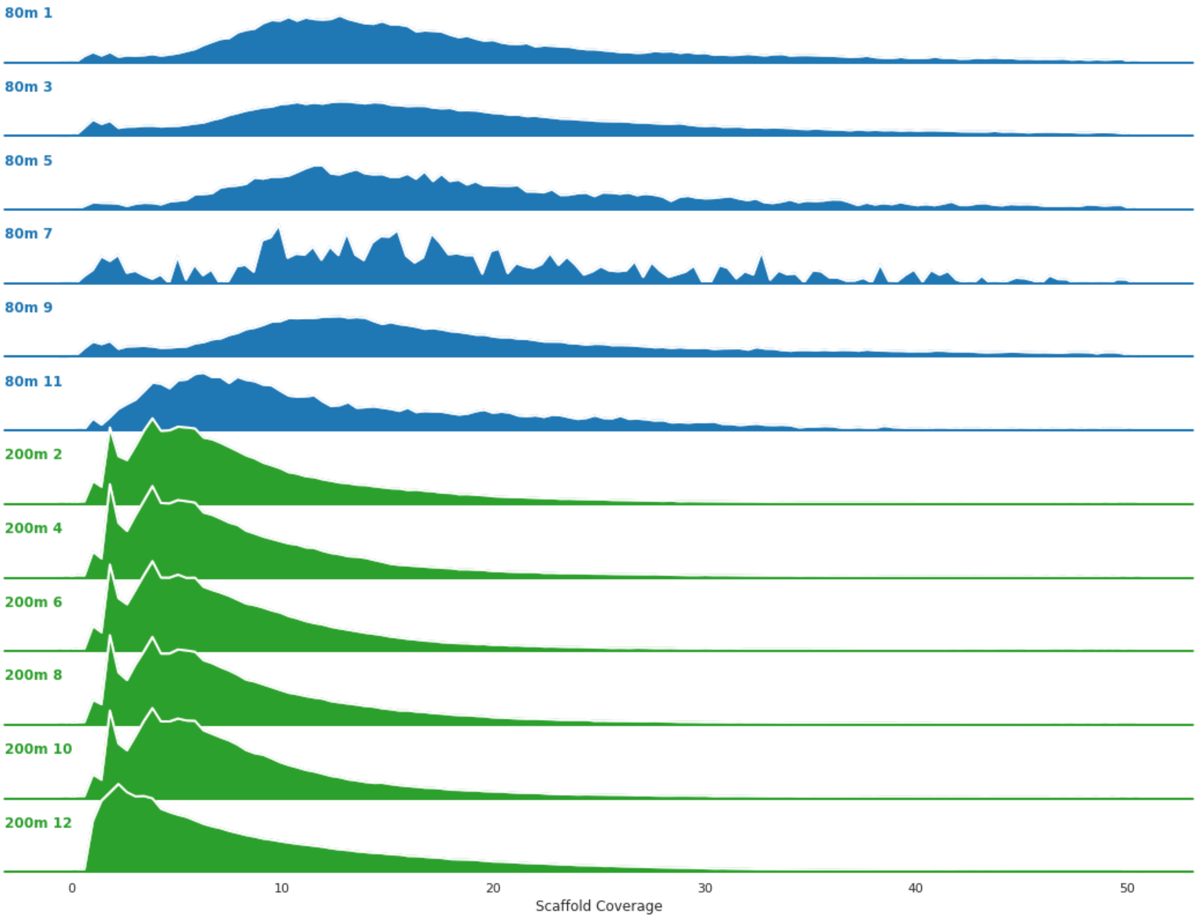


Figure 2.2 Ridgeline plot of scaffold coverage derived from recruited reads for each marine metagenome from (**Table 2.1**). Scaffolds with coverage sizes over 50 are not shown.

2.3.2 Taxonomic relative abundance

Reads from each metagenome were taxonomically classified against the nr database (NCBI, 2019) using kaiju (Menzel, Ng, and Krogh 2015). Only the ten most abundant Orders for each metagenome are shown (**Fig 2.3**). About 60% of all 80m samples are unable to be classified by at least an order level, rising to 80% in 200m samples. The Pelagibacterales order, consisting of all members of the SAR11 clade, are present in all metagenomic samples (25-20% 80m, 20-30% 200m), confirming previous findings (Craig A. Carlson et al. 2009; Morris et al. 2002; Becker et al. 2019), and suggesting recovery of SAR11 MAGs from this data was feasible. Above order classification made up 20-25% of all reads within metagenomes, indicating Pelagibacterales and other order level classifications are likely to be underestimated. Stacked bar plots do not reach 100% due to other order classifications being excluded due to low abundances.

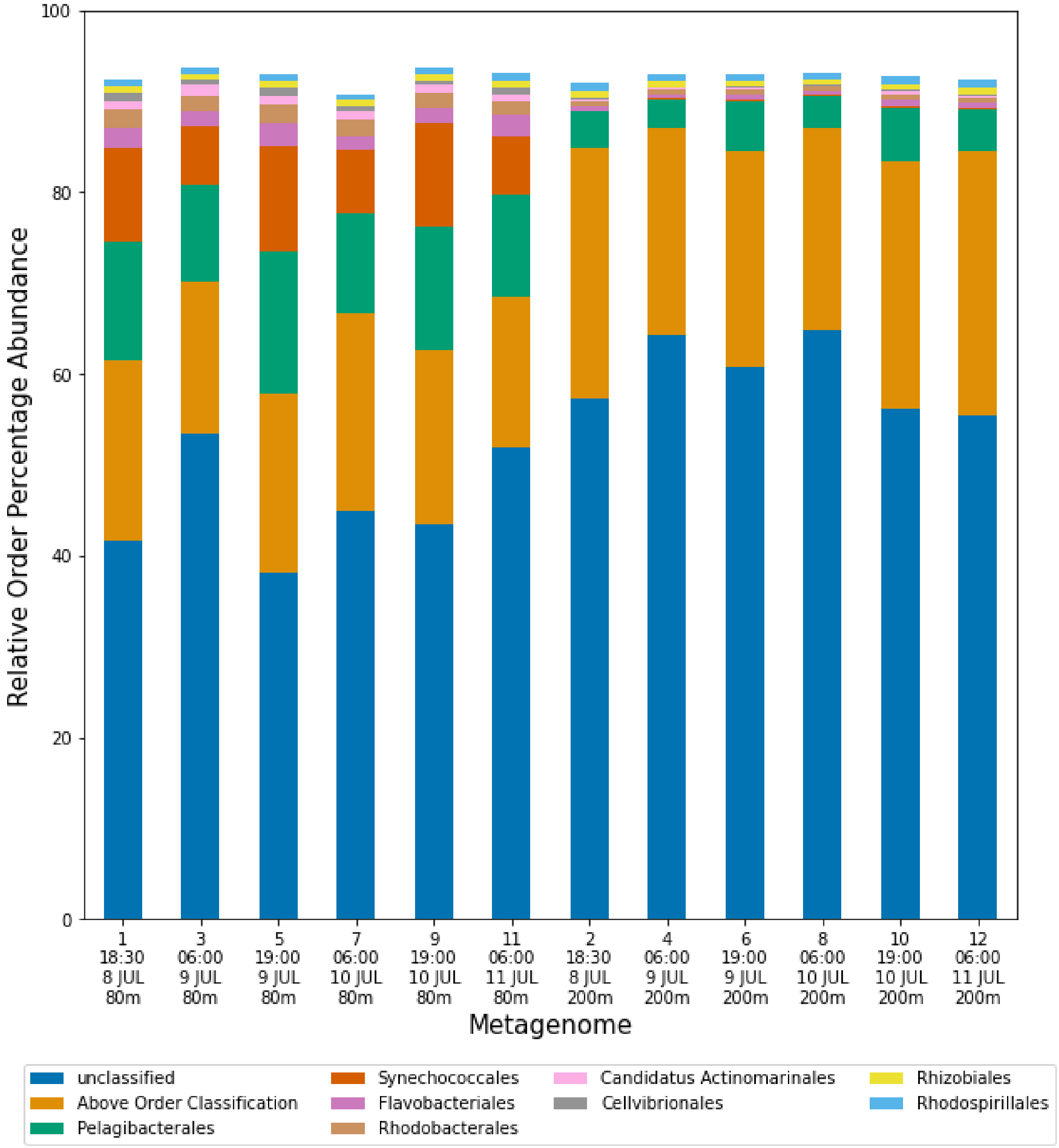


Figure 2.3 Stacked barplot of the relative abundance of marine organisms within twelve BATS metagenomic samples (**Table 2.1**). Taxonomic classification is grouped by order and displayed in colour. Only the top ten most abundant orders per metagenome are shown. Unclassified reads are defined as reads where no taxonomic classification could be deduced. Above order classification

is defined as only having a classification higher than order, for example, Phylum: Bacteria, Class Alphaproteobacteria.

2.3.2 Binning by binning software

BinSanity was used to group contigs from a metagenome into bins to provide MAGs. Taxonomic classification of contigs along with completeness and contamination percentage was calculated for each resulting bin based on single-copy marker genes through CheckM's (D. H. Parks et al. 2015) lineage_wf algorithm. BinSanity produced 159 bins, plotted on a scatter plot based on their completion and contamination percentages, with an estimation of taxa based on single-marker copy genes, along with MAG size represented through point size. (Bowers et al. 2017) describes high quality MAGs having ≥90% completion and <5% contamination (green box) with medium quality genome ≥50% completion and <10% contamination (yellow box). BinSanity produced no high-quality MAGs and one medium quality MAG. The taxonomic classification made no clear indication of any SAR11 MAGs.

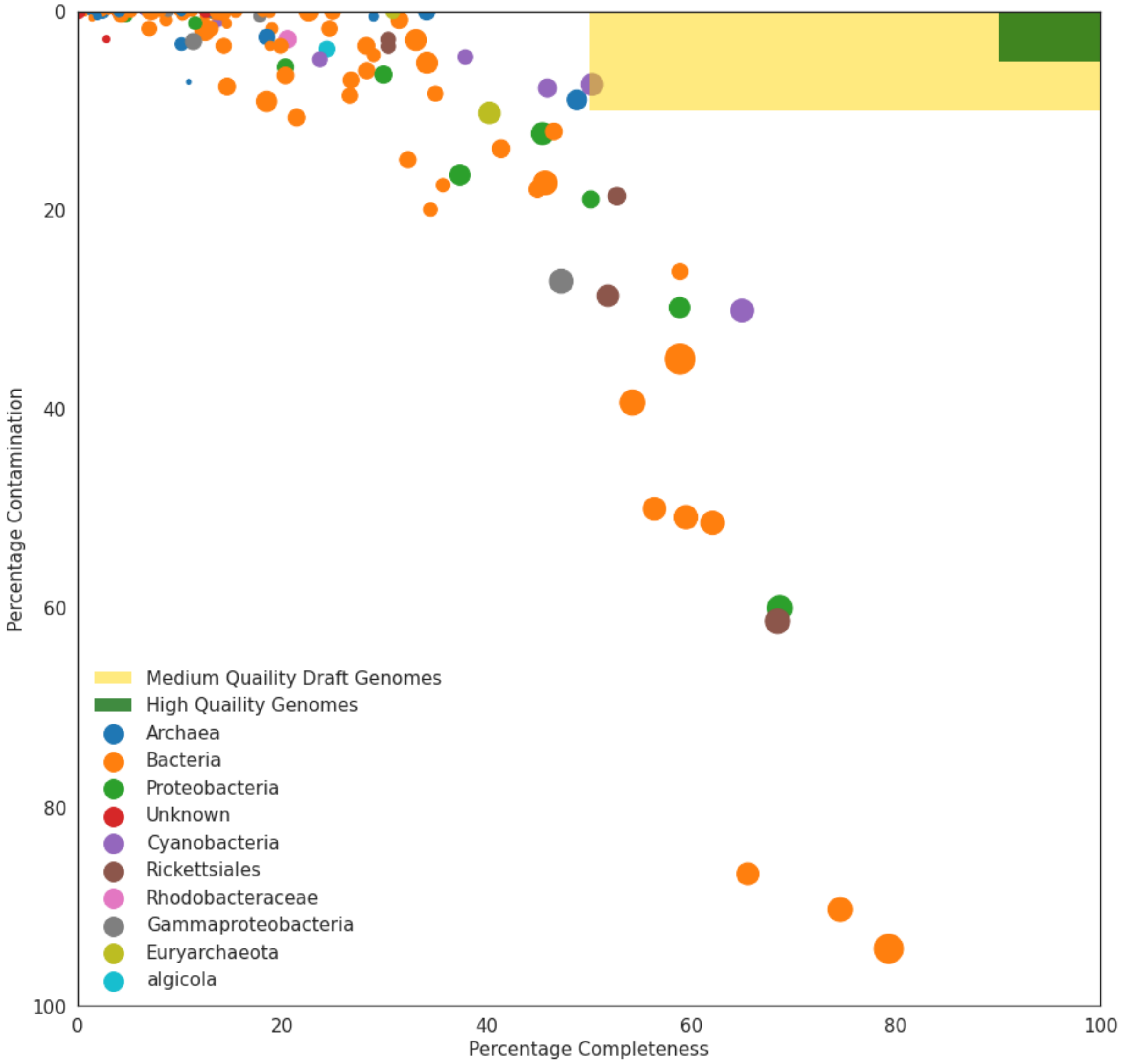


Figure 2.4 Bins produced by BinSanity from an 80m marine metagenome (No. 3) are plotted on a scatter plot based on their completion and redundancy values produced from CheckM. Taxonomic identification is produced through CheckM's lineage_wf pipeline. Scatter plot point sizes are representative of total bin size, with the legend indicating a 1 Mb genome. The green box indicates the region high-quality MAGs should reside along with the yellow box for medium quality MAGs.

Binning, taxonomic classification and genome quality determination of the same metagenome were repeated with MetaBAT2 producing twelve bins. No MAGs were of high and two of medium quality.

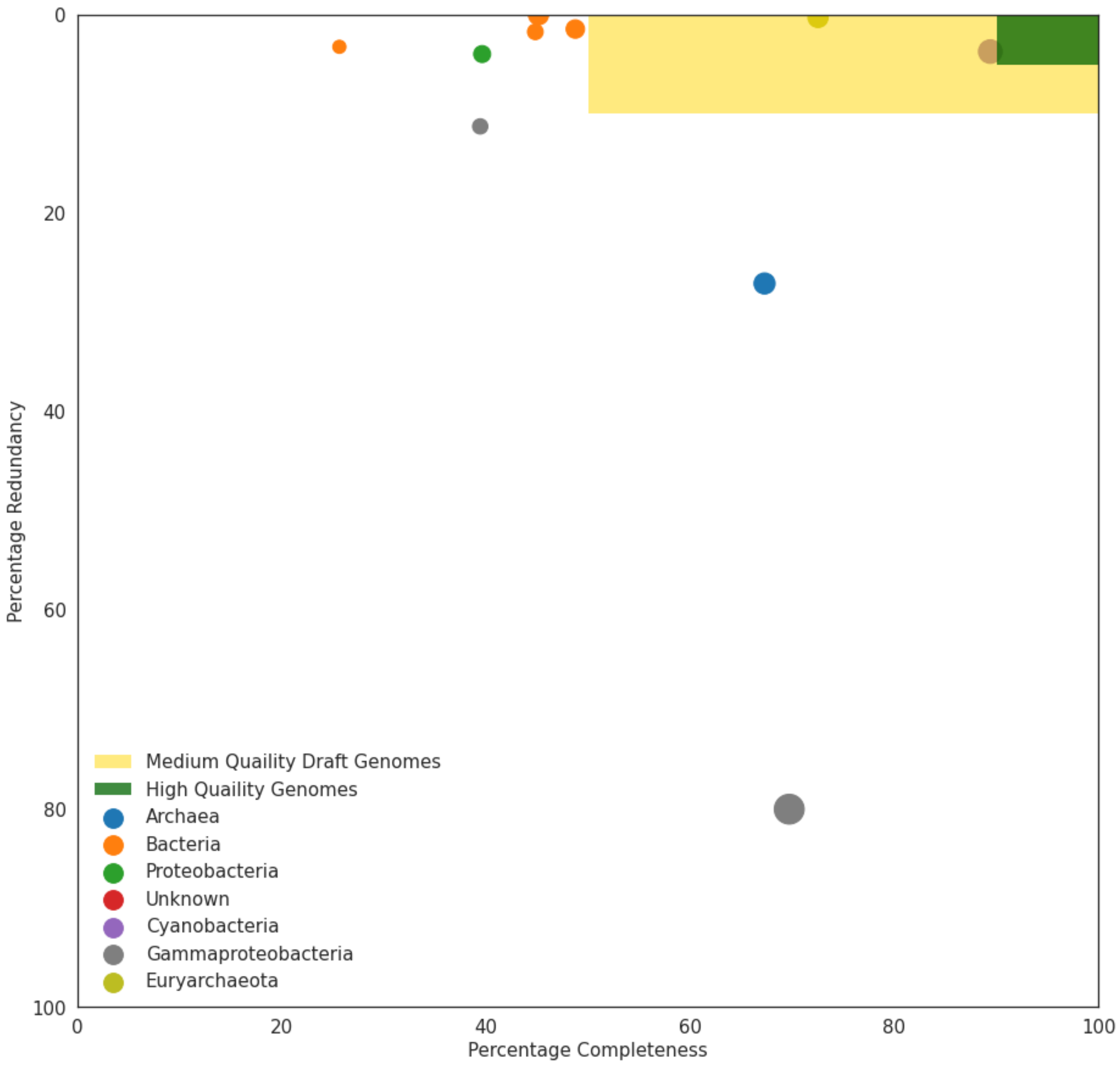


Figure 2.5 Bins produced by MetaBAT2 from an 80m marine metagenome (No. 3) are plotted on a scatter plot based on their completion and redundancy values produced from CheckM. Taxonomic identification is produced through CheckM's lineage_wf pipeline. Scatter plot point sizes are representative of total bin size, with the legend indicating a 1 Mb genome. Green box indicates the region high-quality MAGs should reside along with the yellow box for medium quality MAGs.

Binning, taxonomic classification and genome quality determination of the same metagenome was repeated with MaxBin producing 69 bins. No MAGs were of high and one of medium quality.

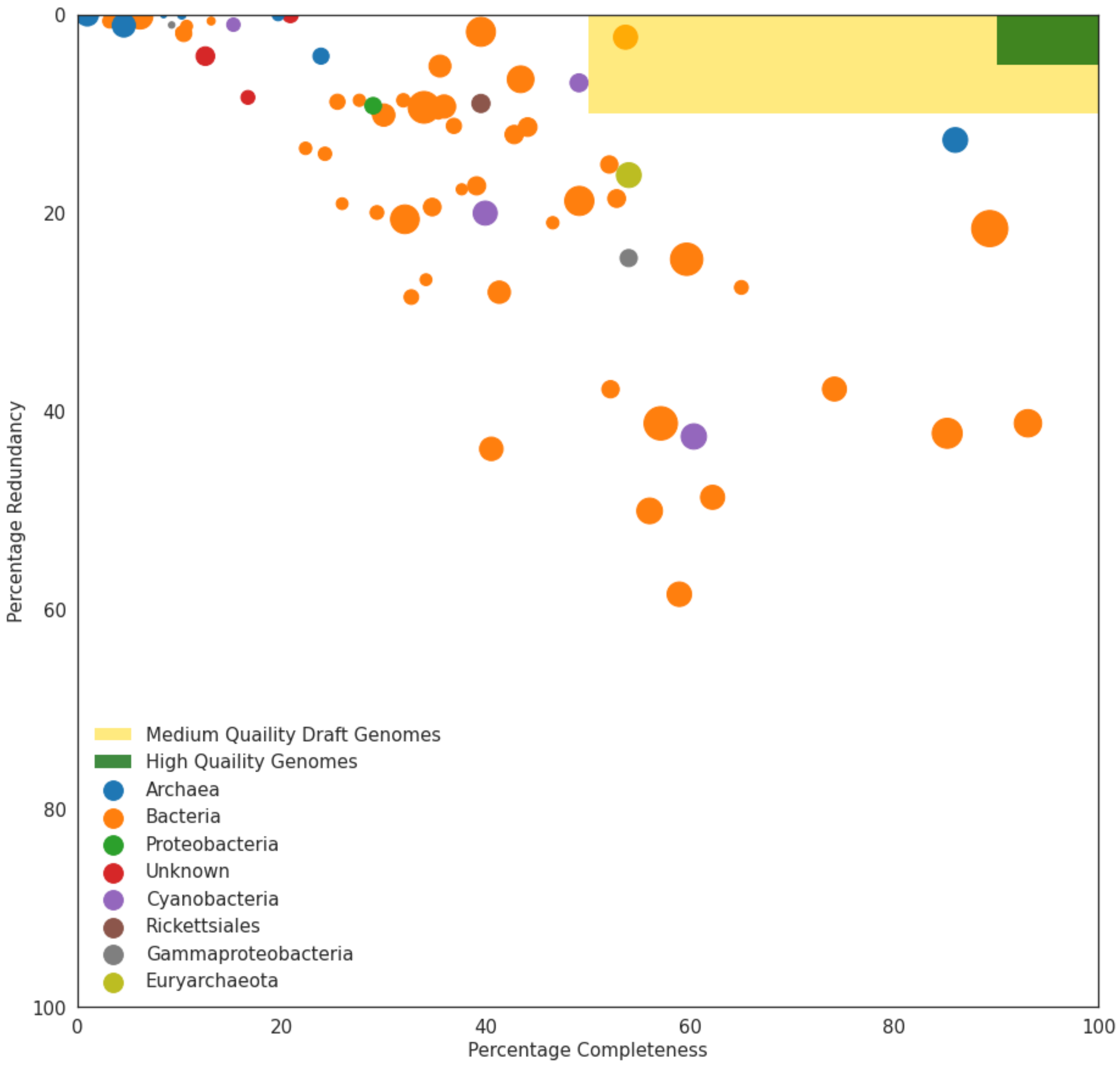
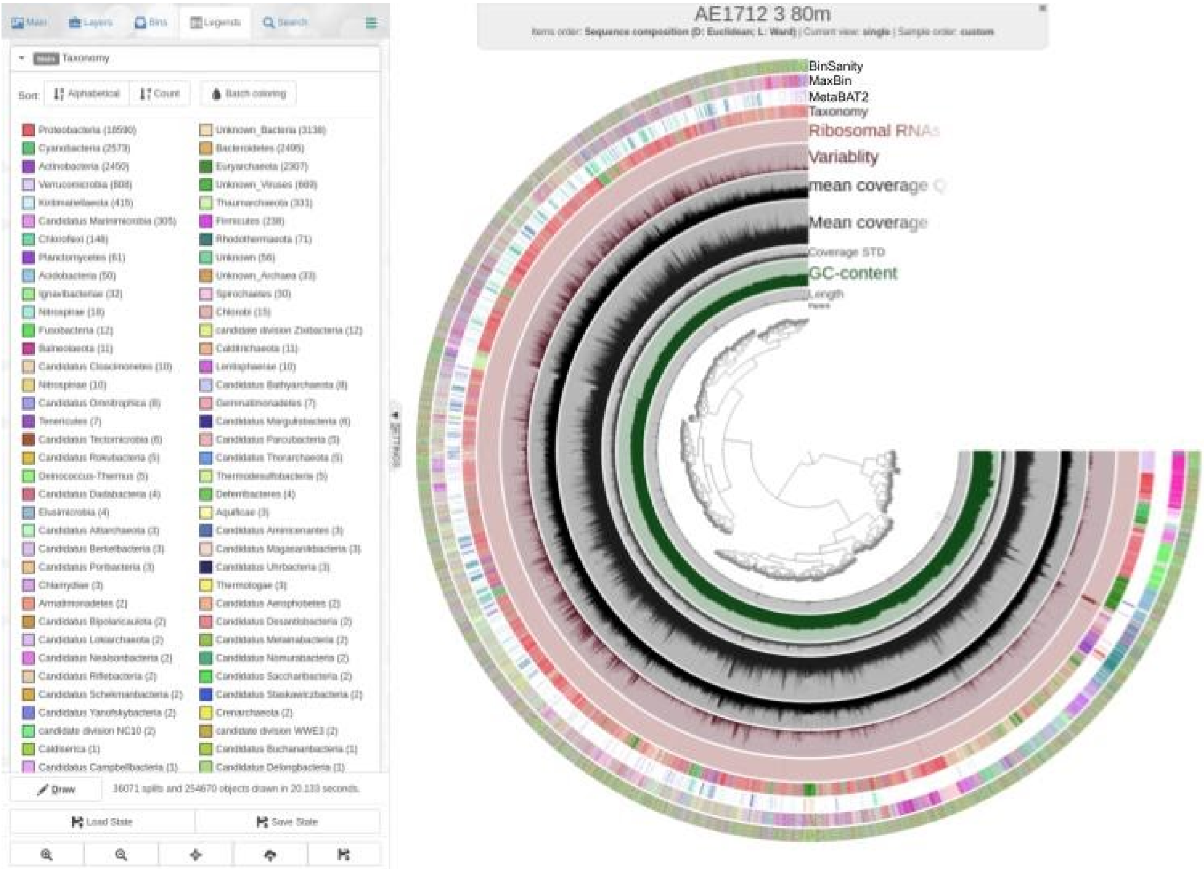


Figure 2.6 Bins produced by MaxBin from an 80m marine metagenome (No. 3) are plotted on a scatter plot based on their completion and redundancy values produced from CheckM. Taxonomic identification is produced through CheckM's lineage_wf pipeline. Scatter plot point sizes are representative of total bin size, with the legend indicating a 1 Mb genome. The green box indicates the region high-quality MAGs should reside along with the yellow box for medium quality MAGs.

2.3.3 Visualisation of metagenomes

Anvi'o, a visualisation platform for 'omics data, provides a through tool efficient in its ability to visualise a whole metagenome and bins concisely. Binning algorithms allow for the classification of contigs into MAGs. However, they are imperfect as each different binning algorithm performs better in different scenarios. Here, multiple different binning algorithms were used to compensate for weakness in each algorithm. Should all algorithms agree that certain contigs should all be binned together, this would improve the likelihood that MAG creation is correct. The process of using multiple binning software is here referred to as consensus binning. Should two binning algorithms agree on a contig identity to a bin and the others not, this would reduce mis-bining of this MAG that otherwise may occur if only one binning software is used. Additionally, the taxonomic classification of genes would provide an extra consensus method that may help binning which anvi'o employs. Should consensus binning work, identified bins would have one bin identity for each binning algorithm. For example, bin 1 from BinSanity would only contain contigs from bin X from Maxbin and not consist of contigs from multiple bins like Maxbins X, Y and Z. This is seen effective in the binning of an infantile gut microbiome performed by anvio's creators seen here with figures here.

Unfortunately, combining three different binning software and taxonomy did not provide a consensus as hundreds of different genes were classified from different species. Additionally, bins from different binning software did not agree with each other to form any kind of consensus.



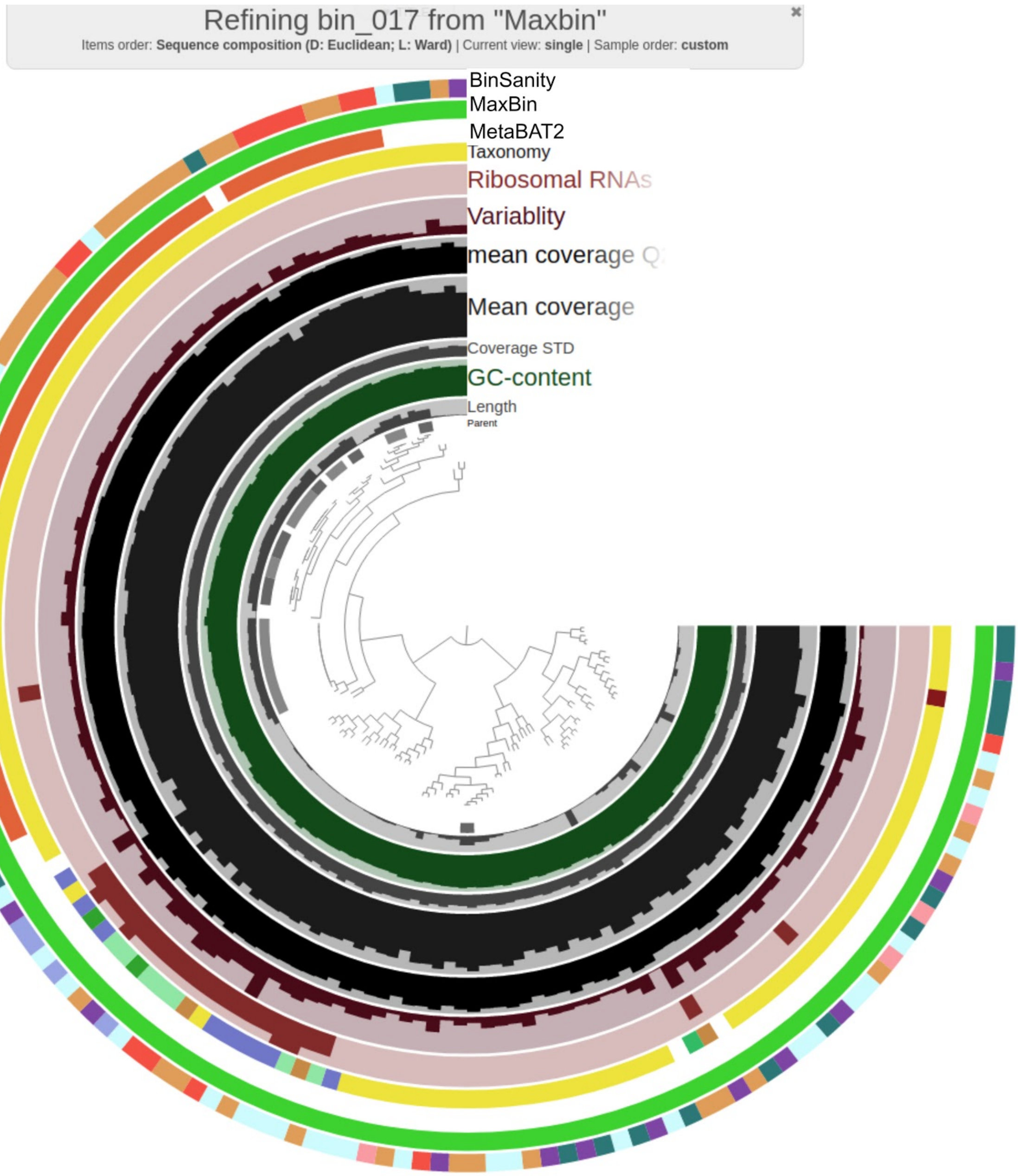
Items order: Sequence composition (D: Euclidean; L: Ward) | Current view: single | Sample order: custom

Proteobacteria (16590)
Cyanobacteria (2573)
Actinobacteria (2450)
Verrucomicrobia (808)
Kiritimatiellaeota (415)
Candidatus Marinimicrobia (305)
Chloroflexi (148)
Planctomycetes (61)
Acidobacteria (50)
Ignavibacteriae (32)
Nitrospirae (18)
Fusobacteria (12)
Balneolaeota (11)
Candidatus Cloacimonetes (10)
Nitrospinae (10)
Candidatus Omnitrophica (8)
Tenericutes (7)
Candidatus Tectomicrobia (6)
Candidatus Rokubacteria (5)
Deinococcus-Thermus (5)
Candidatus Dadabacteria (4)
Elusimicrobia (4)
Candidatus Altiarchaeota (3)
Candidatus Berkelbacteria (3)
Candidatus Portnoybacteria (3)
Chlamydiae (3)
Armatimonadetes (2)
Candidatus Bipolaricaulota (2)
Candidatus Lokiarchaeota (2)
Candidatus Nealsonbacteria (2)
Candidatus Riflebacteria (2)
Candidatus Schekmanbacteria (2)
Candidatus Yanofskybacteria (2)
candidate division NC10 (2)
Caldiserica (1)
Candidatus Campbellbacteria (1)
Unknown_Bacteria (3138)
Bacteroidetes (2409)
Euryarchaeota (2307)
Unknown_Viruses (689)
Thaumarchaeota (331)
Firmicutes (238)
Rhodothermaeota (71)
Unknown (56)
Unknown_Archaea (33)
Spirochaetes (30)
Chlorobi (15)
candidate division Zixibacteria (12)
Calditrichaeota (11)
Lentisphaerae (10)
Candidatus Bathyarchaeota (8)
Gemmatimonadetes (7)
Candidatus Margulisbacteria (6)
Candidatus Parcubacteria (5)
Candidatus Thorarchaeota (5)
Thermodesulfobacteria (5)
Deferribacteres (4)
Aquificae (3)
Candidatus Aminicenantes (3)
Candidatus Magasanikbacteria (3)
Candidatus Uhrbacteria (3)
Thermotogae (3)
Candidatus Aerophobetes (2)
Candidatus Desantisbacteria (2)
Candidatus Melainabacteria (2)
Candidatus Nomurabacteria (2)
Candidatus Saccharibacteria (2)
Candidatus Staskawiczbacteria (2)
Crenarchaeota (2)
candidate division WWE3 (2)
Candidatus Buchananbacteria (1)
Candidatus Delongbacteria (1)

36071 splits and 254670 objects drawn in 20.133 seconds.

Figure 2.7 Metagenome of a marine cellular fraction at a depth of 80m visualised in anvi'o. Hierarchical clustering of contigs is represented via the dendrogram in the centre. Contigs are split into 2500 base-pairs fragments for analysis. Tracks from the centre out include parent, indicating a grouping of fragments larger than 2500 base-pairs, length indicating the size of contigs, GC content as a shaded scatter plot in green, coverage statistics, the presence of ribosomal RNAs, taxonomy produced by kaiju, and various binning software's bin identity. Taxonomic identity is displayed as a key on the left ordered by the highest abundance.

To reduce redundancy and create higher quality MAGs, a random high completeness bin was chosen to see if consensus binning and taxonomy could be used to remove redundant genes. Removal of redundant genes from a high completeness and high redundancy bin may result in just a high completeness bin. This is the basis of redundancy removal from MAGs by human input (see **Figure 7**). A lack of consensus at a bin level made it difficult to improve MAGs quality without removing genes related to that organism.

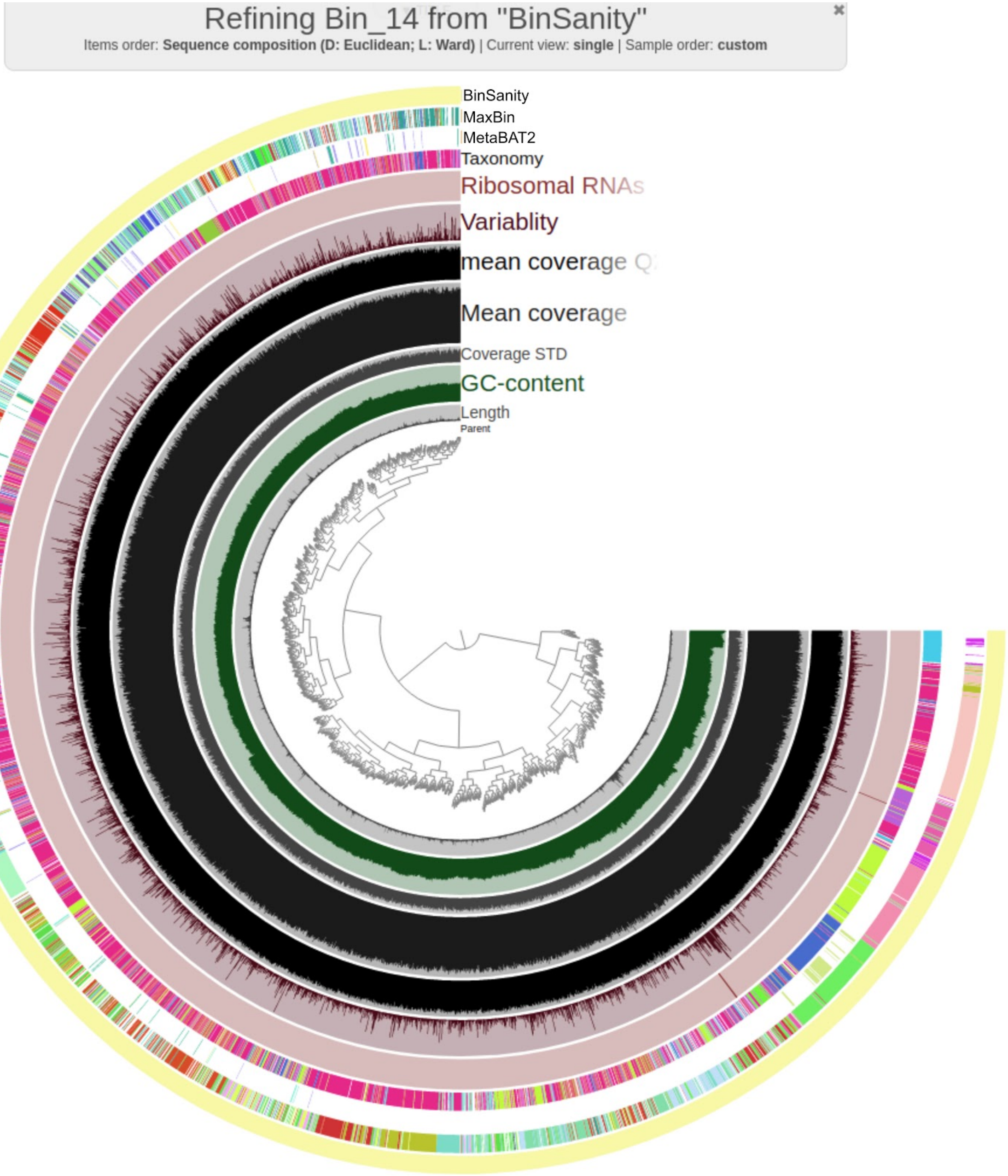


Refining bin_017 from "Maxbin"

Items order: **Sequence composition (D: Euclidean; L: Ward)** | Current view: **single** | Sample order: **custom**

Figure 2.8 A random high completeness and high redundancy bin from Maxbin taken from a metagenome of an 80m sample, visualised using anvi'o. Coloured tracks on the outside indicate the bin identity of different binning software and taxonomy of contigs provided by Kaiju.

The same process was repeated with BinSanity to reduce redundancy percentage and create higher quality MAGs. A random high completeness and low redundancy bin from BinSanity was chosen to remove redundant genes (see **Figure 2.8**). If multiple genomes are present within a bin, a consensus of binning algorithms would help indicate where each genome would start and end. However, no consensus was shown at any capacity within any anvi'o bin.



Refining Bin_14 from "BinSanity"

Items order: **Sequence composition (D: Euclidean; L: Ward)** | Current view: **single** | Sample order: **custom**

Figure 2.9 Bin 14 from BinSanity taken from a metagenome of an 80m sample, visualised using anvi'o. Coloured tracks on the outside indicate the bin identity of different binning software and taxonomy of contigs provided by Kaiju.

2.3.4 Exploration of UMAP n_neighbours for data structure interpretation

Consensus binning using different algorithms proved difficult to obtain high-quality MAGs, therefore an alternative approach was used. Here a script by (Beaulaurier et al. 2019) was used showing success with viral genomes. Contigs were 5-mer counted to obtain a 5-mer frequency for each contig. These were plotted on a scatter plot using UMAP, a dimension reduction technique. Cluster determination was performed by HDBSCAN with different colours indicating different clusters. The script was applied with default settings which produced 64 bins (**Fig 2.10**). Quantification of MAGs determined from UMAP and HDBSCAN was assessed with CheckM (**Fig 2.11**).

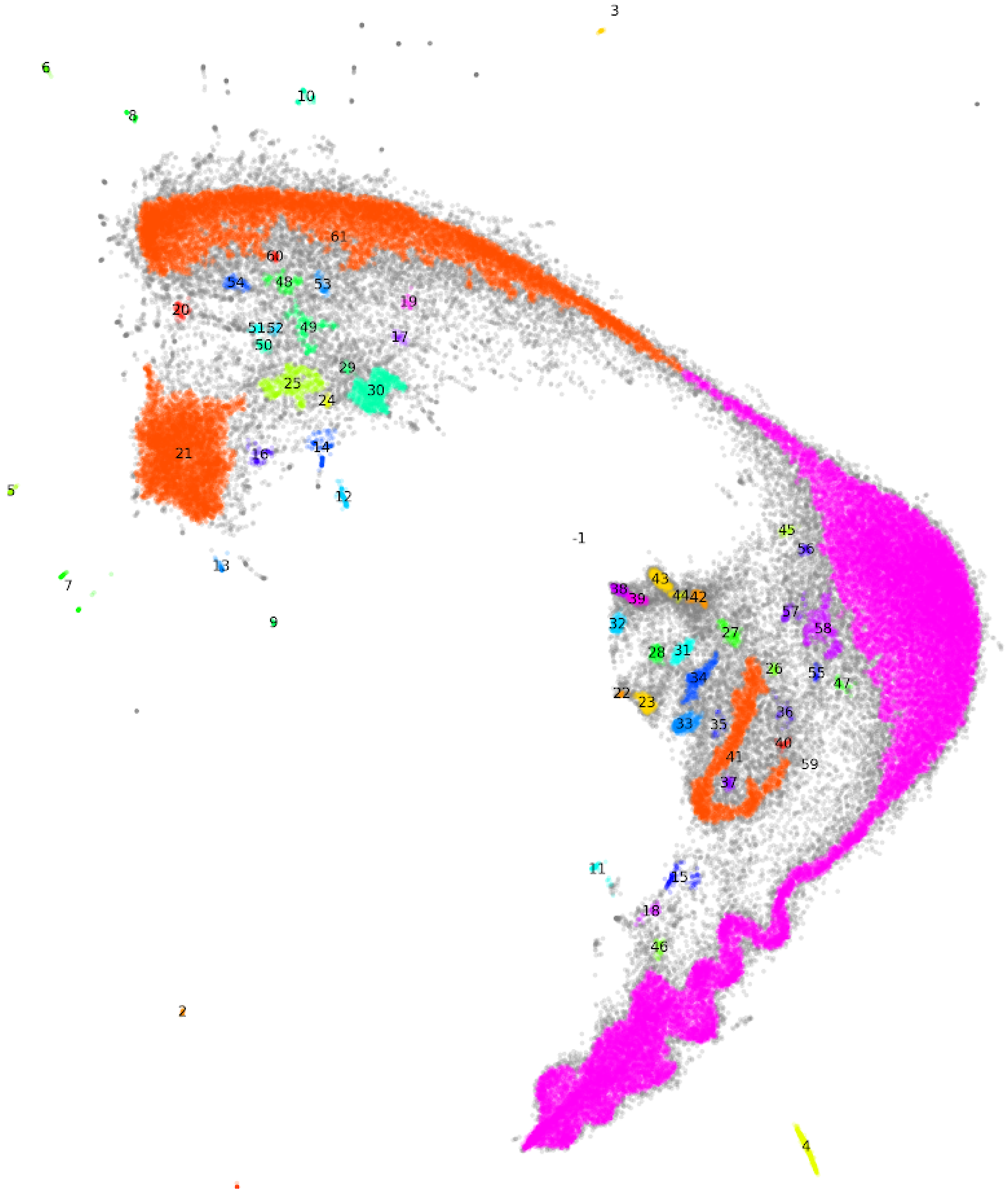


Figure 2.10 Scatterplot of **all *5*-mer frequencies contigs** dimensionally reduced using default settings from UMAP from an 80m marine metagenome (No. 3). Cluster determination is displayed in colour using HDBSCAN.

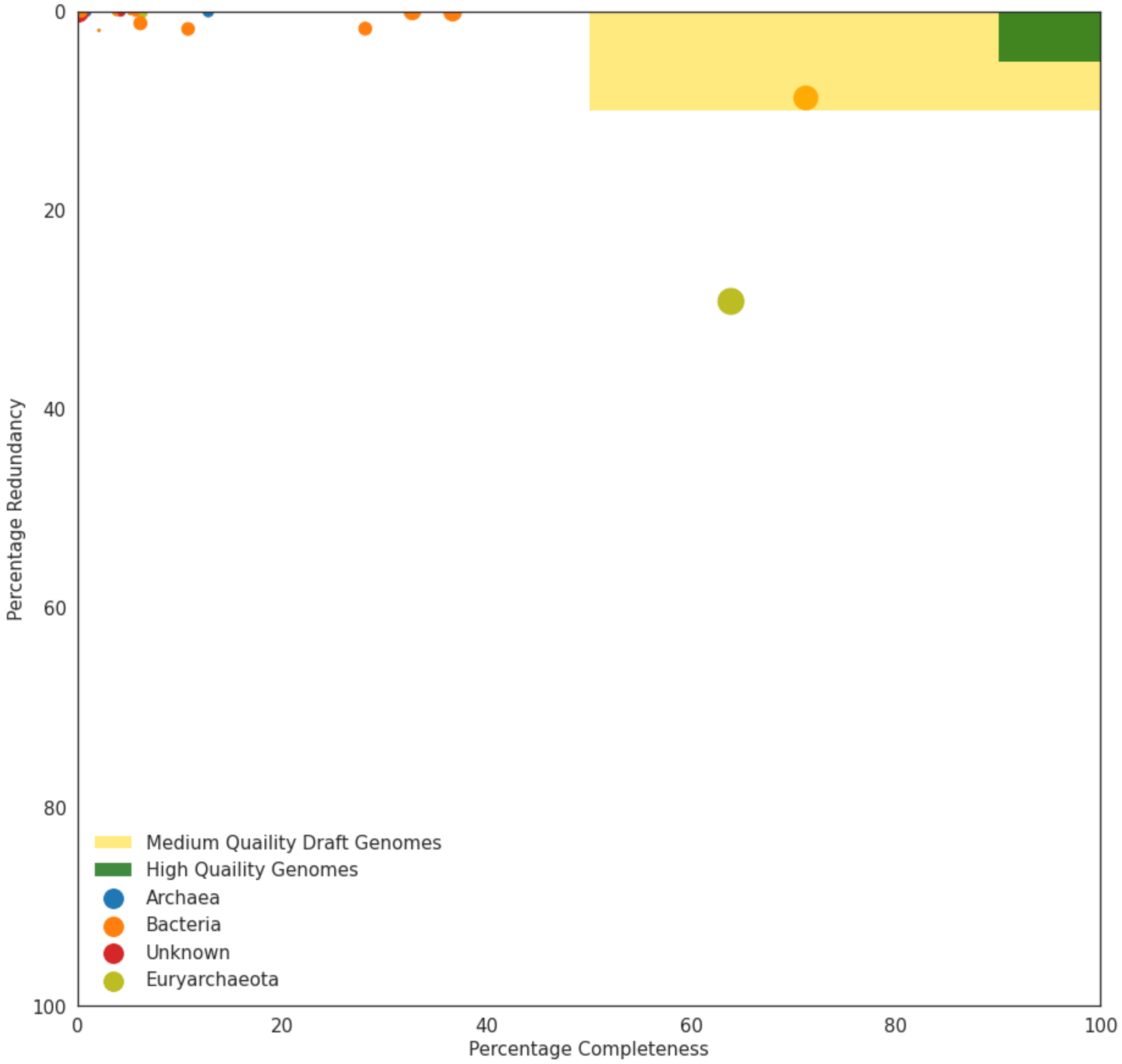


Figure 2.11 Bins produced by *k*-mer counting **all contigs** from an 80m marine metagenome (No. 3). Dimension reduction technique was performed by UMAP with cluster determination and binning was performed by HDBSCAN using **default settings**. Datapoints are MAGs plotted on a scatter plot based on their completion and redundancy values produced from CheckM. Taxonomic identification is produced through CheckM's lineage_wf pipeline. Scatter plot point sizes are representative of total bin size, with the legend indicating a 1 Mb genome. The green box indicates the region high-quality MAGs should reside along with the yellow box for medium quality MAGs.

Various other informatics parameters were trialled. Increasing the contig cutoff size may reduce noise as seen in **Fig 2.9** due to smaller contigs having outlier *k*-mer frequencies. This arises due to shorter contigs not having enough sequence data to provide a representative sample of expected unique kmer frequency an organism may have. This prevents mis-binning of unrelated genetic sequences within the same bin. Results indicated trimming a metagenome to remove contigs below 2500 base pairs as recommended by anvi'o's developers resulted in 16 bins **(Fig 2.11)**. However, this resulted in the removal of 91.40% of all contigs and consisted of 61.21% of all nucleotides.

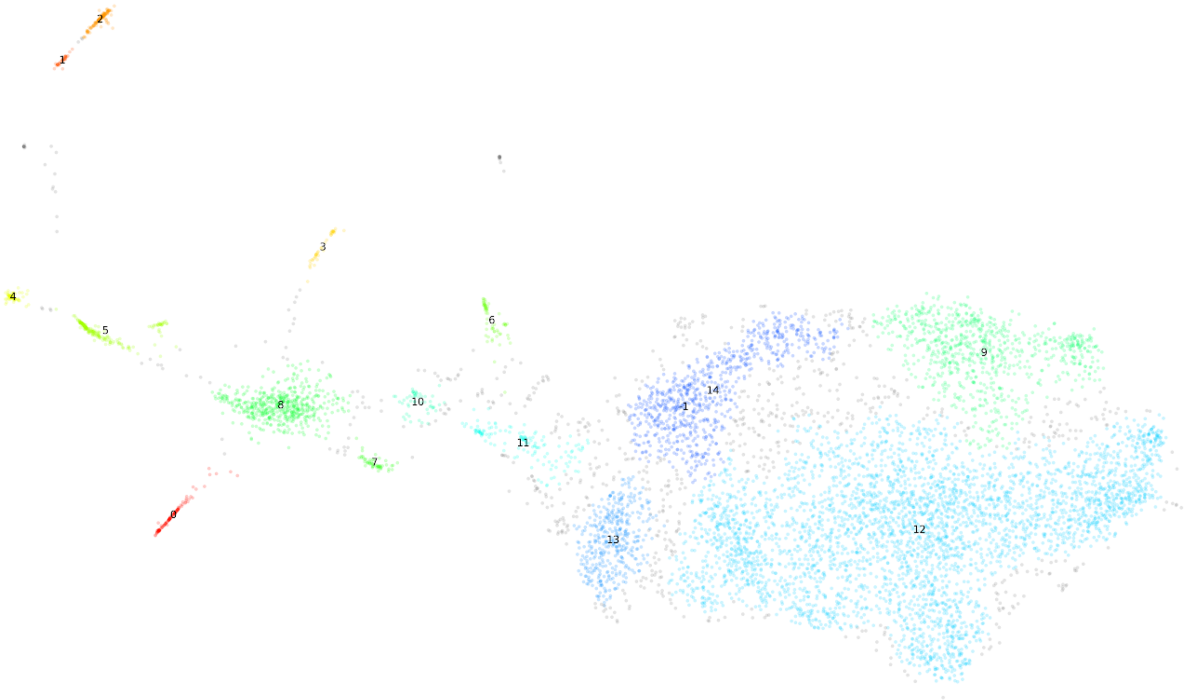


Figure 2.12 Scatterplot of contig *5*-mer frequencies **above 2500 base pairs** in length are dimensionally reduced using default settings from UMAP from an 80m marine metagenome (No. 3). Cluster determination is displayed in colour using HDBSCAN.

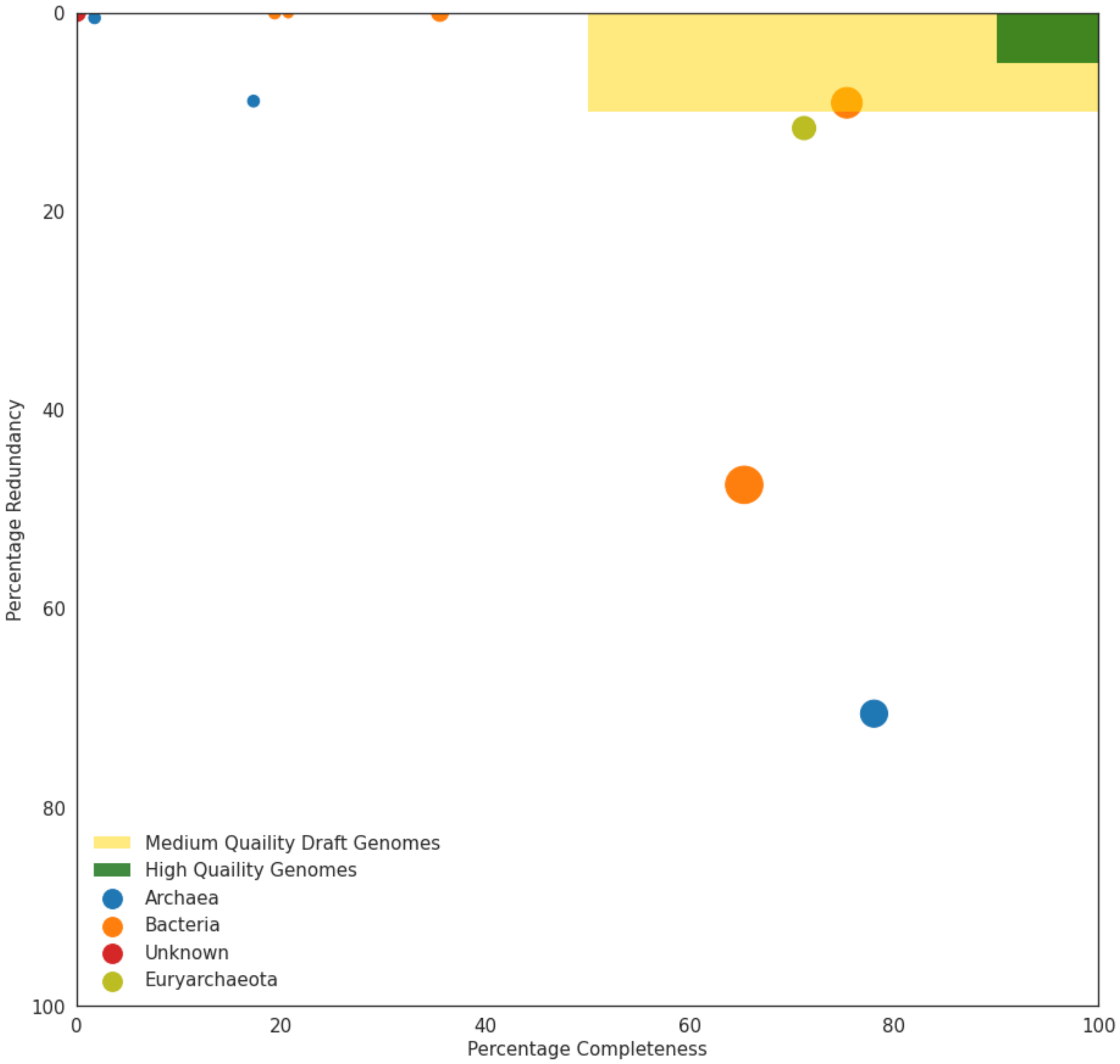


Figure 2.13 Bins produced by *k*-mer counting **contigs above 2500 base pairs** from an 80m marine metagenome (No. 3). Dimension reduction technique was performed by UMAP with cluster determination and binning was performed by HDBSCAN using **default settings**. Datapoints are MAGs plotted on a scatter plot based on their completion and redundancy values produced from CheckM. Taxonomic identification is produced through CheckM's lineage_wf pipeline. Scatter plot point sizes are representative of total bin size, with the legend indicating a 1 Mb genome. The green box indicates the region high-quality MAGs should reside along with the yellow box for medium quality MAGs.

With increasing contig cutoff not affecting genome quality, n_neighbours setting determining UMAP scatter plot plotting was explored (default n_neighbours = 15). Increasing n_neighbours (N_neighbours = 100) only resulted in one bin with 100% completeness and over 6000 in redundancy (not shown). N_neighbours = 2 (**Fig 2.14**) was plotted and bin quality determined. This resulted in 65 bins with one MAG being of medium quality. No SAR11 MAGs were detected taxonomically.

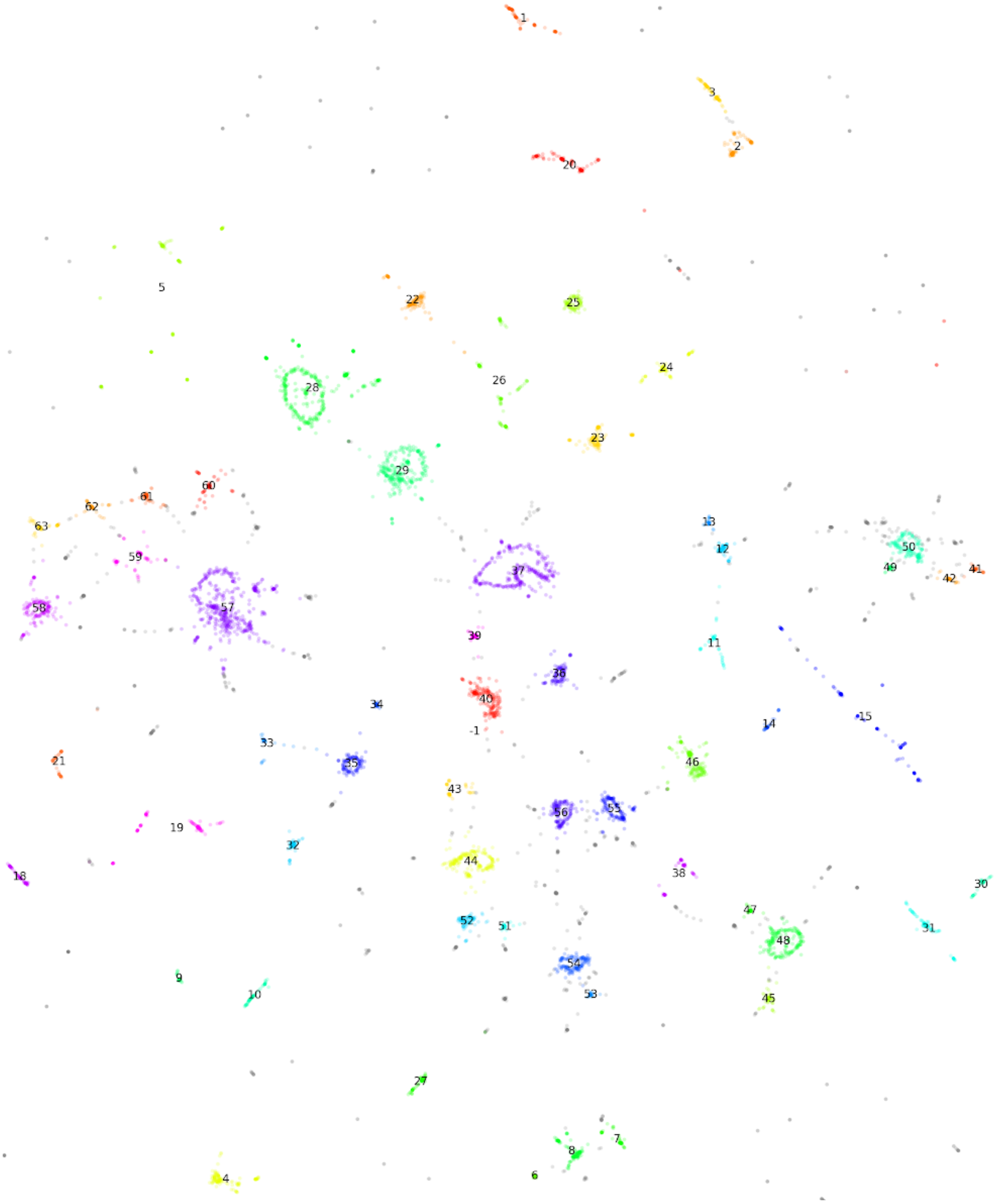


Figure 2.14 Scatterplot of **all contig 5-mer frequencies** are dimensionally reduced using UMAP with **n_neighbours = 2** from an 80m marine

metagenome (No. 3). Cluster determination is displayed in colour using HDBSCAN.

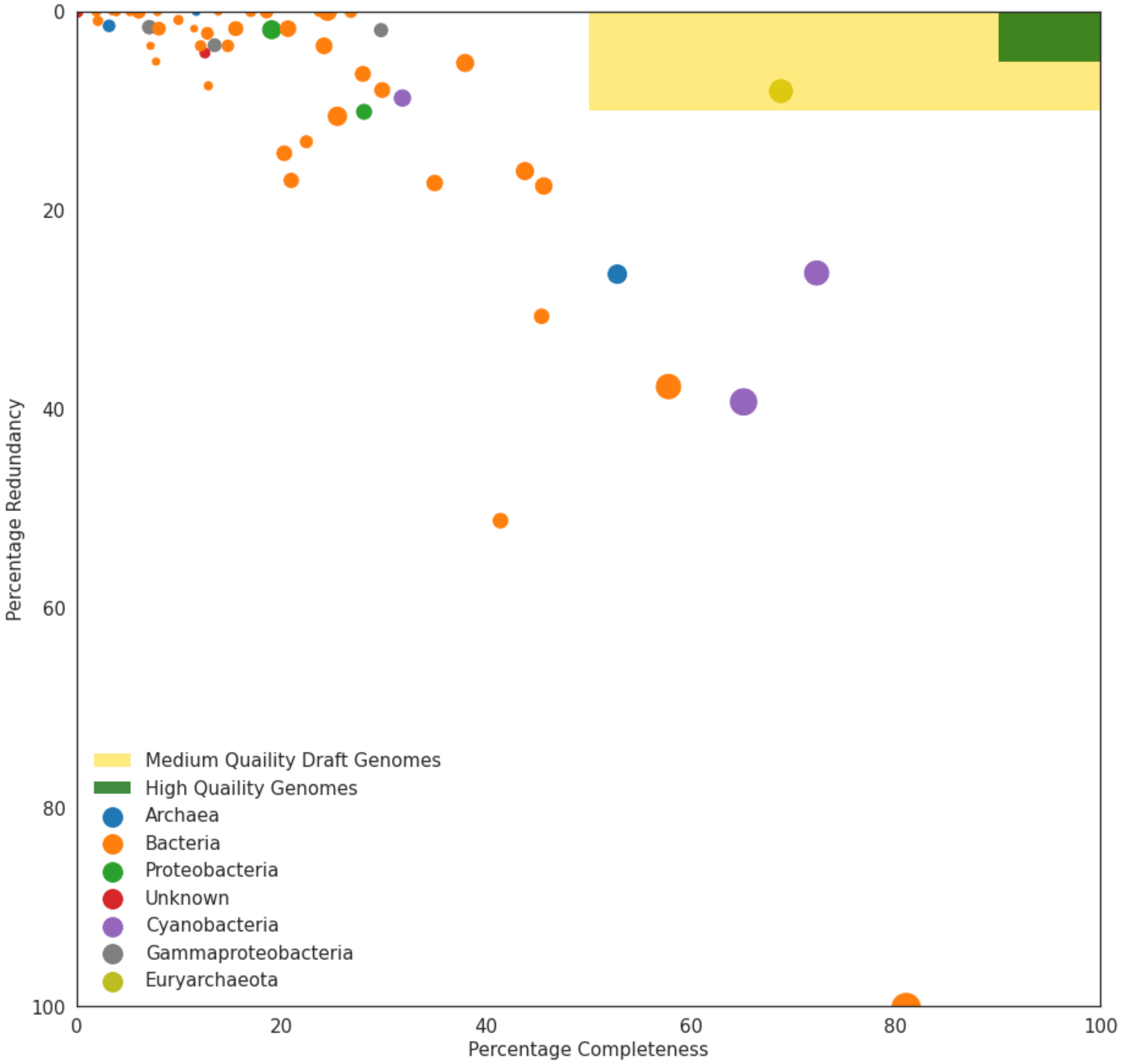


Figure 2.15 Bins produced by *k*-mer counting **all contigs** from an 80m marine metagenome (No. 3). Dimension reduction technique was performed by UMAP with **n_neighbours = 2**. Cluster determination and binning were performed by HDBSCAN. Datapoints are MAGs plotted on a scatter plot based on their completion and redundancy values produced from CheckM. Taxonomic identification is produced through CheckM's lineage_wf pipeline. Scatter plot point sizes are representative of total bin size, with the legend indicating a 1 Mb genome. The green box indicates the region high-quality MAGs should reside along with the yellow box for medium quality MAGs.

2.3.5 Mapping by reference genome

Reads from metagenomes were mapped to over 730 SAR11 genomes belonging to all members of the SAR11 clades (I-III). Mapped reads were sub selected and used in reference-based assembly. Once assembled with a single-cell assembler, shown possible with the SPAdes single-cell assembler in previously uncultivable phyla TM6 (McLean et al. 2013), SAR11-like reads were mapped back to the assembly to produce coverage values. 80.53% of reads mapped back to the assembly with a total “mini” metagenome size of 127Mb with 194,346 scaffolds. With the SAR11 genome being 1.3 Mb in size, this was assumed to contain multiple SAR11 genomes and possible contaminants. Therefore, the *k*-mer frequency binning method along with the BinSanity algorithm was used to derive SAR11 genomes from a reduced metagenomic dataset. Established lower n_neighbours was more effective at producing higher quality bins therefore UMAP parameter n_neighbours = 2 was performed on the SAR11-like metagenome **(Fig 2.15)**. Overall, 419 bins were produced with no SAR11 bins identified through taxonomy or high/medium quality MAGs produced.

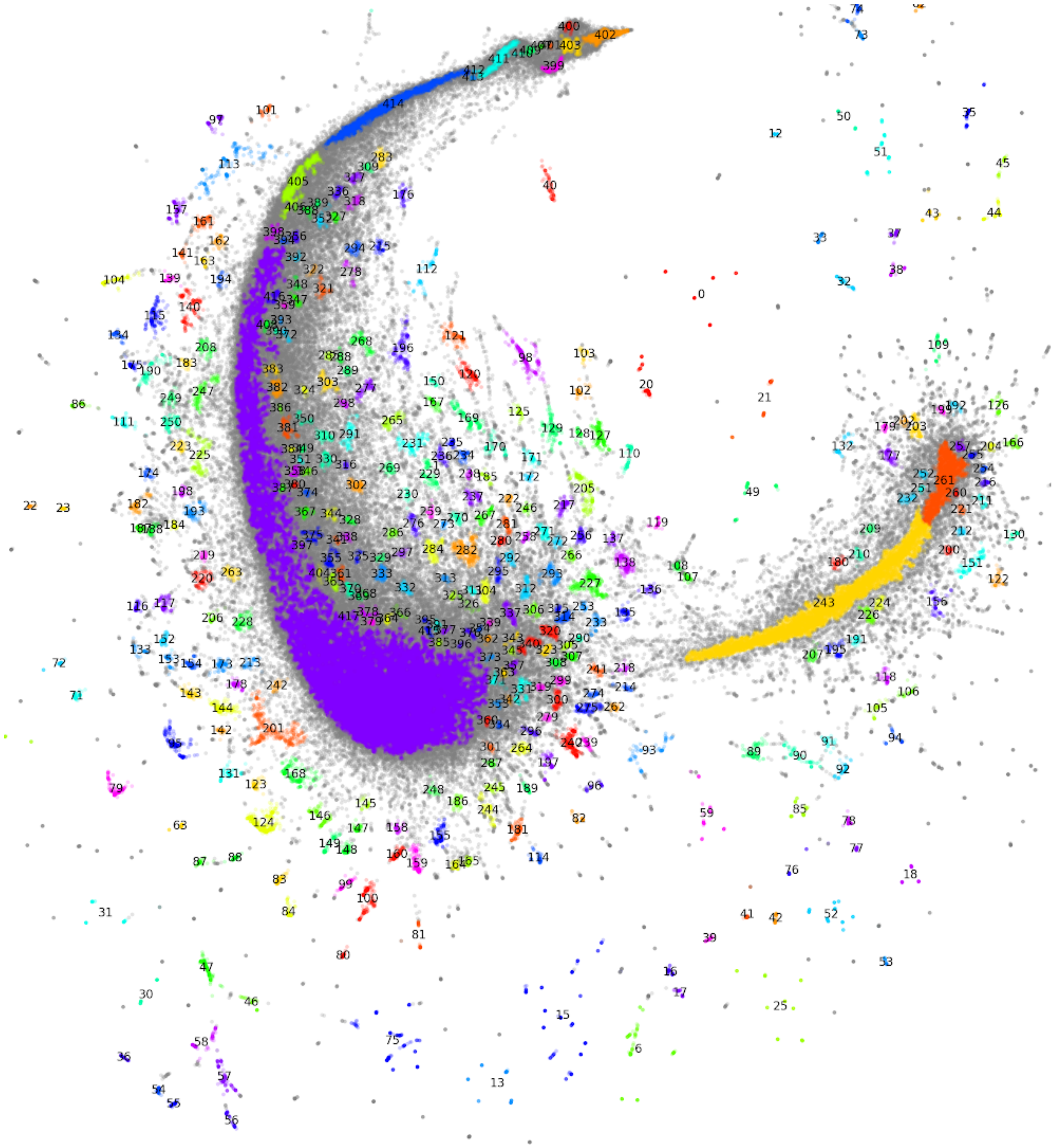


Figure 2.16 Scatterplot of 5-mer frequencies **of all SAR11-like contigs** is dimensionally reduced using UMAP with **n_neighbours = 2** from a mini-metagenome derived from an 80m marine metagenome (No. 3). Cluster determination is displayed in colour using HDBSCAN.

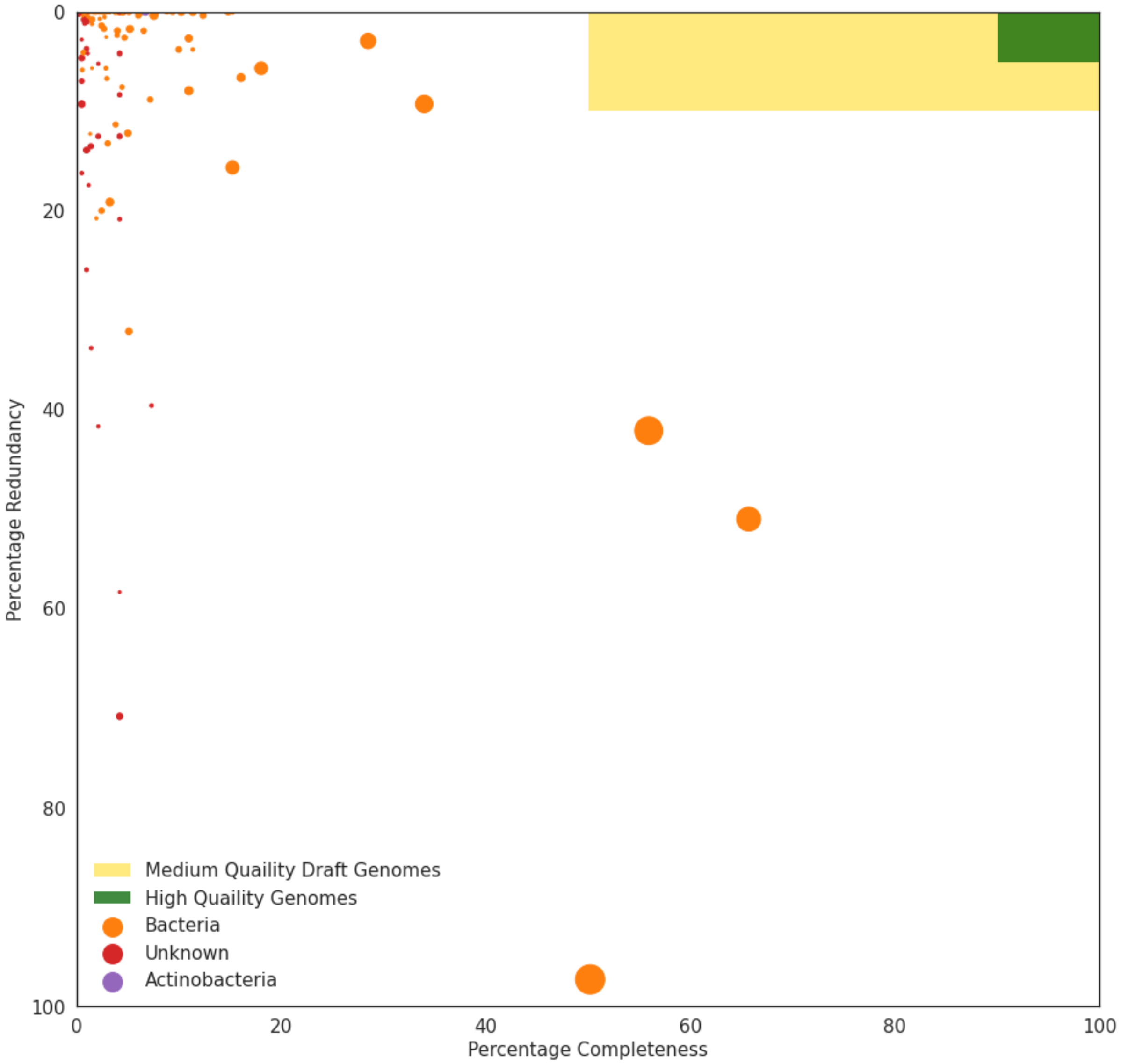


Figure 2.17 Bins produced by *k*-mer counting **all contigs** from a **SAR11-like "mini-metagenome"**. Dimension reduction technique was performed by UMAP with **n_neighbours = 2**. Cluster determination and binning were performed by HDBSCAN. Datapoints are MAGs plotted on a scatter plot based on their completion and redundancy values produced from CheckM. Taxonomic identification is produced through CheckM's lineage_wf pipeline. Scatter plot point sizes are representative of total bin size, with the legend indicating a 1 Mb genome. The green box indicates the region high-quality MAGs should reside along with the yellow box for medium quality MAGs.

BinSanity was used to bin the SAR11-like mini metagenome to improve on MAG recovery. Overall, 112 bins were produced with MAGs of high or medium quilty or SAR11 like MAGs from taxonomic classification.

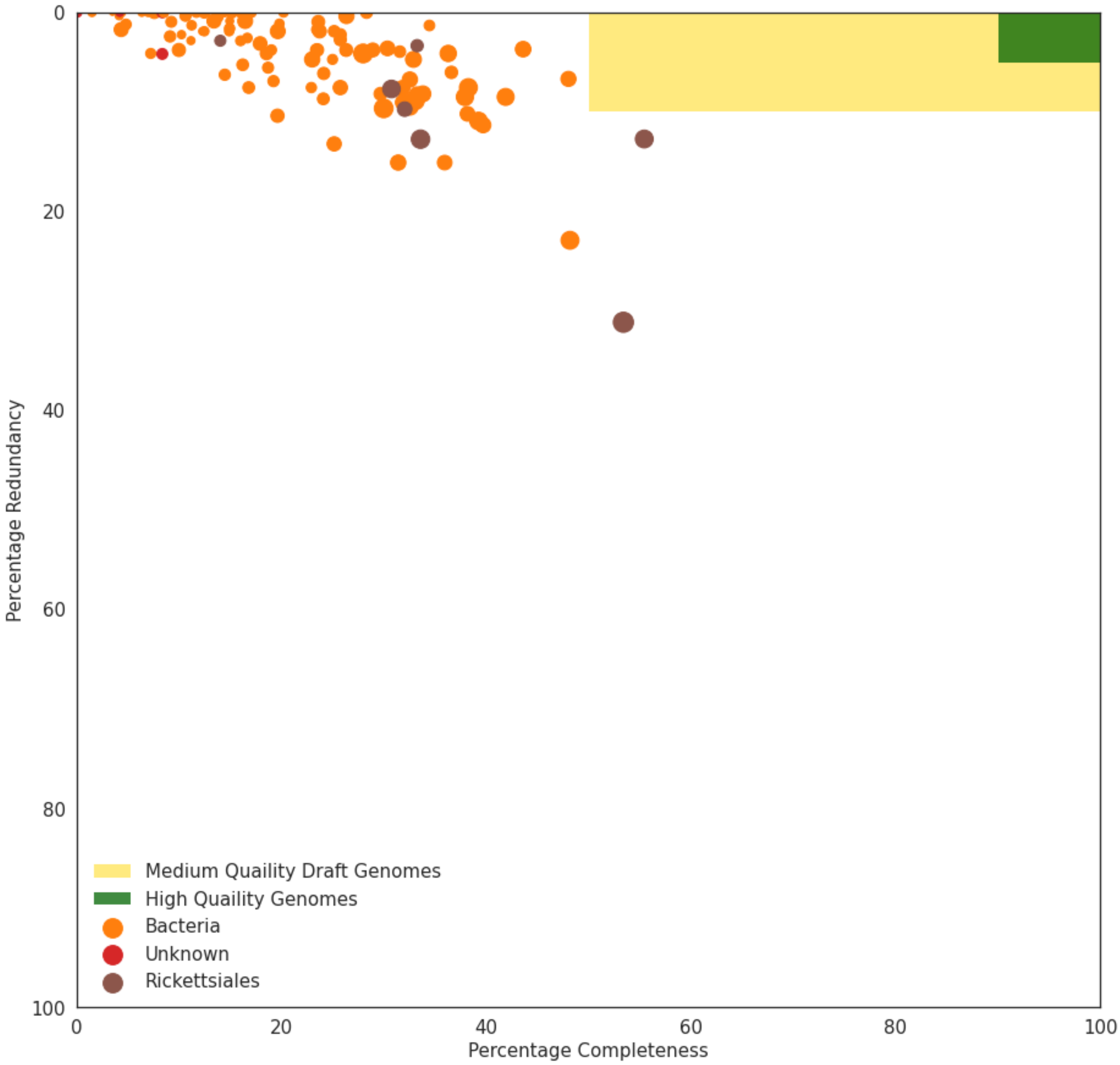


Figure 2.18 Bins produced by the BinSanity algorithm workflow using contigs above 1 kb in size from a **SAR11-like "mini-metagenome"**. Datapoints are resulting MAGs plotted on a scatter plot based on their completion and redundancy values produced from CheckM. Taxonomic identification is produced through CheckM's lineage_wf pipeline. Scatter plot point sizes are representative of total bin size, with the legend indicating a 1 Mb genome. The green box indicates the region high-quality MAGs should reside along with the yellow box for medium quality MAGs.

2.3.6 *K*-mer frequencies validation

A lack of effective binning from *k*-mer counting methods led to methods testing using artificial datasets in order to confirm if the same organisms group together when genomic fragments are *k*-mer counted using previously established methods. Ten random bacteria from across the tree of life were taken and their genomes split into 10 kbp fragments. Plotting of 10 kbp fragments of other bacterial genomes resulted in cluster formation, but it was not 100% accurate. This showed that *k*-mer counting is only somewhat effective at clustering different species' genomes.

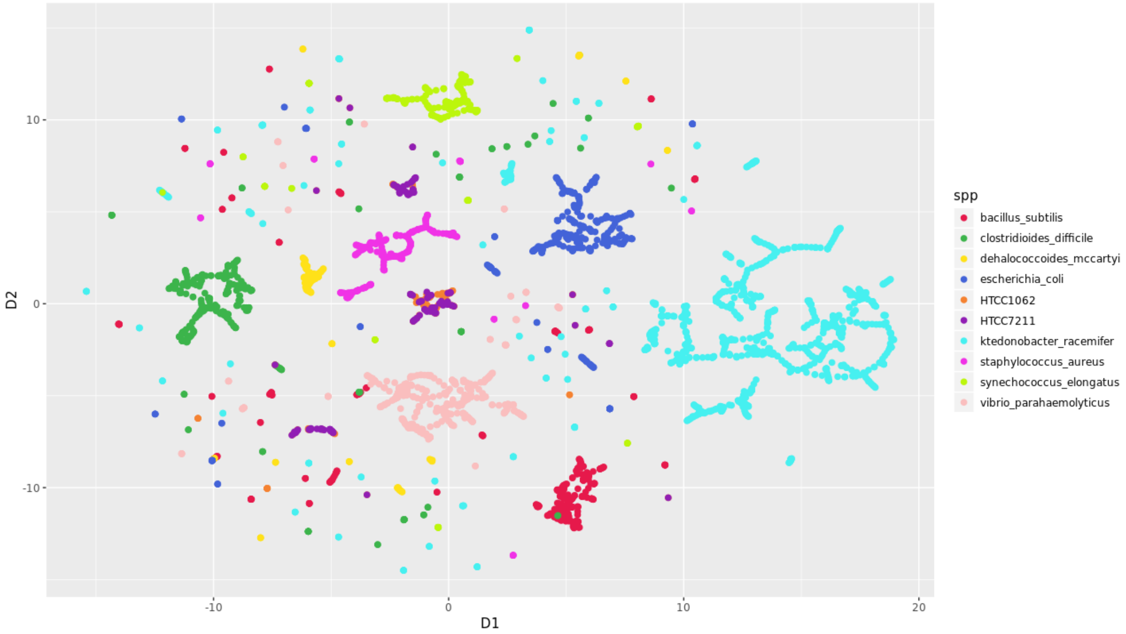


Figure 2.19 BH-tSNE plot of *k*-mer frequencies of ten bacteria across the tree of life. Genomes were split into 10 kbp fragments, *k*-mer frequency counted and plotted as a scatter plot using dimension reduction technique.

Plotting of SAR11 species HTCC7211 and HTCC1062 showed fragmentation of their genomes across the BH-tSNE plot. However, there was also overlap between them, indicating that BH-tSNE is not effective at differentiating between closely related fragments of the closely related species.

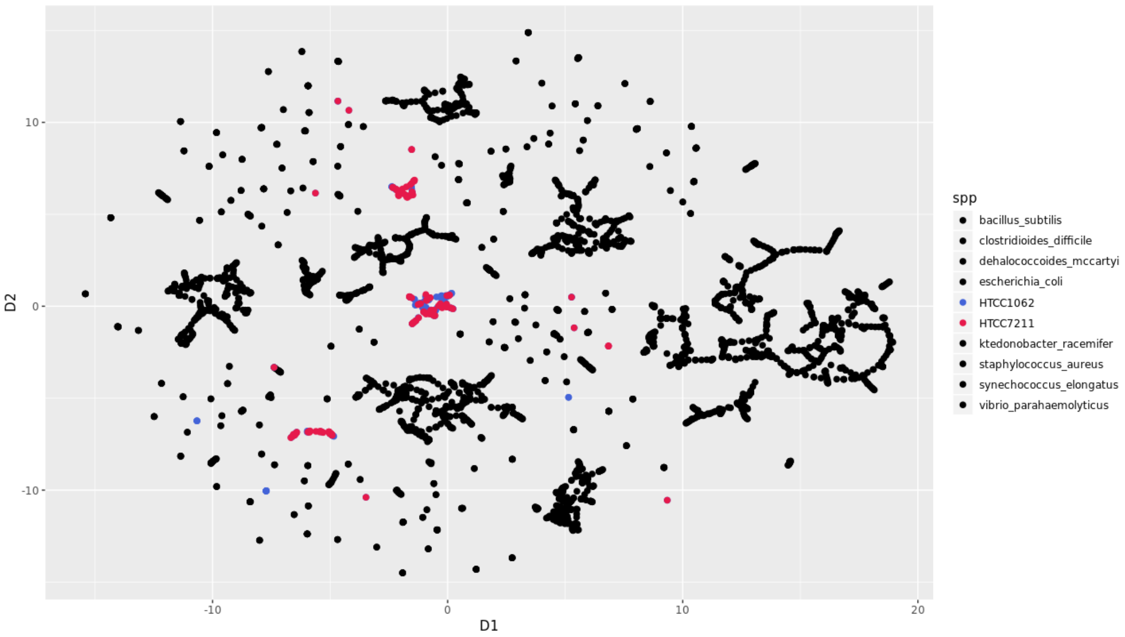


Figure 2.20 BH-tSNE plot of *k*-mer frequencies of ten bacteria across the tree of life. Genomes were split into 10 kbp fragments, *k*-mer frequency counted and plotted as a scatter plot using dimension reduction techniques BH-tSNE. SAR11 species are in colour.

HDBSCAN was used to automate cluster determination and produce MAGs. Results were mixed with a majority of bins being unable to be resolved into species-level taxonomy. Overall, this produced no high-quality genomes and five medium-quality genomes from 11 bins.

Bin Id	Marker lineage	% Completeness	% Contamination	Number of reads as a percentage (1dp) and Identities
8	g__Staphylococcus (UID301)	67.68	1.09	*63% Staphylococcus aureus*
3	s__difficile (UID1169)	84.75	1.62	*76% Clostridioides difficile*
1	f__Vibrionaceae (UID4865)	87.2	19.55	*86% Vibrio parahaemolyticus* 30% HTCC7211 19% HTCC1062 *3% Bacillus subtilis* *5% Dehalococcoides mccartyi* *1% Ktedonobacter racemifer* *1% Clostridioides difficile*
4	p__Cyanobacteria (UID2143)	78.01	1.04	*75% Synechococcus elongatus* *1% Clostridioides difficile*
7	k__Bacteria (UID1452)	79.21	7.26	*55% Dehalococcoides mccartyi* *2% Staphylococcus aureus*
11	k__Bacteria (UID203)	91.38	131.49	*23% Ktedonobacter racemifer* *20% Bacillus subtilis* *23% Escherichia coli* *35% Staphylococcus aureus* *22% Clostridioides difficile* *14% Vibrio parahaemolyticus* *25% Synechococcus elongatus* *39% Dehalococcoides mccartyi* 34% HTCC7211 28% HTCC1062
10	k__Bacteria (UID203)	71.55	1.72	*37% Bacillus subtilis* *1% Clostridioides difficile*
2	k__Bacteria (UID203)	88.36	17.87	*77% Ktedonobacter racemifer* *3% Bacillus subtilis* *1% Escherichia coli* 2% HTCC7211 1% HTCC1062
5	k__Bacteria (UID203)	59.83	0	*76% Escherichia coli* *1% Ktedonobacter racemifer*
6	k__Bacteria (UID203)	72.18	27.65	52% HTCC1062 34% HTCC7211
9	k__Bacteria (UID203)	71.55	1.72	*37% Bacillus subtilis* *1% Clostridioides difficile*

Table 2.2 CheckM bin identities, contamination, completeness and actual read identities as a percentage of the genome. Each read is 10 kbp in size. SAR11 spp. are highlighted in red.Table 2.2 CheckM bin identities, contamination, completeness and actual read identities as a percentage of the genome.

2.3.6 Assessment of MAG quality by different bioinformatic methods

The results from the previously described bioinformatic method are tabulated for easy referencing. Results are displayed in the table below. No method was effective at producing SAR11 genomes from metagenomic data for downstream analysis.

Metagenome 3	BinSanity	MaxBin	MetaBAT2	HDBSCAN	Reference Mapping with HDBSCAN	Reference Mapping with BinSanity
Figure Number	2.3	2.4	2.5	2.14	2.16	2.17
Bins produced	159	69	12	65	419	112
HQ MAGs*	0	0	0	0	0	0
MQ MAGs**	1	1	2	1	0	0
SAR11 Genomes	0	0	0	0	0	0

Table 2.3 Summary table of different bioinformatic methods of obtaining MAGs from 80m BATS metagenomic (No.3). * High-Quality MAGs are defined having ≥90% completion and <5% contamination with 23S, 16S, 5S and at least 18 tRNAs. **Medium quality genomes are defined as having ≥50% completion and <10% contamination (Bowers et al. 2017).

The above process was repeated for metagenome twelve from this dataset consisting of a 200m sample. This dataset did not produce any MAGs taxonomically identified as SAR11 or any high-quality MAGs. This pipeline was not repeated for the other samples due to their similarity in read coverage and the percentage of reads recruited (**Table 2.1, Fig 2.3**).

Metagenome 12	BinSanity	HDBSCAN	Reference Mapping with HDBSCAN	Reference Mapping with BinSanity
Bins produced	318	58	4	14
HQ MAGs*	1	0	0	0
MQ MAGs**	8	0	0	0
SAR11 Genomes	0	0	0	0

Table 2.4 Summary table of different bioinformatic methods of obtaining MAGs from 200m BATS metagenomic (No.12). * High-Quality MAGs are defined having ≥90% completion and <5% contamination with complete 23S, 16S, 5S rRNA sequences and at least 18 tRNAs. **Medium-Quality genomes are defined as having ≥50% completion and <10% contamination (Bowers et al. 2017).

Low sequencing depth and read coverage were likely parameters that resulted in zero recoveries of SAR11 genomes. To combat this, all metagenomic samples from 80 metres were co-assembled into one large metagenomic assembly to assess if additional data provided additional coverage depth. The above protocol is repeated and results tabulated. This was not repeated for the 200m depths due to memory requirements exceeding computational resources, even when performed with low memory assembler Megahit (D. Li et al. 2015).

Additional sequencing material in a large co-assembled dataset helped marginally in producing more medium quality MAGs but did not produce any clear SAR11 MAGs. Scaffold coverage was still poor indicating additional genetic material did not improve scaffold coverage. Although higher numbers of bins were produced indicating a possible higher diversity, without a higher sequencing depth the combined assemblies could not be successfully separated into MAGs.

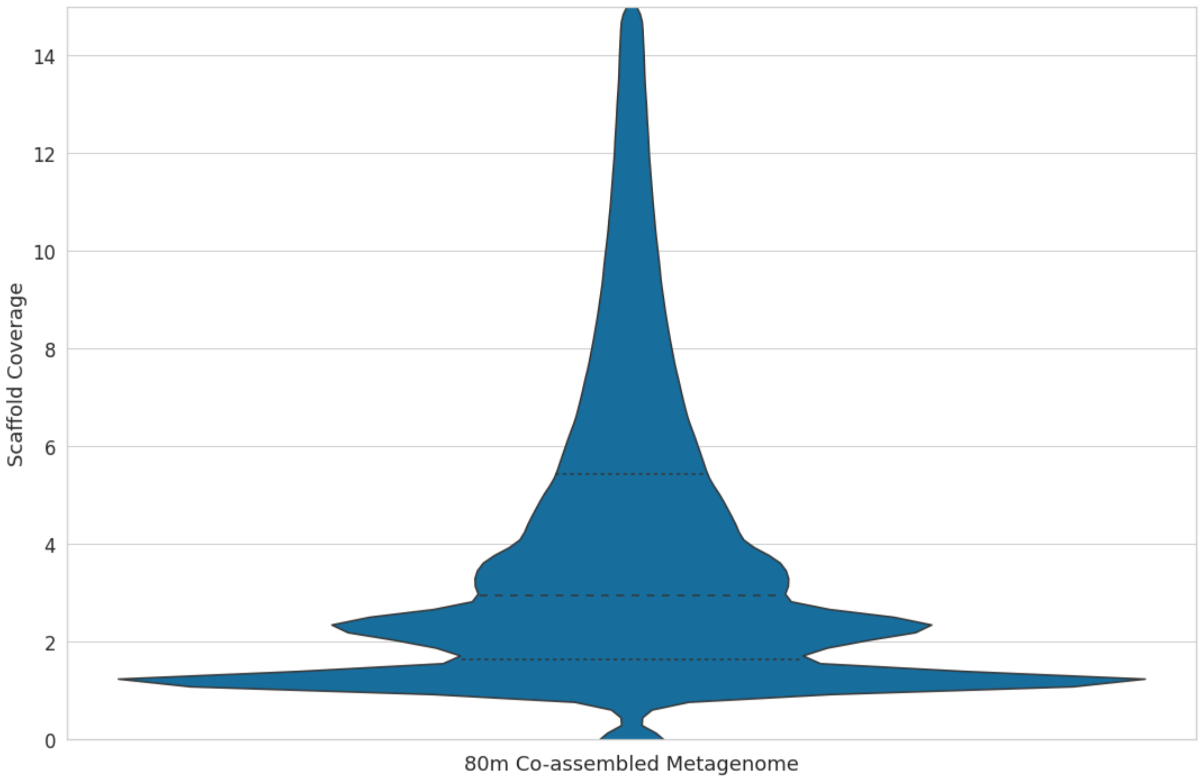


Figure 2.21 Violin plot of scaffold coverage derived from recruited reads from each 80m marine metagenome against a co-assembled 80m metagenome. Scaffolds with coverage sizes over 15 are not shown. The interquartile range of coverage dataset is indicated with dotted lines.

All 80m samples	BinSanity	HDBSCAN	Reference Mapping with HDBSCAN	Reference Mapping with BinSanity
Bins produced	1245	987	56	37
HQ MAGs*	0	0	0	0
MQ MAGs**	4	2	0	0
SAR11 Genomes	0	0	0	0

Table 2.5 Summary table of different bioinformatic methods of obtaining MAGs from a combined dataset of 80m BATS metagenomes. * High-Quality MAGs are defined having ≥90% completion and <5% contamination with 23S, 16S, 5S and 18 out of 20 tRNAs (not tested). **Medium quality genomes are defined as having ≥50% completion and <10% contamination (Bowers et al. 2017).

2.4 Discussion

2.4.1 Improving assembly quality

Overall, **Table 2.1** and **Figure 2.3** show assemblies were low in coverage and consist of hundreds of thousands of scaffolds. A median coverage depth around seven for 80m metagenomes and between 15-20 for 200m depths indicate rare taxa would be unlikely to be assembled. Assemblies also resulted in more than half of genetic material not included within assemblies as read recruitment back to assemblies ranged from 10 - 45%. Compared to the other marine metagenomic assemblies with on average two million contigs per metagenome ((Biller et al. 2018) and 2.4 million contigs per metagenome (Tully, Graham, and Heidelberg 2018), scaffold numbers were similar to other studies. However, these numbers still suggest highly fragmented assemblies which are difficult to bin and produce high-quality MAGs. An additional metric such as the average cellular metagenome size is around 750 Mb based on 610 metagenomes (Biller et al. 2018) or about 1.5 Gb from 248 metagenomes (Tully, Graham, and Heidelberg 2018). This was not reached with any of the samples but was hugely reduced in the 80m samples. Assembled metagenome size varied from 0.0003 to 0.385 Gb in size in comparison to the 0.75 - 1.5 Gb range, indicating a huge loss in expected size. Only when assemblies were co-assembled together based on the same depth did metagenome sizes match those seen in other studies (80m combined = 1.5 Gb). This could be the result of the low biological diversity at the BATS, but is more likely a result of shallow sequencing depth. Read quality pipelines and assembly follow well established and successful assembly pipelines, therefore, this likely points towards a shallow sequencing depth over the bioinformatics processes.

2.4.2 Binning software results

Despite numerous different methods to bin MAGs from the metagenomes, no method produced more than one MAG of high quality (**Table 2.4-6**). Both metaBAT2 and Maxbin produced no MAGS of medium or high quality, nor indicated that MAGs may contain SAR11 species. BinSanity produced one medium quality MAG in an 80m metagenome and eight in a 200m one, all of unknown taxa. With binning algorithms suggesting at least 60 bins if not more (**Table 2.4, 2.5**), this indicates at maximum ~1.5% high-quality genomes and ~15% of medium quality genomes were returned per metagenome. Additionally, a lack of any kind of consensus between binning software and taxonomy made it impossible to improve assemblies when trying to remove contaminants via anvi'o (**Fig 2.7-9**). Nor would this process be automatable due to the human input requirements and would likely be time-consuming, subjective and non-reproducible. With between 60 to 300 bins being produced per metagenome, substantial time would be needed to refine each bin with likely few improvements to create higher quality MAGs.

In comparison, a TARA Ocean assembly study assembled 420 high-quality genomes and 2028 medium quality genomes from 248 marine metagenomes using an iterative process of BinSanity (Tully, Graham, and Heidelberg 2018). This equates to roughly 1.5 high-quality MAGs and eight medium quality MAGs per metagenome. This is in line with binning results from a 200m metagenome, indicating pipelines used within this report achieved similar results to other research groups. However, data for low-quality bins and other unsuccessful bins were not provided and so could not be compared against. No binning of (Biller et al. 2018) metagenomes were performed in any study to compare. All metagenomic binning followed identical pipelines, with a substantial difference in MAG recovery rates between 80m and 200m metagenomes. This likely points towards a non-bioinformatic issue that led to low numbers of MAG recovery.

2.4.3 Binning by *k*-mer counting

Both the default (**Fig 2.10**) and high settings of n_neighbours did little to increase the number of bins within the UMAP plot. This is likely due to data points producing a similar global structure and would indicate that *k*-mer counts were unlikely to be highly dissimilar from each other. Therefore, clustering data based on higher n_neighbours would not be effective. Although there would be some clustering, this would likely be from distinct phyla, for example, archaea and bacteria. A lower n_neighbours (**Fig 2.14**) was more effective at producing a larger number of bins with lower contamination values. Additionally, a lower n_neighbours revealed clear cluster boundaries than with a higher n_neighbours, although without any large clusters. However, it should be noted that a low n_neighbours only produced one genome of medium quality, in-line with the best binning algorithm BinSanity that had the same results within a low coverage 80m metagenome (**Table 2.3**). Binning by *k*-mer counting was less effective in high coverage 200m metagenomes (**Table 2.4**). This would suggest this method is effective in datasets where coverage is not used to further bin MAGs.

Methods for *k*-mer counting assemblies were further explored for recovering MAGs (**Table 2.3-5**). Bacteria from across the tree of life were randomly chosen and genomes *k*-mer counted. This was performed to verify the *k*-mer counting methodology and indicate if *k*-mer counting is more applicable for certain bacterial metagenomes. Initial results indicated methods involving *k*-mer counting are ineffective at binning MAGs compared to specialised binning algorithms. CheckM of clusters (**Table 2.2**) showed that *k*-mer counting is only somewhat effective at clustering different species genomes. It is clear that *k*-mer counting is not the perfect solution as there exist variations within *k*-mer counts throughout the genome with even fragments 10 kb in size. Recovery of MAGs was also dependant on species as SAR11 organisms HTCC7211 and HTCC1062 had a more varied *k*-mer count in comparison with *Clostridium difficile,* the genome of which had a distinctly different *k*-mer count and clustered mostly within Bin 3 (**Table 2.2**). Therefore, *k*-mer counting is more effective for bacteria that are distantly related, suggesting that *k*-mer counting may not be a suitable metric for separation of SAR11 species from marine metagenomes. It should also be noted here that CheckM produced contamination scores for Bin 8 and 3 (**Table 2.2**), despite containing pure isolates of only one organism, and therefore should have contamination values of zero. Therefore, metrics that rely on CheckM should be used as an estimate for genome completion and redundancy rather than accurate metrics.

K-mer counting involving higher numbers of k should be explored if computational resources allow, along with an iterative approach to binning. Success has been shown in viromes where Average Nucleotide Identity (ANI) is used to separate bins with multiple MAGs (Beaulaurier et al. 2019). Additionally, *k*-mer counting is likely to be more effective with longer contigs and reads as *k*-mer frequencies become inaccurate with smaller contigs. This is not recommended by most bioinformatics software with anvi'o recommending 2500 base-pairs and (Beaulaurier et al. 2019) recommends 15,000 base-pairs. Therefore, with long-read metagenomes, *k*-mer counting is likely to be a more effective method of binning compared to short-read assemblies.

2.4.4 Binning by reference mapping

Recruiting reads against organisms of interest and subsequent assembly and binning did not improve MAG recovery (**Fig 2.17, 2.18, Table 2.3**). This was effective in low diversity metagenomes (Sharon et al. 2013) but likely to be ineffective with a metagenome majority of unknown (**Fig 2.3**) and genetically similar species (**Fig 2.11**). BinSanity or binning by *k*-mer counting was unsuccessful as large numbers of bins were produced with no MAGs of high or medium quality. This may indicate an order level of metagenome mapping may still result in many genetic similarities (**Fig 2.16**). This process could be further explored by mapping based on taxonomy, but with 40-60% of a metagenome having no taxonomic classification and an additional 20% with no taxonomy beyond an order level, this would result in only 20 - 40% of a metagenome with a known taxon and would discard a majority of a metagenome.

2.4.5 Combining similar depth samples to increase sequencing coverage depth

Co-assemblies of an 80m metagenome to increase sequencing depth produce metagenomes of a similar size to other studies (Biller et al. 2018). However, binning performed only marginally better with the recovery of four medium quality MAGs (**Table 2.5**) over one medium quality MAG (**Table 2.3**). Additionally, compared to a metagenome 12 (**Table 2.4**) of higher sequence depth but less material, the 80m co-assembled metagenome performed more poorly as one high quality and eight medium quality genomes were produced. Therefore, I concluded that co-assemblies of low sequencing depth can improve MAG recovery, but only marginally. Higher sequencing depth metagenomes should be prioritised instead.

2.4.6 MAG creation improvements

Overall, the Pelagibacterales order was present and abundant within marine metagenomes within this study yet current state of the art algorithms were unable to produce any MAGs belonging to the Pelagibacter taxonomic group. As a result, analysis of SAR11 MAGs for SAR11 phages was not possible within this chapter. A lack of recovery of SAR11 MAGs may suggest that their high abundance and microdiversity inhibits effective bining, a problem in the assembly of SAR11 phage genomes (Warwick-Dugdale et al. 2019). This would indicate that co-occurrence of closely related strains leads to multiple paths when assembling reads (Olson et al. 2019) resulting in fragmentation of poor representation in marine MAG recovery studies (Delmont et al. 2018; Tully, Graham, and Heidelberg 2018; Chen et al. 2020). Longer contigs may assist assembly of MAGs and derive a higher consensus between binning software (Suzuki et al. 2019; Pearman, Freed, and Silander 2019; Quick 2019; Somerville et al. 2019), shown effective with viral genomes in both a hybrid (Warwick-Dugdale et al. 2019) and long-read only strategy (Beaulaurier et al. 2019). Long read technology should be explored as an alternative method for MAG assembly, assuming the DNA extraction method did not fragment DNA, common in physical extraction protocols (Quick 2019). Obtaining longer reads should be prioritised over coverage. Higher coverage of short reads may assist but, with indications of marine samples being highly similar, this may lead to reads being misassembled into chimeric contigs. Instead, it is recommended for long reads to be used to establish contigs and short reads used to polish long read assemblies for a higher consensus accuracy (Vaser et al. 2017; Kundu, Casey, and Sung 2019). This would prevent mis-binning due to abnormal *k*-mer frequencies or erroneous reads. Longer reads may also allow for binning of reads instead of contigs, preserving the complexity of a metagenome without data loss through the assembly process. This is a much less computationally intensive step and may produce better quality assemblies.

Overall the evidence suggests that sequencing depth is most probably responsible for being unable to produce any high-quality MAGs from this dataset. Instead, attempts should be focused on producing more sequence data for the binning of a metagenome into MAGs. Long reads are also a promising avenue where error-corrected long reads could be binned without prior assembly. Binning software like BinSanity and MaxBin are also more effective at producing prokaryotic MAGs when coverage data is available compared to *k*-mer counting methods. However, even with a high sequencing depth of short-reads, binning does not improve with current bioinformatics algorithms. Therefore, new algorithms or wet-lab methods, such as single-cell amplified genomes, are still needed to differentiate between organisms in highly diverse and low coverage metagenomes. Single-cell amplified genomes allow for sequencing of cells individually, removing the need for bioinformatic binning algorithms. This was problematic within this chapter and allows for the study of organisms without the fear of contamination. This method was pursued throughout the rest of this project to analyse SAR11 genomes and their phages.

3 SAR11 Single-cell Amplified Genome Phylogenetics and Ecological analysis

3.1 Introduction

3.1.1 Abstract

SAR11 are one of the most ubiquitous microorganisms in the ocean, yet population-wide studies remain elusive due to their difficulty in culturing and extraction from metagenomic data resulting in relatively few genetic sequences being publically available. Here I describe 451 novel SAR11 SAGs obtained from various TARA Ocean cruises allowing for population-wide studies. Phylogenetics revealed two new SAR11 clades, broadening our understanding of the genetic variation within the SAR11 clade. Metagenomic analysis revealed these two novel clades inhabit the bathypelagic and abyssopelagic zone similar to that of the Ic clade of SAR11 - previously the only SAR11 clade to inhabit the bathypelagic. These 451 novel SAR11 SAGs provide the largest contribution of SAR11 SAGs as a community resource for additional studies.

3.1.2 Where were the SAGs obtained from

451 SAR11 SAGs were generated as part of the Tara Oceans and Tara Oceans Polar Cruises (Pesant et al. 2015), a scientific research expedition sampling ecosystems of the world's oceans at differing times and depths. Confirmation of SAR11 phylogeny was performed through 16S rRNA screening performed with primers 27F-907RM at Bigelow's Single Cell Genomics Center. SAGs confirmed as SAR11 were whole genomes sequenced at Genoscope producing 2 x 100 bp paired-end reads.

Plate ID	Station	Expedition	Region	Layer	Number of SAR11 SAGs
AAA-536	TARA_023	Tara Oceans	Mediterranean sea	DCM	30
AD-623	TARA_078	Tara Oceans	South Atlantic Ocean	SUR	17
AD-625	TARA_084	Tara Oceans	Southern Ocean	SUR	9
AD-627	TARA_085	Tara Oceans	Southern Ocean	SUR	37
AG-943	TARA_163	Tara Polar	Atlantic Arctic	SUR	26
AG-946	TARA_175	Tara Polar	Atlantic Arctic	SUR	34
AG-948	TARA_201	Tara Polar	Arctic Archipelago	SUR	31
AG-984	TARA_102	Tara Oceans	Pacific Ocean	DCM	44
AG-988	TARA_102	Tara Oceans	Pacific Ocean	MES	30
AG-989	TARA_102	Tara Oceans	Pacific Ocean	MES	20
AG-997	TARA_111	Tara Oceans	Pacific Ocean	MES	26
AG-987	TARA_102	Tara Oceans	Pacific Ocean	MES	27
AG-998	TARA_111	Tara Oceans	Pacific Ocean	MES	28
AG-993	TARA_111	Tara Oceans	Pacific Ocean	DCM	59
AG-996	TARA_111	Tara Oceans	Pacific Ocean	MES	33

Table 3.1 SAR11 SAG metadata table detail location each SAG was isolated during the TARA Oceans and Polar cruises. DCM = Deep Chlorophyll Maximum, SUR = Surface, MES = Mesopelagic

3.1.3 Phylogenetics

Population studies allow characterisation of a group of organisms based on their relatedness and provide insight into how the selection pressures from environments impact populations. This is done through using phylogenetics - the measure of how related a species is to another represented on a phylogenetic tree (Semple, Steel, and Both in the Department of Mathematics and Statistics Mike Steel 2003). Similarly grouped species sit within the same “branch” of a phylogenetic tree, and more distantly related species are then located further away (Nei and Kumar 2000). To establish relatedness, conserved genes within all species, for example, the 16/18S rRNA coding regions are used (Fox et al. 1977). Without conserved genes, it would be difficult to measure species relatedness to compare similarity on a genomic level.

More closely related organisms are likely to have multiple similar if not the same genes, as such genes are required to perform a similar function within an organism. This provides additional genes to assess species relatedness which can be used in conjunction to provide a broader assessment of relatedness (J. H. Campbell et al. 2013; Alneberg et al. 2014; Dupont et al. 2012; Creevey et al. 2011). However, where one clade of organisms may exist on multiple branches due to a large genetic diversity, characterisation of clade boundaries may prove difficult and are assisted with the use of ingroups and outgroups.

To confirm the taxonomic identity of samples within a phylogenetic tree, ingroups can be used as positive controls and represent the species as a whole (Semple, Steel, and Both in the Department of Mathematics and Statistics Mike Steel 2003). Ingroups are genetically similar to the expected samples and act as 'anchors' within the group. They are known quantities within the tree of life and samples would be expected to form on the same branch or in close proximity. These are used to confirm sample identity and perform an inverse function to outgroups.

Outgroups are organisms defined as distantly related to the samples in question and are used to visualise the tree in the existing tree of life (Lyons-Weiler, Hoelzer, and Tausch 1998). They can be used as reference points, linking the tree within the bigger picture of existing life forms. These can also be used as negative controls where organisms would not be expected to sit within the same branch. Outgroups are picked on some related genetic similarity but would not be regarded as the same taxa of tested individuals. For example, a phylogenetic tree of a species taxonomic level, an outgroup would be an organism of the same genus but different species. Organisms that are located on the same branch or similar to outgroups indicate samples may be contaminated, or less closely related to samples than previously thought. This acts as quality control for sequences and helps to ‘anchor’ trees. However, uncertainty can arise in the position of organisms within a tree. How sure are we that organisms have been assigned to the correct branch on a phylogenetic tree? Bootstrapping can be used as a confidence metric to ascertain how certain a branch within a tree is correctly placed.

Bootstrapping is performed by replicating subsections of the data and recording how often a branch is placed in the same location (Felsenstein 1985; Efron 1979). Generally, a smaller proportion of the data is taken and the branch location is recorded. This can be repeated as many times as possible and is normally expressed as a percentage of times the branch occurred in the same place using a subset of the data. This provides a metric of the confidence of the location of the branch based on smaller variations with the data. Branches below a certain percentage can be collapsed to reflect low confidence in a branch positioning. Accepted bootstrapping values expressing confidence of a branch position are subjective and mostly depend on the data provided. With data of low resolution or with multiple gaps, a bootstrapping value of 33 would indicate that only a third of replicates indicate this structure. However, with a multi-gene phylogeny with thousands of amino acid alignments (Chaumeil et al. 2019), a bootstrapping value of greater than 95 would indicate confidence that a phylogenetic branch is correctly placed. A bootstrapping value below 50 is generally regarded as ambiguous and collapsed as only 50% of the time subsetted data confirmed branch placement. However, trees could be constructed where all species could sit on their own branch and therefore have high bootstrapping values and show no relatedness. This would not be an accurate reflection of a population as it would be expected that some species would be related. Therefore maximum parsimony (MP) or maximum likelihood (ML) is used to determine the most likely tree structure.

Parsimony is a principle similar to Occam’s razor, - if all else is equal, the simplest explanation is more likely to be correct in comparison to more complex solutions (Edwards 1996). MP in phylogenetics follows this principle where the tree with the simplest branching solutions is probably the most correct one (Steel and Penny 2000). Since it is challenging to calculate the most parsimony tree or the maximum parsimony for a dataset, multiple trees using different models are made and the parsimony calculated for each. The tree with the highest parsimony score is then selected and used as the final tree (Nguyen et al. 2015). However, because parsimony rewards the smallest and simplest of trees, this may not accurately represent the true relatedness of a clade as there may exist a more complex relationship. Other models exist like ML where the selection of the best tree coincides with the variations that occur in a set of aligned sequences (Edwards, n.d.). It calculates the probability of the aligned sequences resulting in the final tree structure, where the highest probability results in the most probable tree model. However, both methods are estimations to the behaviour of genetic evolution and provide only approximations to the phylogenetic structures of biological entities.

3.1.4 Multiple Sequence Alignments

A Multiple Sequence Alignment (MSA) is a string of characters aligned according to each other's order. This is usually done to allow the comparison of the same gene from multiple samples (Phillips, Janies, and Wheeler 2000). For example, 16S rRNA coding regions are isolated from samples and its nucleotide composition is aligned against all other sample's 16S rRNA (Fox et al. 1977). Variations and similarities within the nucleotide sequence would reveal relatedness, where samples with a similar composition of nucleotides expected to be increasingly similar and vice versa. Alternatively, amino acid sequences can be used instead as protein-coding regions which are translated into their amino acid composition and aligned against each other (Sievers et al. 2011). This normally results in less variation within closely related species as variations within nucleotides compositions can still result in the same amino acid translation. This is because amino acids have more than one translated codon. However, smaller changes within nucleotides, such as silent mutations, are not captured and may indicate slight differences between similar strains or artefacts of sequencing error. Therefore, to reduce variation MSA can include multiple different protein-coding regions to allow for additional comparisons across multiple genes and a broader analysis (Chaumeil et al. 2019). Although this requires larger computational resources, it is more sensitive and will allow for the detection of smaller genetic changes that may result in a more accurate representation of phylogenetic groups. Lastly, 16S rRNA sequencing is well established with multiple primers available to amplify specific locations of the gene. Whole-genome sequencing is reliant on the sequencing of all conserved genes between all tested species to provide comparison points.

2.1.5 Average Nucleotide / Amino Acid Identity Clustering

Average Nucleotide Identity (ANI) or Average Amino acid Identity (AAI) is a metric for determining how similar two genetic sequences are based on either nucleotides or their translated amino acids (Konstantinidis and Tiedje 2005). ANI/AAI are useful metrics for clustering or delineating strains from each other. They work using *k*-mer frequencies, calculating the frequency of nucleotides or amino acids within a sequence and producing a percentage relatedness to other compared sequences (D. Parks 2018; Medlar, Törönen, and Holm 2018). Sequences are clustered if they have similar percentage relatedness to all members of the cluster. This becomes an alternative method to phylogenetics to establish strain relatedness and may help to support branch splitting (Alnajar and Gupta 2017; Santos-Garcia et al. 2017; C. Jain et al. 2018b). AAI is the preferred method of cluster determination, shown effective to at least 50% similarity (Rodriguez-R and Konstantinidis 2014; Qin et al. 2014) with ANI only effective at a species level at least 95% (Rodriguez-R and Konstantinidis 2014) and 85% at a genus level (Rodriguez-R and Konstantinidis 2014; Qin et al. 2014; Richter and Rosselló-Móra 2009; Chung et al. 2018). However, AAI is reliant on the accurate translation of protein sequences and relies on accurate gene calling algorithms.

3.1.5 Metagenomic mapping for geographical range

Ecological mapping of the host range can be performed by mapping of genetic material against metagenomic samples. Here, metagenomes of the (Biller et al. 2018) study are recruited against SAGs. This comprises of five terabases of metagenomic data from 610 sampling sites across a range of depths and times within the ocean over roughly a four year period, 2003-04 and 2009-2011. Each metagenome is a representation of the microbial community at that location and time. Mapping of metagenomic reads against SAGs would indicate if the mapped genetic content was present within that metagenome, and if so, provide evidence for the species' existence within that ecological niche. Using this methodology, it would then be possible to see if different phylogenetic clades inhabit different areas of the ocean and therefore infer that these belong to different ecotypes.

3.2 Materials and Methods

3.2.1 Read preprocessing

Resulting reads were checked for quality using an adapted quality control script. This script is located within bbmap/pipelines/assemblyPipeline.sh was developed by Brian Bushnell from JGI using the BBTools suite of bioinformatics tools for preprocessing of Illumina reads before assembly. Briefly, this script removes duplicate reads and low-quality regions, trims adaptors and synthetic artefacts and performs error correction. All settings follow the “Usage Examples” within the BBMap online manual.

3.2.2 SAG assembly and quality assessment

After quality control, reads were assembled using SPAdes (Nurk et al. 2013) in single-cell mode. *k*-mers of intervals of 10 from 11 to 99 were used in conjunction with mismatch error correction mode. Resulting assemblies were assessed for quality via CheckM (D. H. Parks et al. 2015). The single marker gene set of class Alphaproteobacteria was chosen over the family Pelagibacteraceae for single-copy marker genes as only 15 representative genomes of Pelagibacteraceae were available compared to 648 for Alphaproteobacteria. The 15 representatives genomes are dominated by Clade I SAR11 and may not encompass the entire pangenome of the SAR11 clade. Comparing 451 SAR11 to these reference genomes may, therefore, provide false negatives. No preferred choice of the order Pelagibacterales was available. Percentage completeness was used as the main metric to define the quality of each assembly as well as percentage contamination, GC percentage deviation and N50.

```
# Run SPAdes in single cell mode.
spades.py --sc -t 16 -m 128 --careful \
-k 11,21,33,43,55,65,77,87,99 \
--pe1-1 <fwd_read> --pe1-2 <rev_read> -o <output_folder>
```

```
# Run CheckM with custom marker set
checkm taxon_set class Alphaproteobacteria \
Alphaproteobacteria_markers

checkm analyze Alphaproteobacteria_markers \
<folder_with_SAGs_assemblies> checkm_output -x fasta -t 16

checkm qa --out_format 2 --file \
checkm_output/completeness_stats.txt --tab_table -t 16 \
Alphaproteobacteria_markers checkm_output
```

3.2.3 Construction of a 16S rRNA phylogenetic tree

Phylogenetics was used to first describe the relationship of each SAG to each other. 16S rRNA sequences were extracted using Barrnap (T. Seemann 2015) with default settings and parameters. An MSA file was created with MAFFT (Katoh et al. 2002) a multiple sequence alignment algorithm based on progressive alignment, starting by comparing similar alignments and progressively adding more distantly related sequences. The options --globalpair and --maxiterations 1000 were used as this was predicted to be similarly related species and therefore global over a local alignment was performed. 1000 iterations were also preferred to ensure the best MSA outcome was used whilst still being computationally reasonable.

TrimAl (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) was used to remove poorly aligned regions within an MSA which can help with the accuracy of future phylogenetic tree analysis and reduce computational time. TrimAl was used with the -automated1 flag to computationally deduce its best options. Lastly, MSA files were plotted into phylogenetic trees through IQ-TREE (Nguyen et al. 2015). Settings for the usage of IQ-TREE include -bb 1000 to allow for 1000 rounds of bootstrapping and -m MFP which allows for IQ-TREE to find the most parsimonious tree using its inbuilt ModelFinder algorithm. Resulting trees were generated into the Newick format and visualised in iTOL (Letunic and Bork

2007), an online phylogenetic tree viewer that allows detailed customisation of branch colours and trimming of lengths.

3.2.4 Construction of a multi-gene MSA

For a multi-gene phylogeny, the GTDB-Tk pipeline (Chaumeil et al. 2019) was used to produce a multi-gene MSA. GTDB-Tk identifies conserved marker genes by all bacteria/archaea through prodigal (Hyatt et al. 2010), aligning marker genes to Hidden Markov Models (HMM) using HMMER (Eddy 1998) and concatenation of these marker genes into an amino acid MSA. Construction of trees via IQ-TREE was performed allowing for selection of the most parsimony tree. TrimAl was used with the -automated1 flag to computationally deduce its best options. IQ-TREE was used to deduce the phylogenetic tree from the trimmed MSA. Settings for the usage of IQ-TREE include -bb 1000 and -m MFP.

Reference genomes were taken from existing SAR11 clade literature (Grote et al. 2012) and included in this analysis to anchor branches within existing phylogenetic trees acting as ingroups and outgroups. Branches with lower bootstrapping values of less than 90 are collapsed to represent areas of ambiguous phylogeny.

3.2.5 Usage of ANI and AAI for fine-scale resolution of genome similarities

Average Nucleotide Identity (ANI) is a metric where two or more nucleotide sequences are compared and similarity deduced based on their composition (Konstantinidis and Tiedje 2005). An all-vs-all ANI search was performed on all SAGs and all publically available SAR11 genomes using FastANI. FastANI (C. Jain et al. 2018a) is a bioinformatics tool that calculates the ANI of protein-coding regions using *k*-mer counting. FastANI was used with default settings and plotted by a heatmap using this script.

```
#Run FastANI with ref_list.txt as path to FASTA files
```

```
fastANI --rl <ref_list.txt> --ql <ref_list.txt> -o
fastANI.output -t 16 --matrix
```

Average Amino acid Identity (AAI) uses translated protein sequences instead of nucleotides to provide a similarity metric. AAI was calculated with only 451 SAR11 SAGs due to computation limits using CompareM (D. Parks 2018) with the default settings. Pairwise AAI results were compared to WGS phylogenetic trees of SAGs and all publically available SAR11 genomes.

```
#Run AAI using CompareM
comparem --cpus 32 aai_wf <all_sags.fasta> <outdir>
```

3.2.6 *k*-mer counting

All 451 SAR11 SAGs and reference genomes from (Grote et al. 2012) were *k*-mer counted and their *k*-mer frequency normalised against contig length using scripts developed by (Beaulaurier et al. 2019). Resulting *k*-mer counting genomes were plotted after dimension reduction technique BH-tSNE via UMAP (McInnes, Healy, and Melville 2018). This allowed for *k*-mer counted genomes to be plotted on a 2D scatter plot using matplotlib (Hunter 2007) to determine clade clustering based on *k*-mer frequency.

3.2.7 Metagenomic mapping for ecological identities

All SAR11s that were categorised into a clade were mapped against the (Biller et al. 2018) dataset using Bowtie2 (Ben Langmead and Salzberg 2012). Only reads mapping at greater than 95% ANI were kept as 95% is generally regarded as the species delimitation for genetic samples (Qin et al. 2014; Thompson et al. 2013). The script for this process can be found here. The average percentage coverage of the SAR11 genome was used as the statistic to infer relative abundance with a metagenome. To infer clade coverage, the median percentage of clade coverage was used to infer clade wide presence-absence. Results were displayed in Ocean Data Viewer (Schlitzer 2002).

3.3 Results

3.3.1 SAR11 SAG genome quality

451 SAR11 SAGs were assembled with SPAdes in single-cell mode and checked for completeness using CheckM. Genome quality metric was suggested from (Bowers et al. 2017).

Genome Quality	Number of SAR11 SAGs
High-quality >90% complete, <5% contaminated	4
Medium Quality ≥50% complete, <10% contamination	272
Low-quality <50% complete, <10% contamination	175

Table 3.2 Completion table of 451 SAR11 SAGs according to completion statistics by (Bowers et al. 2017).

3.3.2 SAR11 Phylogenetics

Extraction of 16S rRNA sequences from SAGs and reference genomes from (Grote et al. 2012) correlated with existing SAR11 phylogenetic trees. Only 373 SAR11 SAGs tested positive for the 16S rRNA gene, losing 78 SAR11 SAGs from this phylogenetic tree. All SAGs were closely associated with existing SAR11 reference genomes. Interestingly, very few SAGs were associated on the same branch as Clade IV and none on Clade V (**Fig 3.1, 3.2**).

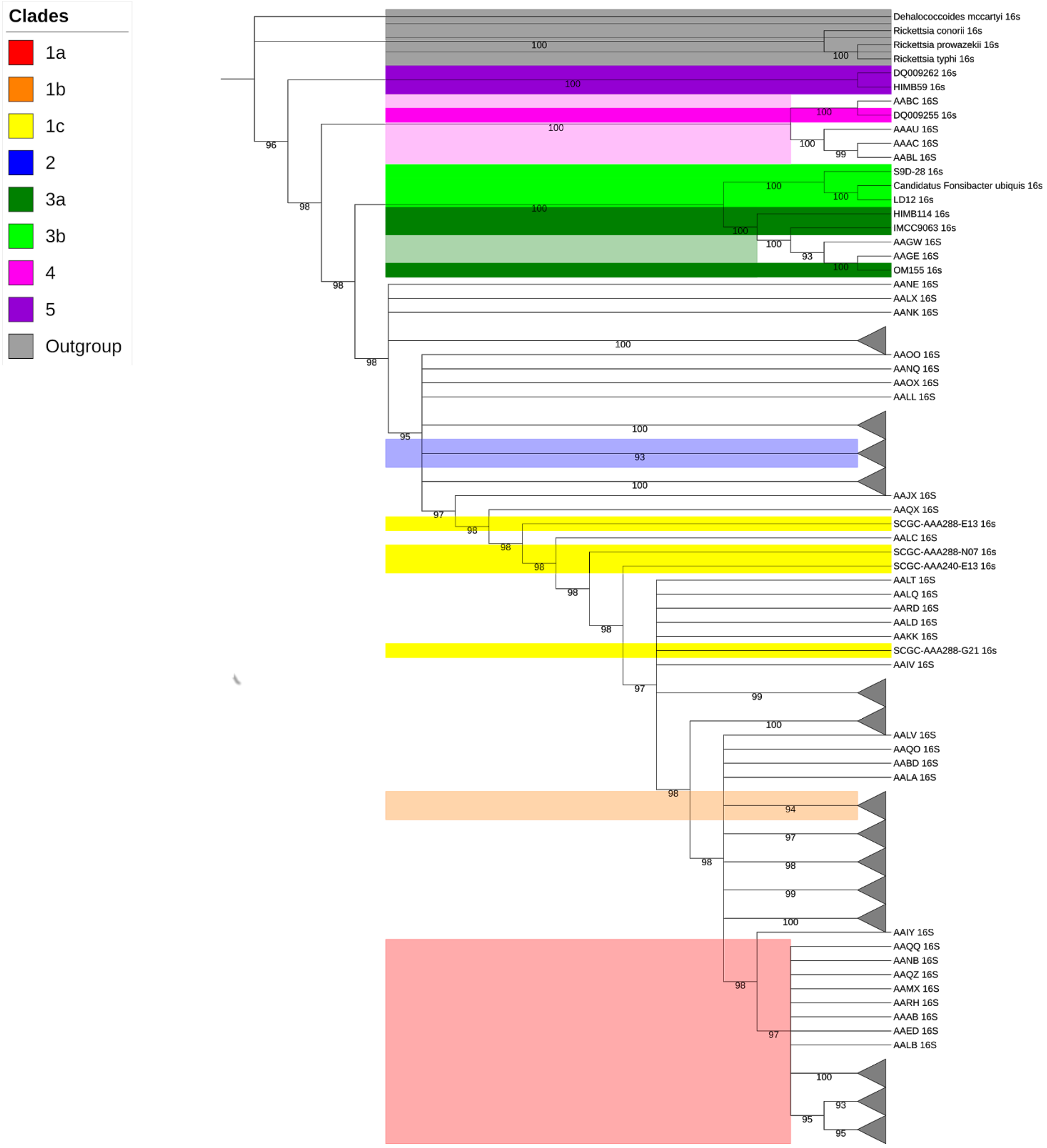


Figure 3.1 A collapsed 16S rRNA phylogenetic tree of 373 SAR11 SAGs. Clades are labelled in colour with bold colours indicating a reference 16S rRNA from the NCBI database. Lighter colours indicate SAR11 SAGs that are probable members of their respective coloured clade. Nodes are collapsed to aid visualisation, with grey triangles indicating a collapsed node containing additional branches. Nodes with bootstrap values under 90 are collapsed to reduce ambiguity. The outgroups are coloured in grey.

WGS phylogenetics with all SAR11 SAGs and publically available SAR11 genomes correlate with existing literature (Grote et al. 2012; Thrash et al. 2014; Tsementzi et al. 2016; Kraemer et al. 2019). Five SAR11 SAGs did not have enough genomic data for a WGS phylogenetic tree and are not included. Clades Id and Ie are previously undescribed clades.

High bootstrapping numbers on the WGS SAR11 tree provide confidence about its structure, but without an ingroup for clade IV it becomes difficult to determine if SAR11 genomes are from an existing group or form a new one. However, when combined with a 16S phylogenetic tree with a clade IV ingroup, I can compare the two and confirm the identity of the clade. In this case, clade IV is highlighted in brown (**Fig 3.1**). Areas of white indicate areas of the phylogenetic tree that cannot be assigned to a group as splits occur upstream from reference genomes. This probably indicates novel or new clades as existing phylogenetic tree clusters cannot explain splits at this location. Therefore it is likely that two new clades are present, clade Id with 43 members and clade Ie with 34 members.

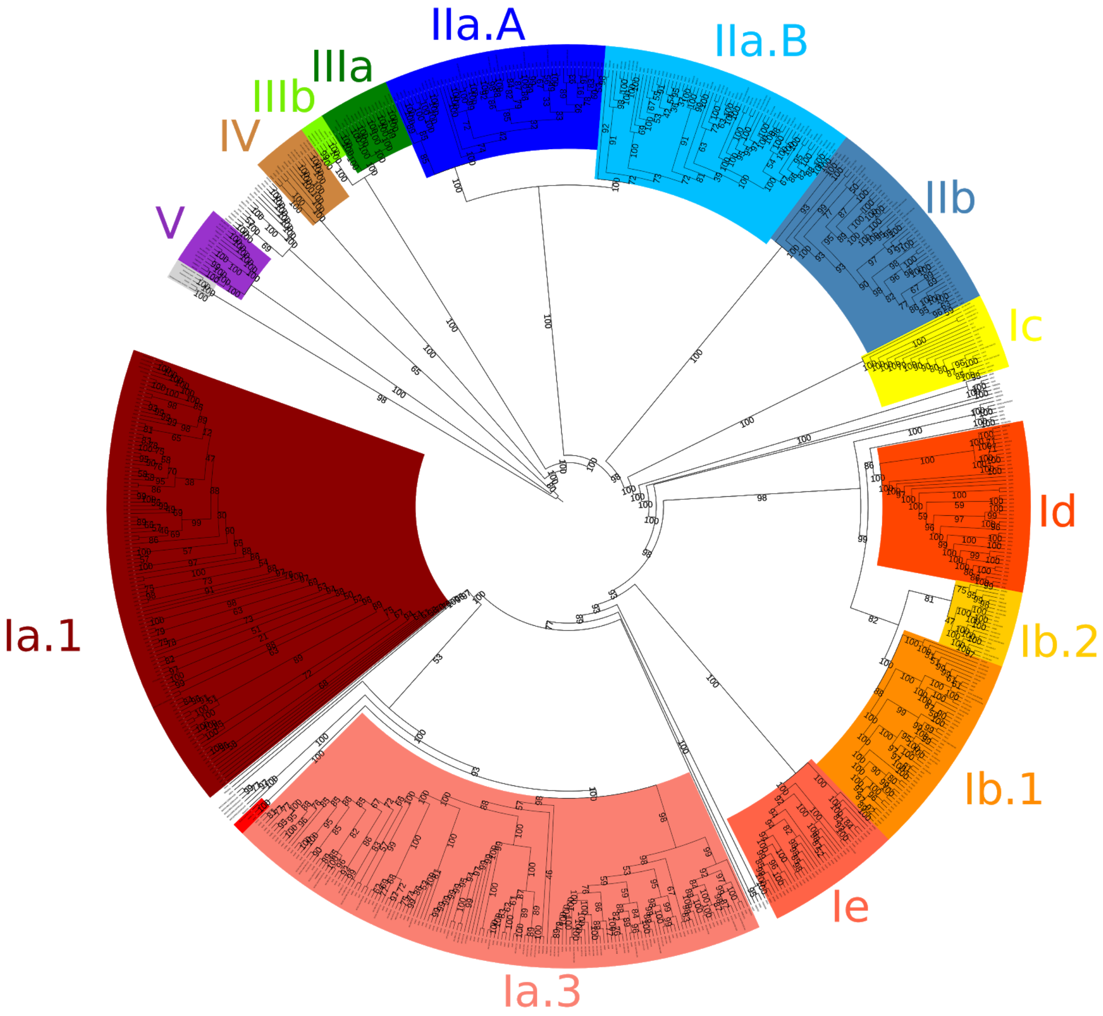


Figure 3.2 A WGS phylogenetic tree of 736 SAR11 genomes which include SAR11 SAGs from this study as well as from public databases (NCBI). Clades are displayed in colour with bootstrapping numbers and clade numbers.

3.3.3 SAR11 ANI and AAI

An all-vs-all ANI heat map provided indications that the data is structured, revealing clustered groups. However, differentiating clusters (representative of different clades) from each other by eye using visual representation proved difficult and ambiguous.

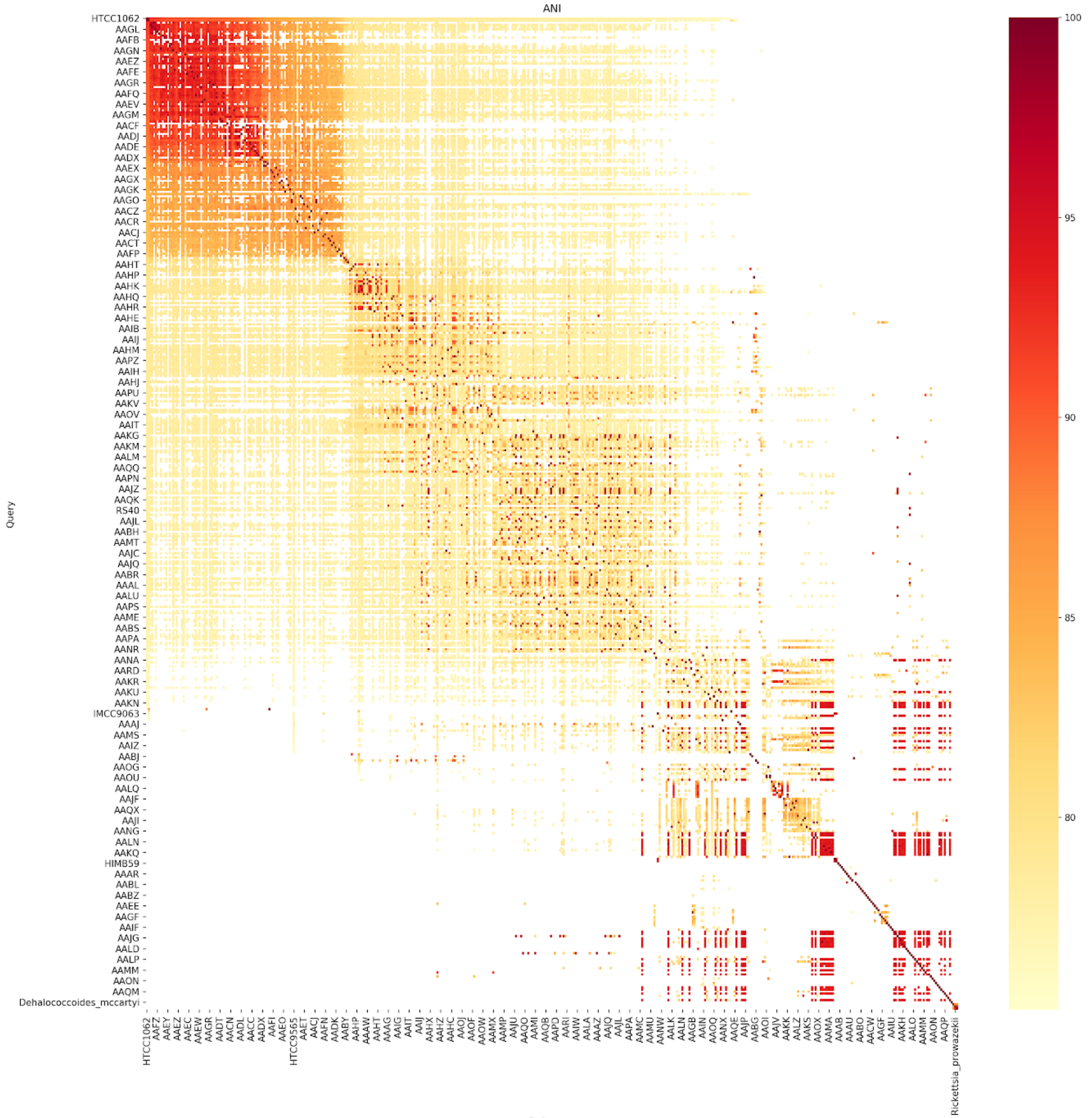


Figure 3.3 A condensed selection of an ANI heatmap plot of an all vs all SAR11 SAGs and reference genomes showing the overall structure of the data. A darker colour indicates a higher ANI and lighter colour a lower ANI on a discontinuous scale. ANI values below 75 are not shown and are displayed in white. Query genomes are labelled on the x-axis, where reference genomes are on the y-axis. Clustering of a darker colour indicates a grouping of similar SAR11 SAGs ANI values. Reference genomes are used to provide identity to clusters. Enlarged plot.

AAI of 451 SAR11 SAGs revealed clustering of SAR11 AAI values based on phylogenetic splits. Reference genomes do not have AAI calculated due to computational limits and therefore have no values. Clear clusters can be seen confirming clade boundaries based on differing genome coding regions.

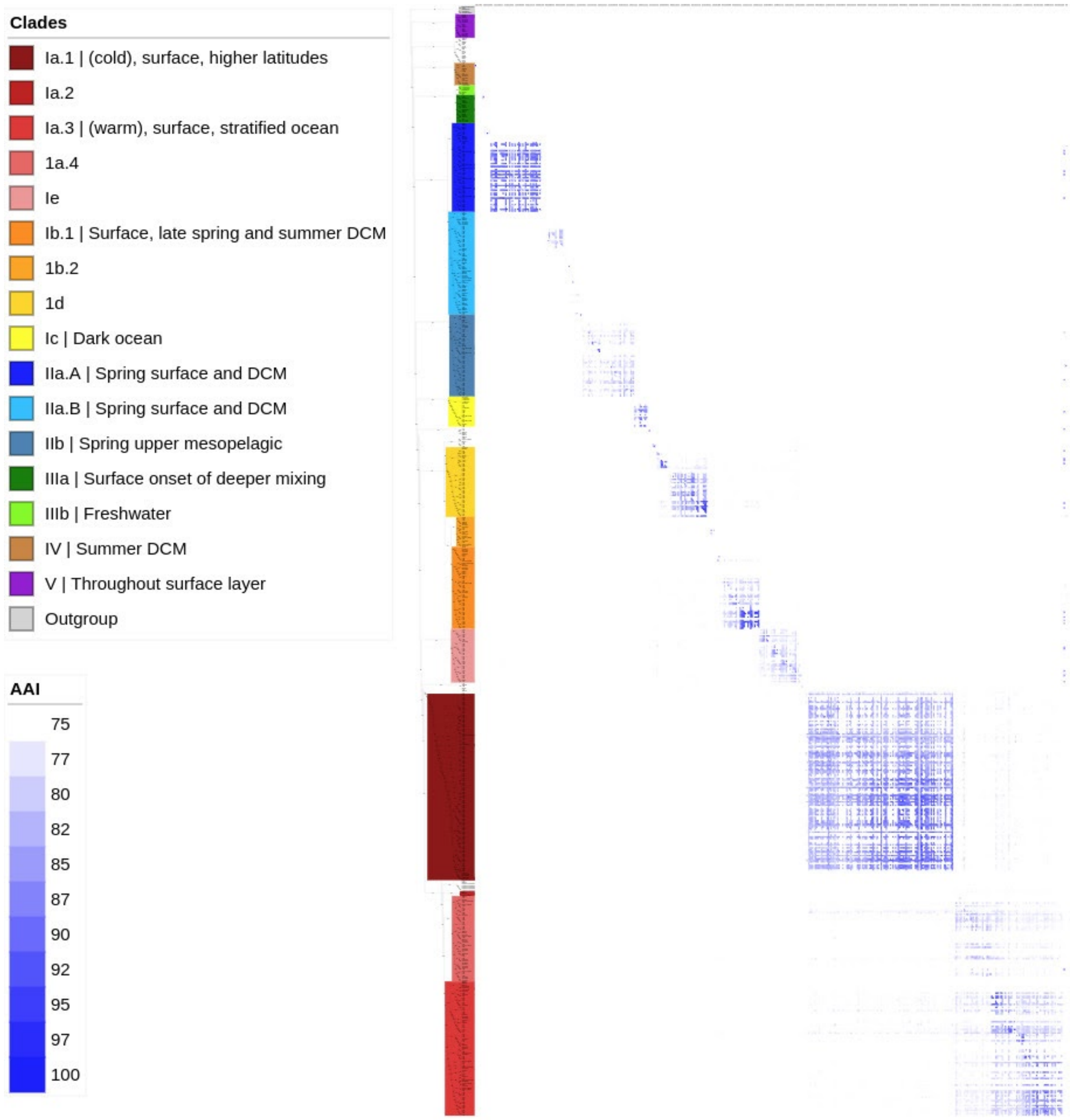


Figure 3.4 AAI of SAR11 SAGs ordered by their position with the phylogenetic tree reveals clustering at existing branches.

To confirm AAI was effective at cluster determination and therefore branch support, a small clade such as SAR11 Clade Ic was examined. AAI plot of clade Ic confirmed that SAGs closely related to reference SAR11s from clade Ic have higher AAI values when compared to each other. Lower more distant AAI values are present when compared to reference SAR11s outside of clade Ic. This confirms the viability of AAI as a method for determining clade identity. A darker blue colour indicates strains are more closely related. A clear darker blue colour in reference members of SAR11 clade Ic and supposed SAG members indicate membership to the same cluster and not to other SAR11 clades.

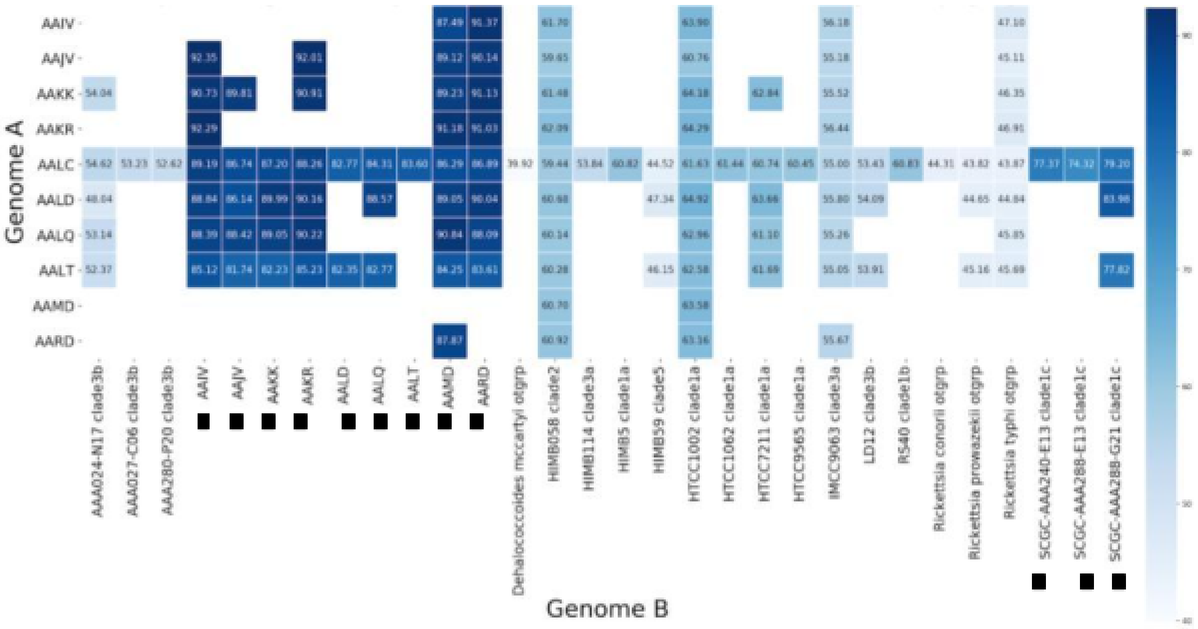


Figure 3.5 AAI percentage heatmap of SAR11 clade Ic determined by WGS phylogenetics. SAR11 SAGs shown to be associated with clade Ic are labelled with a black square. SAR11 reference genomes within clade Ic are labelled with an Ic suffix and a black square.

3.3.4 *k*-mer counting SAR11 genomes for clade boundaries

WGS of SAR11 SAGs and reference genomes are *5*-mer counted and plotted on a BH-tSNE scatter plot. This is normalised, allowing for genomes of shorter or longer lengths to be compared, as differing length can lead to differing numbers of *k*-mer counts. Numerical counts allow for genomes to be plotted in relation to each other to produce clusters where normalised *k*-mer counts of the same composition can be grouped together. Colours indicate SAR11 reference genomes. SAR11 reference genomes did not cluster together, although there is an overall structure to the data. This is surprising due to AAI and *k*-mer counting both being derived from the same genetic data. Six clusters appear clearly and are well separated from different reference genomes. However, clustering also shows reference genomes are separated from each other and don't cluster together. This is unusual as organisms that are similarly related would be expected to have similar *k*-mer counts.

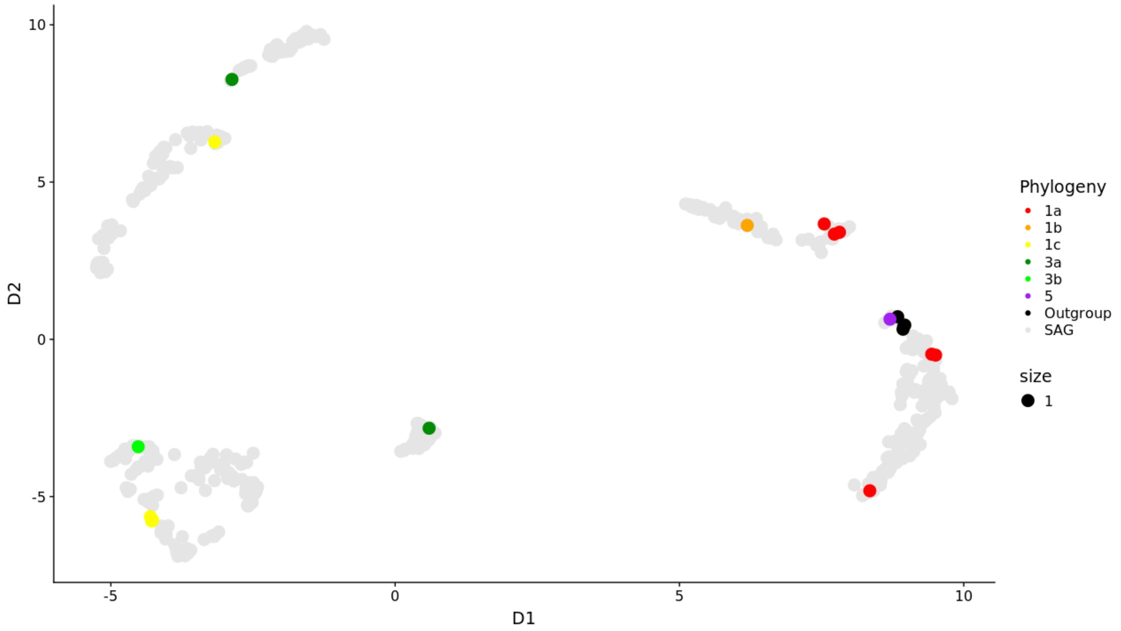


Figure 3.6 BH-tSNE scatter plot of WGS of SAR11 SAGs. SAG contigs were concatenated together forming a single contig FASTA file. Kmer counting was performed and dimension reduction techniques in a BH-tSNE plot were performed via UMAP. SAGs were plotted via their BH-tSNE values in the form of a scatter plot. Reference genomes are plotted in coloured dots.

3.3.5 Ecological mapping of SAR11 clades

Clades established in **Figure 3.4** were grouped together and mapped against the (Biller et al. 2018) metagenome dataset to establish if different SAR11 clades inhabit different ecological clades. Clade I can clearly be seen to inhabit different areas of the ocean based on depth and the time of the year at Hawaii Ocean Time-series (HOTs). Individual plots for each clade are included in the appendices.

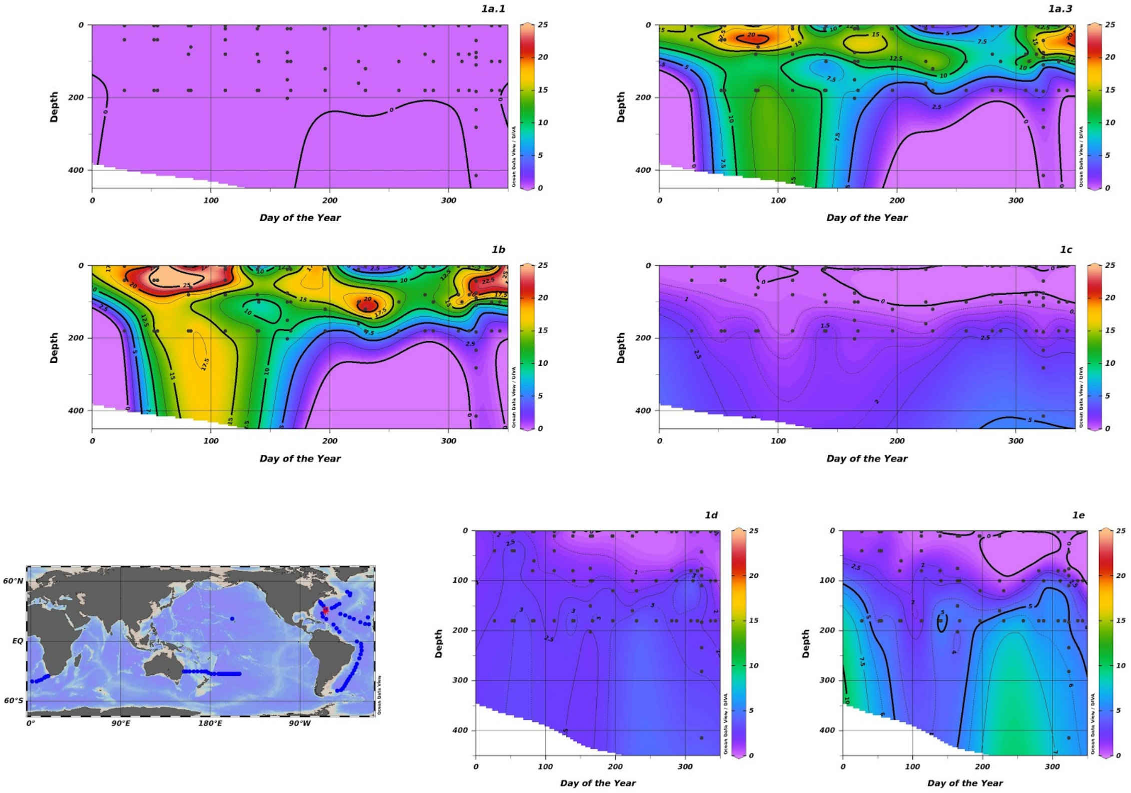


Figure 3.7 Ocean Data View plots of all SAR11 clade 1 at the Hawaii Oceanographic Time Series. Heat plots are plotted with depth against the day of the year. Units are in percentage median clade coverage at spatial-temporal points (black dots).

Each SAR11 clade's global ecological location is summarised to characterise its niche from individual clade heat plots. Barcharts detail SAR11 clade abundance over depth and time of the year using the median tpmeans (Imelfort and Lamberton 2015) for each clade. Tpmeans is defined as the average coverage of each mapped contig after the top and bottom 10% coverage values have been excluded. This is done to reduce bias to highly conserved regions like 16S. Median clade coverage is defined as the median coverage derived from all members of a clade. Overall, clades generally conform to established SAR11 clade locations (Stephen J. Giovannoni 2017).

The proteorhodopsin genes are proteins involved in light-mediated functionality, providing energy for actions such as a proton pump (Beja, Pinhassi, and Spudich 2013). These genes are responsive to light and therefore can indicate if an organism's ecological niche contains light, such as within shallower waters.

Organisms whose sole niche is devoid of light is unlikely to contain a proteorhodopsin gene as they would be inactive in such conditions. Clade Ic, although most abundant in bathypelagic regions where there is an absence of light, still encodes for the proteorhodopsin gene (Thrash et al. 2014), indicating that they still use light for some metabolic function. Indeed, this may suggest a migratory nature where SAR11 from Clade Ic may periodically enter shallower waters and therefore activate proteorhodopsin related activity. Genes encoding for proteorhodopsin were searched for within Clade Id and Ie genomes using Prokka (Torsten Seemann 2014), a gene annotation algorithm. Proteorhodopsin genes were found in a majority of Clade Id and Ie SAR11, indicating they may exhibit a similar migratory pattern as Clade Ic. This is further reinforced by their equal or higher abundance within mesopelagic regions over bathypelagic and abyssopelagic regions of the ocean (**Fig 3.8**).

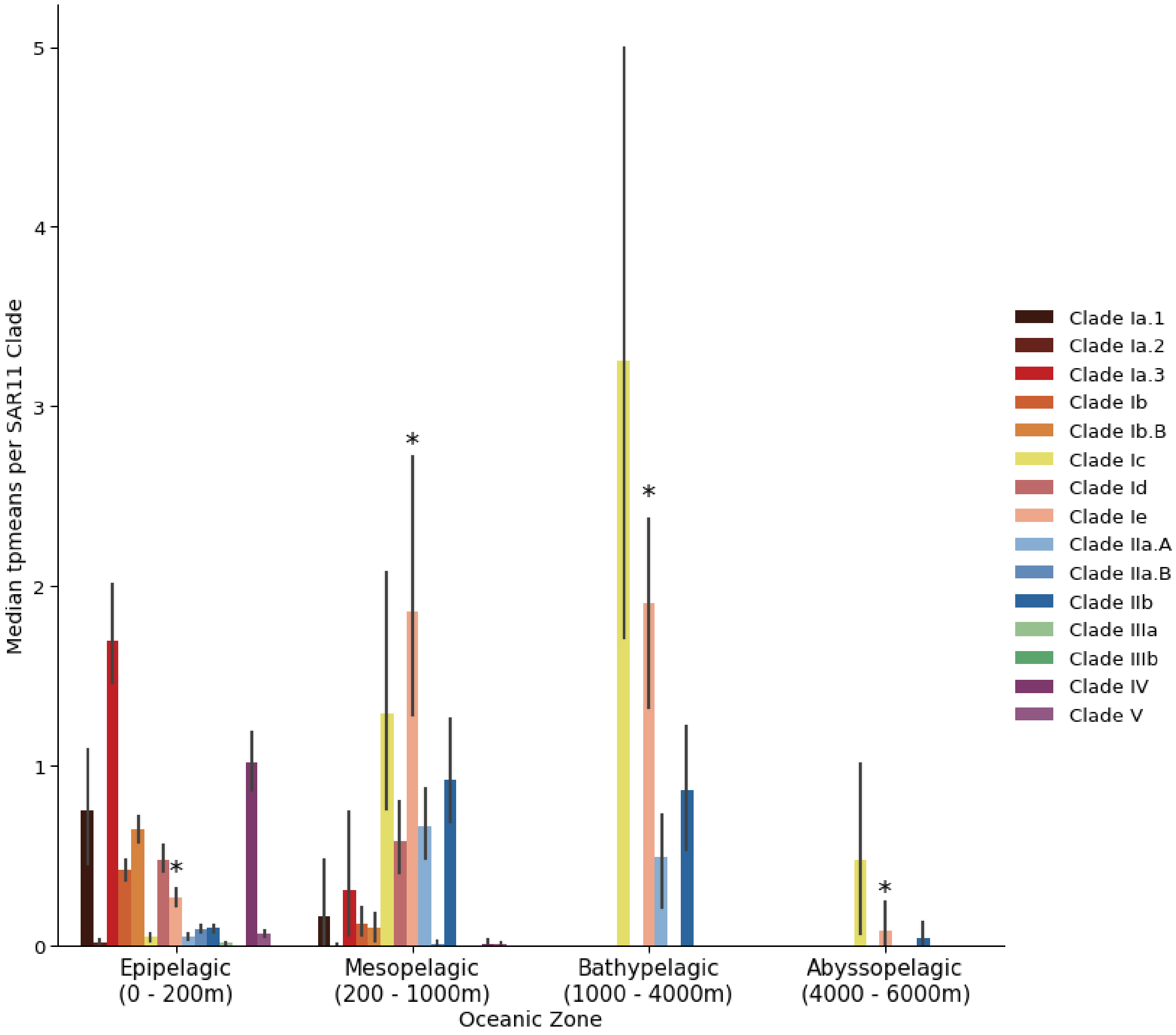


Figure 3.8 SAR11 clades mapped against metagenomic datasets by the (Biller et al. 2018) study to reveal global abundance by depth. Abundance is shown as the median tpmeans of each clade by depth category within the ocean. Asterisks indicate clade Ie.

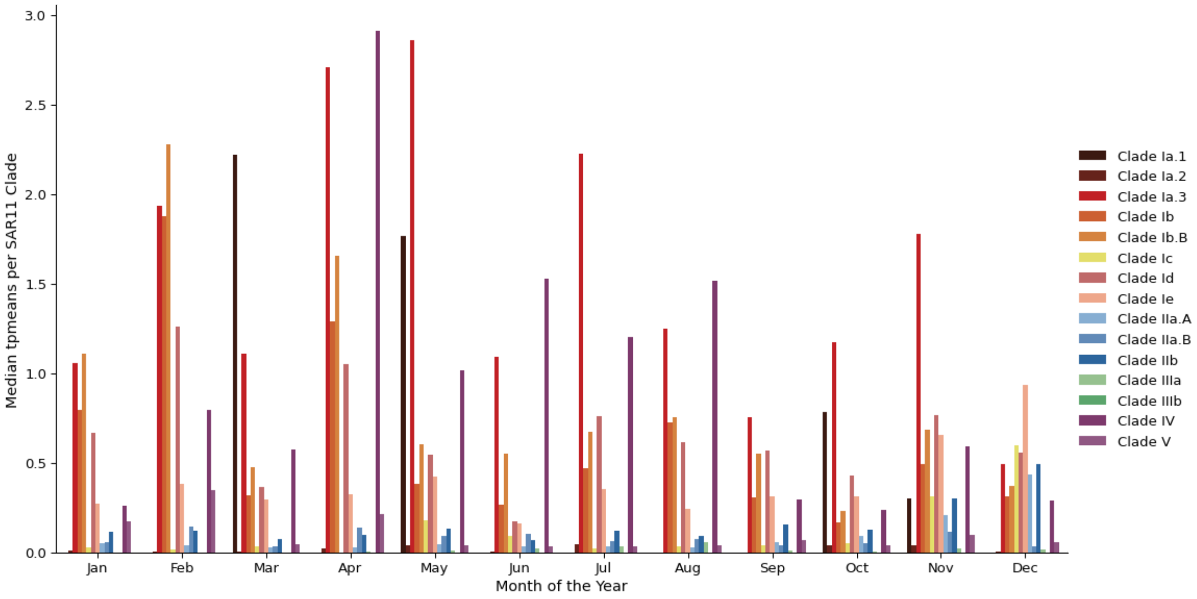


Figure 3.9 SAR11 clades mapped against metagenomic datasets by the (Biller et al. 2018) study to reveal global abundance by month. Abundance is shown as the median tpmeans of each clade.

3.4 Discussion

SAR11 SAGs were assembled and grouped by phylogenetics into clades. However, without reference sequences from SAR11 clades IV and V, it is challenging to confirm that splits within the WGS phylogenetic tree are true members of clades IV and V and not new clades. A variety of different methods like ANI and *k*-mer counting were used to cluster clades to confirm phylogenetic identities. Branching structures aligned with AAI clusters confirming splits within the phylogenetic tree. Ecotype mappings confirmed splits into clades also coincided with ecological niches. Branching structures of SAR11 clades allowed the discovery of two new clades, Id and Ie.

3.4.1 Improving SAR11 SAG assembly completeness

The total assembly completion of a SAR11 SAG varied drastically from 4 to 95% complete, with an average of 60%. This means it is not guaranteed that the isolated organism will have a complete contiguous genome. Additionally, chimeric artefacts created in the branching step of the MDA reaction leads to chimeric sequences (Nurk et al. 2013). Chimeric sequences are an inaccurate representation of an organism's genome leading to splicing of different areas of the genome together, resulting in the loss of data and/or truncated genes. Bias in the MDA reaction produces uneven coverage of SAGs which becomes problematic in the assembly process (Lasken 2009). Additionally, short read sequencing struggles to assemble long repeat regions. Even with sequencing depths of x1000, on average less than 50% of the genome was sequenced (Stepanauskas et al. 2017). Instead, other methods of genome amplification should be explored to avoid the MDA step, such as PicoPLEX (Kurihara et al. 2011) and MALBAC (Chapman et al. 2015), which reduce chimeric contigs by making circular DNA amplified fragments. Overall, the technology of genome amplification needs to be further developed in order to amplify the full genome sequences of an isolated organism.

Bioinformatics methods that could be used to obtain a higher completeness score of SAR11 genomes from SAGs include pooling of similar SAR11 genomes reads to form combined assemblies. This may provide more sequence data that covers areas of low coverage and can indicate the presence of chimeric reads, allowing for consensus error correction of these regions. This may increase the completeness of SAR11 SAGs assemblies at the loss of the number of SAR11 SAGs, shown to be effective in a study from (Kogawa et al. 2018). This would instead provide a representation of all SAR11 genomes with the same similarity score at the loss of any strain-level genetic difference in these “pooled” samples. In addition, there are diminishing returns where higher numbers of SAR11 SAGs are required to span across unsequenced regions. It may require only a couple of SAGs to get a completion value of 50% but many more to get over a 95% completion due to the chance of obtaining a read spanning an unsequenced region. This also runs the risk of increased contamination where a consensus is not achieved and multiple contigs of the same regions are produced.

Lastly, a combined method with metagenomics in which genetic material from an environment from which a SAG is isolated can be used as “scaffolds” between contigs or even as hybrid assemblies. If reads or contigs map at high percentage levels, these can be used to close gaps within an assembly. This would improve the completeness scores and provide an ordering of contigs within an assembly at the expense of strain-level differences using scaffolding. Both the co-assembly of multiple SAGs and usage of metagenomic scaffolds can improve completeness within an assembly but at the risk of increased contamination and loss of individual genetic differences.

Overall, bioinformatics methods can produce more complete SAGs assemblies at the cost of the loss of strain-specific genetic differences, instead of providing a representation of a group of organisms. If the goal of the study is to look at strain level divergence, bioinformatic methods would not help with the assembly and wet-lab methods should instead be used to improve the recovery of genetic material from the amplification and DNA sequencing process.

3.4.2 Phylogenetics

It is well-established that multi-gene phylogeny is likely to be more accurate in its placement of organisms within a phylogenetic tree than amplicon studies (Devulder, Pérouse de Montclos, and Flandrois 2005; Kamali et al. 2014; Gontcharov 2003; Tarasov and Dimitrov 2016). Additionally, it has the added advantage that the 16S gene is not required, only a combination of conserved genes within each organism. This allows for longer MSA files where smaller local alignment differences can assist in further differentiation of each organism. The main drawback to this method is the large computational time required to produce trees where MSA files of potential 5000 amino acids are used (Chaumeil et al. 2019). Although trimming of such regions can be used to filter out poor quality regions with little or no variation, these still are computationally intensive. Within this study, a 16S phylogenetic tree would take a couple of hours to complete, whereas a multigene phylogenetic tree took almost four days, both on a 16 core HPC node.

An advantage of 16S phylogeny is the highly conserved nature of the 16S gene, found within all bacteria and archaea. This allows for comparisons of distantly related species but loses its advantage within closely related organisms. However, with multiple studies acknowledging that bacteria have between 30 - 120 conserved gene markers (D. H. Parks et al. 2015; Rinke et al. 2013; B. J. Campbell et al. 2011), computation processing power becomes the limiting factor in discerning these differences. However, only 16S sequences are available due to the well-established practice of using 16S sequences as a genetic marker resulting in multiple amplicon studies. This makes performing more accurate WGS phylogenetic trees with ingroups more difficult. An example of this is clade IV of the SAR11 phylogeny, which are only represented as a 16S sequence within the NCBI database. Hence, within this study both the 16S rRNA phylogenetics and multi-gene phylogenetics are used for classification of SAR11 SAGs. Lastly, the main advantages of 16S sequencing are the relatively cheap monetary cost in comparison to whole-genome sequencing, and 16S sequencing does not require sorting in MAGs, commonly used in metagenomic studies, as each 16S sequence is presumed to come from a different individual.

Therefore, this study combines the power of WGS and 16S phylogenetics to accurately determine the position of each SAR11 SAG. Where 16S branches align with the same ingroups in the WGS branches, with a high bootstrapping of greater than 90, it would be reasonable to assume that these organisms are correctly classified. From this combination study of phylogenetic trees, we can also be confident in the discovery of two new clades, clade Id and Ie, which were all previously undescribed in other literature (Tsementzi et al. 2016; Kraemer et al. 2019; Stephen J. Giovannoni 2017).

3.4.3 ANI and AAI

Clustering of ANIs within a heatmap is a good indicator to confirm branch splits within a phylogenetic tree. Within an ANI heatmap, clusters should be identified with a reference genome within each cluster, providing identities of each grouping. This allows for the identification of groups and can reinforce existing categorisations from phylogenetic trees. However, with a discontinuous spectrum of ANIs, this makes it difficult to see clade boundaries as these are not clearly defined. Additionally, ANI is not recommended to be used if species are less than 75% similar (C. Jain et al. 2018a). Instead, AAI is recommended to look at species divergence beyond this value. However, due to the increased numbers of *k*-mers with amino acids over nucleotides, computational limits become the limiting factor in AAI studies. But when AAI was calculated for SAGs and aligned against a WGS phylogenetic tree, clear clusters become apparent along phylogenetic clade boundaries, confirming that splits on phylogenetic trees are accurate. This additionally allowed for determination if subclades were present within phylogenetic structures and led to the confirmation of clade IIb.b within clade II.

3.4.4 *k*-mer counting

Although plotting AAI is very useful to delineate species boundaries on a discontinuous scale, it is highly subjective where boundaries occur between very closely related organisms. Arbitrary groupings exist, with >95% dictating an intra-species relationship and <83% an interspecies relationship at a nucleotide level (Qin et al. 2014). However, with SAGs where multiple genes are missing, these results are deduced based on small amounts of common genes and can exacerbate bias in certain genes for and against these boundaries. *K*-mer counting provides an alternative method for categorisation and delineation of SAGs taxonomic groups, although there is limited success with closely related individuals.

3.4.5 Ecological identities of SAR11 Clades

Phylogenetics allows for the identification of clades within a group of organisms. This can help differentiate organisms based on their genetics and infer groupings of organisms. Isolation of an organism from a location provides some evidence of where an organism may exist within an environment, but not the extent of its geographical range. To examine if different SAR11 clades are of different ecotypes and occupy different niches with the ocean, clades were mapped to metagenomic data within the ocean. Overall, results aligned closely with previous studies (Stephen J. Giovannoni 2017), validating the methodology. However, with new clades Id and Ie, this led to the discovery that they are deeper waters specialists thought to be the sole domain of Clade Ic. Clade Id seems to operate in the mesopelagic region between 200 and 1000m, with Clade Ie present in even deeper waters of the abyssopelagic of up to 4500m. However, clades Id and Ie are not seen to exclusively reside within deeper waters as they contain proteorhodopsin genes, genes that interact directly with light. It is therefore likely that these SAR11 clades migrate to shallower depths, in line with existing findings with clade Ic (Thrash et al. 2014). Interestingly, clade Ib.2 inhabits the same region as clade Ib throughout the year, yet has distinctive enough genetics to be classified as a separate cluster within AAI plots and branch off from within the Ib clade. This may indicate that these two clades interact with nutrients within their environment in a different manner to each other, or variations are the result of the Kill the Winner hypothesis by phage predation.

With the phylogenetics of new SAR11 clades, it becomes clear that only a small number of SAR11 have been isolated and cultured predominantly from: clade Ia.1 (HTCC1062) (Rappé et al. 2002) Ia.2 (HIMB5) (Grote et al. 2012) Ia.3 (HTCC7211) (Stingl, Tripp, and Giovannoni 2007) Ib.1 (RS40) (Jimenez-Infante et al. 2017) IIIa (IMCC9063) (Oh et al. 2011) and (HIMB114) (Grote et al. 2012) IIIb (LD12) (Henson et al. 2018) V (HIMB59) (Grote et al. 2012). SAGs provide a unique opportunity to study these organisms, especially Clade IV which has no known isolate, but here I provide five additions to this phylogenetic group.

SAGs therefore provide a valuable resource to study such organisms without the need for the challenging task of isolating rarer clades from this abundant taxa. Indeed, studies into SAR11’s abnormally low viral abundance (**Chapter 4**) and the presence of a conserved region involved in phage defence (**Chapter 5**) would be challenging without culture-independent methods.

4 SAR11 Single-cell Amplified Genomes viral signatures

4.1 Introduction

4.1.1 Abstract

Within marine microorganisms communities, viruses moderate long-term carbon storage and increase community metabolism through the viral shunt. Viruses infecting SAR11, known as Pelagiphages have been identified as some of the most abundant viruses within marine environments. However, relatively few Pelagiphage genomes are publically available for analysis despite their ubiquity. This is due in part to the difficulty of culturing their host. Here, I use bioinformatic methods to derive viral sequences from SAR11 SAGs from this study and from other public sources. I compare rates of infections across a range of marine microorganisms and conclude that members of the SAR11 clade contain three-fold fewer viral sequences than other taxa. This conflicts with the current ecological theory Piggyback-the-Winner predicting higher lysogeny in abundant microorganisms. I highlight the discrepancies between different viral identifying algorithms and suggest an improved method. The lack of successful gene annotations of Pelagiphages is attributed to poor gene calling and databases with improved annotation pipelines are suggested. Overall, low Pelagiphages infection rates are unexplained with Kill/Piggyback-the-Winner hypotheses and I instead suggest support for the King-of-the-Mountain hypothesis that SAR11 undergoes positive rather than negative density-dependent selection.

4.1.2 Identifying and sequencing viral sequences

There are two main methods for obtaining viral sequences - experimentally or computationally. To obtain high quality and complete viral sequences experimentally, viruses are propagated in a host culture. Once a high density of viruses has been reached, filtration, extraction and sequencing of viral genetic material provides viral genomes (Hyman 2019). Temperate phages are phages that have the additional capacity to integrate themselves within a host genomes or exist as lytic phages (Howard-Varona et al. 2017). They are rarely exclusive to one lifestyle and excise themselves from the host genome depending on environmental triggers (Lang, Pleška, and Guet 2020; Fortier and Sekulovic 2013). Temperate phages that have integrated themselves into the host genome are called prophages and proliferate when the host genome is replicated (Canchaya et al. 2003; Du Toit 2019). Theoretically, the exposure to UV or other environmental stressors can trigger the prophage to excise itself from the host (Nanda, Thormann, and Frunzke 2015). The virus then switches to a lytic lifestyle where the previously described method can be used to capture its viral sequence. This relies on the prophage being viable and able to excise itself in the presence of environmental triggers. This allows for the cultivation and sequencing of prophages experimentally.

Alternatively, bioinformatic algorithms allow for the identification of viral sequences without culturing (Roux et al. 2015; Zhou et al. 2011; Ren et al. 2017; Akhter, Aziz, and Edwards 2012; Fouts 2006; Gipsi Lima-Mendez et al. 2008; Arndt et al. 2017; Bose and Barber 2006; Arndt et al. 2016). However, this is an estimation of viral signatures based on databases, previously identified viral hallmark characteristics and *k*-mer frequencies. Generally, a lytic phage is an isolated viral sequence usually located on a single contig. This is separate from the organism's assembly and represents an ongoing infection captured during the DNA sequencing process, or is environmental DNA captured through unsterile techniques or ineffective isolation protocols. Conversely, prophages are viral sequences that have been integrated into the host genome usually through tRNAs and integrases (A. Campbell 2003), thus becoming part of the host genome.

Bioinformatically obtaining viral signatures from genetic sequences usually relies on the detection of properties established as viral (Adriaenssens and

Cowan 2014). Some basic “viral-like” properties are identified such as the proportion of “AT” and “GC” nucleotides within a proportion of the genome (Akhter, Aziz, and Edwards 2012; Van Hemert et al. 2018; Grigoriev 1999). A skew in the frequencies of these nucleotides can be an indicator of foreign genes (Elhai, Liu, and Taton 2012; Xiong 2006). Additionally, attachment sites and disrupted or shorter genes can indicate the start and end of a prophage sequence from which the virus has integrated (Ramisetty and Sudhakari 2019). Lower numbers of characterised Protein families (Pfam) domains and uncharacterised genes are also indicative of viral sequences (Breitbart et al. 2002). Circular sequences can be of phage origin, as they indicate a circular DNA structure and a complete viral sequence (Dion, Oechslin, and Moineau 2020).

4.1.3 Extraction of viral signatures by VirSorter

VirSorter (Roux et al. 2015), a bioinformatic software for extracting viral signatures uses MetaGeneAnnotator (Noguchi, Taniguchi, and Itoh 2008) and hmmsearch (Finn, Clements, and Eddy 2011) to look for Pfam and viral domains using Hidden Markov Models (HMMs). Briefly, it uses a sliding window, to look for areas enriched with viral hallmark genes like capsids and a tail fibre protein. Based on these variables, a category is determined representing the confidence that a sequence is viral. If 80% of a contig contains viral genes, then the whole contig is deemed viral. Otherwise, it is classified as a prophage. Identified viral sequences are split into six categories: one to three are lytic phage sequences and four to six are prophages. A lower category number represents higher confidence in the categorisation. For example, categories one and four are regarded as the “most confident” prediction, category two and five “likely” predictions and category three and six as “possible” predictions. Extraction of these viral sequences allows for further studies, for example, viral genetics and phylogeny (Dion, Oechslin, and Moineau 2020; Gorbalenya 2008).

4.1.4 Assessing viral-like *k*-mer frequencies by VirFinder

VirFinder (Ren et al. 2017) is a *k*-mer frequency based algorithm for viral contig identification. Unlike VirSorter, it does not rely on upstream gene annotation algorithms to accurately predict protein location and function. It uses machine learning techniques based on the different *k*-mer frequencies a host and viral contig have to identify viral sequences. It returns a score from zero to one, with a higher score indicating the sequence is similar to viral sequences in the default viral sequences training dataset, and therefore most likely viral. It also returns a *p*-value indicating how distinct the *k*-mer frequency of identified viral sequences is from prokaryotic host contigs within the default training dataset. It then becomes up to the user to determine cutoffs for which sequences they consider viral.

4.1.5 Viral Ecological niches

Chapter 3 of this project establishes ecological niches of each of these projects SAR11 SAGs along with publicly available SAR11 genomes (NCBI). It would be interesting to see if phages derived bioinformatically from SAR11 SAGs matched their host ranges. Localisation of phages in areas absent of their host may indicate SAR11 phages have a broader host range, able to infect SAR11 of other clades. Although not conclusive of broad host specificity (the exact same phage would need to be found in different hosts to confirm this), it could indicate an area of future exploration.

4.1.6 Gene-sharing network map

Common genes exist between prokaryotic and eukaryotic organisms like 16/18S rRNA. However, viral phylogeny is difficult to conclusively characterise as these do not share any specific genes in common. Although capsids are the most promising target, not all viruses have them (Koonin 2009). An alternative method looks at proteins within viral sequences and compares these to existing viruses, with the proportion of shared proteins between two viruses serving as a proxy for evolutionary relatedness (G. Lima-Mendez et al. 2008). The presence and absence of these functional proteins enable clustering of closely related viruses similar to branches on a phylogenetic tree.

4.1.7 Viral signatures present in marine microorganisms

It has long been established that 20% of marine microorganisms are virally infected at any one time. However, evidence for this number is based on theoretical calculations (Curtis A. Suttle 2005, [a] 2007) and counting virocells (Proctor and Fuhrman 1990) (cells undergoing infection) within seawater samples. Rates of viral infection of SAR11 were noticed to be lower than the expected 20% infection rate (**Fig 4.5**), therefore a comparative study of other marine SAGs was conducted to explore this statistic. Combining data from multiple sources as well as data from this study helped establish if this biological anomaly was by chance or statistically significant.

4.2 Materials and Method

4.2.1 Extraction of viral signatures by VirSorter

All 451 SAGs were analysed for the presence of viral sequences using VirSorter with default settings. Only categories 1, 2, 4 and 5 were retained as these represented “likely” confidence that sequences produced were viral. Additionally, sequence lengths below 10 kbps were removed as VirSorter has been documented to be less accurate within those ranges (Roux et al. 2015).

```
#Run VirSorter
wrapper_phage_contigs_sorter_iPlant.pl -f <SAGs.fasta> --db 2
--ncpu 16 --data-dir <virsorter-data-dir> --diamond
```

4.2.2 Gene annotation of viral signatures

Extraction of viral sequences from VirSorter allows for annotation of phage genes; obtaining locations of protein-coding regions, gene function and COG category. This is performed using prodigal (Hyatt et al. 2010) and diamond (Buchfink, Xie, and Huson 2015). Gene annotation maps were performed in R using the gggenes library (David Wilkins 2019).

```
# Run Prodigal
prodigal -p meta -i <viral_input> -d <viral_output>

# Run Diamond
diamond blastx -d <diamond_nr_database> \
-q <prodigal_genecalls> \
-o <diamond_output> \
--outfmt 6 qseqid evalue bitscore stitle -k 1 \
--more-sensitive
```

4.2.3 Phylogeny of Viral Protein Clusters

vConTACT2 (Bolduc et al. 2017; Jang et al. 2019) was used to create gene sharing networks to produce genome-based viral taxonomy and cluster viral genomes into ICTV-recognised phage genera (Bolduc et al. 2017). vConTACT2 was performed on Cyverse (Goff et al. 2011; Merchant et al. 2016), a publically available computational infrastructure with default settings and input files from VirSorter. Protein clusters were visualised in Cytoscape (Shannon et al. 2003) according to recommendations suggested by the creators of vContact2 described here on protocols.io.

4.2.4 Ecological mapping of viral signatures

Metagenomic samples from the (Biller et al. 2018) dataset were mapped against phage sequences. Mappings were filtered for contigs with 95% identity against metagenomic reads by BamM (Imelfort and Lamberton 2015). 95% was chosen as this was regarded as the delimiter for determining species (C. C. Thompson et al. 2013; Konstantinidis and Tiedje 2005; Richter and

Rosselló-Móra 2009). Percentage genome coverage was chosen as the visualisation metric and displayed using ODV (Schlitzer 2002) produced using pileup.sh from BBTools.

```
bowtie2-build <genome> <genome_name>
bowtie2 --threads 16 -x <genome_name> \
-1 <fwd_read> \
-2 <rev_read> \
--no-unal -S <output_sam_file>

#sort and index the bam files to order them
for i in *.sam; do
samtools view -F 4 -@ 16 -buSh $i | samtools sort -@ 16 - -o
$(basename $i .sam).srt.bam;
samtools index -@ 16 $(basename $i .sam).srt.bam;
done

#filter mappings below >95% identity
for i in *.srt.bam; do
bamm filter -b $i --percentage_id 0.95 2>&1 | tee $(basename
$i .bam).srt.fltr.log;
samtools sort -@ 16 $(basename $i .bam)_filtered.bam -o
$(basename $i .bam).srt.fltr.bam;
samtools index $(basename $i .bam).srt.fltr.bam;
rm $(basename $i .bam)_filtered.bam ${i}*;
done
```

4.2.5 Viral signature abundance in Marine Organisms

All additional SAGs were obtained from JGI IMG/M repository, March 2020 with the following criteria: (Cultivation Metadata -- Uncultured Type [Single Cell]) AND (Environmental Classification -- Ecosystem Type [Marine]).

Name	# of SAGs obtained	Taxonomy
Synechococcus	65	Genus [*Synechococcus*]
Prochlorococcus	531	Genus [*Prochlorococcus*]
Other Marine Taxa	240	NOT (Taxonomy -- Order [Synechococcales, Pelagibacterales]
Public SAR11	153	Order [Pelagibacterales]
Our SAR11	451	This Study

Table 4.1 Number of SAGs obtained from this study or the JGI IMG/M repository used for this Chapter's analysis

VirSorter was run with the following code:

```
#Run VirSorter
wrapper_phage_contigs_sorter_iPlant.pl -f <SAGs.fasta> --db 2
--ncpu 16 --data-dir <virsorter-data-dir> --diamond
```

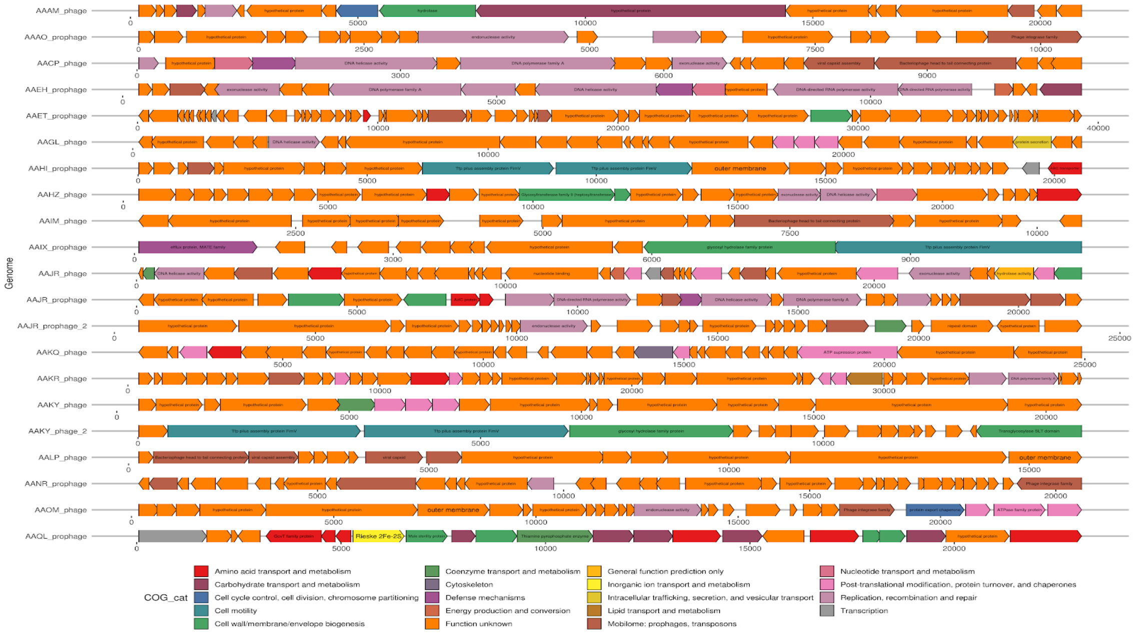
VirFinder was run in R with the following code with score amended depending on the analysis. Multithreaded version of this can be seen in full here.

```
#Run VirFinder
require(Biostrings)
require(parallel)
require(VirFinder)
[...]
predResult <- parVF.pred("cat123456.fasta", cores=16)
virfinder = subset(predResult, pvalue <= 0.05 & score >= 0.7 &
length >= 1000)
write.csv(virfinder, "VirFinder_cat123456_trm.csv", row.names
= FALSE)
```

4.3 Results

4.3.1 Gene annotation of viral signatures

VirSorter confidently identified 21 viral sequences and genome annotation identified genes consistent with expected viral genes. Large proportions of viral genes had no known function. No phages were identified as circular and therefore probable complete phage sequences. With the overwhelming amount of protein sequences (~90%) having a function of “hypothetical” indicates automated protein function determination for these viral signatures is best performed manually. This is recommended by the SEA-PHAGES protocol and is not further explored within this study due to time constraints with a recommendation of two to three days per phage for accurate gene annotation (Salisbury and Tsourkas 2019).



Genome
AAAM_phage
AAAO_prophage
AACP_phage
AAEH_prophage
AAET_prophage
AAGL_phage
AAHI_prophage
AAHZ_phage
AAIM_phage
AAIX_prophage
AAJR_phage
AAJR_prophage
AAJR_prophage_2
AAKQ_phage
AAKR_phage
AAKY_phage
AAKY_phage_2
AALP_phage
AANR_prophage
AAOM_phage
AAQL_prophage
COG_cat
Amino acid transport and metabolism
Carbohydrate transport and metabolism
Cell cycle control, cell division, chromosome partitioning
Cell motility
Cell wall/membrane/envelope biogenesis
Coenzyme transport and metabolism
Cytoskeleton
Defense mechanisms
Energy production and conversion
Function unknown
General function prediction only
Inorganic ion transport and metabolism
Intracellular trafficking, secretion, and vesicular transport
Lipid transport and metabolism
Mobilome: prophages, transposons
Nucleotide transport and metabolism
Post-translational modification, protein turnover, and chaperones
Replication, recombination and repair
Transcription

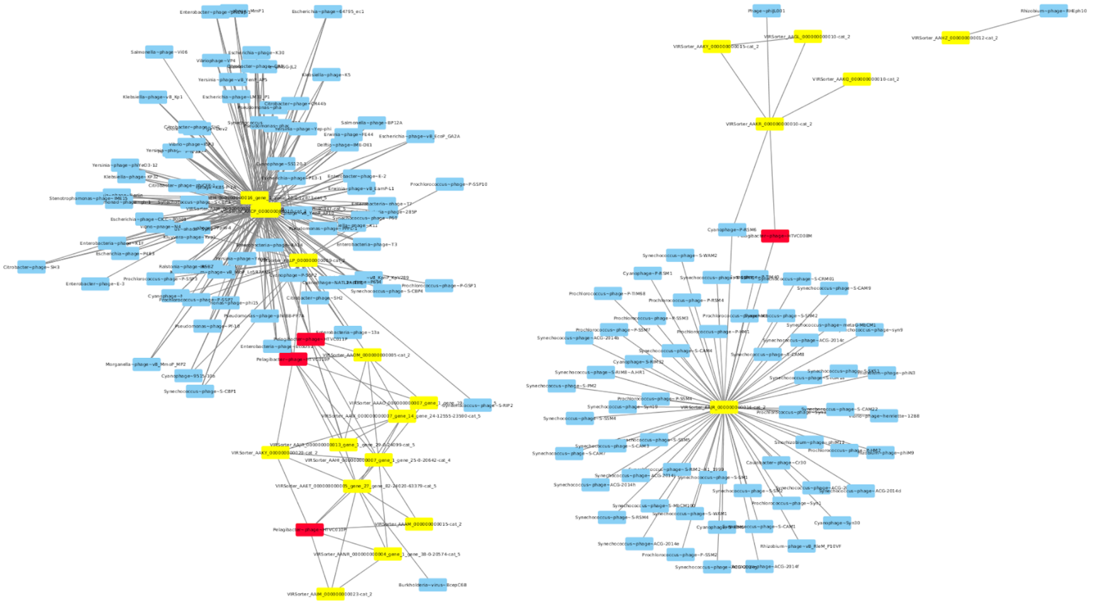
Figure 4.1 Genome annotation of SAR11 SAG viral signatures. Colours indicate COG grouping.

Classification	Genes	Length in base-pairs	Viral taxonomy	Host Clade
AAAM-cat2	19	20929	None found	Ib.1
AAAO-cat5	20	10506	None found	Ib.2
AACP-cat2	19	11081	None found	Ia.1
AAEH-cat5	20	12874	None found	Ia.1
AAET-cat5	58	39360	None found	Ia.1
AAGL-cat2	30	27196	None found	Ia.1
AAHI-cat4	25	20643	None found	Ia.3
AAHZ-cat4	28	23476	None found	Id
AAIM-cat2	15	10482	None found	Id
AAIX-cat5	12	11036	None found	IIa.A
AAJR-cat5	26	21483	None found	Ib.1
AAJR-cat5_2	29	24100	None found	Ib.1
AAJR-cat2	36	25764	None found	Ib.1
AAKQ-cat2	32	24962	None found	IIa.A
AAKR-cat2	42	38061	None found	Ic
AAKY-cat2	18	20815	None found	Ia.1
AAKY-cat2_2	17	14055	None found	Ia.1
AALP-cat2	15	15954	None found	IIa.A
AANR-cat5	37	20575	None found	Ib.1
AAOM-cat2	27	23547	None found	Id
AAQL-cat5	23	23166	None found	Ib.2

Table 4.2 SAR11 phages derived bioinformatically using VIRSorter. Basic statistics are listed about each phage. None of the viral genomes were classified as circular and thus may represent genomic fragments.

4.3.2 Phylogeny of Viral Protein Clusters

Viral sequences identified from VirSorter were uploaded into vConTACT2 v0.9.5 using database RefSeq-85 (14-Nov-2017), gene-based viral taxonomy was determined and visualised with Cytoscape. Gene-based taxonomy using gene-clusters revealed an association of isolated phage sequences with other common marine viruses within two distinct clusters with other known SAR11 phages. Nodes with no edges connected to any SAR11 SAG viral sequences were removed. SAR11 SAG phage sequences (yellow, **Fig 4.2**) were all shown to share genes with existing SAR11 phages (red, **Fig 4.2**) within the vConTACT2 database. No SAR11 SAG phage had only unique genes unrelated to any other SAR11 phage. Additional phages with related genes are displayed (blue, **Fig 4.2**). These are predominantly known *Synechococcus* spp. phages. vConTACT2 was unable to provide a taxonomic classification of viruses for any of the viral sequences.



Escherichia~phage~64795_ec1
Salmonella~phage~Vi06
Vibriophage~VP4
Escherichia~phage~K30
Klebsiella~phage~K5
Yersinia~phage~vB_YenP_AP5
Klebsiella~phage~vB_Kp1
Escherichia~phage~LM33_P1
Citrobacter~phage~CR44b
Salmonella~phage~BP12A
Erwinia~phage~FE44
Escherichia~phage~vB_EcoP_GA2A
Delftia~phage~IME-DE1
Yersinia~phage~phiYeO3-12
Klebsiella~phage~KP32
Cyanophage~SS120-1
Enterobacter~phage~E-2
Erwinia~phage~vB_EamP-L1
Prochlorococcus~phage~P-SSP10
Stenotrophomonas~phage~IME15
Escherichia~phage~CICC_80001
Enterobacteria~phage~K1F
Escherichia~phage~P483
Citrobacter~phage~SH3
Enterobacter~phage~E-3
Enterobacteria~phage~T3
Enterobacteria~phage~T7
Prochlorococcus~phage~P-SSP3
Prochlorococcus~phage~P-GSP1
Synechococcus~phage~S-CBP4
Citrobacter~phage~SH2
Pseudomonas~phage~Pf-10
Enterobacteria~phage~13a
Pelagibacter~phage~HTVC011P
Pelagibacter~phage~HTVC019P
Pelagibacter~phage~HTVC010P
Morganella~phage~vB_MmoP_MP2
Cyanophage~9515-10a
Synechococcus~phage~S-CBP1
Synechococcus~phage~S-RIP2
VIRSorter_AAOM_000000000005-cat_2
VIRSorter_AAIX_000000000007_gene_14_gene_24-12555-23590-cat_5
VIRSorter_AAJR_000000000013_gene_1_gene_29-0-24099-cat_5
VIRSorter_AAKY_000000000020-cat_2
VIRSorter_AAHI_000000000007_gene_1_gene_25-0-20642-cat_4
VIRSorter_AAET_000000000005_gene_27_gene_82-24020-63379-cat_5
VIRSorter_AAAM_000000000015-cat_2
VIRSorter_AANR_000000000004_gene_1_gene_38-0-20574-cat_5
Burkholderia~virus~BcepC6B
VIRSorter_AAIM_000000000023-cat_2
Phage~phiJL001
VIRSorter_AAKY_000000000015-cat_2
VIRSorter_AAGL_000000000010-cat_2
VIRSorter_AAHZ_000000000012-cat_2
Rhizobium~phage~RHEph10
VIRSorter_AAKQ_000000000010-cat_2
VIRSorter_AAKR_000000000010-cat_2
Cyanophage~P-RSM6
Pelagibacter~phage~HTVC008M
Synechococcus~phage~S-WAM2
Cyanophage~P-RSM1
Synechococcus~phage~S-CRM01
Synechococcus~phage~S-CAM9
Prochlorococcus~phage~P-TIM68
Prochlorococcus~phage~P-RSM4
Prochlorococcus~phage~P-SSM3
Synechococcus~phage~S-SHM2
Prochlorococcus~phage~P-SSM7
Synechococcus~phage~metaG-MbCM1
Synechococcus~phage~syn9
Synechococcus~phage~ACG-2014b
Prochlorococcus~phage~P-RM1
Synechococcus~phage~ACG-2014c
Synechococcus~phage~S-CAM4
Synechococcus~phage~S-CAM8
Cyanophage~S-RIM32
Synechococcus~phage~S-RIM8~A.HR1
Rhizobium~phage~phiN3
Synechococcus~phage~S-PM2
Prochlorococcus~phage~P-SSM4
Synechococcus~phage~Syn19
VIRSorter_AAJR_000000000014-cat_2
Synechococcus~phage~S-CAM22
Vibrio~phage~henriette~1288
Synechococcus~phage~S-SSM4
Synechococcus~phage~S-SSM5
Sinorhizobium~phage~phiM12
Synechococcus~phage~S-CAM3
Synechococcus~phage~S-CAM7
Rhizobium~phage~phiM9
Caulobacter~phage~Cr30
Synechococcus~phage~S-SM1
Synechococcus~phage~ACG-2014h
Synechococcus~phage~ACG-2014d
Synechococcus~phage~S-MbCM100
Prochlorococcus~phage~Syn1
Synechococcus~phage~S-WAM1
Synechococcus~phage~S-RSM4
Synechococcus~phage~S-CAM1
Cyanophage~Syn30
Synechococcus~phage~ACG-2014e
Rhizobium~phage~vB_RleM_P10VF
Prochlorococcus~phage~P-SSM2
Synechococcus~phage~ACG-2014f

Figure 4.2 vConTACT2 protein cluster graph of a gene-sharing network of 21 SAR11 SAG viral sequences (yellow), four known SAR11 phages (red) and other reference phages (blue). Visualised using Cytoscape, showing linkage of protein clusters between existing viral groups.

4.3.3 Ecological mapping of viral signatures

Phage sequences were mapped to metagenomic datasets from the (Biller et al. 2018) dataset to determine ecological niches. Only phages infecting SAR11 Clade Ib are shown within this chapter due to ODV plots being large in size. Instead, a summary table is provided for easy viewing (**Table 4.2**). All other heat maps are located in the appendices. The host clade from where the SAR11 SAG phage was isolated is also included to compare host and phage ecological ranges. Prophage abundance was recorded as much lower in comparison to host and phage abundance.

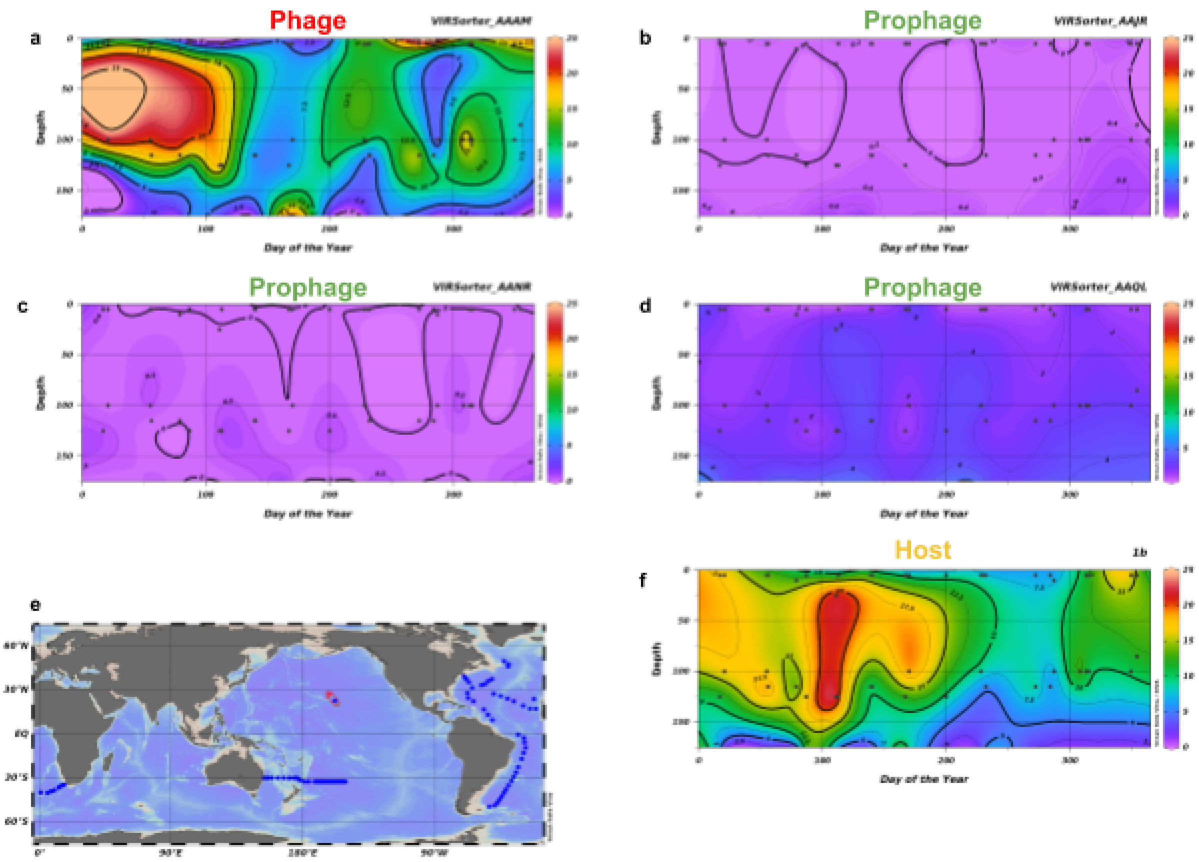


Figure 4.3 ODV plots of Pelagiphage sequences from the SAR11 clade Ib and their host using ecological data from the **Hawaii Ocean Time-series**. Pelagiphage sequences are deduced bioinformatically using VirSorter from SAR11 SAGs belonging to the Ib clade. Resulting viral contigs are mapped against cellular metagenomic reads from the Hawaii Ocean Time-series. Resulting viral contig abundance data is sorted according to time of the year

metagenomic sampling took place and represented as the percentage of the total viral contig that was successfully mapped to reads within a metagenome. **A.** A Pelagiphage sequence identified as a phage. **B,C,D.** Pelagiphages identified as prophages. **E.** Location of Hawaii Ocean Time-series (red). **F.** SAR11 Clade Ib ecological mapping against the Hawaii Ocean Time-series.

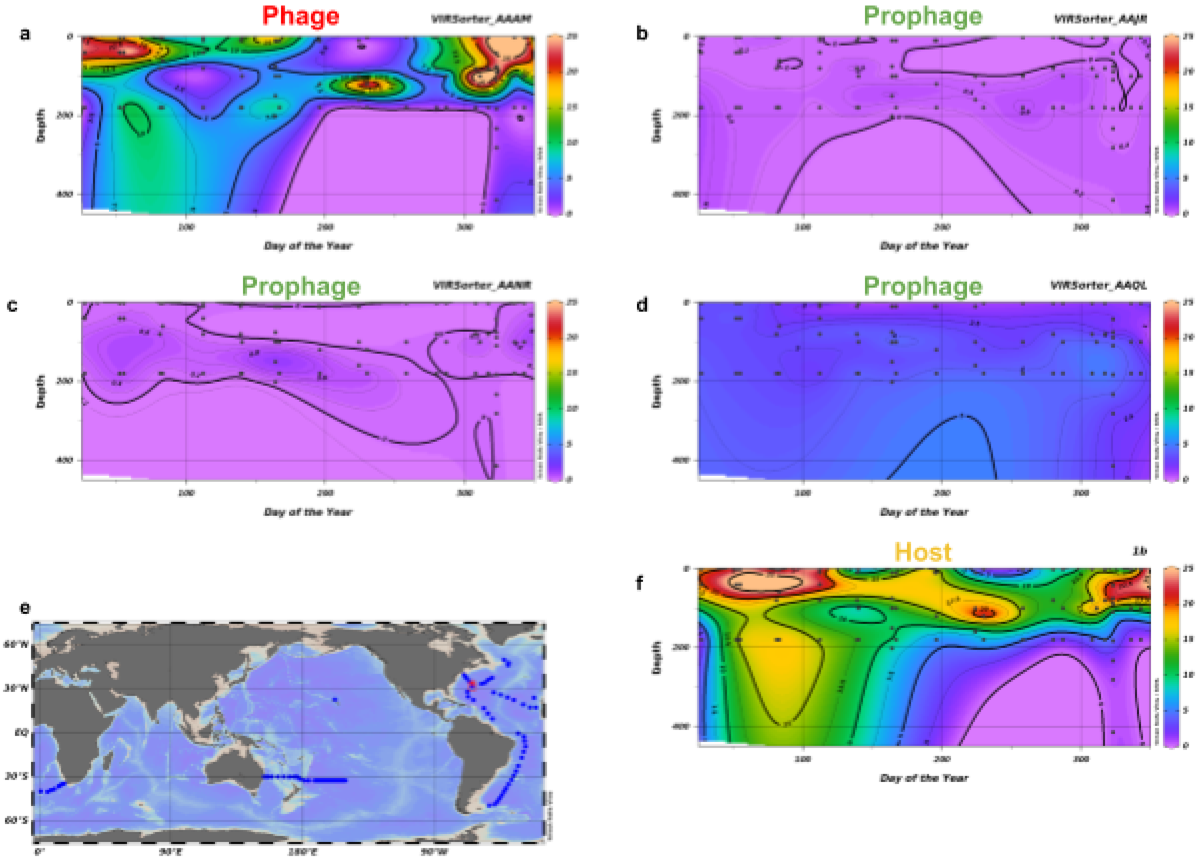


Figure 4.4 ODV plots of Pelagiphage sequences from the SAR11 clade Ib and their host using ecological data from the **Bermuda Atlantic Time-series Study**. Pelagiphage sequences are deduced bioinformatically using VirSorter from SAR11 SAGs belonging to the Ib clade. Resulting viral contigs are mapped against cellular metagenomic reads from the Bermuda Atlantic Time-series Study. Resulting viral contig abundance data is sorted according to time of the year metagenomic sampling took place and represented as the percentage of the total viral contig that was successfully mapped to reads within a metagenome. **A.** A Pelagiphage sequence identified as a phage. **B,C,D.** Pelagiphages identified as prophages. **E.** Location of Bermuda Atlantic Time-series Study (red). **F.** SAR11 Clade Ib ecological mapping against the

Bermuda Atlantic Time-series Study.

ODV plots are highly subjective to categorise phage and host ecological niches and location, therefore a summary table attempting to detail host and phage range is given. Tables are my interpretation of ODV plots to allow for easy comparisons against host phage ranges. Interestingly, phages infecting hosts clade Ib had largely different ranges to their host occupying much deeper depths. Also, phages from host clade IIa.A had a much shallower preferred depth in comparison to their host range.

Phage	Host Clade	Estimated Host Range	Estimated Viral Range
AACP_Phage	Ia.1	0 - 200	0 - 200
AAEH_Prophage	Ia.1	0 - 200	0 - 200
AAET_Prophage	Ia.1	0 - 200	0 - 200
AAGL_Phage	Ia.1	0 - 200	0 - 200
AAIM_Phage	Ia.3	0 - 200	0 - 200
AAOM_Phage	Ia.3	0 - 200	0 -100
AAHI_Prophage	Ia.3	0 - 200	0 - 200
AAHZ_Phage	Ia.3	0 - 200	0 - 50
AAQL_Prophage	Ib	0 - 200	0 - 200
AAAM_Phage	Ib	0 - 200	0 - 150
AAAO_Prophage	Ib	0 - 200	0 - 200
AANR_Prophage	Ib	0 - 200	250 - 500
AAJR_Phage	Ib	0 - 200	200 - 4500
AAJR_Prophage1	Ib	0 - 200	150 - 400
AAJR_Prophage2	Ib	0 - 200	200 - 5000
AAKR_Phage	Ic	200 - 5500	150 - 6000
AAKQ_Phage	Ie	500 - 2500	200 - 6000
AAIX_Prophage	IIa.A	200 - 5000	100 - 400
AAKY_Phage1	IIa.A	200 - 5000	100 - 1000
AALP_Phage	IIa.A	200 - 5000	100 - 500
AAKY_Phage2	IIa.A	200 - 5000	200 - 6000

Table 4.3 Summary table of host and phage abundance at depth based on ecological mapping data.

4.3.4 Viral lysogeny in marine microorganisms

Phages are abundant within marine systems, often at an order of magnitude higher than the host they predate upon (Wommack and Colwell 2000). Phages directly affect microbial communities, best described by the Kill-the-Winner hypothesis (Thingstad and Lignell 1997; Winter et al. 2010) describing negative density- ependentlytic predation and Piggyback-the-winner stating lysogeny by temperate phages is common when hosts population abundance is abundant (Knowles et al. 2016). SAR11 is one of the most abundant organisms within the marine environment (Morris et al. 2002; Craig A. Carlson et al. 2009), with their ubiquitous nature suggesting temperate phages that have integrated within SAR11 genomes, called prophages are a common occurrence. Incidence of lysogeny in the cyanobacterium *Prochlorococcus* is indicated as the cause of low viral particle counts when *Prochlorococcus* abundance is high (Sullivan, Waterbury, and Chisholm 2003). To explore this, bioinformatics tools such as VirFinder (Ren et al. 2017) and VirSorter (Roux et al. 2015) were used to quantify the number of lysogenic phage sequences within SAR11 and other marine taxa. Results indicated that contrary to the Piggyback-the-Winner hypothesis, the SAR11 clade does not contain a high percentage of viral sequence within their genomes. This is in contrast to organisms with the genus of *Prochlorococcus* and *Synechococcus* that are highly abundant in surface waters (Flombaum et al. 2013), where expected numbers of viral sequences were found. Additionally, SAR11 have lower numbers of viral signatures at a two to three-fold magnitude when compared to other marine taxa.

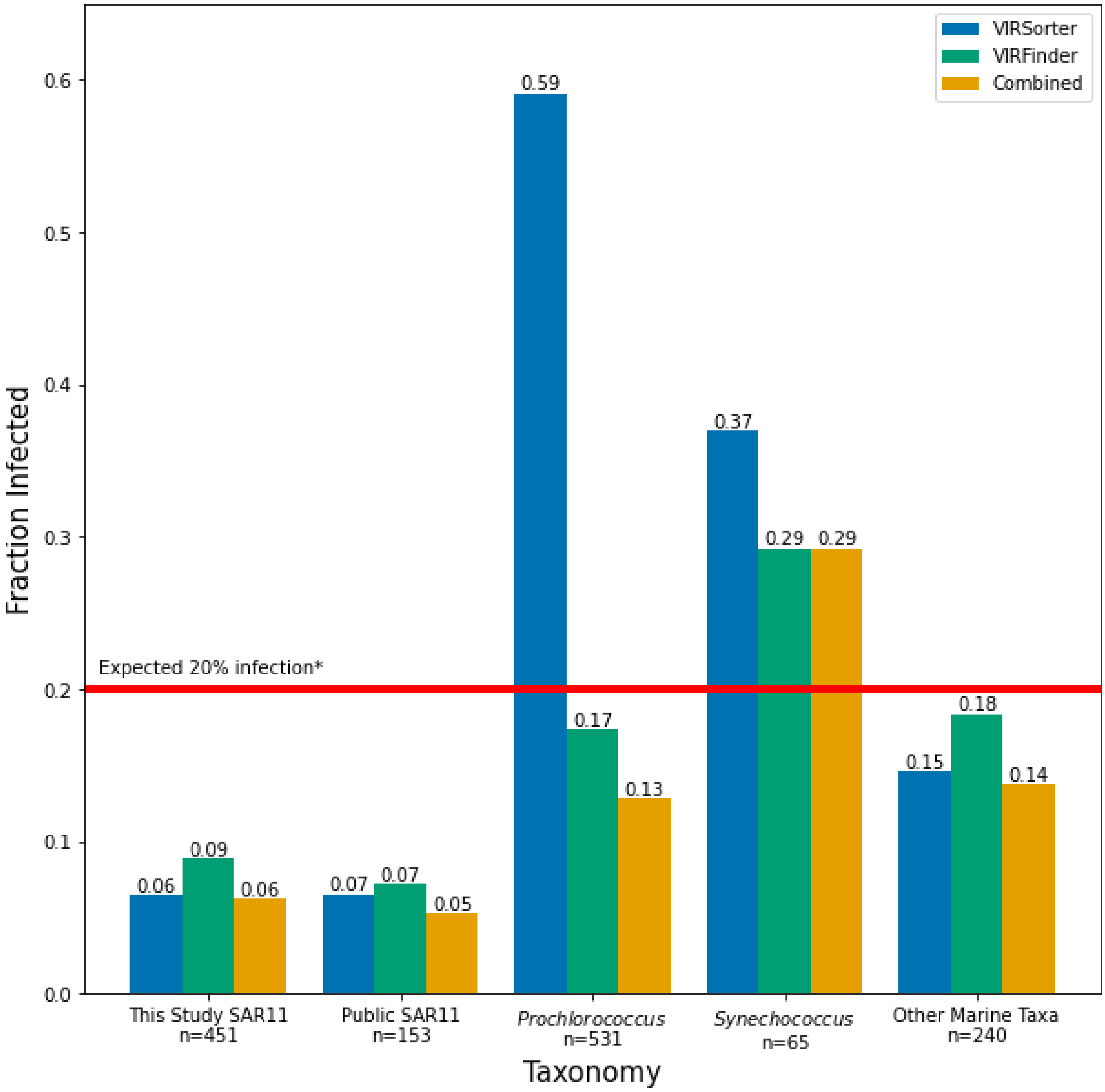


Figure 4.5 Fraction of viral signatures identified as SAGs across the marine environment. Two bioinformatics algorithms were used to determine the presence of viral signatures within a SAG. A SAG was considered infected if it contained a viral sequence within its assembly that is over **1000 bps in length** and: VirSorter (blue) identified a viral sequence as category **1,2,4 or 5.** VirFinder (green) identified a viral sequence if it received a score of **≥0.9** and a p-value of **≤0.05**. A combined approach (yellow) identified a viral sequence if VirSorter identified it as viral (**any category**) and the sequence had a VirFinder score of **≥0.7** and a p-value of **≤0.05**. *20% of all marine microorganisms are expected to be infected at one time (C. A. Suttle 1994; Proctor and Fuhrman 1990; Curtis A. Suttle 2007a).

4.4 Discussion

4.4.1 Pelagiphage ecology

21 SAR11 viral signatures were isolated from 451 SAR11 SAGs using VirSorter, equating to approximately 5% of all hosts containing viral signatures. This is lower than the expected 20% of all bacteria expected to contain viral sequences (C. A. Suttle 1994; Proctor and Fuhrman 1990). This may be due to the incomplete assemblies resulting from the MDA reaction used in the SAGs amplification process. However, it is expected that viral signatures would be enriched due to multiple copies of virions being made in the infection cycle (Labonté et al. 2015). Additionally, this may provide evidence to the ability of SAR11 to escape viral predation (explored more in **Chapter 5).** Genome analysis of SAR11 phages contained many unknown genes, suggesting gene calling processes are ineffective at finding coding regions and/or databases are incomplete. It has been suggested that automated viral gene calling is inaccurate and the SEA-PHAGES protocol includes methods to improve this process (Salisbury and Tsourkas 2019). This involves using multiple gene calling algorithms to provide consensus gene calls (If a gene is found using multiple gene calling algorithms, it is more likely to be correct, removing false positives). This protocol uses multiple additional biological indicators of coding regions such as the presence of operons where gene starting locations are within one, four or eight base pairs from upstream genes and coding potential graphs to determine gene starting locations. These are all laborious processes that require on average two to three days per phage sequences to identify by hand (Salisbury and Tsourkas 2019). This was not pursued within this project due to time constraints and the possibility that identified phages are false positives (phages identified are actual artifacts), but is considered a "gold standard" to phage gene annotation. An attempt to replicate this process bioinformatically can be viewed here, however, due again to time constraints and limited understanding of coding was not pursued further within the timeframe of this project. There is a definite need for such programs to identify

and annotate phage sequences as new bioinformatics programs have been created to serve this process (McNair et al. 2019; Salisbury and Tsourkas 2019; Ecale Zhou et al. 2019). This highlights the inaccuracy of current models that require large 100k base pairs regions of the genome (Hyatt et al. 2010; Delcher et al. 1999; Lomsadze et al. 2018) to train on before providing accurate gene calls. Phages can fall into this category of <100k base-pair genomes with for example, only one of the 48 isolated Pelagiphages having a genome larger than 100k base-pairs (Y. Zhao et al. 2013, 2019; Buchholz et al. 2020). Therefore, bioinformatic algorithms cannot rely on such *a priori* training mechanisms in small hosts such as SAR11.

A lack of annotation of phage proteins is unlikely to be solely due to the inaccurate gene calling methods, and probably also the result of unrepresented proteins within such databases. With roughly 90% of genes returned as unknown, this highlights the large numbers of novel genes identified within Pelagiphage genomes. Accurate deduction of gene function from unknown genes may help identify auxiliary metabolic genes which may help describe Pelagiphage dynamics. Auxiliary metabolic genes have been described to increase the fitness of the host to positively benefit viral replication (L. R. Thompson et al. 2011; Hurwitz and U'Ren 2016; Howard-Varona et al. 2018; Crummett et al. 2016). SAR11 undergoes slow replication rates (Stephen J. Giovannoni 2017) and therefore viral replication may be increased with auxiliary metabolic genes. In addition, current knowledge of how SAR11 is able to evade Pelagiphage infection on a population level is based on genomic regions conferring viral immunity (Y. Zhao et al. 2013; S. Giovannoni, Temperton, and Zhao 2013). Pelagiphages are likely to have mechanisms to counteract viral evasion as postulated by the Red Queen hypothesis (Van Valen 1973; Woolhouse et al. 2002; Lively and Apanius 1995) and detailed gene annotation may highlight them. These points towards the need for databases and gene callers to be improved to facilitate the deduction of phage gene function bioinformatically.

Interestingly, phages isolated from clade Ib show a larger ecological zone than their host. Phages and prophages from SAR11 SAG AAJR isolated from clade Ib show a preferential mapping to deeper waters (200m and below), but the host range for clade Ib is generally in shallower water above 200m. This may indicate that SAR11 phages are able to cross infect different clades of SAR11 bacteria. This trend is also shown in prophage AAQL isolated from a clade Ib host, but which follows clade Ie's ecological mapping preference. This is also seen to a lesser extent in prophage AANR, with its preference for deeper waters akin to phages isolated from AAJR. This could be due to the migratory nature of SAR11 species with the presence of members of deepwater clades in shallower waters (see **Chapter 3**). Isolation of phage sequences may also indicate accidental isolation of a phage inconsistent with its host when sequencing, acting as environmental contamination. However, viral prophages indicate integration into the host genome and would not explain this. Overall, I highlight the possibility of cross-clade infections and non-specificity for SAR11 clade predation from Pelagiphages. To provide evidence for this bioinformatically, chance encounters of the same phage sequence would need to be present in SAR11 from different clades at a significant level. Instead, cross-clade infections could be performed where isolated Pelagiphages are used in infection studies to determine specificity (Buchholz et al. 2020; Y. Zhao et al. 2013, 2019). As the exact predator-prey relationships between SAR11 and its phage remain contentious (Y. Zhao et al. 2013; Våge, Storesund, and Thingstad 2013; S. Giovannoni, Temperton, and Zhao 2013), a broad specificity may help inform ecological models in determining the ecological theories explaining Pelagiphages abundance (Våge et al. 2014; Winter et al. 2010).

4.4.2 Viral predation rates in Marine Microorganisms

Figure 4.4 includes the accuracy of using different bioinformatic algorithms and cutoffs to determine viral sequences from genomic data. *Prochlorococcus* infection rates are identified at 59% if VirSorter is used, but drops to 17% when using VirFinder *k*-mer counting based algorithm. This large variation shows that there is little consensus between different algorithms to what is considered viral. This is particularly obvious in a combined approach that has a more lenient

VirFinder score required to be considered viral (70% vs 90%) but still produces fewer viral hits (**Fig 4.4**, yellow versus green). This highlights the possible false positives when using algorithms searching for viral signatures and why I would recommend a combined approach where both algorithms form a consensus that a sequence is viral.

Other observations indicate that lysogeny events within the SAR11 clade are rare in comparison to other abundant marine taxa like *Prochlorococcus* and *Synechococcus*. With a combination (VirSorter and VirFinder) approach (**Fig 4.4, Yellow**) SAGs obtained within this study had almost identical percentage infection rates compared with publicly available SAR11 from JGI (6% vs 5%), validating that results from this study can be extrapolated to the entire SAR11 clade. This indicates in temperate SAR11 phages, lysogeny is a rare event and below the expected 20% in all marine taxa. Phage lysogeny within SAR11 is almost three to six fold less common than other marine taxa, who have on average a 15% lysogenic rate. (Parsons et al. 2012) showed that virioplankton abundance was negatively correlated with *Synechococcus* and SAR11 abundance, in-line with Piggyback-the-Winner hypothesis that under high abundance, lysogeny is a more common phage lifestyle (Knowles et al. 2016). Here I show that between 29-37% of *Synechococcus* are infected depending on the bioinformatic algorithm used (**Fig 4.4**); above the 20% average infected rate and consistent with a ubiquitous marine organism. However, SAR11 lysogeny rates would be expected to follow this high infection rate, which it does not, with only 6-9% infected. This contributes to the mystery of where SAR11 phages are located if not within SAR11 genomes or as virioplankton, due to their high abundance in metagenomic data (Y. Zhao et al. 2013; Buchholz et al. 2020). SAR11 undergoes slow replication rates compared to other marine microorganisms (~36 hours (Rappé et al. 2002)) which would align with (Knowles et al. 2016) suggesting that high rates of replication are also required for a transition to lysogeny. However, as a population, SAR11 undergoes similar or exceeds population growth rates compared to other marine taxa (Malmstrom et al. 2005). This highlights that there may be additional biological mechanisms involved in Pelagiphage host-prey dynamics we are still unaware of. Within

Chapter 5 I explore a possible biological interaction SAR11 may exhibit to escape viral predation and discuss its impact on current ecological theories in **Chapter 6**.

5 SAR11 Hypervariable regions

5.1 Introduction

5.1.1 Abstract

SAR11 is one of the most ubiquitous marine plankton along with its predatory viruses, the Pelagiphages, yet top-down predation theories are unable to explain their predator-prey relationship. SAR11 is able to resist population decline even when challenged with abundant Pelagiphages, contrary to the Kill-the-Winner hypothesis previously observed in marine plankton. Theories such as defensive specialism seek to explain SAR11's high abundance through defensive mechanisms. Although these mechanisms explain high SAR11 abundance, they are unable to explain high Pelagiphage numbers where SAR11 has a streamlined genome dominated by nutrient acquisition genes. Instead, hypervariable regions have been proposed as an alternative mechanism where small changes within these regions confer viral immunity by preventing viral adsorption. This champions the King-of-the-Mountain theory that through SAR11's superior nutrient acquisition genes and streamlined genome is able to outcompete other heterotrophic organisms. SAR11 resulting abundance suggests that positive density-dependent selection increases contact with other SAR11 species and their DNA. This increases the chances of genetic recombination events, especially in hypervariable regions, allowing for genetic material to be exchanged between members of SAR11 populations. These small changes are hypothesised to confer enough changes on surface glycoproteins in the genomic cassette HVR2 preventing Pelagiphage recognition as prey. This is coined "defence on a budget" allowing SAR11 to use a majority of its resources towards nutrient acquisition while still acquiring viral immunity as a population. However, the HVR2 - predicted to be the region that allows this mechanism, has not been studied extensively. Here I show that the HVR2 is present in all SAR11 clades I-III and consistently found between the

23S and 5S rRNA coding region. I highlight that they are variable in length but significantly enriched in surface related proteins. However, I contest that mutations within HVR2 are from replication events, not recombination and highlight bioinformatic evidence which suggests this. Overall, this suggests that more work is needed to study the evolution of these regions to determine the mechanism in which they produce variation. Either way, mutations within HVR2 would explain how SAR11 is able to retain its global abundance and suggest a mechanism of defence against abundant Pelagiphages.

5.1.1 The impact of Hypervariable Regions on viral ecology

Planktonic cells are predated upon by viruses and other microorganisms, returning abundant organisms to its environmental carrying capacity (an example of top-down predation), yet SAR11 - a clade of planktonic cells, is one of the most ubiquitous aquatic microorganisms (Morris et al. 2002; Becker et al. 2019; Biers, Sun, and Howard 2009; Herlemann et al. 2014). This conflicts with current ecological theories such as the Kill-the-Winner hypothesis (negative density dependant selection) (Thingstad and Lignell 1997) states that rising population density increases chance encounters with predatory phages and accelerates predation events (Marston et al. 2012), eventually returning an organism back to equilibrium (Fuhrman and Schwalbach 2003). This is a common strategy in “bloom-bust” dynamics of marine phytoplankton where increased nutrients cause phytoplankton blooms (Boyd et al. 2000) and virally caused mortality returns a population to its normal abundance (Guixa-Boixereu, Lysnes, and Pedros-Alio 1999; Alarcón-Schumacher et al. 2019; Matteson et al. 2012). In contrast, SAR11 comprises at least 20 to 40% of all marine microorganism biomass (Morris et al. 2002), conflicting with the KtW ecological strategy. Initially, SAR11’s global dominance was thought to be explained by cryptic escape, avoiding predation due to its small cell size (Yooseph et al. 2010) or slow replication rates making viral infection inefficient (Curtis A. Suttle 2007a; Parsons et al. 2012). However, *Prochlorococcus*, an autotrophic cyanobacteria has a similar surface area to volume ratios to SAR11 and is clearly affected by phage predation, following the Kill-the-Winner ecological model (Sullivan, Waterbury, and Chisholm 2003; Sullivan et al. 2005, 2009, 2010). Lastly, the SAR11 clade equaled or exceeded growth rates compared to the total prokaryotic community, suggesting that SAR11 population does not exhibit a slower growth rate (Malmstrom et al. 2005).

The high abundance of SAR11 can be explained by superior resource acquisition (Sowell et al. 2009; Malmstrom et al. 2005) or defensive specialism (Curtis A. Suttle 2007a; Våge et al. 2014). The idea of a defensive specialist suggests SAR11's high population abundance is due to a majority of resources expended on survivorship against top-down loss over replication and other functions. These take the form of defensive systems and have been identified in SAR11 which include a possible CRISPR system (Thrash et al. 2014) and a phosphorothioate system (Wang et al. 2011). However, SAR11 phages also exist at high abundances (Y. Zhao et al. 2013; Buchholz et al. 2020; Y. Zhao et al. 2019) which would indicate a high rate of predation events and therefore SAR11 survivorship is not effective. Pangenomics studies concluded the most abundant proteins within a SAR11 metaproteome were substrate-binding proteins related to nutrient transport (Grote et al. 2012), with no proteins associated with defence against predation (Sowell et al. 2009). This concluded that SAR11 probably devotes a majority of resources towards nutrient acquisition over survivorship, and therefore unlikely to be a defensive specialist. Instead, it is proposed that SAR11's dominance as the most ubiquitous marine microorganism is due to its superior resource acquisition, especially in nutrient-poor environments (Y. Zhao et al. 2013). Once a high abundance is achieved, it provides protection against loss by viral predation. Although a high population density increases the encounter rate with predatory phage particles, it also increases contact with other SAR11 and their DNA. This allows for increased opportunities for genetic recombination events to occur. The King-of-the-Mountain hypothesis was proposed (S. Giovannoni, Temperton, and Zhao 2013; Y. Zhao et al. 2013), stating SAR11 are able to resist a population decline caused by its predatory phages by recombination of viral defensive genes and collectively “out evolving” its predatory phages. This would be described as Positive Density Dependant selection (John Wiley & Sons, Ltd 2001), where increased contact with other SAR11 due to their high abundances allow for recombination events to occur at higher rates. This concept was further supported by high DNA recombination rates observed in SAR11 (Vergin et al. 2007) allowing for the horizontal transfer of DNA elements. Recombination allows for genomic regions to undergo higher rates of evolution in populations compared to clonal replication in asexual organisms. A particular region of interest is the genomic region HVR2, located between the 23S and 5S rRNA region in SAR11. HVR2 is a cassette of genes enriched for genes targeting the cell wall or membrane (Grote et al. 2012; L. J. Wilhelm et al. 2007). These types of genes contain external facing glycoprotein receptors that are often associated with phages adsorption. Variations of these receptors may prevent viral adsorption and therefore prevent predation events (Stephen J. Giovannoni 2017; Clément et al. 1983). Although it may seem counterintuitive for a streamlined genome prioritising only core genes to include HVRs, (a region of high gene variation and size) it is thought to be an advantage in SAR11 as significant parts of its genome are required for nutrient uptake. HVRs are regions of the genome undergoing high evolutionary pressure and variation, allowing for genes to become highly efficient and change rapidly according to their surrounding environment. This can be seen as a “defence on a budget” where small changes within an HVR can prevent viral adsorption without the dedication of large resources otherwise needed for nutrient transport.

5.2 Materials and Methods

5.2.1 Confirming existence and location of HVRs

Fifteen high quality published SAR11 genomes were chosen from the NCBI database (January 2019) across a range of clades were chosen to identify HVR regions with their genomes. Reads from the (Biller et al. 2018) metagenomic dataset were recruited against each SAR11 genome to find areas with little or no coverage against any marine prokaryotic metagenomes. This would identify unique areas of the SAR11 genome with high microdiversity. Uncovered regions would be dissimilar to any other marine prokaryote and would provide candidates for HVRs. The mapping process was performed using bowtie2 (B. Langmead and Salzberg 2013) recruiting metagenomic reads against the SAR11 genomes and subsequent filtering, including only reads with a 95% identity for downstream analysis with BamM (Imelfort and Lamberton 2015). Coverage of each base pair is plotted in the form of a barplot (**Fig 5.2**). Gaps in the barplot, signifying unique areas of the SAR11 genome were identified in this way.

Name	RefSeq assembly accession
LD12	GCF_002688585.1
HIMB083	GCF_000504225.1
HIMB114	GCF_000163555.2
HIMB1321	GCF_900177485.1
HIMB5	GCF_000299095.1
HIMB59	GCF_000299115.1
HTCC1040	GCF_000384455.1
HTCC1062	GCF_000012345.1
HTCC7211	GCF_000155895.1
HTCC7214	GCF_000701385.1
HTCC7217	GCF_000702645.1
HTCC8051	GCF_000472605.1
IMCC9063	GCF_000195085.1
RS39	GCF_002101315.1
RS40	GCF_002101295.1

Table 5.1 List of complete SAR11 genomes from the NCBI database

5.2.2 Usage of pangenomics to identify flanking genes

Three members of each SAR11 Clade I-IV from publically available genomes and 451 SAR11 SAGs isolated in this study were chosen based on maximising genome size and percentage completeness ascertained in **Chapter 3**. Clade V was not included due to evidence suggesting it is not a true SAR11 (Viklund et al. 2013). Anvi'os pangenomic workflow was performed on these samples to produce an anvi'o pangenomic circos plot (**Fig 5.3**). All settings and parameters were performed as recommended in their online workflow (Murat Eren 2016).

5.2.3 Extraction and analysis of HVR flanking genes

HVRs identified using anvi'o's pangenomics workflow were extracted along with flanking genes. These regions were gene called using prodigal (Hyatt et al. 2010) and gene annotation performed by diamond (Buchfink, Xie, and Huson 2015). Results are displayed in a gene annotation format using gggenes (David Wilkins 2019) and overlaid with coverage bar plots from methods described in **5.2.1.**

```
# Run Prodigal and diamond
prodigal -p meta -i <input> -d <output>
diamond blastx -d <diamond_nr_database> -q
<prodigal_genecalls> \
-o <diamond_output> --outfmt 6 qseqid evalue bitscore stitle
-k 1
--more-sensitive


# Run gggenes in Rscript
library(ggfittext)
library(gggenes)
library(ggplot2)
library(ggrepel)
library(RColorBrewer)
virus <- read.csv("<VIRSorter_diamond_output>", sep = "\t")
colourCount = length(unique(virus$COG_cat))
getPalette = colorRampPalette(brewer.pal(9, "Set1"))
ggplot(virus, aes(xmin = Start, xmax = End, y = Genome, fill =
COG_cat, forward = Strand, label = role)) +
  geom_gene_arrow(arrowhead_height = unit(10, "mm"),
arrowhead_width = unit(3, "mm"), arrow_body_height = unit(10,
"mm") ) +
  geom_gene_label(align = "middle") +
  facet_wrap(~ Genome, scales = "free", ncol = 1) +
  scale_fill_manual(values = colorRampPalette(brewer.pal(9,
"Set1"))(colourCount)) +
  theme_genes() +
  theme(legend.position="bottom") +
  guides(fill=guide_legend(nrow=5))
ggsave("<gggenes_output>", width = 420, height = 297, units =
"mm", limitsize = FALSE, dpi = 150)
```

5.2.4 Extraction and analysis of HVR2

All publically available SAR11s from the NCBI database and 451 SAR11 SAGs in this study from clade I-III (generated using previously established phylogenetic trees in **Chapter 3**) were used to obtain their HVR2 region, located between the 23S and 5S rRNA regions (Grote et al. 2012; L. J. Wilhelm et al. 2007). Therefore any SAR11 genomes with a 23S and 5S rRNA gene present on the same contig were identified using barrnap (T. Seemann 2015). The HVR2 region was extracted from each contig using seqtk (H. Li 2012). Both HVR2 region and the whole genome sequence for each organism were gene called with prodigal, as indicated in methods in **5.2.3** and gene annotated using the webserver based eggNOG-MAPPER (Huerta-Cepas et al. 2016) to establish Clusters of Orthologous Groups (COGs) (Tatusov et al. 2000) for each gene. The proportion of COGs in HVR2 and the whole genome were compared to assess if HVR2 is enriched in any COG category. Enrichments were tested for significance with a hypergeometric distribution test (Fog 2008).

```
# Establish rRNA locations with barrnap
barrnap <fasta_file>

# Extract HVR2 using seqtk
seqtk subseq <concatiated_fasta_all_SAR11> reg.bed >
ITS_reg.fasta


# Run hypergeometric distribution test with scipy in python
# Function from article "Hypergeometric Distribution Explained
#With Python" from towardsdatascience.com
import numpy as np
import matplotlib.pyplot as plt
from scipy.special import comb

def hypergeom_pmf(N, A, n, x):

    '''
    Probability Mass Function for Hypergeometric Distribution
    :param N: population size
    :param A: total number of desired items in N
    :param n: number of draws made from N
    :param x: number of desired items in our draw of n items
    :returns: PMF computed at x
    '''
    Achoosex = comb(A,x)
    NAchoosenx = comb(N-A, n-x)
    Nchoosen = comb(N,n)

    return (Achoosex)*NAchoosenx/Nchoosen
```

5.2.5 Bootstrapping of HVR2

The sample sizes of HVR2 in Clade IV and V was 11, making the calculation of its size distribution difficult to determine accurately. Here bootstrapping methods were used to resample HVR2 distributions to provide a larger sample size. Bootstrapping consisted of randomly sampling existing HVR2 datasets over 100,000 iterations to produce additional data points.

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

def bootstrap(sample_list):
    # perform bootstrapping from sample list
    boot_means = []
    for _ in range(100000):
        bootsample = np.random.choice(sample_list,
size=len(sample_list), replace=True)
        boot_means.append(bootsample.mean())
    print("Bootstrapped Mean genes = {}, 95% CI =
{}".format(np.mean(boot_means), np.percentile(boot_means,
[2.5, 97.5])))
    return boot_means

# plot histogram
sns.set(rc={'figure.figsize':(18,12)})
sns.set_style("whitegrid")
sns.distplot(bootstrap(cld123_gene), color='#0173b2', label=
"SAR11 Clades I,II,III")
sns.distplot(bootstrap(cld45_gene), color='#882D72',
label="SAR11 Clades IV, V")
plt.xlabel("Number of genes in HVR2")
plt.ylabel("Frequency")
plt.xlim(10, 70)
plt.legend()
```

5.3 Results

5.3.1 Coverage of SAR11 reference genomes against metagenomic datasets

Mapping of 15 complete SAR11 genomes against the (Biller et al. 2018) dataset revealed multiple candidate sites for HVRs, defined as regions over 10k base pairs and greater than a ten-fold coverage as a low level of read recruitment was still observed in some areas of HVRs (**Fig 5.1**). Areas of shallow or no mapping are highlighted as potential HVR regions (**Fig 5.2**). Multiple HVR regions were identified in SAR11 genomes.

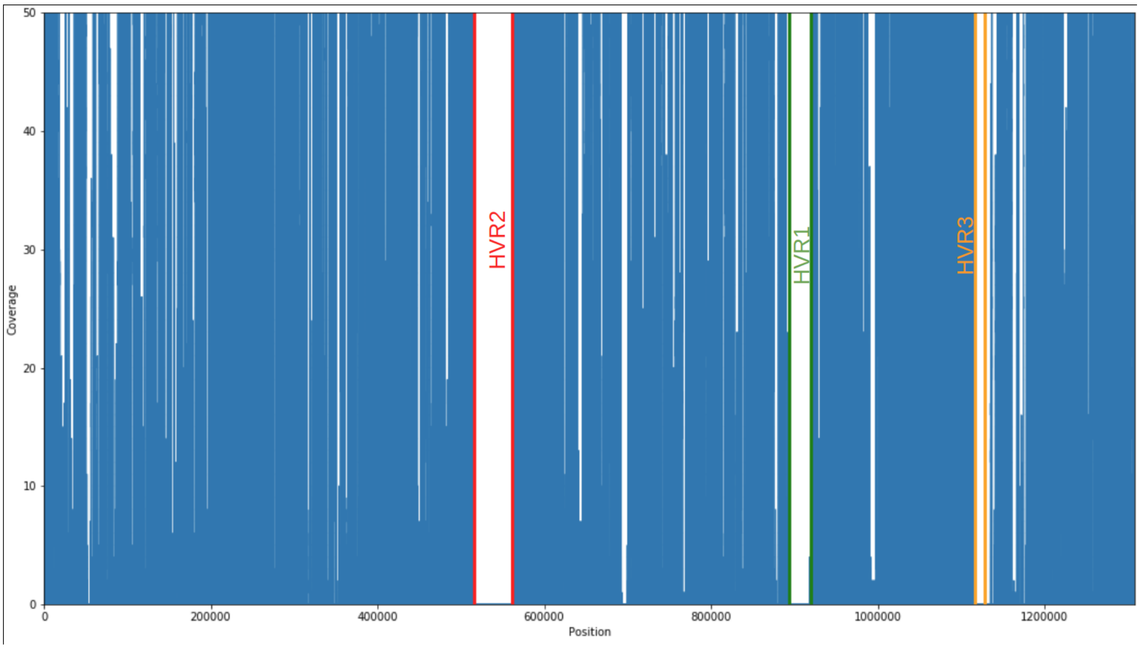


Figure 5.1 Coverage plot of HVRs identified in SAR11 HTCC1062. Proportions of the genome with identical genetic material present within metagenomic samples are represented with higher coverage in blue. HVRs, regions where genetic material is explicitly different from any metagenomic data, are identified as low coverage regions and highlighted.

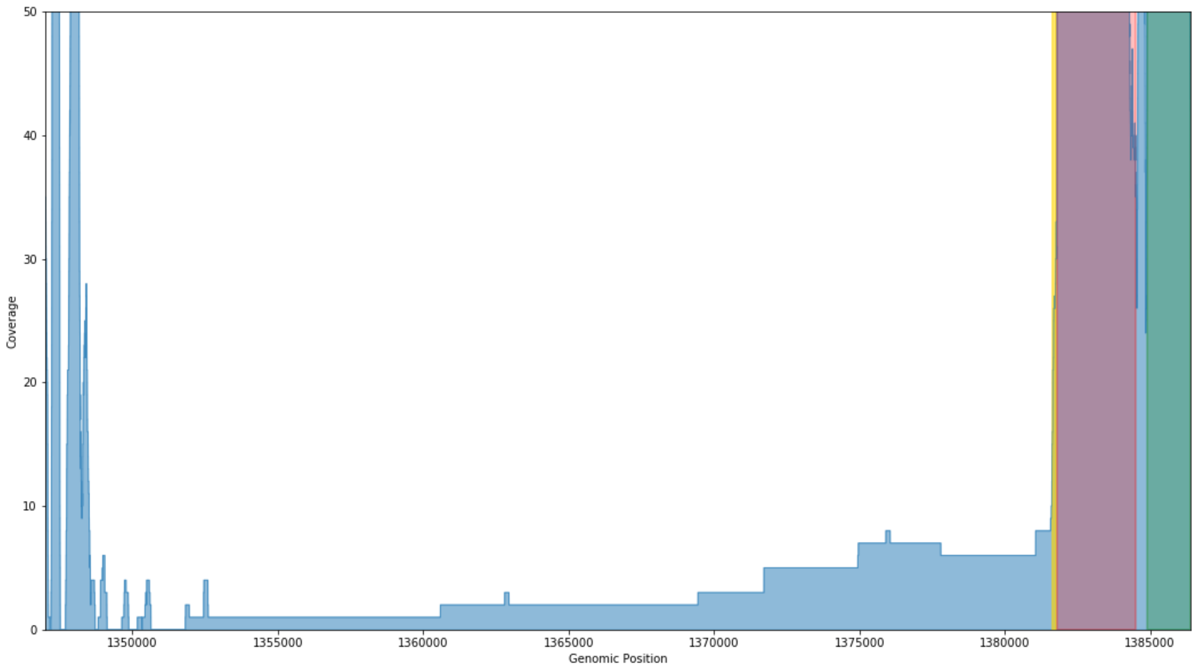


Figure 5.2 Coverage plot of HVR2 in AG-333-C14 a member of SAR11 clade IV mapped against metagenomic data from the (Biller et al. 2018) dataset. Yellow indicates the 5S, red 23S and green 16S rRNA coding regions. A "stepped" coverage is observed from the 5S rRNA region.

5.3.2 Pangenomics of SAR11 *spp.*

HVRs, by their nature, are different from any other genetic material with the same taxa (Langille, Hsiao, and Brinkman 2010; Johnson and Grossman 2015). Therefore to deduce if all SAR11 have the same HVR located in the same location on their genome or whether they perform the same function, an analysis of these regions is needed. This is because HVRs can arise from other genetic events such as artefacts of deactivated prophages through horizontal gene transfer events (Lindell et al. 2004; Sullivan et al. 2005). Organisms are likely to locate their HVRs with the same flanking genes to allow for recombination events between HVRs (Juhas et al. 2009; Johnson and Grossman 2015). These flanking genes act as "anchors", acting as areas of homology where other HVRs can bind to allowing the HVR region to recombine. The location of each HVR was first analysed with flanking genes - genes immediately upstream and downstream of HVRs, through a pangenomic analysis. A pangenomic analysis would identify identical genes between different genomes, establishing a "core" genome of a group of organisms.

Should organisms have the same HVR, they may have the same flanking genes around this HVR.

Pangenomic analysis of three representatives of each clade of SAR11, totalling 28 tested genomes confirmed HVRs within all clades I-III of SAR11.

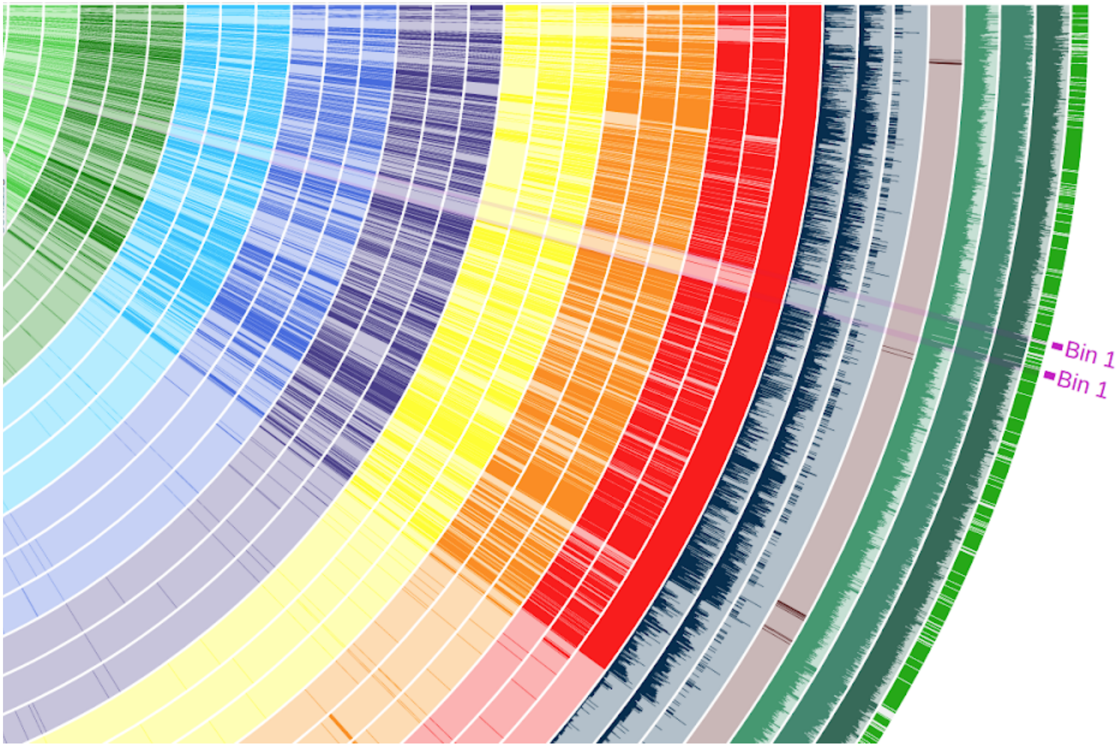


Figure 5.3 Pangenome of three representatives of each SAR11 clade. Ia (red), Ib (orange), Ic (yellow), IIa (Navy), IIb (steel blue), IIc (light blue), IIIa (dark green), IIIb (light green). Bin 1 represents flanking genes identified around a HVR in SAR11 spp. HTCC1062.

5.3.3 Extraction and analysis of HVR2 flanking genes

HVRs and flanking genes extracted from anvi'o were analysed for their genomic content. Only flanking genes around HVR2 were found to be identical in tested SAR11 spp. HVR2 is contained between the 23S to 5S region within SAR11 clades I-III (**Fig 5.4-5.5**), corroborating findings from (Grote et al. 2012)across 75 genomes. HVR2 was located in a different location (after the 5S rRNA) in clade IV-V and therefore included in a separate downstream analysis. Analysis of flanking genes around HVR2 was performed to find if more than one gene, rRNA or tRNA was consistently outside of HVR2, looking for conserved genes present when comparing Clades I-III against IV-IV (**Fig 5.4-5.5**).

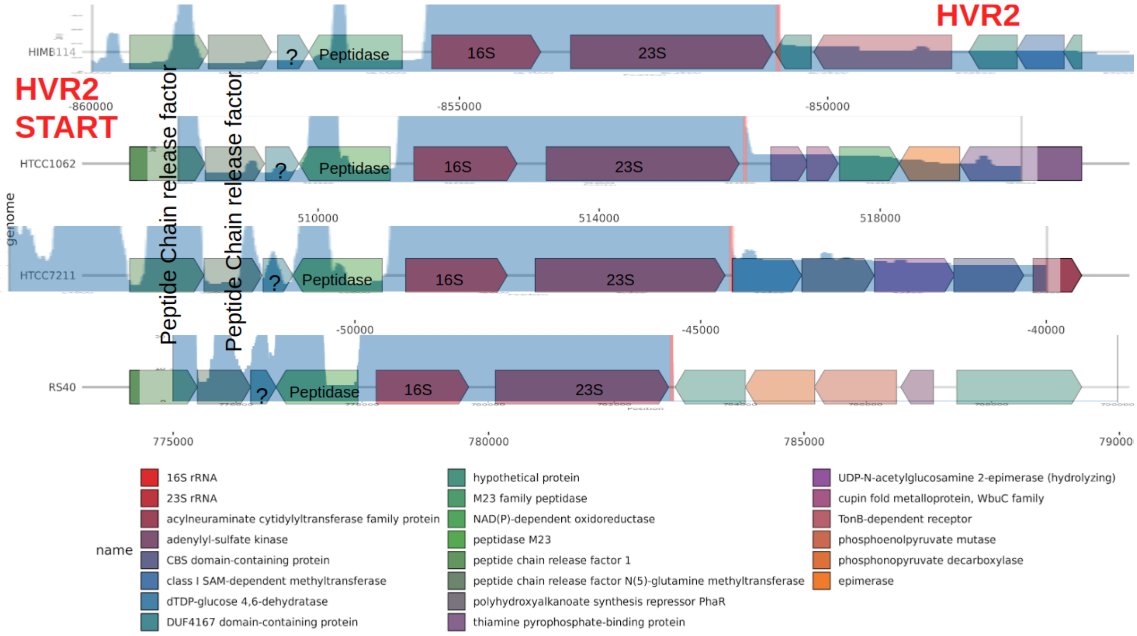


Figure 5.4 Start location of HVR2 located between the 23S and 5S rRNA coding regions with various SAR11 clade I-III organisms used as a reference. (Top row: HIMB114 Clade IIIb, second row: HTCC1062 Clade Ia.1, third row: HTCC7211 Clade Ia.3, Bottom row RS40 Clade Ib).

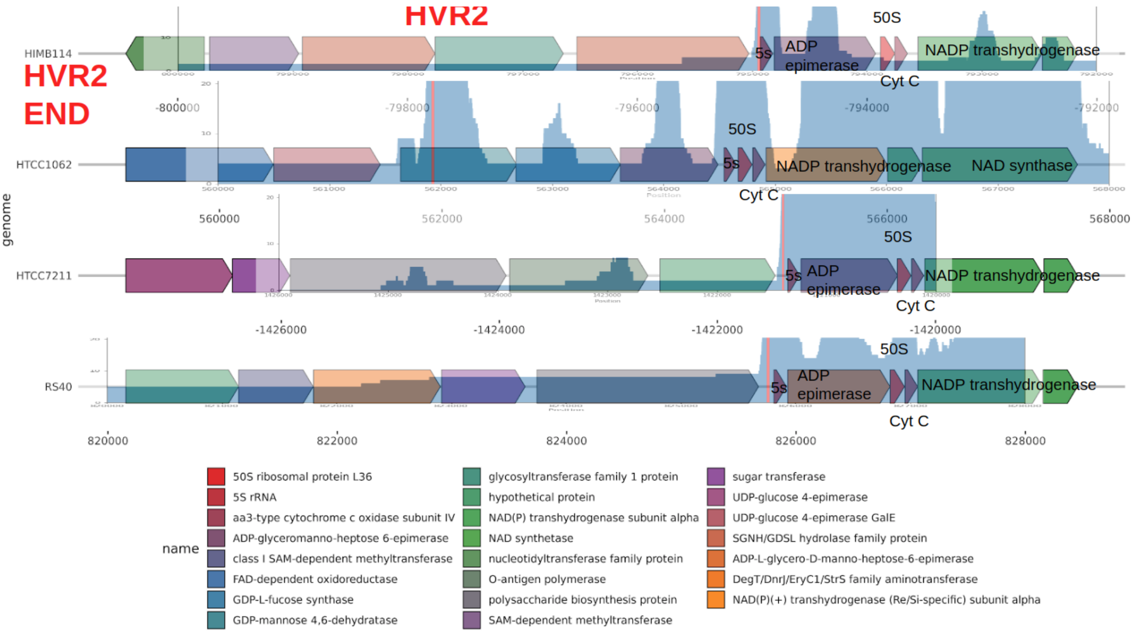


Figure 5.5 End location of HVR2 located between the 23S and 5S rRNA coding regions with various SAR11 clade organisms used as a reference.

5.3.4 HVR analysis and Hypergeometric distribution test of enrichment of COG categories

SAR11 HVR2 was confirmed to exist between the 23S and 5S rRNA region in all complete publically available SAR11 in clade **I-III**. 75 SAGs were confirmed to have this HVR2 region on the same contig in SAR11 clades **I-III**, and their length and genes were analysed (**Fig 5.8**). However, in SAR11 clades IV-V, HVR2 was located outside of the 23S-5S rRNA and analysed separately to assess if their content had any variation. Only 11 genomes were identified with HVR2 on the same contig in SAR11 clades IV-V (**Fig 5.8**). BLAST analysis of gene functions revealed multiple unknown genes and proved difficult to analyse due to large numbers. Instead, Clusters of Orthologous Groups (COG) category was determined and a hypergeometric test was performed to look for over-representation of COG function in HVRs. No COG category was shown to have significant enrichment of COG categories with the exception of category M for cell wall/membrane/envelope biogenesis. All produced p-values under 0.05 and therefore significant.

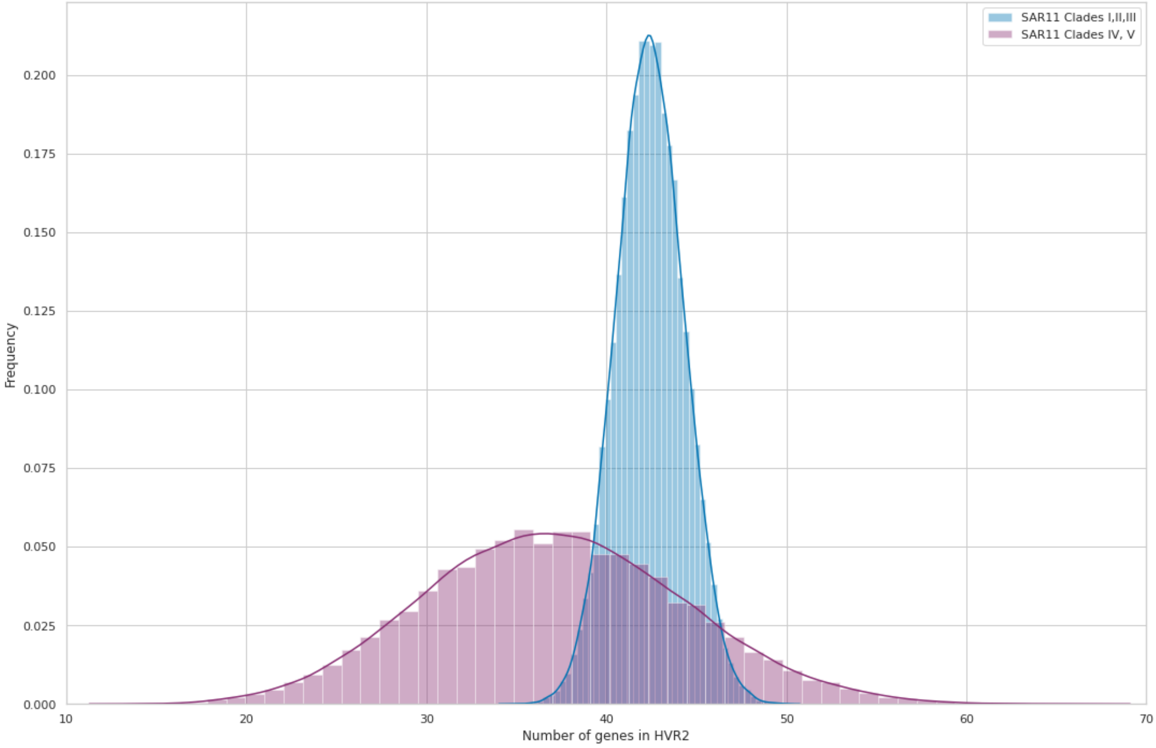


Fig 5.6 Bootstrapped histogram (n=100,000) of the number of genes identified within the HVR2 region in SAR11 clades.

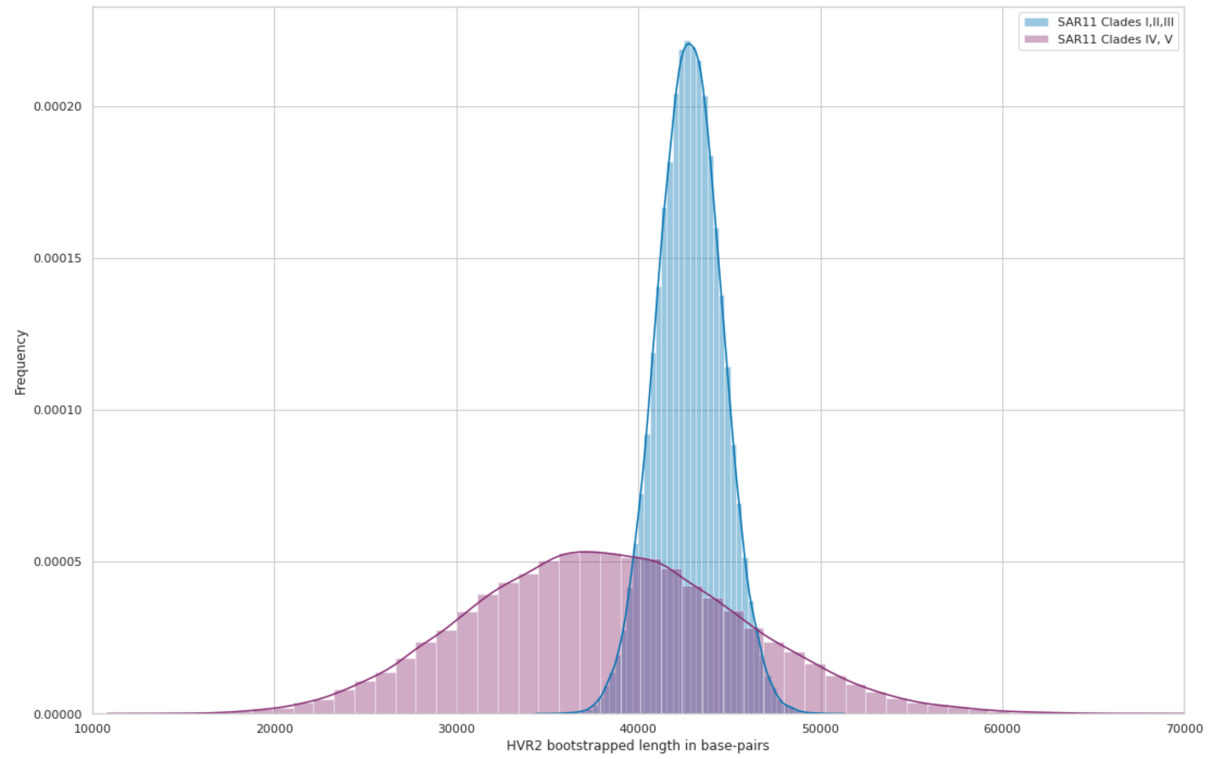


Fig 5.7 Bootstrapped histogram (n=100,000) of the lengths of HVR2 in base pairs in SAR11 clades.

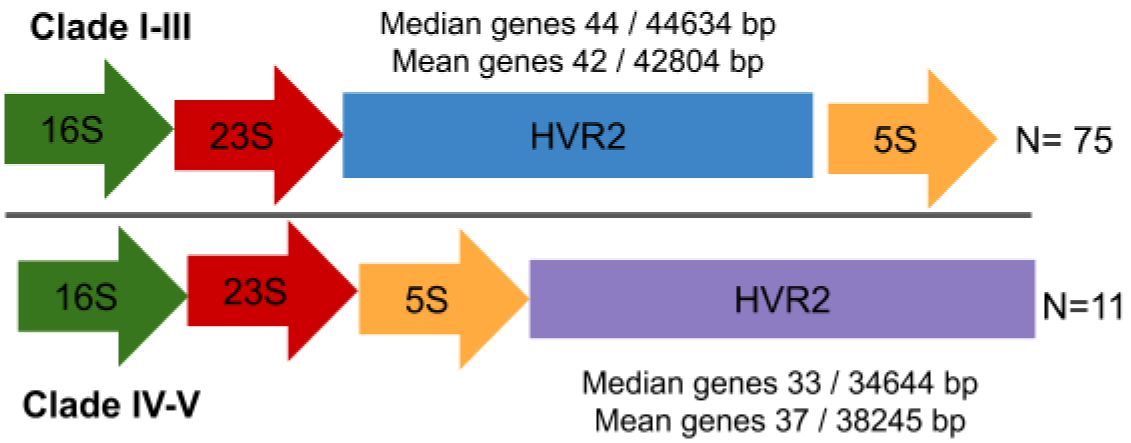


Fig 5.8 Figurative diagram of the location of HVR2 within SAR11 clades with median and mean size of the region.

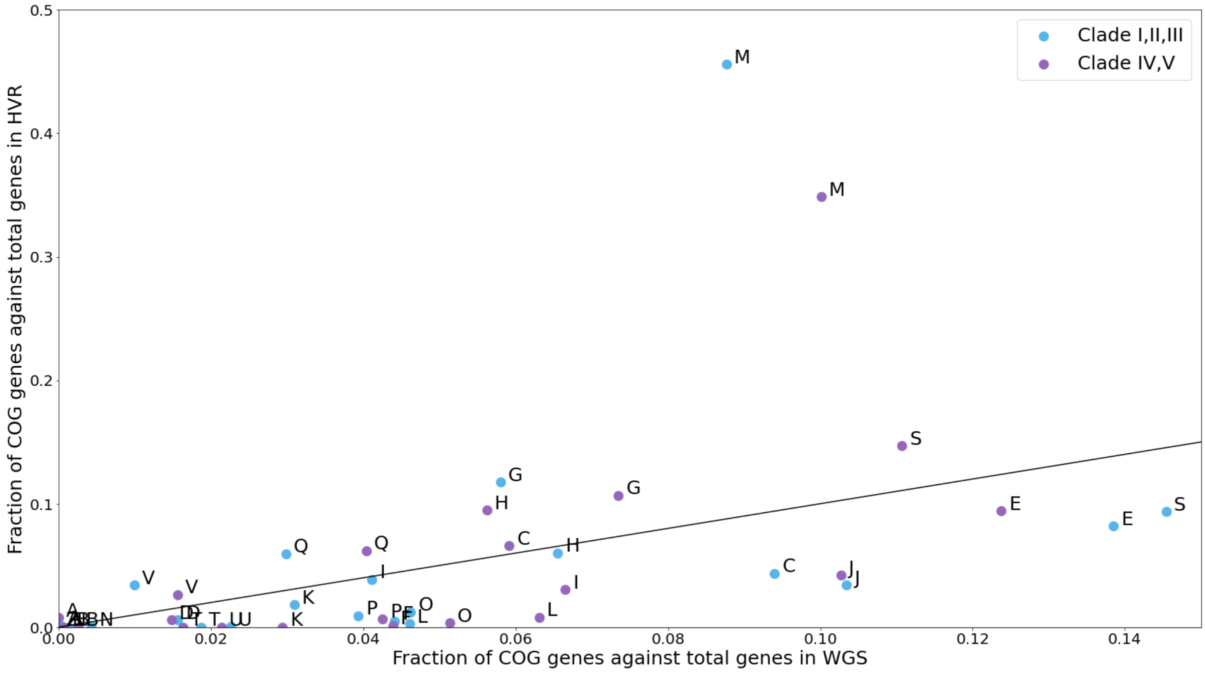


Figure 5.9 COG distribution of HVR regions against host genomes in SAR11 Clades I-III (blue) and IV-V (purple). Expected 1:1 ratio of the distribution of COG genes in HVR versus WGS is shown as a black line where y=x. Hypergeometric distribution test shows significant enrichment in COG M within HVR2 in all clades, $p < 0.05$.

5.3.5 The RecF Pathway

In order for bacteria to undergo homologous recombination, several genes are required. This is normally performed by the RecBCD pathway (not present in SAR11) or the alternative and less frequent RecF pathway (Hiom 2009). The RecF pathway comprises of seven proteins (RecF, RecO, RecR, RecA, RecJ, RecQ and the single-strand-binding protein SSB) of which critical for joint molecule formation include RecO, RecR, RecA and RecJ (Handa et al. 2009). RecJ an endonuclease, cleaves ssDNA with RecQ, a helicase assisting (Lovett and Sutera 1995; Harmon, Brockman, and Kowalczykowski 2003). SSBs coat ssDNA preventing self-complementation (Kowalczykowski et al. 1987) and are displaced by the RecOR complex along with RecF (Inoue et al. 2008). The RecOR complex is required for loading of the RecA that performs recombination and strand invasion (Stasiak, DiCapua, and Koller 1983; Umezu, Chi, and Kolodner 1993). RecF pathway genes are all found in reference SAR11 from **Table 5.1** (which all contain RecJ, RecG, RecO, RecR, RecA and SSB). Although they lack RecQ, RecG encodes for a helicase and is proposed to replace RecQ's helicase function (Z. Sun et al. 2015). Additionally, SAR11 does not have a RecF but is not critical for recombination to occur (Sakai and Cox 2009). Therefore, SAR11 encodes the RecF family of genes and it is likely that it undergoes homologous recombination in some capacity, perhaps at HVRs.

5.3 Discussion

5.3.1 The impact of HVR2 on viral ecology and SAR11 phylogeny

HVRs have been previously identified in SAR11 spp. (Grote et al. 2012; L. J. Wilhelm et al. 2007) but their distribution or depth across the breadth of the SAR11 phylogenetic tree nor its contents have not been previously studied. This study shows that all SAR11s within clades I-III have HVR2 between the 23S and 5S rRNA region, and SAR11 clades IV-V after the 5S region (**Fig 5.8**). No other HVR was identified between all SAR11 clades. Although there may be other SAR11 HVRs, no common flanking genes, rRNA or tRNA coding positions were identified that would encompass these regions. HVR2 was highly variable in length, ranging from 10k to almost 90k base pairs and exists as a normal distribution of lengths centred around ~44k base pairs. This is consistent with findings from (Grote et al. 2012) of about ~48k base pairs. It is unclear if (Grote et al. 2012) included rRNA coding regions in their analysis, but if included unlike in this study, sizes would be closely consistent. As previously stated by (Y. Zhao et al. 2013) and in a review by (Stephen J. Giovannoni 2017), HVR2 is statistically enriched in cell membrane related proteins which may include proteins related to Pelagiphage recognition and subsequent adsorption to the cell. Here I show supportive evidence of this statement otherwise briefly mentioned in previous studies (Grote et al. 2012; L. J. Wilhelm et al. 2007). Interestingly, no enrichment for genes related to nutrient uptake was found. It would be imagined a nutrient specialist would benefit from a region that undergoes high evolutionary rates containing genes related to resource acquisition. Reasons for this were not explored further within this study.

An alternative theory of defensive specialism may also be valid here, as it can clearly be seen that SAR11 does have a defensive mechanism against phage predation, - an HVR region. However, (S. Giovannoni, Temperton, and Zhao 2013) suggests that a specialist requires a majority of resources dedicated towards a single purpose, and when comparing the ~44 genes within this HVR2 region compared to the ~1100 other genes a SAR11 organism has, it would be a fair assumption that if resources are distributed based on the number of genes, SAR11 is not a defensive specialist. SAR11 have 678 genes unique within clade I-III, (Grote et al. 2012) with ~44 being a minority fraction. Instead, the HVR2 region provides enough variation to escape phage predation when challenged with limited resources.

There is some debate to the phylogenetic placement of SAR11 clade V as a true member of the SAR11 clade, with studies indicating it is not a true member (Viklund et al. 2013), or others suggesting its membership as a sister clade (Thrash et al. 2011; Ferla et al. 2013) and other studies omitting clades IV and V entirely (Kraemer et al. 2019; Haro-Moreno et al. 2020). Within this study, I show support that SAR11 clades IV and V should be classified differently to Clades I-III due to the different location of HVR2. Although conserved in function, their locations differ slightly enabling events such as homologous recombination impossible as they would lack identical flanking genes required to anchor such mechanisms. This is especially noted in clade IIIa, which are freshwater SAR11 containing HVR2 in the same location as their marine clades I, II, IIIb. Freshwater SAR11's inhibit an entirely different environment, yet combined with phylogenetic evidence in **Chapter 3** this indicates they are more closely related to clades I-III compared to other marine SAR11s clade IV and V. However, it should be noted that with a conserved function and almost identical location, I suggest that the HVR2 mechanisms of recombination come from a common ancestor and therefore SAR11 clades I-III and IV-V should still be closely related.

Additional differences include the size of SAR11 Clade IV and V's HVR2, of roughly ~38k base pairs compared to Clade I-III's of ~44k may suggest a smaller size, but with their HVR2 regions showing an overlapping distribution on bootstrapped histogram plots (**Fig 5.6-5.7**), I suggest this is a product of small sample size (n=11). Based on this, I would still suggest they contain a similar size and distribution as HVR2 from SAR11 Clade I-III. The lack of conserved flanking genes around HVR2 may be an example of a speciation event, providing evidence that SAR11 clade I-III and IV-V should be classified as taxonomically different. Lastly, a flanking gene could not be found to establish the end of the genomic cassette of HVR2 in SAR11 clade IV-V and instead metagenomic mapping data were required to establish this location (**Fig 5.2**). This may lend evidence that HVR2 does not undergo recombination as its main form of genetic diversity and is explored later on.

Overall, these points support the King-of-the-Mountain hypothesis that SAR11 is able to exchange advantageous genes in HVR2 by recombination events with other SAR11, conferring resistance to phage infection and allowing it to retain its ubiquity in the presence of abundant Pelagiphages. The transferred genes would undergo high levels of recombination in other SAR11s, an example of positive density-dependent selection and allow for parallel evolution of additional advantageous genes. This allows the SAR11 population to collectively "out evolve" its predatory phages by recombining genes that encode for cell receptors that result in phage evasion. This mechanism is conserved in all SAR11 clades and may be one of the defensive mechanisms that allow SAR11 to retain its ubiquity with minimal resource cost. This allows it to dedicate a majority of its resources to nutrient acquisition, consistent with previous findings (Sowell et al. 2009; Malmstrom et al. 2005). SAR11 is not the only species with HVRs, as these are present in *Prochlorococcus* as genomic islands (Avrani et al. 2011). Further studies into these regions may help to inform the exact mechanistic process of host-phage defence and the genetic arms race between host and phage, encapsulated by the Red Queen Hypothesis.

5.3.2 Methods of Genetic variation in HVR2

There are three main mechanisms of horizontal gene transfer: transformation, the uptake of DNA from the environment; transduction, DNA transfer via bacteriophages; and conjugation, the transfer of DNA by mobile genetic elements (Thomas and Nielsen 2005). SAR11 is an extremely streamlined cell without plasmids or any known transposable elements (Stephen J. Giovannoni

2017), reducing the likelihood DNA is transferred through conjugation, although still possible with a type IV pilus (Y. Zhao et al. 2013; X. Zhao, Schwartz, and Pierson 2017) and the RecF pathway required in homologous recombination. Transduction is unlikely to be the main mechanism of DNA transfer as resistance to viral predation is unlikely to be propagated by viruses themselves, although perhaps as a strategy to prevent additional infections in an infected host (Bondy-Denomy et al. 2016). Therefore horizontal gene transfer by transformation seems the most probable, with the intake of genetic material by SAR11's type IV pilus system. DNA fragments from previously lysed SAR11 may be recycled in this way. Homologous recombination acts as a mechanism in which this can occur, exchanging genetic material from the donor strand to the host. This requires surrounding areas of homology between DNA fragments to facilitate the exchange of non-homologous genetic material. HVR2 is flanked by both the 23S and 5S rRNA coding regions to allow for possible recombination events. However, unless used as a method of repair where donor strands are used as templates to repair gaps and loss of genetic material, recombination events are a one to one exchange. It is unlikely that free environmental DNA would be an advantage to exchange for due to the likelihood they came from conspecifics that have not escaped predation. Nor would conjugation events be beneficial for the host, should they encode for genes able to escape viral predation and exchanging them with another conspecific which may not encode the beneficial genes. Instead, recombination events would need to include a method of partial recombination, where this partial change is enough to create variation and therefore viral evasion.

Natural transformation of chromosomal DNA fragments of large sizes can occur at high frequencies if two flanking regions of high similarity are present (Nielsen, Bones, and Van Elsas 1997). This can also lead to the addition of DNA sequences within the host where recombination and strand exchange results in substitutions and integrations of DNA sequences. This process is named homology-facilitated illegitimate recombination (Meier and Wackernagel 2003; de Vries, Herzfeld, and Wackernagel 2004) where integrations of over 1000 base pair DNA fragments into the host's genome have been observed (de Vries and Wackernagel 2002; Prudhomme, Libante, and Claverys 2002). Therefore I think it is more likely that SAR11 undergoes nonhomologous recombination with the insertion and deletion of genetic material. This is further supported in this work with the highly variable length of HVR2. However, genetic recombination is a complicated mechanism with transduction events usually in response to specific environmental signals and requiring up to 50 different proteins (Thomas and Nielsen 2005). This is further exacerbated in requiring its host's ability to take up extracellular DNA, exhibited by about 1% of bacterial species (Jonas et al. 2001). SAR11 is highly abundant in nutrient-poor environments and with its highly streamline genome, suggest resources are spent predominantly on resource acquisition. It would be unlikely that such a ubiquitous process was not highlighted in metaproteomic studies (Sowell et al. 2009).

An additional mechanism of achieving variation within HVR2 is increasing mutation rates within such regions. This allows for daughter cells to have variations in their HVR compared to their mother cell and possibly create immunity. This could be an area where DNA polymerase makes more mistakes than normal when making copies of the (mother) template strand during replication. There are several points of evidence that may suggest SAR11 exhibit this mechanism. When sequencing DNA from a SAR11 culture, coverage of HVR2 is extremely low indicating that DNA sequences in this area are highly variable and provide no consensus. This can be seen in strain NP1 Extended Data **Fig. 1** (Morris et al. 2020) where HVR2 is located between ~600,000 and ~640,000. This area contains a drop in coverage depth of ~50% and elsewhere else remains roughly consistent throughout the rest of the genome. This could be evidence of recombination of conspecifics within an axenic culture, but if generations are derived from the same mother cell, there would arguably be little to no genetic variation in daughter cells. This suggests when sequencing an axenic culture, recombination events would allow recombination of the same genetic material providing no overall net change. Here I suggest that the HVR2 region undergoes a higher rate of mutation due to higher rates of DNA polymerase mistakes, allowing daughter cells to have slight changes in their HVR2 region, creating this area of low coverage in an axenic population of SAR11.

Additional evidence is the “stepped” coverage within HVR2 (**Fig 5.2**), starting from the 23S rRNA (or 5S in Clade IV-V). This is the direction transcription is performed and gradual losses in sequence coverage are observed the further the distance from the 23S rRNA region. This “jarring” loss of coverage could be the result of point mutations occurring within the HVR2 region, losing homology with other HVR2 regions. This would suggest that gradually enough point mutations lead to a distinctive HVR2 region, but require several of these point mutations to occur. This would result in coverage data to appear as “steps”. If recombination of entire genes were to occur, there would be a sudden loss of coverage with no steps, only seen at the start and end of HVR2 and not throughout.

It would be interesting to test these hypotheses through evolution studies achieved with long-read sequencing studies of the SAR11 genome. SAR11 grown over time could be periodically harvested and the HVR2 region sequenced to determine as a population whether or not large amounts of DNA are transferred and/or small point mutations observed. Should SAR11 recombine their HVR2 regions, this would indicate that SAR11 requires large populations in order to create viral immunity, or that given enough time any SAR11 will create its own immunity through point mutations, assuming HVR2 undergoes variations faster than Pelagiphage evolution. This would explain how SAR11 initially gains its ubiquity currently explained by its superior nutrient acquisition solely outcompeting viral predation. Another interesting observation would be how SAR11 is able to generate changes to its HVR2 region, as the current assumption is that recombination of genetic material is a one to one exchange where there would be no net benefit to the individual or population. Sequencing of HVR2 would reveal how these regions change and suggest a recombination method.

6 Project Discussion

Single-cell Amplified Genomes (SAGs) and metagenomics both provide valid ways of studying marine populations without the need to culture organisms however, both have their strengths and drawbacks. It becomes clear that SAGs are more suited for describing genetic and clade wide dynamics but, without metagenomics, ecological predictions and abundance data would be impossible. Metagenomics allows for assessment of presence-absence of individuals but, without reference genomes like SAG data, individual clades would be difficult to determine as only higher taxonomic levels can be consistently discerned.

It appears it is difficult to extract high-quality SAR11 genomes from marine metagenomic data (**Chapter 2**, (Tully, Graham, and Heidelberg 2018). This is suspected to be due to a mixture of reasons, mainly short reads being unable to assemble into long contiguous regions due to shallow sequencing depth of samples (**Fig 2.2**). Within this study, it would appear that marine samples are very closely related, likely due to their high microdiversity (**Fig 2.12, 2.14, 2.16**) and therefore separating out closely related reads or contigs from SAR11 genetic material is challenging (**Fig 2.19, 2.20, Table 2.2**). Therefore, metagenomic experiments could be repeated using long-read technology where the goal is to obtain longer reads and therefore longer contiguous regions over higher coverage. Longer reads would provide more unique sequences allowing for more distinctive *k*-mer frequencies and coverage depths. This should lead to the extraction of a few high-quality MAGs over large numbers of low quality and undetermined marine microorganism genomes. This has been shown successful in wastewater treatment plant metagenomes recovering 1045 high quality MAGs with 37 as circularised and complete (Singleton et al. 2020). This would be advantageous as genomes of low completion or of unknown taxonomy are challenging to translate into meaningful biological data. Increasing SAR11 MAG recovery rates would also benefit from larger databases to allow for assembly by reference mapping, shown to work well in low diversity microbiome projects (Sharon et al. 2013). Although SAR11 MAG recovery was unsuccessful with metagenomes provided within this study (**Table 2.3, 2.4, 2.5**), they still provide valuable species presence-absence data and can contribute to our understanding of marine microorganisms ecological niches and population dynamics.

SAG technology allowed for the extraction of 451 SAR11 genomes (**Chapter 3**) of varying quality (**Table 3.2**). Although some additional bioinformatics methods to pool together reads from similar strains (Kogawa et al. 2018) could be explored in order to improve the completeness of these SAGs, it was decided that the loss of strain-level variations outweighed any additional genome completeness. Assembly of SAR11 genomes and subsequent phylogenetics led to the discovery of two new uncharacterised clades of SAR11 bacteria, Id and Ie (**Fig 3.1, 3.2, 7.1**). The usage of Average Nucleotide Identity supported branching structure within phylogenetic trees, and the characterisation of new clades (**Fig 3.4, 3.5**). Mapping of these SAR11 clades to marine metagenomic data led to the discovery of each clade inhabiting different ecological niches (**Fig 3.9**). Clade Id and Ie's presence in the bathypelagic and abyssopelagic respectively (**Fig 3.8**) extend our current knowledge of SAR11 clades that inhabit these deep-sea regions. This confirmed that the discovery of new SAR11 clades was consistent with new ecotypes but also highlights the migratory nature of SAR11 conspecifics, driven by ocean currents (**Fig 7.2-7.15**). Comparative population genomics should also be pursued within this dataset, assessing how different SAR11 species at different depths and geographical locations differ in their gene content. This would help to expand our understanding of the core SAR11 genome, but also unique genes required in each ecological niche. This was not pursued within this study due to time constraints. These assessments of the SAR11 genome would help to inform streamlining theories in the minimum required genes needed for survival and replication (Stephen J. Giovannoni, Cameron Thrash, and Temperton 2014).

Analysis of SAR11 SAG genomes with viral signature determining algorithms (**Chapter 4**) led to the identification of 21 novel phage genomes. All 21 phage genomes shared gene homology with at least one existing pelagiphage from the (Y. Zhao et al. 2013) study in vContact2 (**Fig 4.2**). Gene annotation of phages revealed a host of novel genes of unknown function highlighting the need for improved databases to ascertain protein functions (**Fig 4.1**). Phages were also mapped to metagenomic data and compared to host abundance and ecological niche. Generally, phage and host mapping were closely related (**Fig 7.16-7.21**), indicating narrow phage specificity however, three phages isolated from clade Ib were shown to inhabit different areas of the ocean (**Fig 4.3, 4.4**). This provided some evidence that some SAR11 phages may have a broader specificity, allowing infection of other clades. However, this could also be the result of infection of SAR11 from clade Ib outside of their normal habitat range due to ocean currents. Wet lab cross infectivity tests of other SAR11 clades with the same phage would be more conclusive to determine Pelagiphage specificity.

When comparing SAR11 infection rates with other marine microorganisms by viruses, SAR11 undergoes a three-fold reduction in Pelagiphage infections (**Fig 4.5**). Two of the most popular viral signature determining algorithms were compared to ensure results were not false positives and a combined approach was implemented to ensure this. Overall, rates of lysogeny in SAR11 conflicted with the Piggyback-the-Winner ecological theory, suggesting that SAR11 have a different mechanism to escape viral predation. A hypervariable region (HVR) was suggested as the mechanism that SAR11 possesses to enable viral evasion as a population (Y. Zhao et al. 2013; S. Giovannoni, Temperton, and Zhao 2013). Exchange and recombination of this region with other conspecifics would allow variation to be achieved in surface relation proteins, critical in viral adsorption into the host. Therefore, assessment of the presence of HVRs on all SAR11 species was required to confirm this.

Mapping of SAR11 clades to metagenomic data led to the confirmation that all SAR11 clades have HVRs (**Chapter 5**). However, the variation of genetic content in SAR11 HVRs made it difficult to determine if HVR regions within one conspecific are associated with another. Flanking genes were used as a method of determining similar HVRs: genes observed on the outskirts of HVR regions are assumed to be the same in all SAR11 species in order to allow homologous recombination. Flanking genes would also be critical in recombination events to allow for the exchange of non-homologous genetic material. Pangenomic analysis of SAR11 species allowed for grouping of similar proteins in all SAR11 genomes, allowing for confirmation that SAR11 flanking genes around any HVR were the same in all SAR11 species (**Fig 5.3, 5.4, 5.5**). It was determined that SAR11 flanking genes around HVR2 were consistent in SAR11 clades I-III, but not IV and V. HVR2 was determined to exist between the 23S and 5S rRNA coding region in SAR11 clades I-III and after the 5S rRNA coding region in SAR11 clades IV-V (**Fig 5.8**). This supports previous findings that SAR11 clades I-III are genetically different from clades IV-V (Haro-Moreno et al. 2020; Viklund et al. 2013). Without flanking genes, recombination events have no genetic binding regions to exchange genes and therefore it would be unlikely that members of clades I-III and IV-V undergo recombination events. This likely a product of a speciation event. Genomic analysis of HVR2 confirmed significant enrichment of COG category M-related genes related to cell wall/membrane/envelope biogenesis in comparison to the whole genome (**Fig 5.9**). This has been hypothesised to be related to phage defence, where variations in the cell wall of SAR11 species allow for improved phage defence by its alteration of the cell wall structure (Y. Zhao et al. 2013). This may prevent phage binding or detection of SAR11 cell wall receptors, consistent with the King-of-the-Mountain hypothesis (S. Giovannoni, Temperton, and Zhao 2013). SAR11 is able to retain its ubiquitous nature in the presence of highly abundant phages due to its ability to undergo recombination of beneficial genes related to phage defence. This highlights that SAR11 probably undergoes positive rather than negative-density dependant selection.

However, I highlight that recombination events are one-to-one exchanges of genetic material, with no net gain within a population. Without new variations, it is likely that phages would still be able to attach and infect a host cell. Therefore, I suggest that SAR11 HVR2 is able to create variation due to mistakes made in DNA replication. I offer no concrete data to support this mechanism, instead highlight observations of HVR2 that support this idea. This includes the “stepped” like nature of HVR2 observed when mapped against metagenomic datasets (**Fig 5.2**) and the reduction of coverage HVR2 exhibits when sequencing axenic cultures of SAR11 (Morris et al. 2020). This would suggest HVR2s in daughter cells are slight variations compared to their parental HVR2s and external-facing cell wall receptors are altered in this way. It is possible that both mechanisms of recombination and DNA replication mistakes are at play to allow for the increased variation in HVR2 allowing SAR11 to evade viral predation. I suggest wet lab evolution studies on axenic cultures to assess mechanisms of HVR2 variation to provide more substantial evidence of HVR evolution. Due to HVR2’s high variation between individuals, long-read sequencing would be more suitable. Short-read data would be unlikely to provide long contiguous sequence due to multiple edges created when trying to assemble data with small variations, needing amplification of variable regions for contiguous sequences (Morris et al. 2020). A unique molecular identifier (UMI) approach which was effective in a study by (Karst et al. 2020) may also be effective in error correcting long reads. Should evidence be shown to support current theories, this would highlight a novel mechanism for viral evasion in the most ubiquitous marine microorganism.

To conclude, this project has been an assessment of obtaining SAR11 and Pelagiphage sequences from metagenomic and SAGs. I have concluded that current bioinformatics and wet lab methods are unable to consistently and reliably create SAR11 MAGs of high completeness. However, metagenomic data has the most potential to produce high throughput and high-quality MAGs from any environment. It is also critical for determining spatio-temporal data for biological organisms. With current technologies, SAGs provide the most complete and usable genomic data and can be used to effectively describe predator-prey dynamics where I highlight several observations to this effect. Therefore, current studies looking to obtain genetic information should pursue SAG technology in the short-term, but I highlight the greater potential and need for metagenomic data. Overall, research efforts should pursue improvements in long-read metagenomic data as it has a higher potential than SAG technology in producing meaningful biological data on a large scale. Studies into SAR11 are critical to provide insight into the global biogeochemistry of labile dissolved organic carbon and provide a model organism to elucidate the evolution and function of streamlined genomes.

Bibliography

"About | BATS." n.d. Bermuda Institute of Ocean Sciences. Accessed November 29, 2019. http://bats.bios.edu/about/.

Adriaenssens, Evelien M., and Don A. Cowan. 2014. "Using Signature Genes as Tools to Assess Environmental Viral Ecology and Diversity." *Applied and Environmental Microbiology* 80 (15): 4470–80.

Ahlgren, Nathan A., Jie Ren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. 2016. "Alignment-Free Oligonucleotide Frequency Dissimilarity Measure Improves Prediction of Hosts from Metagenomically-Derived Viral Sequences." *Nucleic Acids Research* 45 (1): 39–53.

Akhter, Sajia, Ramy K. Aziz, and Robert A. Edwards. 2012. "PhiSpy: A Novel Algorithm for Finding Prophages in Bacterial Genomes That Combines Similarity- and Composition-Based Strategies." *Nucleic Acids Research* 40 (16): e126.

Alarcón-Schumacher, Tomás, Sergio Guajardo-Leiva, Josefa Antón, and Beatriz Díez. 2019. "Elucidating Viral Communities During a Phytoplankton Bloom on the West Antarctic Peninsula." *Frontiers in Microbiology* 10 (May): 1014.

Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L. Nielsen, Gene W. Tyson, and Per H. Nielsen. 2013. "Genome Sequences of Rare, Uncultured Bacteria Obtained by Differential Coverage Binning of Multiple Metagenomes." *Nature Biotechnology* 31 (6): 533–38.

Allen, Lisa Zeigler, Thomas Ishoey, Mark A. Novotny, Jeffrey S. McLean, Roger S. Lasken, and Shannon J. Williamson. 2011. "Single Virus Genomics: A New Tool for Virus Discovery." *PloS One* 6 (3): e17722.

Alnajar, Seema, and Radhey S. Gupta. 2017. "Phylogenomics and Comparative Genomic Studies Delineate Six Main Clades within the Family Enterobacteriaceae and Support the Reclassification of Several Polyphyletic Members of the Family." *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 54 (October): 108–27.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. 2014. "Binning Metagenomic Contigs by Coverage and Composition." *Nature Methods* 11 (11): 1144–46.

Alneberg, Johannes, Christofer M. G. Karlsson, Anna-Maria Divne, Claudia Bergin, Felix Homa, Markus V. Lindh, Luisa W. Hugerth, et al. 2018. "Genomes from Uncultivated Prokaryotes: A Comparison of Metagenome-Assembled and Single-Amplified Genomes." *Microbiome* 6 (1): 173.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10.

Amann, R. I., W. Ludwig, and K. H. Schleifer. 1995. "Phylogenetic Identification and in Situ Detection of Individual Microbial Cells without Cultivation." *Microbiological Reviews* 59 (1): 143–69.

Amon, Rainer M. W., and Ronald Benner. 1996. "Bacterial Utilization of Different Size Classes of Dissolved Organic Matter." *Limnology and Oceanography*. https://doi.org/10.4319/lo.1996.41.1.0041.

Anantharaman, Karthik, Melissa B. Duhaime, John A. Breier, Kathleen A. Wendt, Brandy M. Toner, and Gregory J. Dick. 2014. “Sulfur Oxidation Genes in Diverse Deep-Sea Viruses.” *Science* 344 (6185): 757–60.

Andersen, Kasper Skytte. 2018. *mmgenome2* (version 2.0). R. Github. https://github.com/KasperSkytte/mmgenome2.

Arndt, David, Jason R. Grant, Ana Marcu, Tanvir Sajed, Allison Pon, Yongjie Liang, and David S. Wishart. 2016. “PHASTER: A Better, Faster Version of the PHAST Phage Search Tool.” *Nucleic Acids Research* 44 (W1): W16–21.

Arndt, David, Ana Marcu, Yongjie Liang, and David S. Wishart. 2017. “PHAST, PHASTER and PHASTEST: Tools for Finding Prophage in Bacterial Genomes.” *Briefings in Bioinformatics*, September. https://doi.org/10.1093/bib/bbx121.

Avrani, Sarit, Daniel A. Schwartz, and Debbie Lindell. 2012. “Virus-Host Swinging Party in the Oceans: Incorporating Biological Complexity into Paradigms of Antagonistic Coexistence.” *Mobile Genetic Elements* 2 (2): 88–95.

Avrani, Sarit, Omri Wurtzel, Itai Sharon, Rotem Sorek, and Debbie Lindell. 2011. “Genomic Island Variability Facilitates Prochlorococcus--Virus Coexistence.” *Nature* 474 (7353): 604–8.

Ayling, Martin, Matthew D. Clark, and Richard M. Leggett. 2019. “New Approaches for Metagenome Assembly with Short Reads.” *Briefings in Bioinformatics*, February. https://doi.org/10.1093/bib/bbz020.

Aylward, Frank O., Dominique Boeuf, Daniel R. Mende, Elisha M. Wood-Charlson, Alice Vislova, John M. Eppley, Anna E. Romano, and Edward F. DeLong. 2017. “Diel Cycling and Long-Term Persistence of Viruses in the Ocean’s Euphotic Zone.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (43): 11446–51.

Azam, Farooq, and Francesca Malfatti. 2007. “Microbial Structuring of Marine Ecosystems.” *Nature Reviews. Microbiology* 5 (10): 782–91.

Azam, F., D. C. Smith, G. F. Steward, and A. Hagström. 1994. “Bacteria-Organic Matter Coupling and Its Significance for Oceanic Carbon Cycling.” *Microbial Ecology* 28 (2): 167–79.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 19 (5): 455–77.

Basu, Sreemanti, Hope M. Campbell, Bonnie N. Dittel, and Avijit Ray. 2010. “Purification of Specific Cell Population by Fluorescence Activated Cell Sorting (FACS).” *Journal of Visualized Experiments: JoVE*, no. 41 (July). https://doi.org/10.3791/1546.

Bates, Nicholas R., Anthony F. Michaels, and Anthony H. Knap. 1996. “Seasonal and Interannual Variability of Oceanic Carbon Dioxide Species at the U.S. JGOFS Bermuda Atlantic Time-Series Study (BATS) Site.” *Deep-Sea Research. Part II, Topical Studies in Oceanography* 43 (2): 347–83.

Bauer, James E., Peter M. Williams, and Ellen R. M. Druffel. 1992. “14C Activity of Dissolved Organic Carbon Fractions in the North-Central Pacific and Sargasso Sea.” *Nature*. https://doi.org/10.1038/357667a0.

Beaulaurier, J., E. Luo, J. Eppley, P. Den Uyl, and X. Dai. 2019. “Assembly-Free

Single-Molecule Nanopore Sequencing Recovers Complete Virus Genomes from Natural Microbial Communities." *bioRxiv*. https://www.biorxiv.org/content/10.1101/619684v1.abstract.

Becker, Jamie W., Shane L. Hogle, Kali Rosendo, and Sallie W. Chisholm. 2019. "Co-Culture and Biogeography of Prochlorococcus and SAR11." *The ISME Journal* 13 (6): 1506–19.

Becraft, Eric D., Jeremy A. Dodsworth, Senthil K. Murugapiran, J. Ingemar Ohlsson, Brandon R. Briggs, Jad Kanbar, Iwijn De Vlaminck, et al. 2016. "Single-Cell-Genomics-Facilitated Read Binning of Candidate Phylum EM19 Genomes from Geothermal Spring Metagenomes." *Applied and Environmental Microbiology* 82 (4): 992–1003.

Beja, Oded, Jarone Pinhassi, and John L. Spudich. 2013. "Proteorhodopsins: Widespread Microbial Light-Driven Proton Pumps," 280–85.

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, et al. 2008. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature* 456 (7218): 53–59.

Bergh, O., K. Y. Børsheim, G. Bratbak, and M. Heldal. 1989. "High Abundance of Viruses Found in Aquatic Environments." *Nature* 340 (6233): 467–68.

Bertrand, Denis, Jim Shaw, Manesh Kalathiyappan, Amanda Hui Qi Ng, M. Senthil Kumar, Chenhao Li, Mirta Dvornicic, et al. 2019. "Hybrid Metagenomic Assembly Enables High-Resolution Analysis of Resistance Determinants and Mobile Elements in Human Microbiomes." *Nature Biotechnology* 37 (8): 937–44.

Biddanda, Bopaiah, and Ronald Benner. 1997. "Carbon, Nitrogen, and Carbohydrate Fluxes during the Production of Particulate and Dissolved Organic Matter by Marine Phytoplankton." *Limnology and Oceanography* 42 (3): 506–18.

Biers, Erin J., Shulei Sun, and Erinn C. Howard. 2009. "Prokaryotic Genomes and Diversity in Surface Ocean Waters: Interrogating the Global Ocean Sampling Metagenome." *Applied and Environmental Microbiology* 75 (7): 2221–29.

Biller, Steven J., Paul M. Berube, Keven Dooley, Madeline Williams, Brandon M. Satinsky, Thomas Hackl, Shane L. Hogle, et al. 2018. "Marine Microbial Metagenomes Sampled across Space and Time." *Scientific Data* 5 (September): 180176.

Binga, Erik K., Roger S. Lasken, and Josh D. Neufeld. 2008. "Something from (almost) Nothing: The Impact of Multiple Displacement Amplification on Microbial Ecology." *The ISME Journal* 2 (3): 233–41.

Blainey, Paul C., and Stephen R. Quake. 2011. "Digital MDA for Enumeration of Total Nucleic Acid Contamination." *Nucleic Acids Research* 39 (4): e19.

Bolduc, Benjamin, Ho Bin Jang, Guilhem Doulcier, Zhi-Qiang You, Simon Roux, and Matthew B. Sullivan. 2017. "vConTACT: An iVirus Tool to Classify Double-Stranded DNA Viruses That Infect Archaea and Bacteria." *PeerJ* 5 (May): e3243.

Bondy-Denomy, Joseph, Jason Qian, Edze R. Westra, Angus Buckling, David S. Guttman, Alan R. Davidson, and Karen L. Maxwell. 2016. "Prophages Mediate Defense against Phage Infection through Diverse Mechanisms." *The ISME Journal* 10 (12): 2854–66.

Bose, M., and Robert D. Barber. 2006. "Prophage Finder: A Prophage Loci Prediction Tool for Prokaryotic Genome Sequences." *In Silico Biology* 6 (3):

223–27.

Bourcy, Charles F. A. de, Iwijn De Vlaminck, Jad N. Kanbar, Jianbin Wang, Charles Gawad, and Stephen R. Quake. 2014. “A Quantitative Comparison of Single-Cell Whole Genome Amplification Methods.” *PloS One* 9 (8): e105585.

Bowers, Robert M., Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, et al. 2017. “Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea.” *Nature Biotechnology* 35 (8): 725–31.

Boyd, P. W., A. J. Watson, C. S. Law, E. R. Abraham, T. Trull, R. Murdoch, D. C. Bakker, et al. 2000. “A Mesoscale Phytoplankton Bloom in the Polar Southern Ocean Stimulated by Iron Fertilization.” *Nature* 407 (6805): 695–702.

Bradnam, Keith R., Joseph N. Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, et al. 2013. “Assemblathon 2: Evaluating de Novo Methods of Genome Assembly in Three Vertebrate Species.” *GigaScience* 2 (1): 10.

Breitbart, Mya, Peter Salamon, Bjarne Andresen, Joseph M. Mahaffy, Anca M. Segall, David Mead, Farooq Azam, and Forest Rohwer. 2002. “Genomic Analysis of Uncultured Marine Viral Communities.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (22): 14250–55.

Breitbart, Mya, Luke R. Thompson, Curtis A. Suttle, and Matthew B. Sullivan. 2007. “Exploring the Vast Diversity of Marine Viruses.” *Oceanography* 20 (2): 135–39.

Brockhurst, Michael A., Tracey Chapman, Kayla C. King, Judith E. Mank, Steve Paterson, and Gregory D. D. Hurst. 2014. “Running with the Red Queen: The Role of Biotic Conflicts in Evolution.” *Proceedings. Biological Sciences / The Royal Society* 281 (1797). https://doi.org/10.1098/rspb.2014.1382.

Brown, Christopher T., Matthew R. Olm, Brian C. Thomas, and Jillian F. Banfield. 2016. “Measurement of Bacterial Replication Rates in Microbial Communities.” *Nature Biotechnology* 34 (12): 1256–63.

Brown, Mark V., Federico M. Lauro, Matthew Z. DeMaere, Les Muir, David Wilkins, Torsten Thomas, Martin J. Riddle, et al. 2012. “Global Biogeography of SAR11 Marine Bacteria.” *Molecular Systems Biology* 8 (July): 595.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60.

Buchholz, Holger H., Michelle L. Michelsen, Michael J. Allen, and Ben Temperton. 2020. “Efficient Dilution-to-Extinction Isolation of Novel Virus-Host Model Systems for Fastidious Heterotrophic Bacteria.” *bioRxiv*. https://doi.org/10.1101/2020.04.27.064238.

Bushnell, B. 2019. “Bbtools: A Suite of Fast, Multithreaded Bioinformatics Tools Designed for Analysis of DNA and Rna Sequence Data; Joint Genome Institute: Berkeley, CA, USA, 2018.” *Jgi. Doe. Gov/data-and-Tools/bbtools (accessed on 20 September 2019)*.

Bushnell, Brian, Jonathan Rood, and Esther Singer. 2017. “BBMerge--Accurate Paired Shotgun Read Merging via Overlap.” *PloS One* 12 (10): e0185056.

Campbell, Allan. 2003. “Prophage Insertion Sites.” *Research in Microbiology* 154 (4): 277–82.

Campbell, Barbara J., Liying Yu, John F. Heidelberg, and David L. Kirchman. 2011. “Activity of Abundant and Rare Bacteria in a Coastal Ocean.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (31): 12776–81.

Campbell, James H., Patrick O’Donoghue, Alisha G. Campbell, Patrick Schwientek, Alexander Sczyrba, Tanja Woyke, Dieter Söll, and Mircea Podar. 2013. “UGA Is an Additional Glycine Codon in Uncultured SR1 Bacteria from the Human Microbiota.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (14): 5540–45.

Canchaya, Carlos, Ghislain Fournous, and Harald Brüssow. 2004. “The Impact of Prophages on Bacterial Chromosomes.” *Molecular Microbiology* 53 (1): 9–18.

Canchaya, Carlos, Caroline Proux, Ghislain Fournous, Anne Bruttin, and Harald Brüssow. 2003. “Prophage Genomics.” *Microbiology and Molecular Biology Reviews: MMBR* 67 (2): 238–76, table of contents.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics* 25 (15): 1972–73.

Carini, Paul, Emily O. Campbell, Jeff Morré, Sergio A. Sañudo-Wilhelmy, J. Cameron Thrash, Samuel E. Bennett, Ben Temperton, Tadhg Begley, and Stephen J. Giovannoni. 2014. “Discovery of a SAR11 Growth Requirement for Thiamin’s Pyrimidine Precursor and Its Distribution in the Sargasso Sea.” *The ISME Journal* 8 (8): 1727–38.

Carlson, C. A., and H. W. Ducklow. 1995. “Dissolved Organic Carbon in the Upper Ocean of the Central Equatorial Pacific Ocean, 1992: Daily and Finescale Vertical Variations.” *Deep Sea Research Part II: Topical Studies in Oceanography*. https://doi.org/10.1016/0967-0645(95)00023-j.

Carlson, Craig A., and Dennis A. Hansell. 2015. “Chapter 3 - DOM Sources, Sinks, Reactivity, and Budgets.” In *Biogeochemistry of Marine Dissolved Organic Matter (Second Edition)*, edited by Dennis A. Hansell and Craig A. Carlson, 65–126. Boston: Academic Press.

Carlson, Craig A., Dennis A. Hansell, Norman B. Nelson, David A. Siegel, William M. Smethie, Samar Khatiwala, Meredith M. Meyers, and Elisa Halewood. 2010. “Dissolved Organic Carbon Export and Subsequent Remineralization in the Mesopelagic and Bathypelagic Realms of the North Atlantic Basin.” *Deep Sea Research Part II: Topical Studies in Oceanography*. https://doi.org/10.1016/j.dsr2.2010.02.013.

Carlson, Craig A., Robert Morris, Rachel Parsons, Alexander H. Treusch, Stephen J. Giovannoni, and Kevin Vergin. 2009. “Seasonal Dynamics of SAR11 Populations in the Euphotic and Mesopelagic Zones of the Northwestern Sargasso Sea.” *The ISME Journal* 3 (3): 283–95.

Casjens, S., N. Palmer, R. van Vugt, W. M. Huang, B. Stevenson, P. Rosa, R. Lathigra, et al. 2000. “A Bacterial Genome in Flux: The Twelve Linear and Nine Circular Extrachromosomal DNAs in an Infectious Isolate of the Lyme Disease Spirochete Borrelia Burgdorferi.” *Molecular Microbiology* 35 (3): 490–516.

Chapman, Alec R., Zi He, Sijia Lu, Jun Yong, Longzhi Tan, Fuchou Tang, and X. Sunney Xie. 2015. “Single Cell Transcriptome Amplification with MALBAC.” *PloS One* 10 (3): e0120889.

Chaumeil, Pierre-Alain, Aaron J. Mussig, Philip Hugenholtz, and Donovan H. Parks. 2019. “GTDB-Tk: A Toolkit to Classify Genomes with the Genome

Taxonomy Database." *Bioinformatics* , November. https://doi.org/10.1093/bioinformatics/btz848.

Chen, Lin-Xing, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield. 2020. "Accurate and Complete Genomes from Metagenomes." *Genome Research* 30 (3): 315–33.

Chen, Yen-Chun, Tsunglin Liu, Chun-Hui Yu, Tzen-Yuh Chiang, and Chi-Chuan Hwang. 2013. "Effects of GC Bias in next-Generation-Sequencing Data on de Novo Genome Assembly." *PloS One* 8 (4): e62856.

Chitsaz, Hamidreza, Joyclyn L. Yee-Greenbaum, Glenn Tesler, Mary-Jane Lombardo, Christopher L. Dupont, Jonathan H. Badger, Mark Novotny, et al. 2011. "Efficient de Novo Assembly of Single-Cell Bacterial Genomes from Short-Read Data Sets." *Nature Biotechnology* 29 (10): 915–21.

Chung, Matthew, James B. Munro, Hervé Tettelin, and Julie C. Dunning Hotopp. 2018. "Using Core Genome Alignments To Assign Bacterial Species." *mSystems* 3 (6). https://doi.org/10.1128/mSystems.00236-18.

Ciais, Philippe, Christopher Sabine, Govindasamy Bala, Laurent Bopp, Victor Brovkin, Josep Canadell, Abha Chhabra, et al. 2014. "Carbon and Other Biogeochemical Cycles." In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 465–570. Cambridge University Press.

Clément, J. M., E. Lepouce, C. Marchal, and M. Hofnung. 1983. "Genetic Study of a Membrane Protein: DNA Sequence Alterations due to 17 lamB Point Mutations Affecting Adsorption of Phage Lambda." *The EMBO Journal* 2 (1): 77–80.

Clingenpeel, S., A. Clum, and P. Schwientek. 2015. "Reconstructing Each Cell's Genome within Complex Microbial Communities—dream or Reality?" *Frontiers in Microbiology* 5: 771.

Collins, Rupert A., Owen S. Wangensteen, Eoin J. O'Gorman, Stefano Mariani, David W. Sims, and Martin J. Genner. 2018. "Persistence of Environmental DNA in Marine Systems." *Communications Biology* 1 (November): 185.

Costello, Mark J., and Chhaya Chaudhary. 2017. "Marine Biodiversity, Biogeography, Deep-Sea Gradients, and Conservation." *Current Biology: CB* 27 (13): 2051.

Creevey, Christopher J., Tobias Doerks, David A. Fitzpatrick, Jeroen Raes, and Peer Bork. 2011. "Universally Distributed Single-Copy Genes Indicate a Constant Rate of Horizontal Transfer." *PloS One* 6 (8): e22099.

Crummett, Lisa T., Richard J. Puxty, Claudia Weihe, Marcia F. Marston, and Jennifer B. H. Martiny. 2016. "The Genomic Content and Context of Auxiliary Metabolic Genes in Marine Cyanomyoviruses." *Virology* 499 (December): 219–29.

David Wilkins, Zachary Kurtz. 2019. *Gggenes: Draw Gene Arrow Maps in "ggplot2"* (version 0.4.0). R. https://cran.r-project.org/web/packages/gggenes/.

Dean, Frank B., John R. Nelson, Theresa L. Giesler, and Roger S. Lasken. 2001. "Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification." *Genome Research* 11 (6): 1095–99.

Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg. 1999. "Improved Microbial Gene Identification with GLIMMER." *Nucleic Acids Research* 27 (23): 4636–41.

Delmont, Tom O., Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny Tm Lee, Michael S. Rappé, Sandra L. McLellan, Sebastian Lücker, and A. Murat Eren. 2018. “Nitrogen-Fixing Populations of Planctomycetes and Proteobacteria Are Abundant in Surface Ocean Metagenomes.” *Nature Microbiology* 3 (7): 804–13.

De Medici, Dario, Luciana Croci, Elisabetta Delibato, Simona Di Pasquale, Emma Filetici, and Laura Toti. 2003. “Evaluation of DNA Extraction Methods for Use in Combination with SYBR Green I Real-Time PCR to Detect Salmonella Enterica Serotype Enteritidis in Poultry.” *Applied and Environmental Microbiology* 69 (6): 3456–61.

Devulder, G., M. Pérouse de Montclos, and J. P. Flandrois. 2005. “A Multigene Approach to Phylogenetic Analysis Using the Genus Mycobacterium as a Model.” *International Journal of Systematic and Evolutionary Microbiology* 55 (Pt 1): 293–302.

Dey, Siddharth S., Lennart Kester, Bastiaan Spanjaard, Magda Bienko, and Alexander van Oudenaarden. 2015. “Integrated Genome and Transcriptome Sequencing of the Same Cell.” *Nature Biotechnology* 33 (3): 285–89.

Dick, Gregory J., Anders F. Andersson, Brett J. Baker, Sheri L. Simmons, Brian C. Thomas, A. Pepper Yelton, and Jillian F. Banfield. 2009. “Community-Wide Analysis of Microbial Genome Sequence Signatures.” *Genome Biology* 10 (8): R85.

Dion, Moïra B., Frank Oechslin, and Sylvain Moineau. 2020. “Phage Diversity, Genomics and Phylogeny.” *Nature Reviews. Microbiology* 18 (3): 125–38.

Direito, Susana O. L., Egija Zaura, Miranda Little, Pascale Ehrenfreund, and Wilfred F. M. Röling. 2014. “Systematic Evaluation of Bias in Microbial Community Profiles Induced by Whole Genome Amplification.” *Environmental Microbiology* 16 (3): 643–57.

Djurhuus, Anni, Jesse Port, Collin J. Closek, Kevan M. Yamahara, Ofelia Romero-Maraccini, Kristine R. Walz, Dawn B. Goldsmith, et al. 2017. “Evaluation of Filtration and DNA Extraction Methods for Environmental DNA Biodiversity Assessments across Multiple Trophic Levels.” *Frontiers in Marine Science* 4 (October): 403.

Doney, Scott C., David M. Glover, and Raymond G. Najjar. 1996. “A New Coupled, One-Dimensional Biological-Physical Model for the Upper Ocean: Applications to the JGOFS Bermuda Atlantic Time-Series Study (BATS) Site.” *Deep-Sea Research. Part II, Topical Studies in Oceanography* 43 (2): 591–624.

Ducklow, Hugh W., Deborah K. Steinberg, and Ken O. Buesseler. 2001. “Upper Ocean Carbon Export and the Biological Pump.” *OCEANOGRAPHY-WASHINGTON DC-OCEANOGRAPHY SOCIETY-* 14 (4): 50–58.

Ducklow, H. W., C. A. Carlson, N. R. Bates, A. H. Knap, A. F. Michaels, T. Jickells, P. J. Le B. Williams, et al. 1995. “Dissolved Organic Carbon as a Component of the Biological Pump in the North Atlantic Ocean.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 348 (1324): 161–67.

Dupont, Chris L., Douglas B. Rusch, Shibu Yooseph, Mary-Jane Lombardo, R. Alexander Richter, Ruben Valas, Mark Novotny, et al. 2012. “Genomic Insights to SAR86, an Abundant and Uncultivated Marine Bacterial Lineage.” *The ISME Journal* 6 (6): 1186–99.

Du Toit, Andrea. 2019. “Phage Induction in Different Contexts.” *Nature Reviews. Microbiology*.

Eberwine, James, Jai-Yoon Sul, Tamas Bartfai, and Junhyong Kim. 2014. “The Promise of Single-Cell Sequencing.” *Nature Methods* 11 (1): 25–27.

Ecale Zhou, Carol L., Stephanie Malfatti, Jeffrey Kimbrel, Casandra Philipson, Katelyn McNair, Theron Hamilton, Robert Edwards, and Brian Souza. 2019. “multiPhATE: Bioinformatics Pipeline for Functional Annotation of Phage Isolates.” *Bioinformatics* 35 (21): 4402–4.

Eddy, Sean. 1998. “HMMER: Profile HMMs for Protein Sequence Analysis.”

Edwards, A. W. F. 1996. “The Origin and Early Development of the Method of Minimum Evolution for the Reconstruction of Phylogenetic Trees.” *Systematic Biology* 45 (1): 79–91.

———. n.d. “Likelihood. 1972.” Cambridge University Press, Cambridge.

Efron, B. 1979. “Bootstrap Methods: Another Look at the Jackknife.” *The Annals of Statistics*. https://doi.org/10.1214/aos/1176344552.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. “Real-Time DNA Sequencing from Single Polymerase Molecules.” *Science* 323 (5910): 133–38.

Eiler, Alexander, Rhiannon Mondav, Lucas Sinclair, Leyden Fernandez-Vidal, Douglas G. Scofield, Patrick Schwientek, Manuel Martinez-Garcia, et al. 2016. “Tuning Fresh: Radiation through Rewiring of Central Metabolism in Streamlined Bacteria.” *The ISME Journal* 10 (8): 1902–14.

Elhai, Jeff, Hailan Liu, and Arnaud Taton. 2012. “Detection of Horizontal Transfer of Individual Genes by Anomalous Oligomer Frequencies.” *BMC Genomics* 13 (June): 245.

Eppley, Richard W., and Bruce J. Peterson. 1979. “Particulate Organic Matter Flux and Planktonic New Production in the Deep Ocean.” *Nature* 282 (5740): 677–80.

Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. “Anvi’o: An Advanced Analysis and Visualization Platform for 'Omics Data.” *PeerJ* 3 (October): e1319.

Felsenstein, Joseph. 1985. “CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP.” *Evolution; International Journal of Organic Evolution* 39 (4): 783–91.

Ferla, Matteo P., J. Cameron Thrash, Stephen J. Giovannoni, and Wayne M. Patrick. 2013. “New rRNA Gene-Based Phylogenies of the Alphaproteobacteria Provide Perspective on Major Groups, Mitochondrial Ancestry and Phylogenetic Instability.” *PloS One* 8 (12): e83383.

Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski. 1998. “Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components.” *Science* 281 (5374): 237–40.

Finn, Robert D., Jody Clements, and Sean R. Eddy. 2011. “HMMER Web Server: Interactive Sequence Similarity Searching.” *Nucleic Acids Research* 39 (Web Server issue): W29–37.

Flombaum, Pedro, José L. Gallegos, Rodolfo A. Gordillo, José Rincón, Lina L. Zabala, Nianzhi Jiao, David M. Karl, et al. 2013. “Present and Future Global Distributions of the Marine Cyanobacteria Prochlorococcus and Synechococcus.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (24): 9824–29.

Fog, Agner. 2008. “Calculation Methods for Wallenius’ Noncentral

Hypergeometric Distribution." *Communications in Statistics - Simulation and Computation* 37 (2): 258–73.

Forouzan, Esmaeil, Parvin Shariati, Masoumeh Sadat Mousavi Maleki, Ali Asghar Karkhane, and Bagher Yakhchali. 2018. "Practical Evaluation of 11 de Novo Assemblers in Metagenome Assembly." *Journal of Microbiological Methods* 151 (August): 99–105.

Fortier, Louis-Charles, and Ognjen Sekulovic. 2013. "Importance of Prophages to Evolution and Virulence of Bacterial Pathogens." *Virulence* 4 (5): 354–65.

Fouts, Derrick E. 2006. "Phage_Finder: Automated Identification and Classification of Prophage Regions in Complete Bacterial Genome Sequences." *Nucleic Acids Research* 34 (20): 5839–51.

Fox, G. E., L. J. Magrum, W. E. Balch, R. S. Wolfe, and C. R. Woese. 1977. "Classification of Methanogenic Bacteria by 16S Ribosomal RNA Characterization." *Proceedings of the National Academy of Sciences of the United States of America* 74 (10): 4537–41.

Fuhrman, J. A. 1999. "Marine Viruses and Their Biogeochemical and Ecological Effects." *Nature* 399 (6736): 541–48.

Fuhrman, J. A., and M. Schwalbach. 2003. "Viral Influence on Aquatic Bacterial Communities." *The Biological Bulletin* 204 (2): 192–95.

Fuller, S. A., M. Takahashi, and J. G. Hurrell. 2001. "Cloning of Hybridoma Cell Lines by Limiting Dilution." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* Chapter 11 (May): Unit11.8.

Fu, Yusi, Chunmei Li, Sijia Lu, Wenxiong Zhou, Fuchou Tang, X. Sunney Xie, and Yanyi Huang. 2015. "Uniform and Accurate Single-Cell Sequencing Based on Emulsion Whole-Genome Amplification." *Proceedings of the National Academy of Sciences of the United States of America* 112 (38): 11923–28.

Gagic, Dragana, Paul H. Maclean, Dong Li, Graeme T. Attwood, and Christina D. Moon. 2015. "Improving the Genetic Representation of Rare Taxa within Complex Microbial Communities Using DNA Normalization Methods." *Molecular Ecology Resources* 15 (3): 464–76.

Galiez, Clovis, Matthias Siebert, François Enault, Jonathan Vincent, and Johannes Söding. 2017. "WIsH: Who Is the Host? Predicting Prokaryotic Hosts from Metagenomic Phage Contigs." *Bioinformatics* 33 (19): 3113–14.

Gawad, Charles, Winston Koh, and Stephen R. Quake. 2016. "Single-Cell Genome Sequencing: Current State of the Science." *Nature Reviews. Genetics* 17 (3): 175–88.

Ghosh, Arpita, Aditya Mehta, and Asif M. Khan. 2019. "Metagenomic Analysis and Its Applications." In *Encyclopedia of Bioinformatics and Computational Biology*, edited by Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, 184–93. Oxford: Academic Press.

Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field. 1990. "Genetic Diversity in Sargasso Sea Bacterioplankton." *Nature* 345 (6270): 60–63.

Giovannoni, S. J., E. F. DeLong, T. M. Schmidt, and N. R. Pace. 1990. "Tangential Flow Filtration and Preliminary Phylogenetic Analysis of Marine Picoplankton." *Applied and Environmental Microbiology* 56 (8): 2572–75.

Giovannoni, Stephen J. 2017. "SAR11 Bacteria: The Most Abundant Plankton in the Oceans." *Annual Review of Marine Science* 9 (January): 231–55.

Giovannoni, Stephen J., J. Cameron Thrash, and Ben Temperton. 2014.

“Implications of Streamlining Theory for Microbial Ecology.” *The ISME Journal* 8 (8): 1553–65.

Giovannoni, Stephen J., H. James Tripp, Scott Givan, Mircea Podar, Kevin L. Vergin, Damon Baptista, Lisa Bibbs, et al. 2005. “Genome Streamlining in a Cosmopolitan Oceanic Bacterium.” *Science* 309 (5738): 1242–45.

Giovannoni, Stephen J., and Kevin L. Vergin. 2012. “Seasonality in Ocean Microbial Communities.” *Science* 335 (6069): 671–76.

Giovannoni, Stephen, Ben Temperton, and Yanlin Zhao. 2013. “Giovannoni et Al. Reply.” *Nature*.

Goff, Stephen A., Matthew Vaughn, Sheldon McKay, Eric Lyons, Ann E. Stapleton, Damian Gessler, Naim Matasci, et al. 2011. “The iPlant Collaborative: Cyberinfrastructure for Plant Biology.” *Frontiers in Plant Science*. https://doi.org/10.3389/fpls.2011.00034.

Goldsmith, D. B., J. R. Brum, M. Hopkins, C. A. Carlson, and M. Breitbart. 2015. “Water Column Stratification Structures Viral Community Composition in the Sargasso Sea.” *Aquatic Microbial Ecology: International Journal* 76 (2): 85–94.

Gontcharov, A. A. 2003. “Are Combined Analyses Better Than Single Gene Phylogenies? A Case Study Using SSU rDNA and rbcL Sequence Comparisons in the Zygnematophyceae (Streptophyta).” *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msh052.

Gorbalenya, A. E. 2008. “Phylogeny of Viruses.” *Encyclopedia of Virology*. https://doi.org/10.1016/b978-012374410-4.00712-3.

Graham, Elaina D., John F. Heidelberg, and Benjamin J. Tully. 2017. “BinSanity: Unsupervised Clustering of Environmental Microbial Assemblies Using Coverage and Affinity Propagation.” *PeerJ*. https://doi.org/10.7717/peerj.3035.

Gregory, E. 1979. “Microbiological Studies of Lake Washington.” Ph.D, University of Washington.

Grigoriev, A. 1999. “Strand-Specific Compositional Asymmetries in Double-Stranded DNA Viruses.” *Virus Research* 60 (1): 1–19.

Grote, Jana, J. Cameron Thrash, Megan J. Huggett, Zachary C. Landry, Paul Carini, Stephen J. Giovannoni, and Michael S. Rappé. 2012. “Streamlining and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade.” *mBio* 3 (5). https://doi.org/10.1128/mBio.00252-12.

Gruber, David F., Jean-Paul Simjouw, Sybil P. Seitzinger, and Gary L. Taghon. 2006. “Dynamics and Characterization of Refractory Dissolved Organic Matter Produced by a Pure Bacterial Culture in an Experimental Predator-Prey System.” *Applied and Environmental Microbiology* 72 (6): 4184–91.

Guixa-Boixereu, N., K. Lysnes, and C. Pedros-Alio. 1999. “Viral Lysis and Bacterivory during a Phytoplankton Bloom in a Coastal Water Microcosm.” *Applied and Environmental Microbiology* 65 (5): 1949–58.

Gupta, Vadakattu V. S. R., Albert D. Rovira, and David K. Roget. 2011. “Principles and Management of Soil Biological Factors for Sustainable Rainfed Farming Systems.” In *Rainfed Farming Systems*, edited by Philip Tow, Ian Cooper, Ian Partridge, and Colin Birch, 149–84. Dordrecht: Springer Netherlands.

Handa, Naofumi, Katsumi Morimatsu, Susan T. Lovett, and Stephen C. Kowalczykowski. 2009. “Reconstitution of Initial Steps of dsDNA Break Repair by the RecF Pathway of E. Coli.” *Genes & Development* 23 (10):

1234–45.

Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. 1998. “Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products.” *Chemistry & Biology* 5 (10): R245–49.

Hara, S., K. Terauchi, and I. Koike. 1991. “Abundance of Viruses in Marine Waters: Assessment by Epifluorescence and Transmission Electron Microscopy.” *Applied and Environmental Microbiology* 57 (9): 2731–34.

Harmon, Frank G., Joel P. Brockman, and Stephen C. Kowalczykowski. 2003. “RecQ Helicase Stimulates Both DNA Catenation and Changes in DNA Topology by Topoisomerase III.” *The Journal of Biological Chemistry* 278 (43): 42668–78.

Haro-Moreno, Jose M., Francisco Rodriguez-Valera, Riccardo Rosselli, Francisco Martinez-Hernandez, Juan J. Roda-Garcia, Monica Lluesma Gomez, Oscar Fornas, Manuel Martinez-Garcia, and Mario López-Pérez. 2020. “Ecogenomics of the SAR11 Clade.” *Environmental Microbiology* 22 (5): 1748–63.

Harrison, Jori B., Jennifer M. Sunday, and Sean M. Rogers. 2019. “Predicting the Fate of eDNA in the Environment and Implications for Studying Biodiversity.” *Proceedings. Biological Sciences / The Royal Society* 286 (1915): 20191409.

Heinemann, Matthias, and Renato Zenobi. 2011. “Single Cell Metabolomics.” *Current Opinion in Biotechnology* 22 (1): 26–31.

Hennes, Kilian P., and Curtis A. Suttle. 1995. “Direct Counts of Viruses in Natural Waters and Laboratory Cultures by Epifluorescence Microscopy.” *Limnology and Oceanography* 40 (6): 1050–55.

Henson, Michael W., V. Celeste Lanclos, Brant C. Faircloth, and J. Cameron Thrash. 2018. “Cultivation and Genomics of the First Freshwater SAR11 (LD12) Isolate.” *The ISME Journal* 12 (7): 1846–60.

Herlemann, Daniel P. R., Jana Woelk, Matthias Labrenz, and Klaus Jürgens. 2014. “Diversity and Abundance of ‘Pelagibacterales’ (SAR11) in the Baltic Sea Salinity Gradient.” *Systematic and Applied Microbiology* 37 (8): 601–4.

Hiom, Kevin. 2009. “DNA Repair: Common Approaches to Fixing Double-Strand Breaks.” *Current Biology: CB* 19 (13): R523–25.

Hooper, Sean D., Jeroen Raes, Konrad U. Foerstner, Eoghan D. Harrington, Daniel Dalevi, and Peer Bork. 2008. “A Molecular Study of Microbe Transfer between Distant Environments.” *PloS One* 3 (7): e2607.

Hopkinson, Charles S., Jr, and Joseph J. Vallino. 2005. “Efficient Export of Carbon to the Deep Ocean through Dissolved Organic Matter.” *Nature* 433 (7022): 142–45.

Hou, Yong, Kui Wu, Xulian Shi, Fuqiang Li, Luting Song, Hanjie Wu, Michael Dean, et al. 2015. “Comparison of Variations Detection between Whole-Genome Amplification Methods Used in Single-Cell Resequencing.” *GigaScience* 4 (August): 37.

Howard-Varona, Cristina, Katherine R. Hargreaves, Stephen T. Abedon, and Matthew B. Sullivan. 2017. “Lysogeny in Nature: Mechanisms, Impact and Ecology of Temperate Phages.” *The ISME Journal* 11 (7): 1511–20.

Howard-Varona, Cristina, Katherine R. Hargreaves, Natalie E. Solonenko, Lye Meng Markillie, Richard Allen White 3rd, Heather M. Brewer, Charles Ansong, Galya Orr, Joshua N. Adkins, and Matthew B. Sullivan. 2018. “Multiple Mechanisms Drive Phage Infection Efficiency in Nearly Identical

Hosts." *The ISME Journal* 12 (6): 1605–18.

Howard-Varona, Cristina, Morgan M. Lindback, G. Eric Bastien, Natalie Solonenko, Ahmed A. Zayed, Hobin Jang, Bill Andreopoulos, et al. 2020. "Phage-Specific Metabolic Reprogramming of Virocells." *The ISME Journal* 14 (4): 881–95.

Huang, Lei, Fei Ma, Alec Chapman, Sijia Lu, and Xiaoliang Sunney Xie. 2015. "Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications." *Annual Review of Genomics and Human Genetics* 16 (June): 79–102.

Huerta-Cepas, Jaime, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, et al. 2016. "eggNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences." *Nucleic Acids Research* 44 (D1): D286–93.

Hugerth, Luisa W., John Larsson, Johannes Alneberg, Markus V. Lindh, Catherine Legrand, Jarone Pinhassi, and Anders F. Andersson. 2015. "Metagenome-Assembled Genomes Uncover a Global Brackish Microbiome." *Genome Biology* 16 (December): 279.

Human Microbiome Jumpstart Reference Strains Consortium, Karen E. Nelson, George M. Weinstock, Sarah K. Highlander, Kim C. Worley, Heather Huot Creasy, Jennifer Russo Wortman, et al. 2010. "A Catalog of Reference Genomes from the Human Microbiome." *Science* 328 (5981): 994–99.

Hunter, John D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95.

Hurwitz, Bonnie L., Jennifer R. Brum, and Matthew B. Sullivan. 2015. "Depth-Stratified Functional and Taxonomic Niche Specialization in the 'Core' and 'Flexible' Pacific Ocean Virome." *The ISME Journal* 9 (2): 472–84.

Hurwitz, Bonnie L., and Jana M. U'Ren. 2016. "Viral Metabolic Reprogramming in Marine Ecosystems." *Current Opinion in Microbiology* 31 (June): 161–68.

Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (March): 119.

Hyman, Paul. 2019. "Phages for Phage Therapy: Isolation, Characterization, and Host Range Breadth." *Pharmaceuticals* 12 (1). https://doi.org/10.3390/ph12010035.

Imelfort, Mike, and Tim Lamberton. 2015. *BamM* (version 1.4.1). Linux. http://ecogenomics.github.io/BamM/.

Inoue, Jin, Masayoshi Honda, Shukuko Ikawa, Takehiko Shibata, and Tsutomu Mikawa. 2008. "The Process of Displacing the Single-Stranded DNA-Binding Protein from Single-Stranded DNA by RecO and RecR Proteins." *Nucleic Acids Research* 36 (1): 94–109.

Iqbal, Zamin, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. 2012. "De Novo Assembly and Genotyping of Variants Using Colored de Bruijn Graphs." *Nature Genetics*. https://doi.org/10.1038/ng.1028.

Iturriaga, R., and H-G Hoppe. 1977. "Observations of Heterotrophic Activity on Photoassimilated Organic Matter." *Marine Biology* 40 (2): 101–8.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018a. "High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries." *Nature*

Communications. https://doi.org/10.1038/s41467-018-07641-9.

———. 2018b. “High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries.” *Nature Communications* 9 (1): 5114.

Jain, Miten, Ian T. Fiddes, Karen H. Miga, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2015. “Improved Data Analysis for the MinION Nanopore Sequencer.” *Nature Methods* 12 (4): 351–56.

Jang, H. B., B. Bolduc, O. Zablocki, J. Kuhn, and S. Roux. 2019. “Gene Sharing Networks to Automate Genome-Based Prokaryotic Viral Taxonomy.” *BioRxiv*. https://www.biorxiv.org/content/10.1101/533240v1.abstract.

Jannasch, Holger W., and Galen E. Jones. 1959. “Bacterial Populations in Sea Water as Determined by Different Methods of Enumeration 1.” *Limnology and Oceanography* 4 (2): 128–39.

Jiao, Nianzhi, Gerhard J. Herndl, Dennis A. Hansell, Ronald Benner, Gerhard Kattner, Steven W. Wilhelm, David L. Kirchman, et al. 2010. “Microbial Production of Recalcitrant Dissolved Organic Matter: Long-Term Carbon Storage in the Global Ocean.” *Nature Reviews. Microbiology* 8 (8): 593–99.

Jimenez-Infante, Francy, David Kamanda Ngugi, Manikandan Vinu, Jochen Blom, Intikhab Alam, Vladimir B. Bajic, and Ulrich Stingl. 2017. “Genomic Characterization of Two Novel SAR11 Isolates from the Red Sea, Including the First Strain of the SAR11 Ib Clade.” *FEMS Microbiology Ecology* 93 (7). https://doi.org/10.1093/femsec/fix083.

Johnson, Christopher M., and Alan D. Grossman. 2015. “Integrative and Conjugative Elements (ICEs): What They Do and How They Work,” December. https://doi.org/10.1146/annurev-genet-112414-055018.

John Wiley & Sons, Ltd, ed. 2001. “Density Dependence and Independence.” In *Encyclopedia of Life Sciences*, 87:1145. Chichester, UK: John Wiley & Sons, Ltd.

Jonas, D. A., I. Elmadfa, K. H. Engel, K. J. Heller, G. Kozianowski, A. König, D. Müller, J. F. Narbonne, W. Wackernagel, and J. Kleiner. 2001. “Safety Considerations of DNA in Food.” *Annals of Nutrition & Metabolism* 45 (6): 235–54.

Juhas, Mario, Jan Roelof van der Meer, Muriel Gaillard, Rosalind M. Harding, Derek W. Hood, and Derrick W. Crook. 2009. “Genomic Islands: Tools of Bacterial Horizontal Gene Transfer and Evolution.” *FEMS Microbiology Reviews* 33 (2): 376–93.

Kamali, Maryam, Paul E. Marek, Ashley Peery, Christophe Antonio-Nkondjio, Cyrille Ndo, Zhijian Tu, Frederic Simard, and Igor V. Sharakhov. 2014. “Multigene Phylogenetics Reveals Temporal Diversification of Major African Malaria Vectors.” *PloS One* 9 (4): e93580.

Kang, Dongwan, Feng Li, Edward S. Kirton, Ashleigh Thomas, Rob S. Egan, Hong An, and Zhong Wang. 2019. “MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies.” e27522v1. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.27522v1.

Karlin, S., J. Mrázek, and A. M. Campbell. 1997. “Compositional Biases of Bacterial Genomes and Evolutionary Implications.” *Journal of Bacteriology* 179 (12): 3899–3913.

Karst, Søren M., Ryan M. Ziels, Rasmus H. Kirkegaard, Emil A. Sørensen, Daniel McDonald, Qiyun Zhu, Rob Knight, and Mads Albertsen. 2020. “Enabling High-Accuracy Long-Read Amplicon Sequences Using Unique Molecular Identifiers with Nanopore or PacBio Sequencing.” *bioRxiv*.

https://doi.org/10.1101/645903.

Katoh, Kazutaka, Kazuharu Misawa, Kei-Ichi Kuma, and Takashi Miyata. 2002. “MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *Nucleic Acids Research* 30 (14): 3059–66.

Kennedy, Nicholas A., Alan W. Walker, Susan H. Berry, Sylvia H. Duncan, Freda M. Farquarson, Petra Louis, John M. Thomson, et al. 2014. “The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing.” *PloS One* 9 (2): e88982.

Kim, Daehwan, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. 2016. “Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences.” *Genome Research* 26 (12): 1721–29.

Klappenbach, J. A., J. M. Dunbar, and T. M. Schmidt. 2000. “rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria.” *Applied and Environmental Microbiology* 66 (4): 1328–33.

Knowles, B., C. B. Silveira, B. A. Bailey, K. Barott, V. A. Cantu, A. G. Cobián-Güemes, F. H. Coutinho, et al. 2016. “Corrigendum: Lytic to Temperate Switching of Viral Communities.” *Nature* 539 (7627): 123.

Kogawa, Masato, Masahito Hosokawa, Yohei Nishikawa, Kazuki Mori, and Haruko Takeyama. 2018. “Obtaining High-Quality Draft Genomes from Uncultured Microbes by Cleaning and Co-Assembly of Single-Cell Amplified Genomes.” *Scientific Reports* 8 (1): 2059.

Konstantinidis, Konstantinos T., and James M. Tiedje. 2005. “Genomic Insights That Advance the Species Definition for Prokaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (7): 2567–72.

Koonin, Eugene V. 2009. “On the Origin of Cells and Viruses: Primordial Virus World Scenario.” *Annals of the New York Academy of Sciences* 1178 (October): 47–64.

Korem, Tal, David Zeevi, Jotham Suez, Adina Weinberger, Tali Avnit-Sagi, Maya Pompan-Lotan, Elad Matot, et al. 2015. “Growth Dynamics of Gut Microbiota in Health and Disease Inferred from Single Metagenomic Samples.” *Science* 349 (6252): 1101–6.

Kowalczykowski, Stephen C., Jennifer Clow, Rahul Somani, and Abraham Varghese. 1987. “Effects of the Escherichia Coli SSB Protein on the Binding of Escherichia Coli RecA Protein to Single-Stranded DNA: Demonstration of Competitive Binding and the Lack of a Specific Protein-Protein Interaction.” *Journal of Molecular Biology* 193 (1): 81–95.

Kraemer, Susanne, Arthi Ramachandran, David Colatriano, Connie Lovejoy, and David A. Walsh. 2019. “Diversity and Biogeography of SAR11 Bacteria from the Arctic Ocean.” *The ISME Journal*, September. https://doi.org/10.1038/s41396-019-0499-4.

Kundu, Ritu, Joshua Casey, and Wing-Kin Sung. 2019. “HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies.” *bioRxiv*. https://doi.org/10.1101/2019.12.19.882506.

Kurihara, T., E. Kamberov, J. M’Mwirichia, T. Tesmer, D. Oldfield, and J. Langmore. 2011. “Rubicon PicoPlex-NGS Kits Available for Sequencing Single Cells Using the Illumina Genome Analyzer.” *Journal of Biomolecular Techniques: JBT* 22 (Suppl): S51.

Labonté, Jessica M., Brandon K. Swan, Bonnie Poulos, Haiwei Luo, Sergey Koren, Steven J. Hallam, Matthew B. Sullivan, Tanja Woyke, K. Eric

Wommack, and Ramunas Stepanauskas. 2015. “Single-Cell Genomics-Based Analysis of Virus-Host Interactions in Marine Surface Bacterioplankton.” *The ISME Journal* 9 (11): 2386–99.

Landry, Zachary C., Stephen J. Giovanonni, Stephen R. Quake, and Paul C. Blainey. 2013. “Optofluidic Cell Selection from Complex Microbial Communities for Single-Genome Analysis.” *Methods in Enzymology* 531: 61–90.

Lane, D. J., B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace. 1985. “Rapid Determination of 16S Ribosomal RNA Sequences for Phylogenetic Analyses.” *Proceedings of the National Academy of Sciences of the United States of America* 82 (20): 6955–59.

Langille, Morgan G. I., William W. L. Hsiao, and Fiona S. L. Brinkman. 2010. “Detecting Genomic Islands Using Bioinformatics Approaches.” *Nature Reviews. Microbiology* 8 (5): 373–82.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Langmead, B., and S. L. Salzberg. 2013. “Langmead. 2013. Bowtie2.” *Nature Methods* 9: 357–59.

Langmore, John P. 2002. “Rubicon Genomics, Inc.” *Pharmacogenomics* 3 (4): 557–60.

Lang, M., M. Pleška, and C. C. Guet. 2020. “Population Dynamics of Decision Making in Temperate Bacteriophages.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.03.18.996918v1.abstract.

Lasken, Roger S. 2007. “Single-Cell Genomic Sequencing Using Multiple Displacement Amplification.” *Current Opinion in Microbiology* 10 (5): 510–16.

———. 2009. “Genomic DNA Amplification by the Multiple Displacement Amplification (MDA) Method.” *Biochemical Society Transactions* 37 (Pt 2): 450–53.

Lasken, Roger S., and Timothy B. Stockwell. 2007. “Mechanism of Chimera Formation during the Multiple Displacement Amplification Reaction.” *BMC Biotechnology* 7 (April): 19.

Laursen, Martin F., Marlene D. Dalgaard, and Martin I. Bahl. 2017. “Genomic GC-Content Affects the Accuracy of 16S rRNA Gene Sequencing Based Microbial Profiling due to PCR Bias.” *Frontiers in Microbiology* 8 (October): 1934.

Letunic, Ivica, and Peer Bork. 2007. “Interactive Tree Of Life (iTOL): An Online Tool for Phylogenetic Tree Display and Annotation.” *Bioinformatics* 23 (1): 127–28.

Leung, Kaston, Hans Zahn, Timothy Leaver, Kishori M. Konwar, Niels W. Hanson, Antoine P. Pagé, Chien-Chi Lo, Patrick S. Chain, Steven J. Hallam, and Carl L. Hansen. 2012. “A Programmable Droplet-Based Microfluidic Device Applied to Multiparameter Analysis of Single Microbes and Microbial Communities.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (20): 7665–70.

Lichter, P., S. A. Ledbetter, D. H. Ledbetter, and D. C. Ward. 1990. “Fluorescence in Situ Hybridization with Alu and L1 Polymerase Chain Reaction Probes for Rapid Characterization of Human Chromosomes in Hybrid Cell Lines.” *Proceedings of the National Academy of Sciences of the United States of America* 87 (17): 6634–38.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam.

2015. “MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph.” *Bioinformatics* 31 (10): 1674–76.

Li, Heng. 2012. “Seqtk Toolkit for Processing Sequences in FASTA/Q Formats.” GitHub.

———. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Lima-Mendez, Gipsi, Jacques Van Helden, Ariane Toussaint, and Raphaël Leplae. 2008. “Prophinder: A Computational Tool for Prophage Prediction in Prokaryotic Genomes.” *Bioinformatics* 24 (6): 863–65.

Lima-Mendez, G., J. Van Helden, A. Toussaint, and R. Leplae. 2008. “Reticulate Representation of Evolutionary and Functional Relationships between Phage Genomes.” *Molecular Biology and Evolution* 25 (4): 762–77.

Lindell, Debbie, Jacob D. Jaffe, Zackary I. Johnson, George M. Church, and Sallie W. Chisholm. 2005. “Photosynthesis Genes in Marine Viruses Yield Proteins during Host Infection.” *Nature* 438 (7064): 86–89.

Lindell, Debbie, Matthew B. Sullivan, Zackary I. Johnson, Andrew C. Tolonen, Forest Rohwer, and Sallie W. Chisholm. 2004. “Transfer of Photosynthesis Genes to and from Prochlorococcus Viruses.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (30): 11013–18.

Lively, C. M., and V. Apanius. 1995. “Genetic Diversity in Host-Parasite Interactions.” *Ecology of Infectious Diseases in Natural Populations* 7: 421.

Loman, Nicholas J., Chrystala Constantinidou, Jacqueline Z. M. Chan, Mihail Halachev, Martin Sergeant, Charles W. Penn, Esther R. Robinson, and Mark J. Pallen. 2012. “High-Throughput Bacterial Genome Sequencing: An Embarrassment of Choice, a World of Opportunity.” *Nature Reviews. Microbiology* 10 (9): 599–606.

Loman, Nicholas J., Joshua Quick, and Jared T. Simpson. 2015. “A Complete Bacterial Genome Assembled de Novo Using Only Nanopore Sequencing Data.” *Nature Methods* 12 (8): 733–35.

Lomsadze, Alexandre, Karl Gemayel, Shiyuyun Tang, and Mark Borodovsky. 2018. “Modeling Leaderless Transcription and Atypical Genes Results in More Accurate Gene Prediction in Prokaryotes.” *Genome Research* 28 (7): 1079–89.

Lovett, S. T., and V. A. Sutera Jr. 1995. “Suppression of recJ Exonuclease Mutants of Escherichia Coli by Alterations in DNA Helicases II (uvrD) and IV (helD).” *Genetics* 140 (1): 27–45.

Luo, Junwei, Mengna Lyu, Ranran Chen, Xiaohong Zhang, Huimin Luo, and Chaokun Yan. 2020. “Correction to: SLR: A Scaffolding Algorithm Based on Long Reads and Contig Classification.” *BMC Bioinformatics* 21 (1): 50.

Lynch, Michael, and John S. Conery. 2003. “The Origins of Genome Complexity.” *Science* 302 (5649): 1401–4.

Lyons-Weiler, James, Guy A. Hoelzer, and Robin J. Tausch. 1998. “Optimal Outgroup Analysis.” *Biological Journal of the Linnean Society. Linnean Society of London* 64 (4): 493–511.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research: JMLR* 9 (Nov): 2579–2605.

Macaulay, Iain C., Wilfried Haerty, Parveen Kumar, Yang I. Li, Tim Xiaoming Hu, Mabel J. Teng, Mubeen Goolam, et al. 2015. “G&T-Seq: Parallel Sequencing of Single-Cell Genomes and Transcriptomes.” *Nature Methods*

12 (6): 519–22.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” *Cell* 161 (5): 1202–14.

Malmstrom, Rex R., Matthew T. Cottrell, Hila Elifantz, and David L. Kirchman. 2005. “Biomass Production and Assimilation of Dissolved Organic Matter by SAR11 Bacteria in the Northwest Atlantic Ocean.” *Applied and Environmental Microbiology* 71 (6): 2979–86.

Marçais, Guillaume, and Carl Kingsford. 2011. “A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers.” *Bioinformatics* 27 (6): 764–70.

Marcy, Yann, Thomas Ishoey, Roger S. Lasken, Timothy B. Stockwell, Brian P. Walenz, Aaron L. Halpern, Karen Y. Beeson, Susanne M. D. Goldberg, and Stephen R. Quake. 2007. “Nanoliter Reactors Improve Multiple Displacement Amplification of Genomes from Single Cells.” *PLoS Genetics* 3 (9): 1702–8.

Marcy, Yann, Cleber Ouverney, Elisabeth M. Bik, Tina Lösekann, Natalia Ivanova, Hector Garcia Martin, Ernest Szeto, et al. 2007. “Dissecting Biological ‘dark Matter’ with Single-Cell Genetic Analysis of Rare and Uncultivated TM7 Microbes from the Human Mouth.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (29): 11889–94.

Marine, Rachel, Coleen McCarren, Vansay Vorrasane, Dan Nasko, Erin Crowgey, Shawn W. Polson, and K. Eric Wommack. 2014. “Caught in the Middle with Multiple Displacement Amplification: The Myth of Pooling for Avoiding Multiple Displacement Amplification Bias in a Metagenome.” *Microbiome* 2 (1): 3.

Marston, Marcia F., Francis J. Pierciey Jr, Alicia Shepard, Gary Gearin, Ji Qi, Chandri Yandava, Stephan C. Schuster, Matthew R. Henn, and Jennifer B. H. Martiny. 2012. “Rapid Diversification of Coevolving Marine Synechococcus and a Virus.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (12): 4544–49.

Martiny, Adam C. 2019. “High Proportions of Bacteria Are Culturable across Major Biomes.” *The ISME Journal* 13 (8): 2125–28.

Mason, Olivia U., Terry C. Hazen, Sharon Borglin, Patrick S. G. Chain, Eric A. Dubinsky, Julian L. Fortney, James Han, et al. 2012. “Metagenome, Metatranscriptome and Single-Cell Sequencing Reveal Microbial Response to Deepwater Horizon Oil Spill.” *The ISME Journal* 6 (9): 1715–27.

Matteson, Audrey R., Star N. Loar, Stuart Pickmere, Jennifer M. DeBruyn, Michael J. Ellwood, Philip W. Boyd, David A. Hutchins, and Steven W. Wilhelm. 2012. “Production of Viruses during a Spring Phytoplankton Bloom in the South Pacific Ocean near of New Zealand.” *FEMS Microbiology Ecology* 79 (3): 709–19.

McDaniel, Lauren D., Elizabeth Young, Jennifer Delaney, Fabian Ruhnau, Kim B. Ritchie, and John H. Paul. 2010. “High Frequency of Horizontal Gene Transfer in the Oceans.” *Science* 330 (6000): 50.

McInnes, Leland, John Healy, and Steve Astels. 2017. “Hdbscan: Hierarchical Density Based Clustering.” *The Journal of Open Source Software* 2 (11): 205.

McInnes, Leland, John Healy, and James Melville. 2018. “UMAP: Uniform

Manifold Approximation and Projection for Dimension Reduction." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426.

McLean, Jeffrey S., Mary-Jane Lombardo, Jonathan H. Badger, Anna Edlund, Mark Novotny, Joyclyn Yee-Greenbaum, Nikolay Vyahhi, et al. 2013. "Candidate Phylum TM6 Genome Recovered from a Hospital Sink Biofilm Provides Genomic Insights into This Uncultivated Phylum." *Proceedings of the National Academy of Sciences of the United States of America* 110 (26): E2390–99.

McNair, Katelyn, Carol Zhou, Elizabeth A. Dinsdale, Brian Souza, and Robert A. Edwards. 2019. "PHANOTATE: A Novel Approach to Gene Identification in Phage Genomes." *Bioinformatics* 35 (22): 4537–42.

Medlar, Alan J., Petri Törönen, and Liisa Holm. 2018. "AAI-Profiler: Fast Proteome-Wide Exploratory Analysis Reveals Taxonomic Identity, Misclassification and Contamination." *Nucleic Acids Research* 46 (W1): W479–85.

Meier, Petra, and Wilfried Wackernagel. 2003. "Mechanisms of Homology-Facilitated Illegitimate Recombination for Foreign DNA Acquisition in Transformable Pseudomonas Stutzeri." *Molecular Microbiology* 48 (4): 1107–18.

Meijenfeldt, F. A. Bastiaan von, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh. 2019. "Robust Taxonomic Classification of Uncharted Microbial Sequences and Bins with CAT and BAT." *Biorxiv*, January. https://doi.org/10.1101/530188.

Mende, Daniel R., Frank O. Aylward, John M. Eppley, Torben N. Nielsen, and Edward F. DeLong. 2016. "Improved Environmental Genomes via Integration of Metagenomic and Single-Cell Assemblies." *Frontiers in Microbiology* 7 (February): 143.

Menzel, Peter, Kim Lee Ng, and Anders Krogh. 2015. "Kaiju: Fast and Sensitive Taxonomic Classification for Metagenomics." *bioRxiv*. https://doi.org/10.1101/031229.

———. 2016. "Fast and Sensitive Taxonomic Classification for Metagenomics with Kaiju." *Nature Communications*. https://doi.org/10.1038/ncomms11257.

Merchant, Nirav, Eric Lyons, Stephen Goff, Matthew Vaughn, Doreen Ware, David Micklos, and Parker Antin. 2016. "The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences." *PLoS Biology* 14 (1): e1002342.

Meyer-Reil, Lutz-Arend. 1978. "Autoradiography and Epifluorescence Microscopy Combined for the Determination of Number and Spectrum of Actively Metabolizing Bacteria in Natural Waters." *Applied and Environmental Microbiology* 36 (3): 506–12.

Michaels, Anthony F., Anthony H. Knap, Rachael L. Dow, Kjell Gundersen, Rodney J. Johnson, Jens Sorensen, Ann Close, et al. 1994. "Seasonal Patterns of Ocean Biogeochemistry at the U.S. JGOFS Bermuda Atlantic Time-Series Study Site." *Deep Sea Research Part I: Oceanographic Research Papers* 41 (7): 1013–38.

Middelboe, Mathias, and Corina P. D. Brussaard. 2017. "Marine Viruses: Key Players in Marine Ecosystems." *Viruses* 9 (10). https://doi.org/10.3390/v9100302.

Mikheenko, Alla, Vladislav Saveliev, and Alexey Gurevich. 2016. "MetaQUAST: Evaluation of Metagenome Assemblies." *Bioinformatics* 32 (7): 1088–90.

Morris, Robert M., Kelsy R. Cain, Kelli L. Hvorecny, and Justin M. Kollman. 2020. “Lysogenic Host–virus Interactions in SAR11 Marine Bacteria.” *Nature Microbiology*, May, 1–5.

Morris, Robert M., Michael S. Rappé, Stephanie A. Connon, Kevin L. Vergin, William A. Siebold, Craig A. Carlson, and Stephen J. Giovannoni. 2002. “SAR11 Clade Dominates Ocean Surface Bacterioplankton Communities.” *Nature* 420 (6917): 806–10.

Mostofa, Khan M. G., Cong-Qiang Liu, M. Abdul Mottaleb, Guojiang Wan, and Fengchang Wu. 2013. “Dissolved Organic Matter in Natural Waters.” In *Photobiogeochemistry of Organic Matter*, 1–137.

Murat Eren, A. 2016. “An Anvi’o Workflow for Microbial Pangenomics.” Meren Lab. November 8, 2016. http://merenlab.org/2016/11/08/pangenomics-v2/.

Mygind, Tina, Lars Østergaard, Svend Birkelund, Jes S. Lindholt, and Gunna Christiansen. 2003. “Evaluation of Five DNA Extraction Methods for Purification of DNA from Atherosclerotic Tissue and Estimation of Prevalence of Chlamydia Pneumoniae in Tissue from a Danish Population Undergoing Vascular Repair.” *BMC Microbiology* 3 (1): 19.

Namiki, Toshiaki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. 2012. “MetaVelvet: An Extension of Velvet Assembler to de Novo Metagenome Assembly from Short Sequence Reads.” *Nucleic Acids Research* 40 (20): e155.

Nanda, Arun M., Kai Thormann, and Julia Frunzke. 2015. “Impact of Spontaneous Prophage Induction on the Fitness of Bacterial Populations and Host-Microbe Interactions.” *Journal of Bacteriology* 197 (3): 410–19.

Navin, Nicholas, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, et al. 2011. “Tumour Evolution Inferred by Single-Cell Sequencing.” *Nature* 472 (7341): 90–94.

Nayfach, Stephen, and Katherine S. Pollard. 2016. “Toward Accurate and Quantitative Comparative Metagenomics.” *Cell* 166 (5): 1103–16.

Nei, Masatoshi, and Sudhir Kumar. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74.

Nielsen, K. M., A. M. Bones, and J. D. Van Elsas. 1997. “Induced Natural Transformation of Acinetobacter Calcoaceticus in Soil Microcosms.” *Applied and Environmental Microbiology* 63 (10): 3972–77.

Nikolenko, Sergey I., Anton I. Korobeynikov, and Max A. Alekseyev. 2013. “BayesHammer: Bayesian Clustering for Error Correction in Single-Cell Sequencing.” *BMC Genomics* 14 Suppl 1 (January): S7.

Noble, R. T., and J. A. Fuhrman. 1998. “Use of SYBR Green I for Rapid Epifluorescence Counts of Marine Viruses and Bacteria.” *Aquatic Microbial Ecology: International Journal* 14: 113–18.

Noguchi, Hideki, Takeaki Taniguchi, and Takehiko Itoh. 2008. “MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes.” *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 15 (6): 387–96.

Nurk, Sergey, Anton Bankevich, Dmitry Antipov, Alexey A. Gurevich, Anton Korobeynikov, Alla Lapidus, Andrey D. Prjibelski, et al. 2013. “Assembling

Single-Cell Genomes and Mini-Metagenomes from Chimeric MDA Products." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 20 (10): 714–37.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. "metaSPAdes: A New Versatile Metagenomic Assembler." *Genome Research* 27 (5): 824–34.

Oh, Hyun-Myung, Ilnam Kang, Kiyoung Lee, Yoonra Jang, Seung-Il Lim, and Jang-Cheon Cho. 2011. "Complete Genome Sequence of Strain IMCC9063, Belonging to SAR11 Subgroup 3, Isolated from the Arctic Ocean." *Journal of Bacteriology* 193 (13): 3379–80.

Oliveira, Caio Fernando de, Thiago Galvão da Silva Paim, Keli Cristine Reiter, Alexandre Rieger, and Pedro Alves D'Azevedo. 2014. "Evaluation of Four Different DNA Extraction Methods in Coagulase-Negative Staphylococci Clinical Isolates." *Revista Do Instituto de Medicina Tropical de Sao Paulo* 56 (1): 29–33.

Olson, Nathan D., Todd J. Treangen, Christopher M. Hill, Victoria Cepeda-Espinoza, Jay Ghurye, Sergey Koren, and Mihai Pop. 2019. "Metagenomic Assembly through the Lens of Validation: Recent Advances in Assessing and Improving the Quality of Genomes Assembled from Metagenomes." *Briefings in Bioinformatics* 20 (4): 1140–50.

Parks, Donovan. 2018. *CompareM* (version 0.0.24). Linux. https://github.com/dparks1134/CompareM.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. "CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes." *Genome Research* 25 (7): 1043–55.

Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. "Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life." *Nature Microbiology* 2 (11): 1533–42.

Parsons, Rachel J., Mya Breitbart, Michael W. Lomas, and Craig A. Carlson. 2012. "Ocean Time-Series Reveals Recurring Seasonal Patterns of Virioplankton Dynamics in the Northwestern Sargasso Sea." *The ISME Journal* 6 (2): 273–84.

Paul, John H. 2008. "Prophages in Marine Bacteria: Dangerous Molecular Time Bombs or the Key to Survival in the Seas?" *The ISME Journal* 2 (6): 579–89.

Pearman, W. S., N. E. Freed, and O. K. Silander. 2019. "The Advantages and Disadvantages of Short-and Long-Read Metagenomics to Infer Bacterial and Eukaryotic Community Composition." *BioRxiv*. https://www.biorxiv.org/content/10.1101/650788v2.abstract.

Peng, Yu, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. 2012. "IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth." *Bioinformatics* 28 (11): 1420–28.

Pesant, Stéphane, Fabrice Not, Marc Picheral, Stefanie Kandels-Lewis, Noan Le Bescot, Gabriel Gorsky, Daniele Iudicone, et al. 2015. "Open Science Resources for the Discovery and Analysis of Tara Oceans Data." *Scientific Data* 2 (May): 150023.

Pevzner, P. A., H. Tang, and M. S. Waterman. 2001. "An Eulerian Path

Approach to DNA Fragment Assembly." *Proceedings of the National Academy of Sciences of the United States of America* 98 (17): 9748–53.

Phillips, A., D. Janies, and W. Wheeler. 2000. "Multiple Sequence Alignment in Phylogenetic Analysis." *Molecular Phylogenetics and Evolution* 16 (3): 317–30.

Pieterse, Corné M. J., Ronnie de Jonge, and Roeland L. Berendsen. 2016. "The Soil-Borne Supremacy." *Trends in Plant Science* 21 (3): 171–73.

Probst, Alexander J., Thomas Weinmaier, Todd Z. DeSantis, Jorge W. Santo Domingo, and Nicholas Ashbolt. 2015. "New Perspectives on Microbial Community Distortion after Whole-Genome Amplification." *PloS One* 10 (5): e0124158.

Proctor, Lita M., and Jed A. Fuhrman. 1990. "Viral Mortality of Marine Bacteria and Cyanobacteria." *Nature* 343 (6253): 60–62.

Prudhomme, Marc, Virginie Libante, and Jean-Pierre Claverys. 2002. "Homologous Recombination at the Border: Insertion-Deletions and the Trapping of Foreign DNA in Streptococcus Pneumoniae." *Proceedings of the National Academy of Sciences of the United States of America* 99 (4): 2100–2105.

Qin, Qi-Long, Bin-Bin Xie, Xi-Ying Zhang, Xiu-Lan Chen, Bai-Cheng Zhou, Jizhong Zhou, Aharon Oren, and Yu-Zhong Zhang. 2014. "A Proposed Genus Boundary for the Prokaryotes Based on Genomic Insights." *Journal of Bacteriology* 196 (12): 2210–15.

Quick, Josh. 2019. "Ultra-Long Read Nanopore Sequencing Methods for Metagenomics." *Journal of Biomolecular Techniques: JBT* 30 (Suppl): S63.

Quince, Christopher, Stephanie Connelly, Sébastien Raguideau, Johannes Alneberg, Seung Gu Shin, Gavin Collins, and A. Murat Eren. 2016. "De Novo Extraction of Microbial Strains from Metagenomes Reveals Intra-Species Niche Partitioning." *bioRxiv*. https://doi.org/, Connelly, S. <http://eprints.gla.ac.uk/view/author/34796.html> <http://orcid.org/0000-0001-5261-2090>, Raguideau, S., Alneberg, J., Shin, S. G. <http://eprints.gla.ac.uk/view/author/29031.html>, Collins, G. <http://eprints.gla.ac.uk/view/author/14103.html> and Eren, A. M. (2016) De novo extraction of microbial strains from metagenomes reveals intra-species niche partitioning. bioRxiv <http://eprints.gla.ac.uk/view/journal_volume/bioRxiv.html>, (doi:10.1101/073825 <http://dx.doi.org/10.1101/073825>) (Submitted) "> Quince, C. <http://eprints.gla.ac.uk/view/author/3392.html>, Connelly, S. <http://eprints.gla.ac.uk/view/author/34796.html> <http://orcid.org/0000-0001-5261-2090>, Raguideau, S., Alneberg, J., Shin, S. G. <http://eprints.gla.ac.uk/view/author/29031.html>, Collins, G. <http://eprints.gla.ac.uk/view/author/14103.html> and Eren, A. M. (2016) De novo extraction of microbial strains from metagenomes reveals intra-species niche partitioning. bioRxiv <http://eprints.gla.ac.uk/view/journal_volume/bioRxiv.html>, (doi:10.1101/073825 <http://dx.doi.org/10.1101/073825>) (Submitted) .

Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. "Shotgun Metagenomics, from Sampling to Analysis." *Nature Biotechnology* 35 (9): 833–44.

Ramisetty, Bhaskar Chandra Mohan, and Pavithra Anantharaman Sudhakari. 2019. "Bacterial 'grounded'prophages: Hotspots for Genetic Renovation and Innovation." *Frontiers in Genetics* 10: 65.

Rappé, Michael S., Stephanie A. Connon, Kevin L. Vergin, and Stephen J. Giovannoni. 2002. “Cultivation of the Ubiquitous SAR11 Marine Bacterioplankton Clade.” *Nature* 418 (6898): 630–33.

Ren, Jie, Nathan A. Ahlgren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. 2017. “VirFinder: A Novel K-Mer Based Tool for Identifying Viral Sequences from Assembled Metagenomic Data.” *Microbiome* 5 (1): 69.

Richter, Michael, and Ramon Rosselló-Móra. 2009. “Shifting the Genomic Gold Standard for the Prokaryotic Species Definition.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (45): 19126–31.

Rinke, Christian, Janey Lee, Nandita Nath, Danielle Goudeau, Brian Thompson, Nicole Poulton, Elizabeth Dmitrieff, Rex Malmstrom, Ramunas Stepanauskas, and Tanja Woyke. 2014. “Obtaining Genomes from Uncultivated Environmental Microorganisms Using FACS–based Single-Cell Genomics.” *Nature Protocols* 9 (5): 1038–48.

Rinke, Christian, Patrick Schwientek, Alexander Sczyrba, Natalia N. Ivanova, Iain J. Anderson, Jan-Fang Cheng, Aaron Darling, et al. 2013. “Insights into the Phylogeny and Coding Potential of Microbial Dark Matter.” *Nature* 499 (7459): 431–37.

Rodriguez-R, Luis M., and Konstantinos T. Konstantinidis. 2014. “Bypassing Cultivation To Identify Bacterial Species.” *Microbe Magazine*. https://doi.org/10.1128/microbe.9.111.1.

Rodriguez-Valera, Francisco, Ana-Belen Martin-Cuadrado, Beltran Rodriguez-Brito, Lejla Pasić, T. Frede Thingstad, Forest Rohwer, and Alex Mira. 2009. “Explaining Microbial Population Genomics through Phage Predation.” *Nature Reviews. Microbiology* 7 (11): 828–36.

Rohwer, Forest, David Prangishvili, and Debbie Lindell. 2009. “Roles of Viruses in the Environment.” *Environmental Microbiology* 11 (11): 2771–74.

Rohwer, Forest, and Rebecca Vega Thurber. 2009. “Viruses Manipulate the Marine Environment.” *Nature* 459 (7244): 207–12.

Rosen, Gail, Elaine Garbarine, Diamantino Caseiro, Robi Polikar, and Bahrad Sokhansanj. 2008. “Metagenome Fragment Classification Using N -Mer Frequency Profiles.” *Advances in Bioinformatics* 2008 (November). https://doi.org/10.1155/2008/205969.

Rothberg, Jonathan M., Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, et al. 2011. “An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing.” *Nature* 475 (7356): 348–52.

Roux, Simon, Francois Enault, Bonnie L. Hurwitz, and Matthew B. Sullivan. 2015. “VirSorter: Mining Viral Signal from Microbial Genomic Data.” *PeerJ* 3 (May): e985.

Roux, Simon, Alyse K. Hawley, Monica Torres Beltran, Melanie Scofield, Patrick Schwientek, Ramunas Stepanauskas, Tanja Woyke, Steven J. Hallam, and Matthew B. Sullivan. 2014. “Ecology and Evolution of Viruses Infecting Uncultivated SUP05 Bacteria as Revealed by Single-Cell- and Meta-Genomics.” *eLife* 3 (August): e03125.

Rubakhin, Stanislav S., Eric J. Lanni, and Jonathan V. Sweedler. 2013. “Progress toward Single Cell Metabolomics.” *Current Opinion in Biotechnology* 24 (1): 95–104.

Rusch, Douglas B., Aaron L. Halpern, Granger Sutton, Karla B. Heidelberg, Shannon Williamson, Shibu Yooseph, Dongying Wu, et al. 2007. “The

Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific.” *PLoS Biology*. https://doi.org/10.1371/journal.pbio.0050077.

Sabine, Christopher L., Richard A. Feely, Nicolas Gruber, Robert M. Key, Kitack Lee, John L. Bullister, Rik Wanninkhof, et al. 2004. “The Oceanic Sink for Anthropogenic CO2.” *Science* 305 (5682): 367–71.

Sakai, Akiko, and Michael M. Cox. 2009. “RecFOR and RecOR as Distinct RecA Loading Pathways.” *The Journal of Biological Chemistry* 284 (5): 3264–72.

Salisbury, Alicia, and Philippos K. Tsourkas. 2019. “A Method for Improving the Accuracy and Efficiency of Bacteriophage Genome Annotation.” *International Journal of Molecular Sciences* 20 (14). https://doi.org/10.3390/ijms20143391.

Salter, Ian. 2018. “Seasonal Variability in the Persistence of Dissolved Environmental DNA (eDNA) in a Marine System: The Role of Microbial Nutrient Limitation.” *PloS One* 13 (2): e0192409.

Salter, Susannah J., Michael J. Cox, Elena M. Turek, Szymon T. Calus, William O. Cookson, Miriam F. Moffatt, Paul Turner, Julian Parkhill, Nicholas J. Loman, and Alan W. Walker. 2014. “Reagent and Laboratory Contamination Can Critically Impact Sequence-Based Microbiome Analyses.” *BMC Biology* 12 (November): 87.

Sangwan, Naseer, Fangfang Xia, and Jack A. Gilbert. 2016. “Recovering Complete and Draft Population Genomes from Metagenome Datasets.” *Microbiome* 4 (March): 8.

Santos-Garcia, Diego, Francisco J. Silva, Shai Morin, Konrad Dettner, and Stefan Martin Kuechler. 2017. “The All-Rounder Sodalis: A New Bacteriome-Associated Endosymbiont of the Lygaeoid Bug Henestaris Halophilus (Heteroptera: Henestarinae) and a Critical Examination of Its Evolution.” *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evx202.

Schlitzer, Reiner. 2002. “Interactive Analysis and Visualization of Geoscience Data with Ocean Data View.” *Computers & Geosciences*. https://doi.org/10.1016/s0098-3004(02)00040-7.

Schmieder, Robert, and Robert Edwards. 2011. “Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets.” *PloS One* 6 (3). https://www.ncbi.nlm.nih.gov/pmc/articles/pmc3052304/.

Scholz, Matthew B., Chien-Chi Lo, and Patrick S. G. Chain. 2012. “Next Generation Sequencing and Bioinformatic Bottlenecks: The Current State of Metagenomic Data Analysis.” *Current Opinion in Biotechnology* 23 (1): 9–15.

Scholz, Matthias, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L. Morrow, and Nicola Segata. 2016. “Strain-Level Microbial Epidemiology and Population Genomics from Shotgun Metagenomics.” *Nature Methods* 13 (5): 435–38.

Schröder, Jan, Heiko Schröder, Simon J. Puglisi, Ranjan Sinha, and Bertil Schmidt. 2009. “SHREC: A Short-Read Error Correction Method.” *Bioinformatics* 25 (17): 2157–63.

Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. “Critical Assessment of Metagenome Interpretation—a Benchmark of Metagenomics Software.”

Nature Methods 14 (11): 1063–71.

Seemann, T. 2015. “Barrnap.” Github.

Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69.

Semple, Charles, Mike Steel, and Both in the Department of Mathematics and Statistics Mike Steel. 2003. *Phylogenetics*. Oxford University Press.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11): 2498–2504.

Shapiro, Howard M. 2005. *Practical Flow Cytometry*. John Wiley & Sons.

Sharon, Itai, Ariella Alperovitch, Forest Rohwer, Matthew Haynes, Fabian Glaser, Nof Atamna-Ismaeel, Ron Y. Pinter, et al. 2009. “Photosystem I Gene Cassettes Are Present in Marine Virus Genomes.” *Nature* 461 (7261): 258–62.

Sharon, Itai, Natalia Battchikova, Eva-Mari Aro, Carmela Giglione, Thierry Meinnel, Fabian Glaser, Ron Y. Pinter, Mya Breitbart, Forest Rohwer, and Oded Béjà. 2011. “Comparative Metagenomics of Microbial Traits within Oceanic Viral Communities.” *The ISME Journal* 5 (7): 1178–90.

Sharon, Itai, Michael J. Morowitz, Brian C. Thomas, Elizabeth K. Costello, David A. Relman, and Jillian F. Banfield. 2013. “Time Series Community Genomics Analysis Reveals Rapid Shifts in Bacterial Species, Strains, and Phage during Infant Gut Colonization.” *Genome Research* 23 (1): 111–20.

Shintaku, Hirofumi, Hidekazu Nishikii, Lewis A. Marshall, Hidetoshi Kotera, and Juan G. Santiago. 2014. “On-Chip Separation and Analysis of RNA and DNA from Single Cells.” *Analytical Chemistry* 86 (4): 1953–57.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. “Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology* 7 (1). https://www.embopress.org/doi/abs/10.1038/msb.2011.75.

Silveira, Cynthia B., and Forest L. Rohwer. 2016. “Piggyback-the-Winner in Host-Associated Microbial Communities.” *Npj Biofilms and Microbiomes*. https://doi.org/10.1038/npjbiofilms.2016.10.

Simpson, Jared T. 2014. “Exploring Genome Characteristics and Sequence Quality without a Reference.” *Bioinformatics* 30 (9): 1228–35.

Simpson, Jared T., and Mihai Pop. 2015. “The Theory and Practice of Genome Sequence Assembly.” *Annual Review of Genomics and Human Genetics* 16 (April): 153–72.

Singleton, Caitlin M., Francesca Petriglieri, Jannie M. Kristensen, Rasmus H. Kirkegaard, Thomas Y. Michaelsen, Martin H. Andersen, Zivile Kondrotaite, et al. 2020. “Connecting Structure to Function with the Recovery of over 1000 High-Quality Activated Sludge Metagenome-Assembled Genomes Encoding Full-Length rRNA Genes Using Long-Read Sequencing.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.05.12.088096v1.full-text.

Sinha, Rahul, Geoff Stanley, Gunsagar S. Gulati, Camille Ezran, Kyle J. Travaglini, Eric Wei, Charles K. F. Chan, et al. 2017. “Index Switching Causes ‘spreading-of-Signal’ among Multiplexed Samples in Illumina HiSeq 4000 DNA Sequencing.” https://doi.org/10.1101/125724.

Somerville, Vincent, Stefanie Lutz, Michael Schmid, Daniel Frei, Aline Moser, Stefan Irmler, Jürg E. Frey, and Christian H. Ahrens. 2019. “Long-Read Based de Novo Assembly of Low-Complexity Metagenome Samples Results in Finished Genomes and Reveals Insights into Strain Diversity and an Active Phage System.” *BMC Microbiology*. https://doi.org/10.1186/s12866-019-1500-0.

Sowell, Sarah M., Larry J. Wilhelm, Angela D. Norbeck, Mary S. Lipton, Carrie D. Nicora, Douglas F. Barofsky, Craig A. Carlson, Richard D. Smith, and Stephen J. Giovanonni. 2009. “Transport Functions Dominate the SAR11 Metaproteome at Low-Nutrient Extremes in the Sargasso Sea.” *The ISME Journal* 3 (1): 93–105.

Spang, Anja, Jimmy H. Saw, Steffen L. Jørgensen, Katarzyna Zaremba-Niedzwiedzka, Joran Martijn, Anders E. Lind, Roel van Eijk, Christa Schleper, Lionel Guy, and Thijs J. G. Ettema. 2015. “Complex Archaea That Bridge the Gap between Prokaryotes and Eukaryotes.” *Nature* 521 (7551): 173–79.

Ståhlberg, Anders, Christer Thomsen, David Ruff, and Pierre Åman. 2012. “Quantitative PCR Analysis of DNA, RNAs, and Proteins in the Same Single Cell.” *Clinical Chemistry* 58 (12): 1682–91.

Staley, J. T., and A. Konopka. 1985a. “Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats.” *Annual Review of Microbiology* 39: 321–46.

———. 1985b. “Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats.” *Annual Review of Microbiology* 39: 321–46.

Stasiak, A., E. DiCapua, and T. Koller. 1983. “Unwinding of Duplex DNA in Complexes with recA Protein.” *Cold Spring Harbor Symposia on Quantitative Biology* 47 Pt 2: 811–20.

Staszewski, R. 1984. “Cloning by Limiting Dilution: An Improved Estimate That an Interesting Culture Is Monoclonal.” *The Yale Journal of Biology and Medicine* 57 (6): 865–68.

Steel, M., and D. Penny. 2000. “Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics.” *Molecular Biology and Evolution* 17 (6): 839–50.

Steen, Andrew D., Alexander Crits-Christoph, Paul Carini, Kristen M. DeAngelis, Noah Fierer, Karen G. Lloyd, and J. Cameron Thrash. 2019. “High Proportions of Bacteria and Archaea across Most Biomes Remain Uncultured.” *The ISME Journal*, August. https://doi.org/10.1038/s41396-019-0484-y.

Stepanauskas, Ramunas. 2012. “Single Cell Genomics: An Individual Look at Microbes.” *Current Opinion in Microbiology*. https://doi.org/10.1016/j.mib.2012.09.001.

Stepanauskas, Ramunas, Elizabeth A. Fergusson, Joseph Brown, Nicole J. Poulton, Ben Tupper, Jessica M. Labonté, Eric D. Becraft, et al. 2017. “Improved Genome Recovery and Integrated Cell-Size Analyses of Individual Uncultured Microbial Cells and Viral Particles.” *Nature Communications* 8 (1): 84.

Stewart, Eric J. 2012. “Growing Unculturable Bacteria.” *Journal of Bacteriology* 194 (16): 4151–60.

Stingl, Ulrich, Harry James Tripp, and Stephen J. Giovannoni. 2007. “Improvements of High-Throughput Culturing Yielded Novel SAR11 Strains and Other Abundant Marine Bacteria from the Oregon Coast and the Bermuda Atlantic Time Series Study Site." *The ISME Journal*. https://doi.org/10.1038/ismej.2007.49.

Strous, Marc, Beate Kraft, Regina Bisdorf, and Halina E. Tegetmeyer. 2012. "The Binning of Metagenomic Contigs for Microbial Physiology of Mixed Cultures." *Frontiers in Microbiology* 3 (December): 410.

Sullivan, Matthew B. 2015. "Viromes, Not Gene Markers, for Studying Double-Stranded DNA Virus Communities." *Journal of Virology* 89 (5): 2459–61.

Sullivan, Matthew B., Maureen L. Coleman, Peter Weigele, Forest Rohwer, and Sallie W. Chisholm. 2005. "Three Prochlorococcus Cyanophage Genomes: Signature Features and Ecological Interpretations." *PLoS Biology* 3 (5): e144.

Sullivan, Matthew B., Katherine H. Huang, Julio C. Ignacio-Espinoza, Aaron M. Berlin, Libusha Kelly, Peter R. Weigele, Alicia S. DeFrancesco, et al. 2010. "Genomic Analysis of Oceanic Cyanobacterial Myoviruses Compared with T4-like Myoviruses from Diverse Hosts and Environments." *Environmental Microbiology*. https://doi.org/10.1111/j.1462-2920.2010.02280.x.

Sullivan, Matthew B., Bryan Krastins, Jennifer L. Hughes, Libusha Kelly, Michael Chase, David Sarracino, and Sallie W. Chisholm. 2009. "The Genome and Structural Proteome of an Ocean Siphovirus: A New Window into the Cyanobacterial 'mobilome.'" *Environmental Microbiology*. https://doi.org/10.1111/j.1462-2920.2009.02081.x.

Sullivan, Matthew B., John B. Waterbury, and Sallie W. Chisholm. 2003. "Cyanophages Infecting the Oceanic Cyanobacterium Prochlorococcus." *Nature* 424 (6952): 1047–51.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, et al. 2015. "Ocean Plankton. Structure and Function of the Global Ocean Microbiome." *Science* 348 (6237): 1261359.

Sun, Jing, Laura Steindler, J. Cameron Thrash, Kimberly H. Halsey, Daniel P. Smith, Amy E. Carter, Zachary C. Landry, and Stephen J. Giovannoni. 2011. "One Carbon Metabolism in SAR11 Pelagic Marine Bacteria." *PloS One* 6 (8): e23973.

Sun, Ying, and Haiwei Luo. 2018. "Homologous Recombination in Core Genomes Facilitates Marine Bacterial Adaptation." *Applied and Environmental Microbiology* 84 (11). https://doi.org/10.1128/AEM.02545-17.

Sun, Zhiqiang, Hui Yin Tan, Piero R. Bianco, and Yuri L. Lyubchenko. 2015. "Remodeling of RecG Helicase at the DNA Replication Fork by SSB Protein." *Scientific Reports* 5 (April): 9625.

Suttle, C. A. 1994. "The Significance of Viruses to Mortality in Aquatic Microbial Communities." *Microbial Ecology* 28 (2): 237–43.

Suttle, Curtis A. 2005. "Viruses in the Sea." *Nature* 437 (7057): 356–61.

———. 2007a. "Marine Viruses—major Players in the Global Ecosystem." *Nature Reviews. Microbiology* 5 (10): 801–12.

———. 2007b. "Marine Viruses—major Players in the Global Ecosystem." *Nature Reviews. Microbiology* 5 (10): 801.

Suttle, Curtis A., Amy M. Chan, and Matthew T. Cottrell. 1990. "Infection of Phytoplankton by Viruses and Reduction of Primary Productivity." *Nature*. https://doi.org/10.1038/347467a0.

Sutton, Thomas D. S., Adam G. Clooney, Feargal J. Ryan, R. Paul Ross, and

Colin Hill. 2019. "Choice of Assembly Software Has a Critical Impact on Virome Characterisation." *Microbiome* 7 (1): 12.

Suzuki, Yoshihiko, Suguru Nishijima, Yoshikazu Furuta, Jun Yoshimura, Wataru Suda, Kenshiro Oshima, Masahira Hattori, and Shinichi Morishita. 2019. "Long-Read Metagenomic Exploration of Extrachromosomal Mobile Genetic Elements in the Human Gut." *Microbiome* 7 (1): 119.

Swan, Brandon K., Manuel Martinez-Garcia, Christina M. Preston, Alexander Sczyrba, Tanja Woyke, Dominique Lamy, Thomas Reinthaler, et al. 2011. "Potential for Chemolithoautotrophy among Ubiquitous Bacteria Lineages in the Dark Ocean." *Science* 333 (6047): 1296–1300.

Tadmor, Arbel D., Elizabeth A. Ottesen, Jared R. Leadbetter, and Rob Phillips. 2011. "Probing Individual Environmental Bacteria for Viruses by Using Microfluidic Digital PCR." *Science* 333 (6038): 58–62.

Takahashi, Taro, Stewart C. Sutherland, Colm Sweeney, Alain Poisson, Nicolas Metzl, Bronte Tilbrook, Nicolas Bates, et al. 2002. "Global Sea–air CO2 Flux Based on Climatological Surface Ocean pCO2, and Seasonal Biological and Temperature Effects." *Deep-Sea Research. Part II, Topical Studies in Oceanography* 49 (9): 1601–22.

Tarasov, Sergei, and Dimitar Dimitrov. 2016. "Multigene Phylogenetic Analysis Redefines Dung Beetles Relationships and Classification (Coleoptera: Scarabaeidae: Scarabaeinae)." *BMC Evolutionary Biology* 16 (1): 257.

Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin. 2000. "The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution." *Nucleic Acids Research* 28 (1): 33–36.

Telenius, H., N. P. Carter, C. E. Bebb, M. Nordenskjöld, B. A. Ponder, and A. Tunnacliffe. 1992. "Degenerate Oligonucleotide-Primed PCR: General Amplification of Target DNA by a Single Degenerate Primer." *Genomics* 13 (3): 718–25.

Thingstad, T. F., and R. Lignell. 1997. "Theoretical Models for the Control of Bacterial Growth Rate, Abundance, Diversity and Carbon Demand." *Aquatic Microbial Ecology: International Journal* 13: 19–27.

Thomas, Christopher M., and Kaare M. Nielsen. 2005. "Mechanisms Of, and Barriers To, Horizontal Gene Transfer between Bacteria." *Nature Reviews. Microbiology* 3 (9): 711–21.

Thompson, Cristiane C., Luciane Chimetto, Robert A. Edwards, Jean Swings, Erko Stackebrandt, and Fabiano L. Thompson. 2013. "Microbial Genomic Taxonomy." *BMC Genomics* 14 (December): 913.

Thompson, Luke R., Qinglu Zeng, Libusha Kelly, Katherine H. Huang, Alexander U. Singer, Joanne Stubbe, and Sallie W. Chisholm. 2011. "Phage Auxiliary Metabolic Genes and the Redirection of Cyanobacterial Host Carbon Metabolism." *Proceedings of the National Academy of Sciences of the United States of America* 108 (39): E757–64.

Thrash, J. Cameron, Alex Boyd, Megan J. Huggett, Jana Grote, Paul Carini, Ryan J. Yoder, Barbara Robbertse, Joseph W. Spatafora, Michael S. Rappé, and Stephen J. Giovannoni. 2011. "Phylogenomic Evidence for a Common Ancestor of Mitochondria and the SAR11 Clade." *Scientific Reports* 1 (June): 13.

Thrash, J. Cameron, Ben Temperton, Brandon K. Swan, Zachary C. Landry, Tanja Woyke, Edward F. DeLong, Ramunas Stepanauskas, and Stephan J. Giovannoni. 2014. "Single-Cell Enabled Comparative Genomics of a Deep Ocean SAR11 Bathytype." *The ISME Journal* 8 (7): 1440–51.

Tørresen, Ole K., Bastiaan Star, Pablo Mier, Miguel A. Andrade-Navarro, Alex Bateman, Patryk Jarnot, Aleksandra Gruca, et al. 2019. “Tandem Repeats Lead to Sequence Assembly Errors and Impose Multi-Level Challenges for Genome and Protein Databases.” *Nucleic Acids Research* 47 (21): 10994–6.

Treangen, Todd J., and Steven L. Salzberg. 2011. “Repetitive DNA and next-Generation Sequencing: Computational Challenges and Solutions.” *Nature Reviews. Genetics* 13 (1): 36–46.

Tringe, Susannah Green, Christian von Mering, Arthur Kobayashi, Asaf A. Salamov, Kevin Chen, Hwai W. Chang, Mircea Podar, et al. 2005. “Comparative Metagenomics of Microbial Communities.” *Science* 308 (5721): 554–57.

Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. “Microbial Strain-Level Population Structure and Genetic Diversity from Metagenomes.” *Genome Research* 27 (4): 626–38.

Tsementzi, Despina, Jieying Wu, Samuel Deutsch, Sangeeta Nath, Luis M. Rodriguez-R, Andrew S. Burns, Piyush Ranjan, et al. 2016. “SAR11 Bacteria Linked to Ocean Anoxia and Nitrogen Loss.” *Nature* 536 (7615): 179–83.

Tully, Benjamin J., Elaina D. Graham, and John F. Heidelberg. 2018. “The Reconstruction of 2,631 Draft Metagenome-Assembled Genomes from the Global Oceans.” *Scientific Data* 5 (January): 170203.

Turnbaugh, Peter J., Ruth E. Ley, Michael A. Mahowald, Vincent Magrini, Elaine R. Mardis, and Jeffrey I. Gordon. 2006. “An Obesity-Associated Gut Microbiome with Increased Capacity for Energy Harvest.” *Nature* 444 (7122): 1027–31.

Tyson, Gene W., Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar, and Jillian F. Banfield. 2004. “Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment.” *Nature* 428 (6978): 37–43.

Umezu, K., N. W. Chi, and R. D. Kolodner. 1993. “Biochemical Interaction of the Escherichia Coli RecF, RecO, and RecR Proteins with RecA Protein and Single-Stranded DNA Binding Protein.” *Proceedings of the National Academy of Sciences of the United States of America* 90 (9): 3875–79.

Våge, Selina, Julia E. Storesund, Jarl Giske, and T. Frede Thingstad. 2014. “Optimal Defense Strategies in an Idealized Microbial Food Web under Trade-off between Competition and Defense.” *PloS One* 9 (7): e101415.

Våge, Selina, Julia E. Storesund, and T. Frede Thingstad. 2013. “SAR11 Viruses and Defensive Host Strains.” *Nature*.

Van Hemert, Formijn, Maarten Jebbink, Andries Van Der Ark, Frits Scholer, and Ben Berkhout. 2018. “Euclidean Distance Analysis Enables Nucleotide Skew Analysis in Viral Genomes.” *Computational and Mathematical Methods in Medicine* 2018. https://www.hindawi.com/journals/cmmm/2018/6490647/abs/.

Van Valen, Leigh. 1973. “A NEW EVOLUTIONARY LAW.” *Evolutionary Theory* 1: 1–30.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. “Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads.” *Genome Research* 27 (5): 737–46.

Venter, J. Craig, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug

Rusch, Jonathan A. Eisen, Dongying Wu, et al. 2004. “Environmental Genome Shotgun Sequencing of the Sargasso Sea.” *Science* 304 (5667): 66–74.

Verdugo, Pedro, Alice L. Alldredge, Farooq Azam, David L. Kirchman, Uta Passow, and Peter H. Santschi. 2004. “The Oceanic Gel Phase: A Bridge in the DOM–POM Continuum.” *Marine Chemistry* 92 (1): 67–85.

Vergin, Kevin L., Bánk Beszteri, Adam Monier, J. Cameron Thrash, Ben Temperton, Alexander H. Treusch, Fabian Kilpert, Alexandra Z. Worden, and Stephen J. Giovannoni. 2013. “High-Resolution SAR11 Ecotype Dynamics at the Bermuda Atlantic Time-Series Study Site by Phylogenetic Placement of Pyrosequences.” *The ISME Journal*. https://doi.org/10.1038/ismej.2013.32.

Vergin, Kevin L., H. James Tripp, Larry J. Wilhelm, Dee R. Denver, Michael S. Rappé, and Stephen J. Giovannoni. 2007. “High Intraspecific Recombination Rate in a Native Population of Candidatus Pelagibacter Ubique (SAR11).” *Environmental Microbiology* 9 (10): 2430–40.

Viklund, Johan, Joran Martijn, Thijs J. G. Ettema, and Siv G. E. Andersson. 2013. “Comparative and Phylogenomic Evidence That the Alphaproteobacterium HIMB59 Is Not a Member of the Oceanic SAR11 Clade.” *PloS One* 8 (11): e78858.

Vollmers, John, Sandra Wiegand, and Anne-Kristin Kaster. 2017. “Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist’s Perspective - Not Only Size Matters!” *PLOS ONE*. https://doi.org/10.1371/journal.pone.0169662.

Vries, Johann de, Thilo Herzfeld, and Wilfried Wackernagel. 2004. “Transfer of Plastid DNA from Tobacco to the Soil Bacterium Acinetobacter Sp. by Natural Transformation.” *Molecular Microbiology* 53 (1): 323–34.

Vries, Johann de, and Wilfried Wackernagel. 2002. “Integration of Foreign DNA during Natural Transformation of Acinetobacter Sp. by Homology-Facilitated Illegitimate Recombination.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (4): 2094–99.

Waldor, Matthew K., and David I. Friedman. 2005. “Phage Regulatory Circuits and Virulence Gene Expression.” *Current Opinion in Microbiology* 8 (4): 459–65.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement.” *PloS One* 9 (11): e112963.

Wang, Lianrong, Shi Chen, Kevin L. Vergin, Stephen J. Giovannoni, Simon W. Chan, Michael S. DeMott, Koli Taghizadeh, et al. 2011. “DNA Phosphorothioation Is Widespread and Quantized in Bacterial Genomes.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (7): 2963–68.

Warwick-Dugdale, Joanna, Natalie Solonenko, Karen Moore, Lauren Chittick, Ann C. Gregory, Michael J. Allen, Matthew B. Sullivan, and Ben Temperton. 2019. “Long-Read Viral Metagenomics Captures Abundant and Microdiverse Viral Populations and Their Niche-Defining Genomic Islands.” *PeerJ* 7 (April): e6800.

Weinbauer, Markus G. 2004. “Ecology of Prokaryotic Viruses.” *FEMS Microbiology Reviews* 28 (2): 127–81.

Weirather, Jason L., Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. 2017. “Comprehensive Comparison of Pacific Biosciences and Oxford Nanopore Technologies and Their Applications to Transcriptome Analysis.” *F1000Research*. https://doi.org/10.12688/f1000research.10571.2.

Wesolowska-Andersen, Agata, Martin Iain Bahl, Vera Carvalho, Karsten Kristiansen, Thomas Sicheritz-Pontén, Ramneek Gupta, and Tine Rask Licht. 2014. “Choice of Bacterial DNA Extraction Method from Fecal Material Influences Community Structure as Evaluated by Metagenomic Analysis.” *Microbiome* 2 (June): 19.

White, Adam K., Michael VanInsberghe, Oleh I. Petriv, Mani Hamidi, Darek Sikorski, Marco A. Marra, James Piret, Samuel Aparicio, and Carl L. Hansen. 2011. “High-Throughput Microfluidic Single-Cell RT-qPCR.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (34): 13999–4.

White, Angelicque E., Stephen J. Giovannoni, Yanlin Zhao, Kevin Vergin, and Craig A. Carlson. 2019. “Elemental Content and Stoichiometry of SAR11 Chemoheterotrophic Marine Bacteria.” *Limnology and Oceanography Letters* 4 (2): 44–51.

Wick, Ryan R., Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. 2017. “Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads.” *PLoS Computational Biology* 13 (6): e1005595.

Wick, Ryan R., Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. 2015. “Bandage: Interactive Visualization of de Novo Genome Assemblies.” *Bioinformatics* 31 (20): 3350–52.

Wilhelm, Larry J., H. James Tripp, Scott A. Givan, Daniel P. Smith, and Stephen J. Giovannoni. 2007. “Natural Variation in SAR11 Marine Bacterioplankton Genomes Inferred from Metagenomic Data.” *Biology Direct* 2 (November): 27.

Wilhelm, Steven W., and Curtis A. Suttle. 1999. “Viruses and Nutrient Cycles in the Sea.” *BioScience*. https://doi.org/10.2307/1313569.

Wilkins, Laetitia G. E., Cassandra L. Ettinger, Guillaume Jospin, and Jonathan A. Eisen. 2020. “Author Correction: Metagenome-Assembled Genomes Provide New Insight into the Microbial Diversity of Two Thermal Pools in Kamchatka, Russia.” *Scientific Reports* 10 (1): 3454.

Winter, Christian, Thierry Bouvier, Markus G. Weinbauer, and T. Frede Thingstad. 2010. “Trade-Offs between Competition and Defense Specialists among Unicellular Planktonic Organisms: The ‘Killing the Winner’ Hypothesis Revisited.” *Microbiology and Molecular Biology Reviews: MMBR* 74 (1): 42–57.

Winter, Christian, Arjan Smit, Gerhard J. Herndl, and Markus G. Weinbauer. 2004. “Impact of Virioplankton on Archaeal and Bacterial Community Richness as Assessed in Seawater Batch Cultures.” *Applied and Environmental Microbiology* 70 (2): 804–13.

Wommack, K. E., and R. R. Colwell. 2000. “Virioplankton: Viruses in Aquatic Ecosystems.” *Microbiology and Molecular Biology Reviews: MMBR* 64 (1): 69–114.

Wommack, K. E., R. T. Hill, M. Kessel, E. Russek-Cohen, and R. R. Colwell. 1992. “Distribution of Viruses in the Chesapeake Bay.” *Applied and Environmental Microbiology* 58 (9): 2965–70.

Wooley, John C., and Yuzhen Ye. 2010. “Metagenomics: Facts and Artifacts, and Computational Challenges." *Journal of Computer Science and Technology* 25 (1): 71–81.

Woolhouse, Mark E. J., Joanne P. Webster, Esteban Domingo, Brian Charlesworth, and Bruce R. Levin. 2002. "Biological and Biomedical Implications of the Co-Evolution of Pathogens and Their Hosts." *Nature Genetics* 32 (4): 569–77.

Woyke, Tanja, Damon Tighe, Konstantinos Mavromatis, Alicia Clum, Alex Copeland, Wendy Schackwitz, Alla Lapidus, et al. 2010. "One Bacterial Cell, One Complete Genome." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0010314.

Woyke, Tanja, Gary Xie, Alex Copeland, José M. González, Cliff Han, Hajnalka Kiss, Jimmy H. Saw, et al. 2009. "Assembling the Marine Metagenome, One Cell at a Time." *PloS One* 4 (4): e5299.

Wu, Yu-Wei, Yung-Hsu Tang, Susannah G. Tringe, Blake A. Simmons, and Steven W. Singer. 2014. "MaxBin: An Automated Binning Method to Recover Individual Genomes from Metagenomes Using an Expectation-Maximization Algorithm." *Microbiome* 2 (August): 26.

Xiong, Jin. 2006. *Essential Bioinformatics*. Cambridge University Press.

Yilmaz, Suzan, Martin Allgaier, and Philip Hugenholtz. 2010. "Multiple Displacement Amplification Compromises Quantitative Analysis of Metagenomes." *Nature Methods* 7 (12): 943–44.

Yooseph, Shibu, Kenneth H. Nealson, Douglas B. Rusch, John P. McCrow, Christopher L. Dupont, Maria Kim, Justin Johnson, et al. 2010. "Genomic and Functional Adaptation in Surface Ocean Planktonic Prokaryotes." *Nature* 468 (7320): 60–66.

Yoshida, Takashi, Yosuke Nishimura, Hiroyasu Watai, Nana Haruki, Daichi Morimoto, Hiroto Kaneko, Takashi Honda, et al. 2018. "Locality and Diel Cycling of Viral Production Revealed by a 24 H Time Course Cross-Omics Analysis in a Coastal Region of Japan." *The ISME Journal*. https://doi.org/10.1038/s41396-018-0052-x.

Yuan, Sanqing, Dora B. Cohen, Jacques Ravel, Zaid Abdo, and Larry J. Forney. 2012. "Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome." *PloS One* 7 (3): e33865.

Zaremba-Niedzwiedzka, Katarzyna, Eva F. Caceres, Jimmy H. Saw, Disa Bäckström, Lina Juzokaite, Emmelien Vancaester, Kiley W. Seitz, et al. 2017. "Asgard Archaea Illuminate the Origin of Eukaryotic Cellular Complexity." *Nature* 541 (7637): 353–58.

Zhang, D. Y., M. Brandwein, T. Hsuih, and H. B. Li. 2001. "Ramification Amplification: A Novel Isothermal DNA Amplification Method." *Molecular Diagnosis: A Journal Devoted to the Understanding of Human Disease through the Clinical Application of Molecular Biology* 6 (2): 141–50.

Zhang, Kun, Adam C. Martiny, Nikos B. Reppas, Kerrie W. Barry, Joel Malek, Sallie W. Chisholm, and George M. Church. 2006. "Sequencing Genomes from Single Cells by Polymerase Cloning." *Nature Biotechnology* 24 (6): 680–86.

Zhang, L., X. Cui, K. Schmitt, R. Hubert, W. Navidi, and N. Arnheim. 1992. "Whole Genome Amplification from a Single Cell: Implications for Genetic Analysis." *Proceedings of the National Academy of Sciences of the United States of America* 89 (13): 5847–51.

Zhang, Wenyu, Jiajia Chen, Yang Yang, Yifei Tang, Jing Shang, and Bairong Shen. 2011. "A Practical Comparison of De Novo Genome Assembly

Software Tools for Next-Generation Sequencing Technologies." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0017915.

Zhao, X., C. L. Schwartz, and J. Pierson. 2017. "Three-Dimensional Structure of the Ultraoligotrophic Marine Bacterium 'Candidatus Pelagibacter Ubique.'" *Appl. Environ*. https://aem.asm.org/content/83/3/e02807-16.short.

Zhao, Yanlin, Fang Qin, Rui Zhang, Stephen J. Giovannoni, Zefeng Zhang, Jing Sun, Sen Du, and Christopher Rensing. 2019. "Pelagiphages in the Podoviridae Family Integrate into Host Genomes." *Environmental Microbiology* 21 (6): 1989–2001.

Zhao, Yanlin, Ben Temperton, J. Cameron Thrash, Michael S. Schwalbach, Kevin L. Vergin, Zachary C. Landry, Mark Ellisman, Tom Deerinck, Matthew B. Sullivan, and Stephen J. Giovannoni. 2013. "Abundant SAR11 Viruses in the Ocean." *Nature* 494 (7437): 357–60.

Zhou, You, Yongjie Liang, Karlene H. Lynch, Jonathan J. Dennis, and David S. Wishart. 2011. "PHAST: A Fast Phage Search Tool." *Nucleic Acids Research* 39 (Web Server issue): W347–52.

Zimmermann, R., R. Iturriaga, and J. Becker-Birck. 1978. "Simultaneous Determination of the Total Number of Aquatic Bacteria and the Number Thereof Involved in Respiration." *Applied and Environmental Microbiology* 36 (6): 926–35.

7 Appendices

7.1.1 Phylogenetic Trees

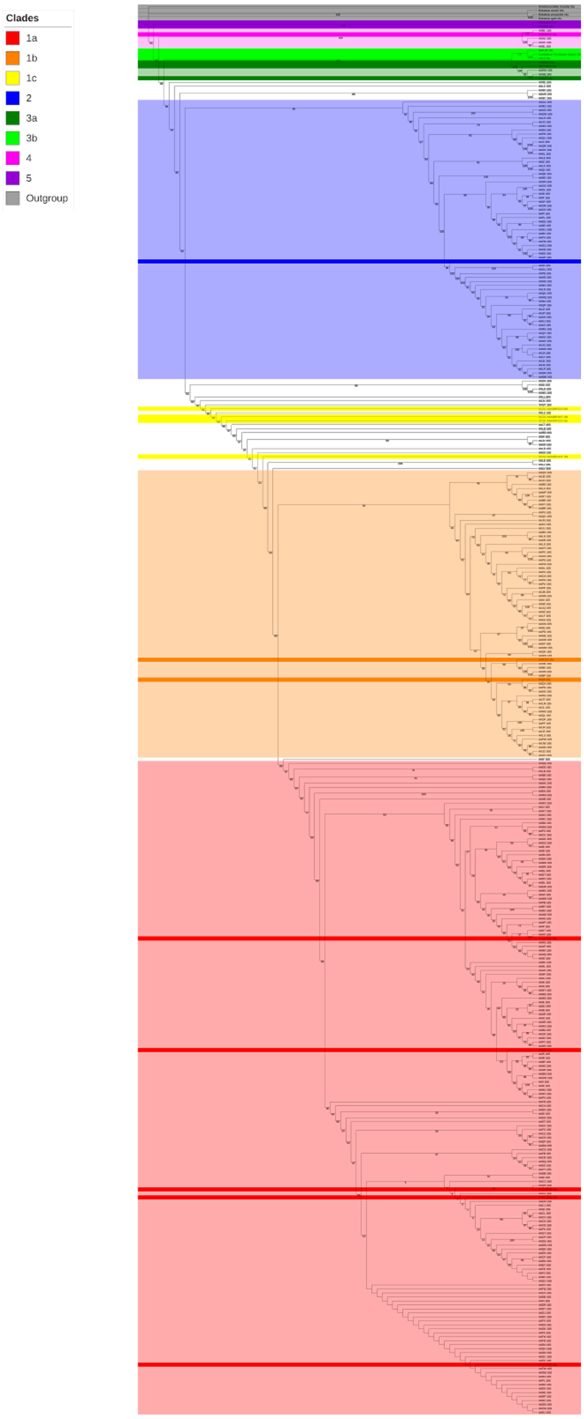


Figure 7.1 Whole-genome sequence phylogenetic tree of SAR11 SAGs with clade colours displayed. Bold colours indicate reference sequences

7.1.2 SAR11 Clade ecological Mappings

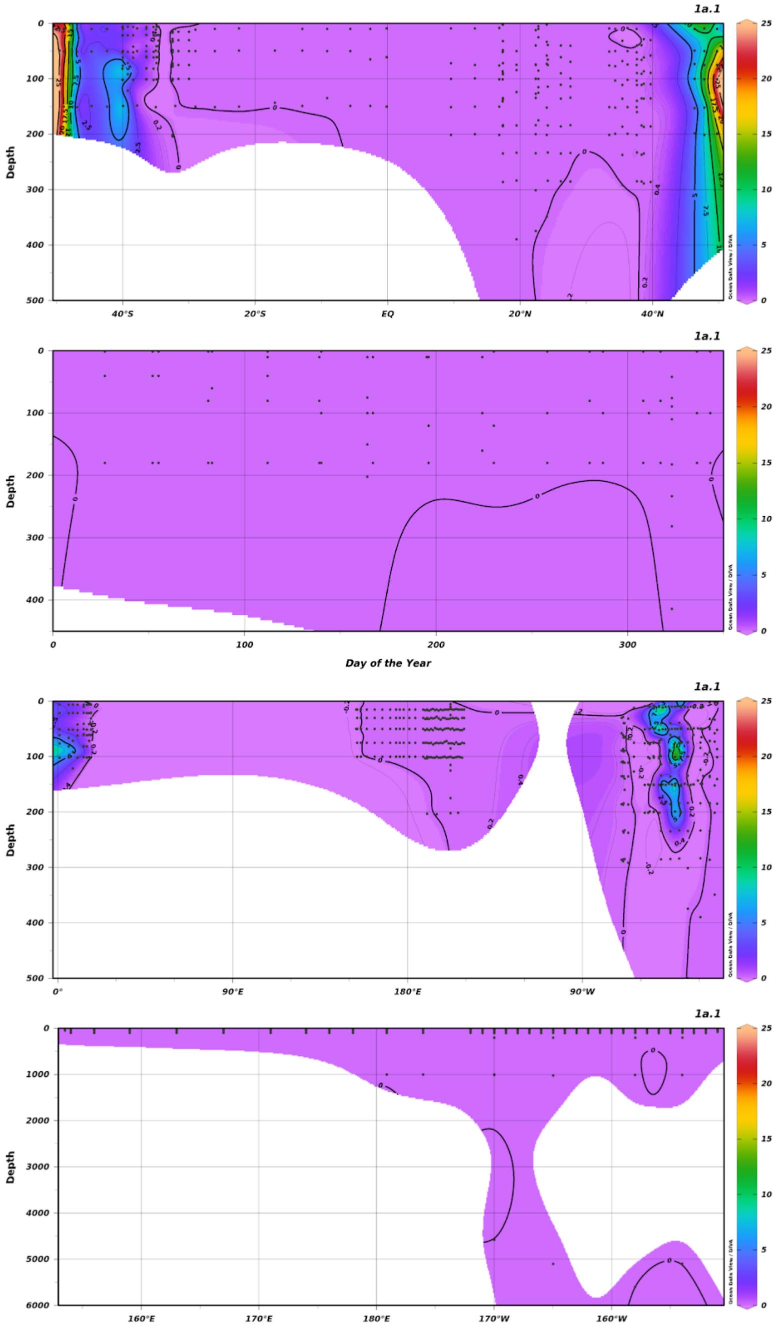


Figure 7.2 Mapping of SAR11 Clade Ia.1 reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade Ia.1 is present at higher and lower latitudes of 40 degrees N or S and above. It is generally present in water depths of 200m or above with a slow decrease of abundance below 200m to 400m.

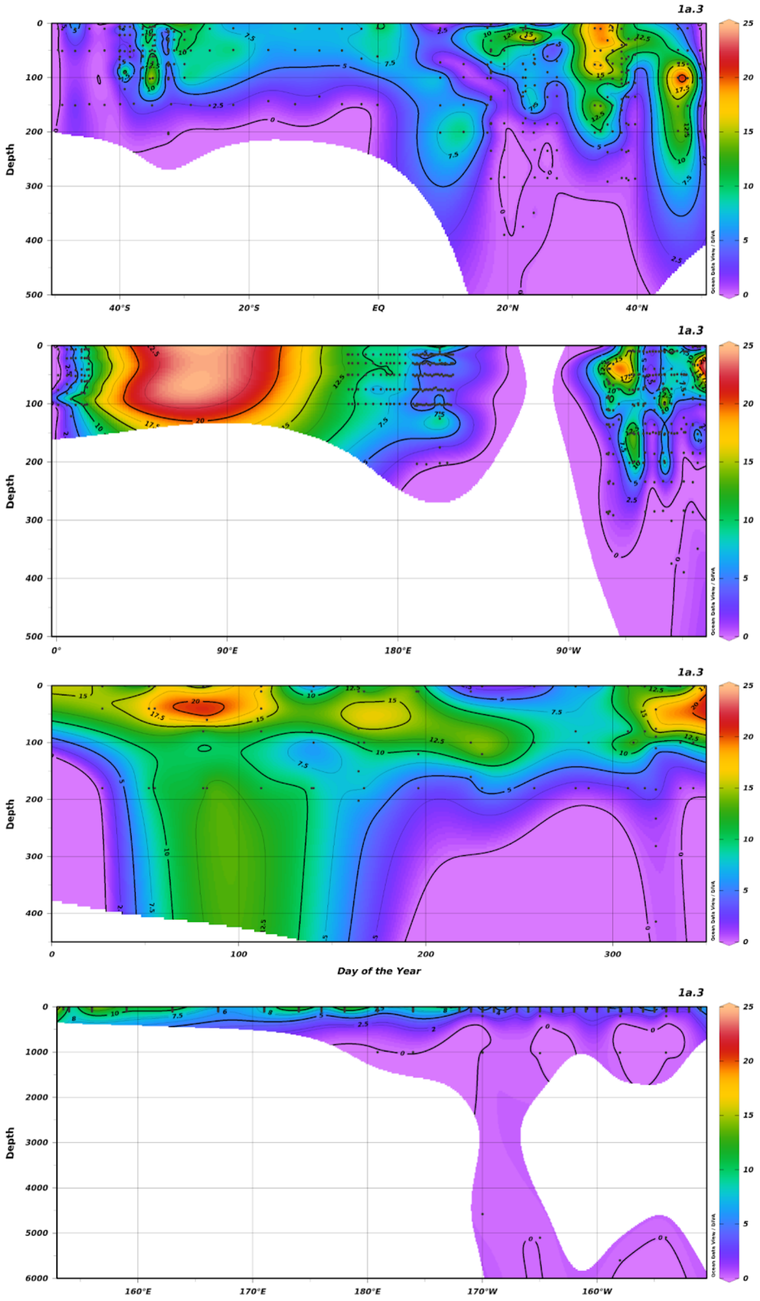


Figure 7.3 Mapping of SAR11 Clade Ia.3 reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Ia.3 is present between 100m and above and observed between the longitudes of 45 - 100 degrees east. It looks to be more generally present in the northern hemisphere over the southern hemisphere. There is a clear seasonal downwelling between the period of February to May and a slight increase in winter months between November and December. It is generally observed to be coastal with a reduction in abundance towards open oceans.

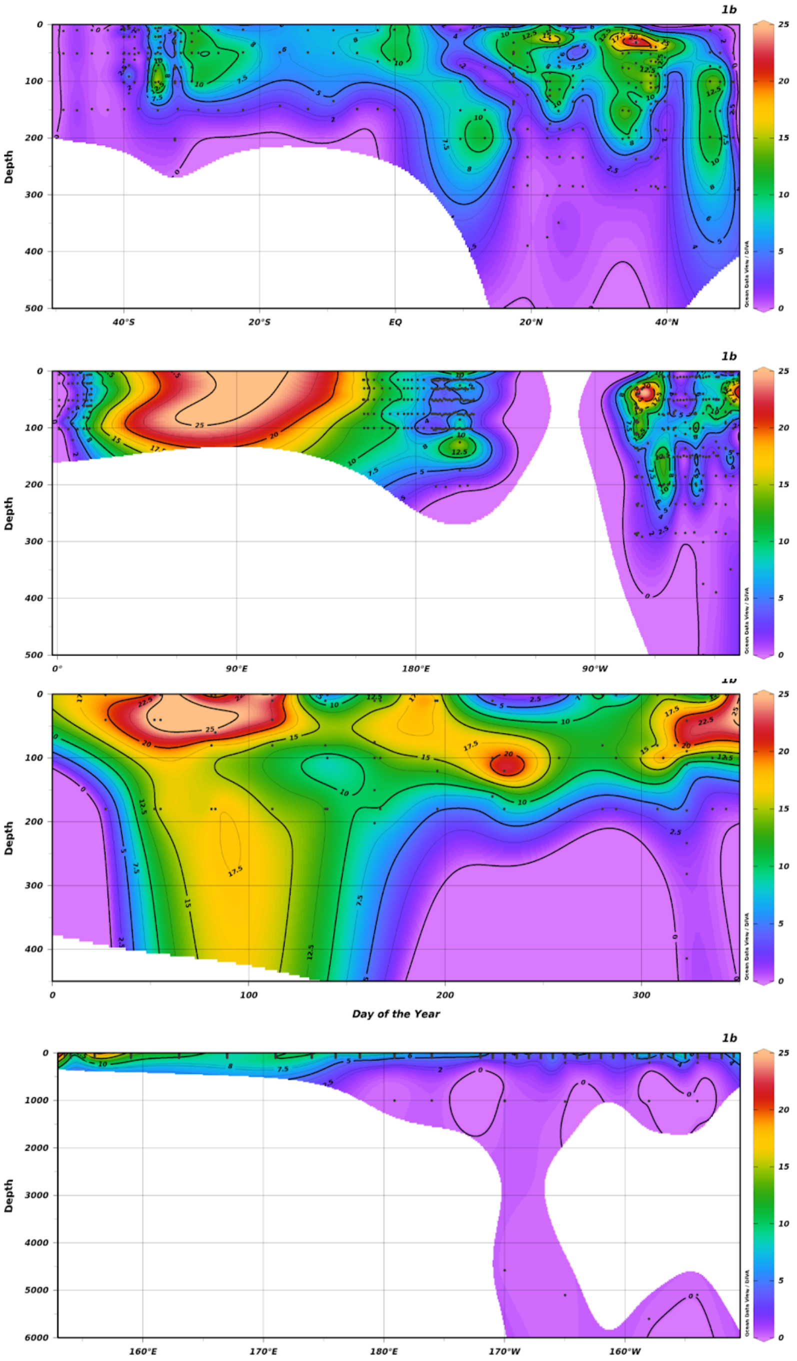


Figure 7.4 Mapping of SAR11 Clade Ib reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade Ib is generally observed in latitudes of 40 degrees north to 40 degrees south. It is particularly abundant between 45 to 110 degrees east and between 0 - 100 m. It has seasonal downwelling events in February to April and present in autumn months of July - August and in winter months from November to December. It is not present in open oceans.

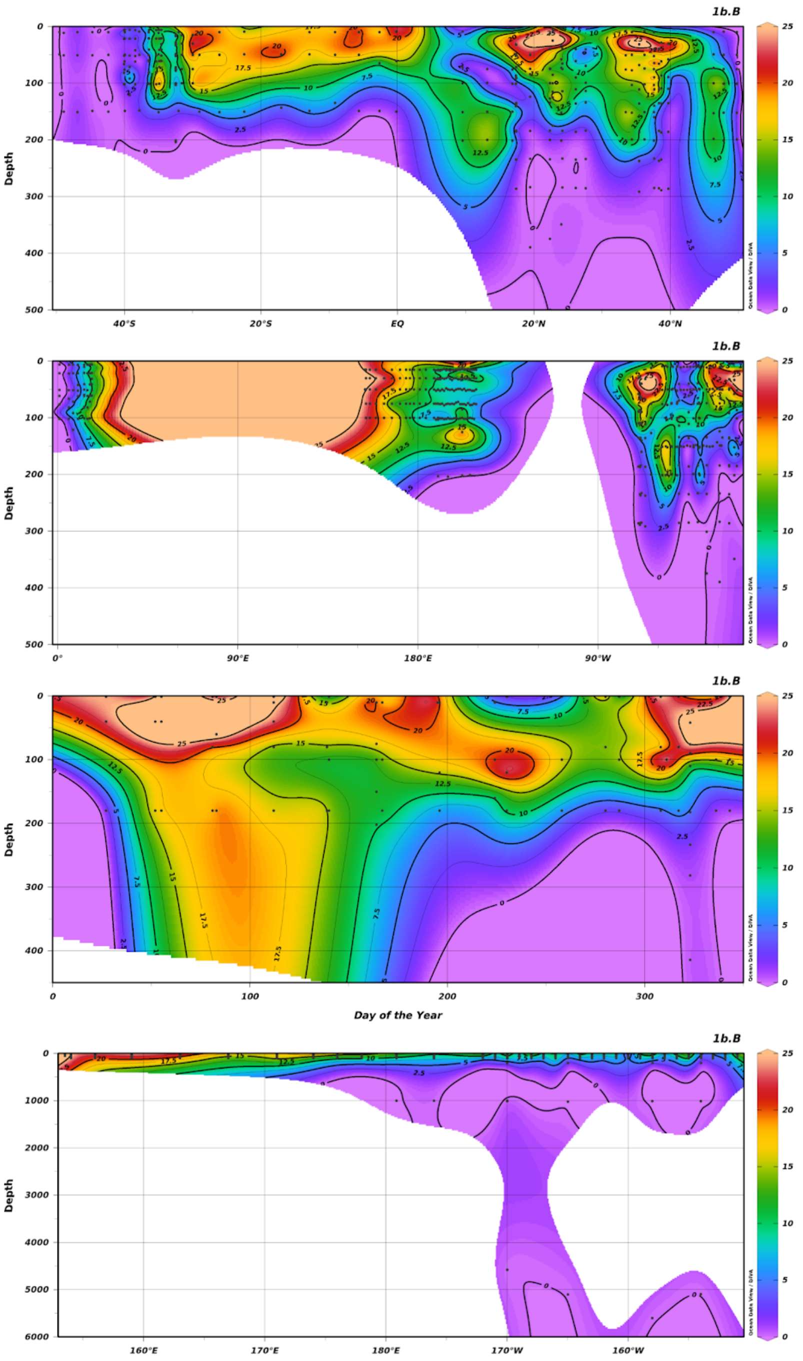


Figure 7.5 Mapping of SAR11 Clade Ib.2 reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade Ib.2 is generally observed in latitudes of 40 degrees north to 40 degrees south. It is particularly abundant between 45 to 110 degrees east and between 0 - 100 m. It has seasonal downwelling events in February to April and present in autumn months of July - August and in winter months from November to December. It is not present in open oceans.

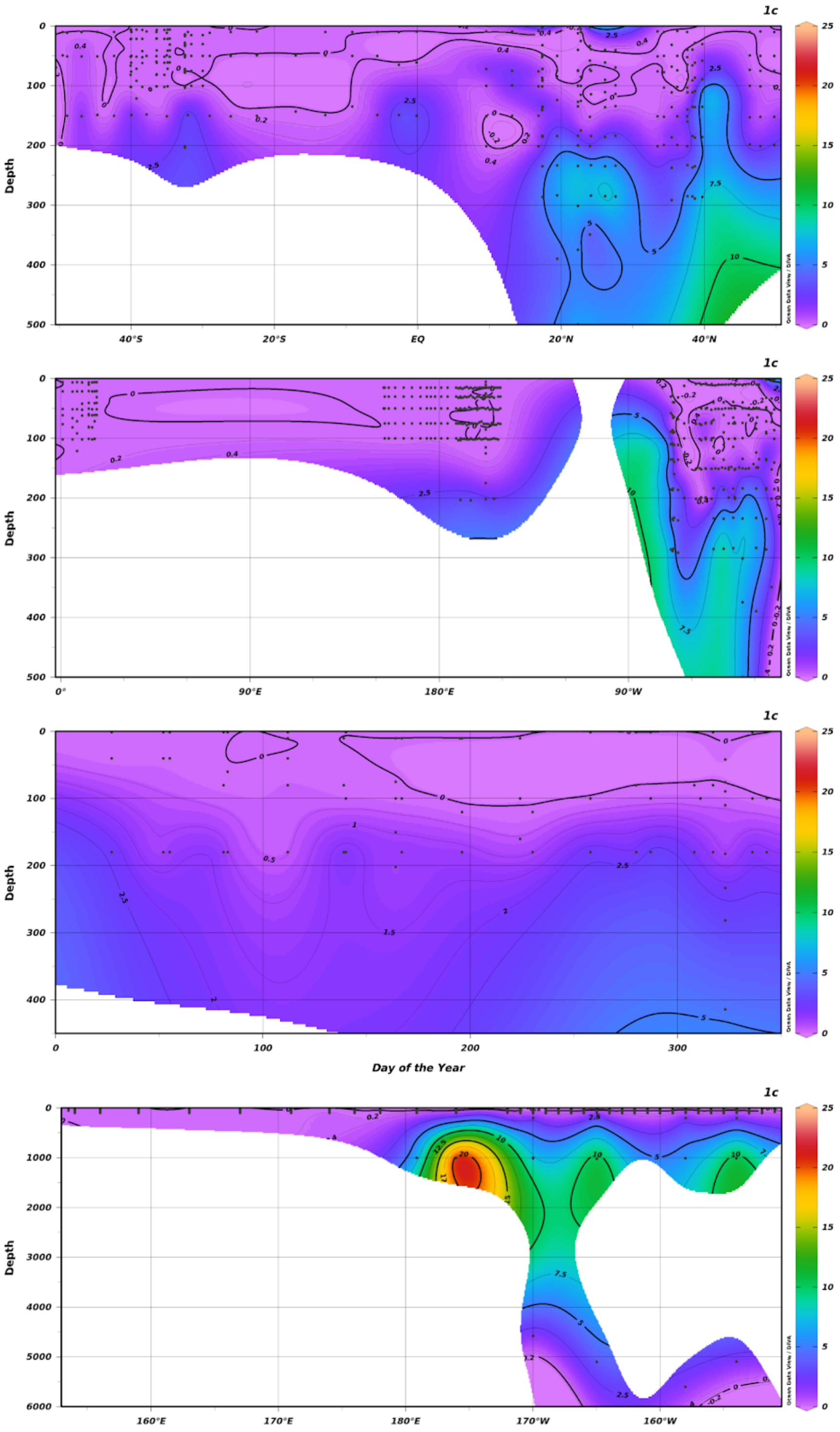


Figure 7.6 Mapping of SAR11 Clade Ic reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade Ic is present from 200m to about 4000m meters in depth and observed throughout all marine habitats. There is a noticeable absence in BATS upwelling events between February and April.

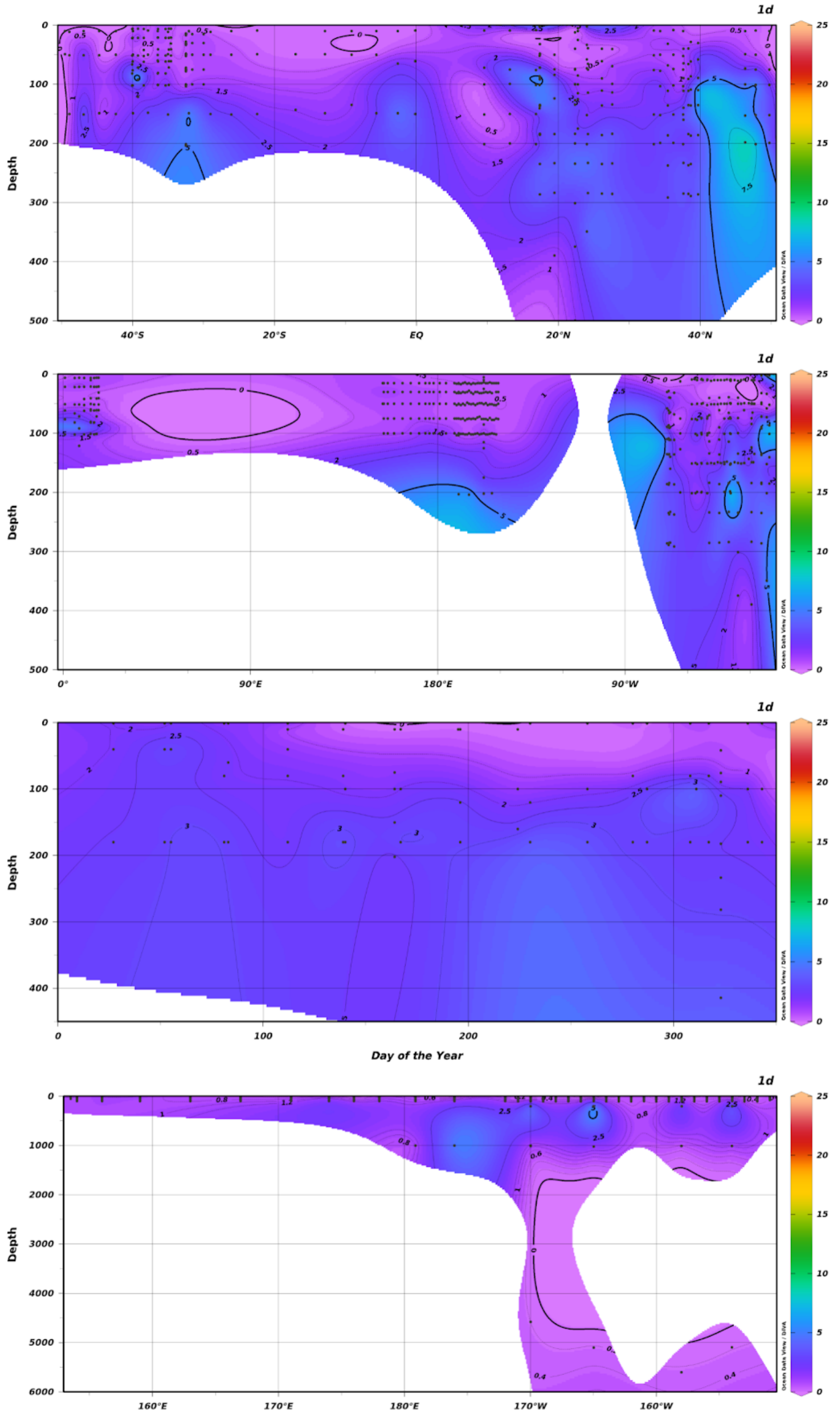


Figure 7.7 Mapping of SAR11 Clade Id reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade Id is present from 100m to 1000m throughout the ocean. At BATS it is noticeably more abundant in spring conditions in surface waters above 100m but decreases in abundance and stratifies to lower depths after April. This cycle repeats from February to April.

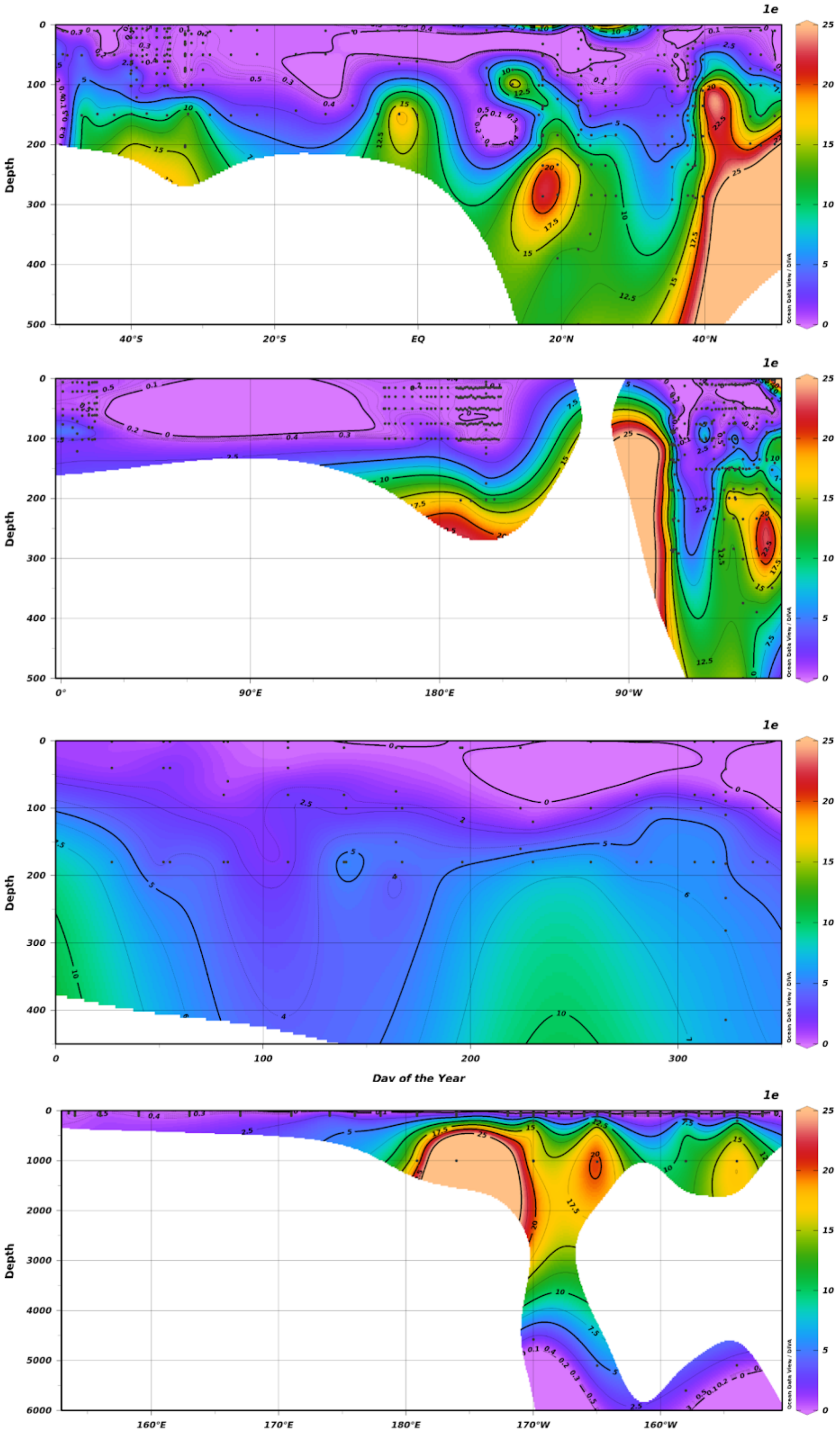


Figure 7.8 Mapping of SAR11 Clade Ie reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade Ie is observed from 100m and below until about 4000m. It is noticeably absent from February to April at lower depths. It is observed throughout the ocean.

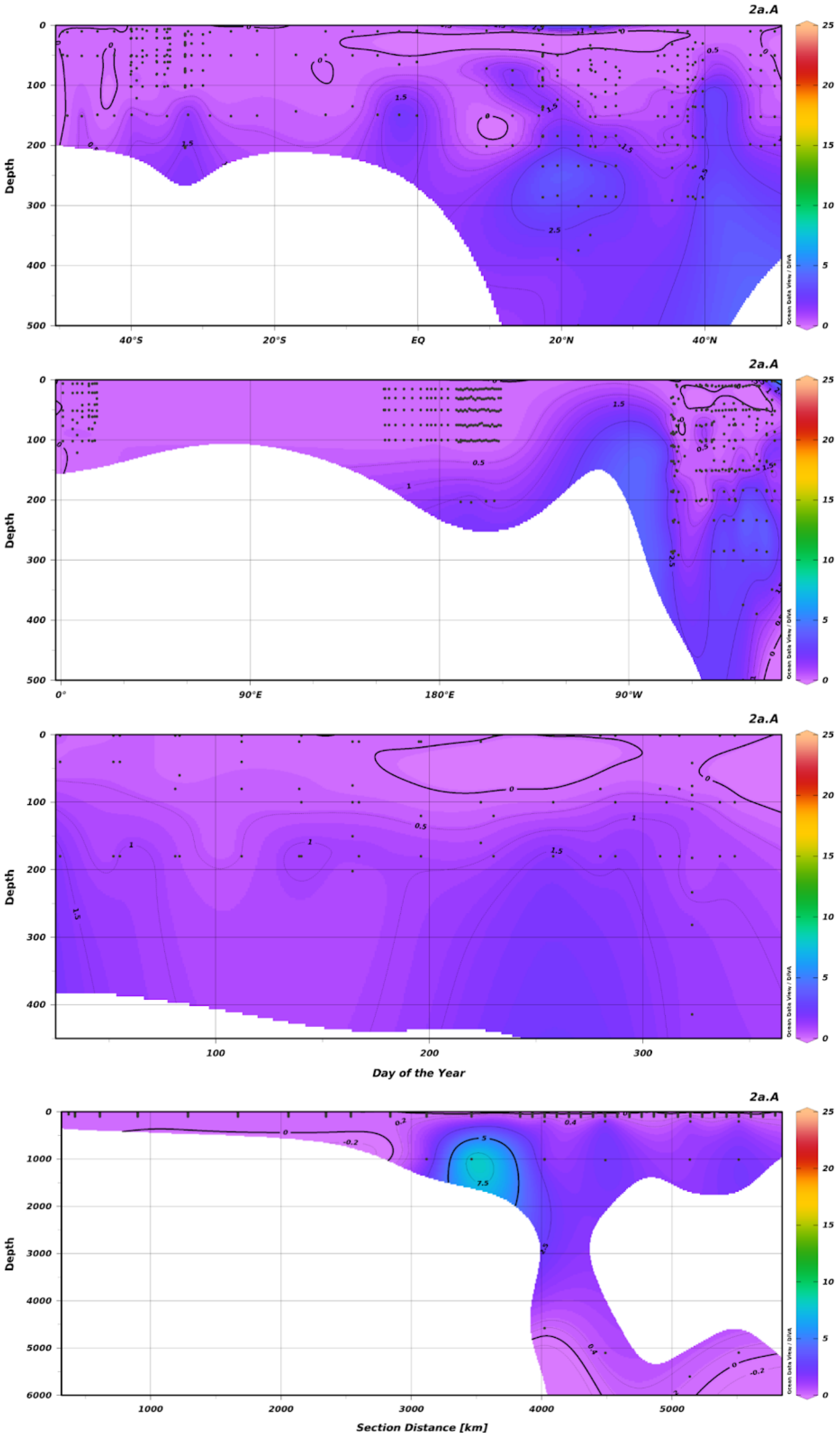


Figure 7.9 Mapping of SAR11 Clade IIa.A reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade IIa.A is present at depths below 200m until 4000m. Loss during February and April period. Present throughout the Ocean.

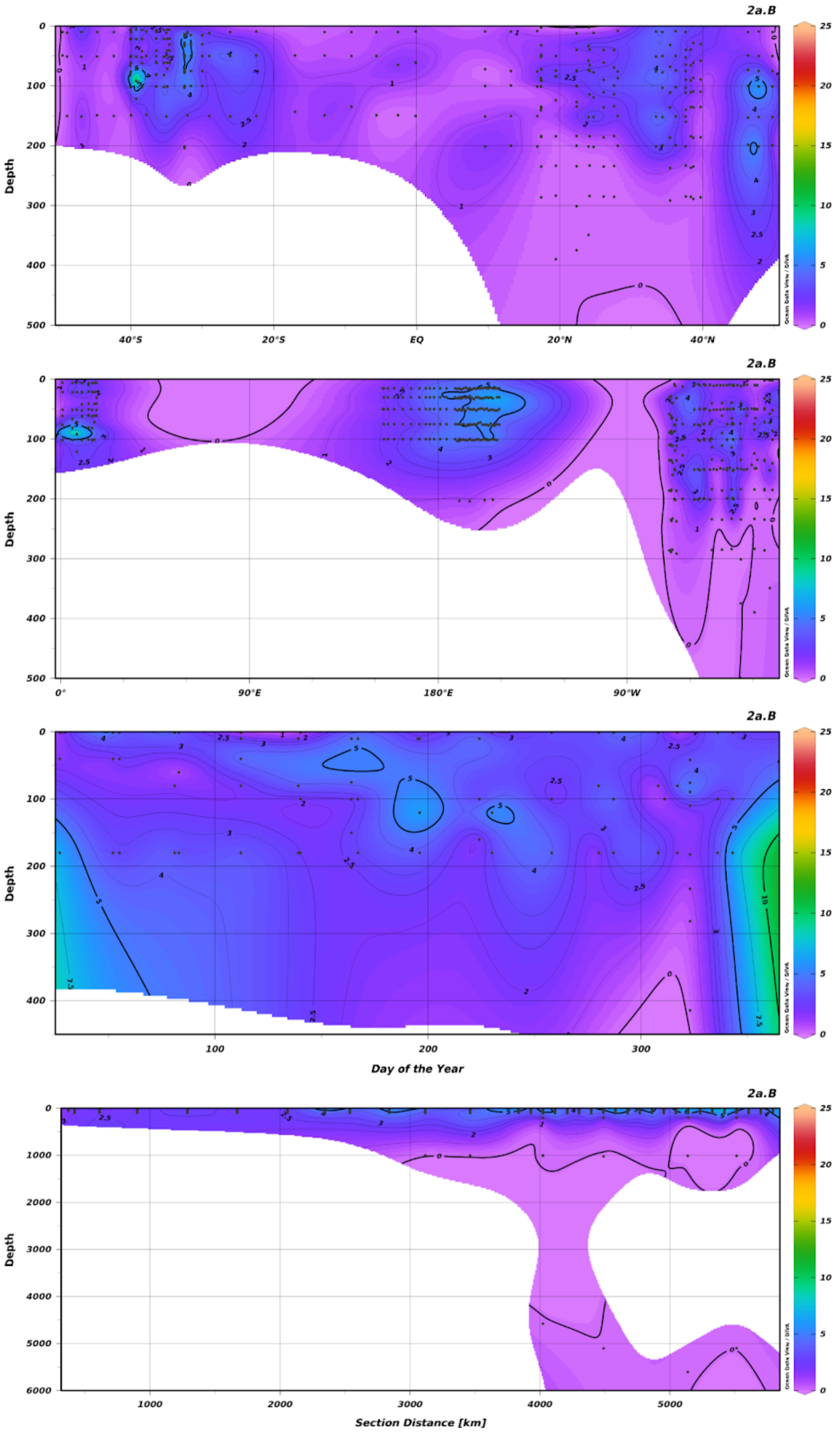


Figure 7.10 Mapping of SAR11 Clade IIa.B reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade IIa.B present throughout the ocean at 200m and above. Not affected by upwelling events from Feb to April. High in abundance from Nov to Jan period at 150m to 400m+ depths at BATS.

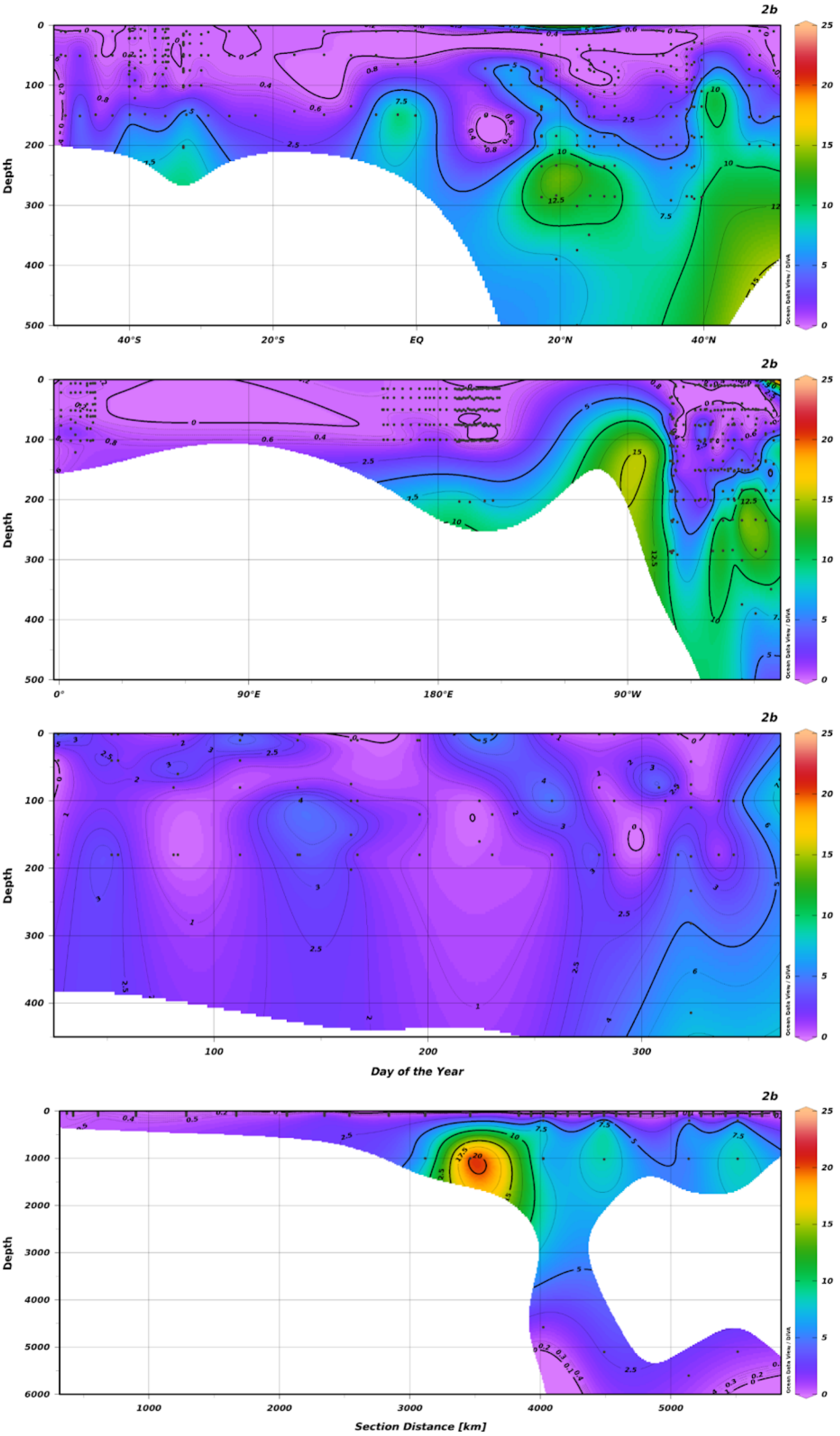


Figure 7.11 Mapping of SAR11 Clade IIb reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade IIb is present at 200m and below throughout the ocean. More abundant in winter months from Nov to Jan with a loss at upwelling events.

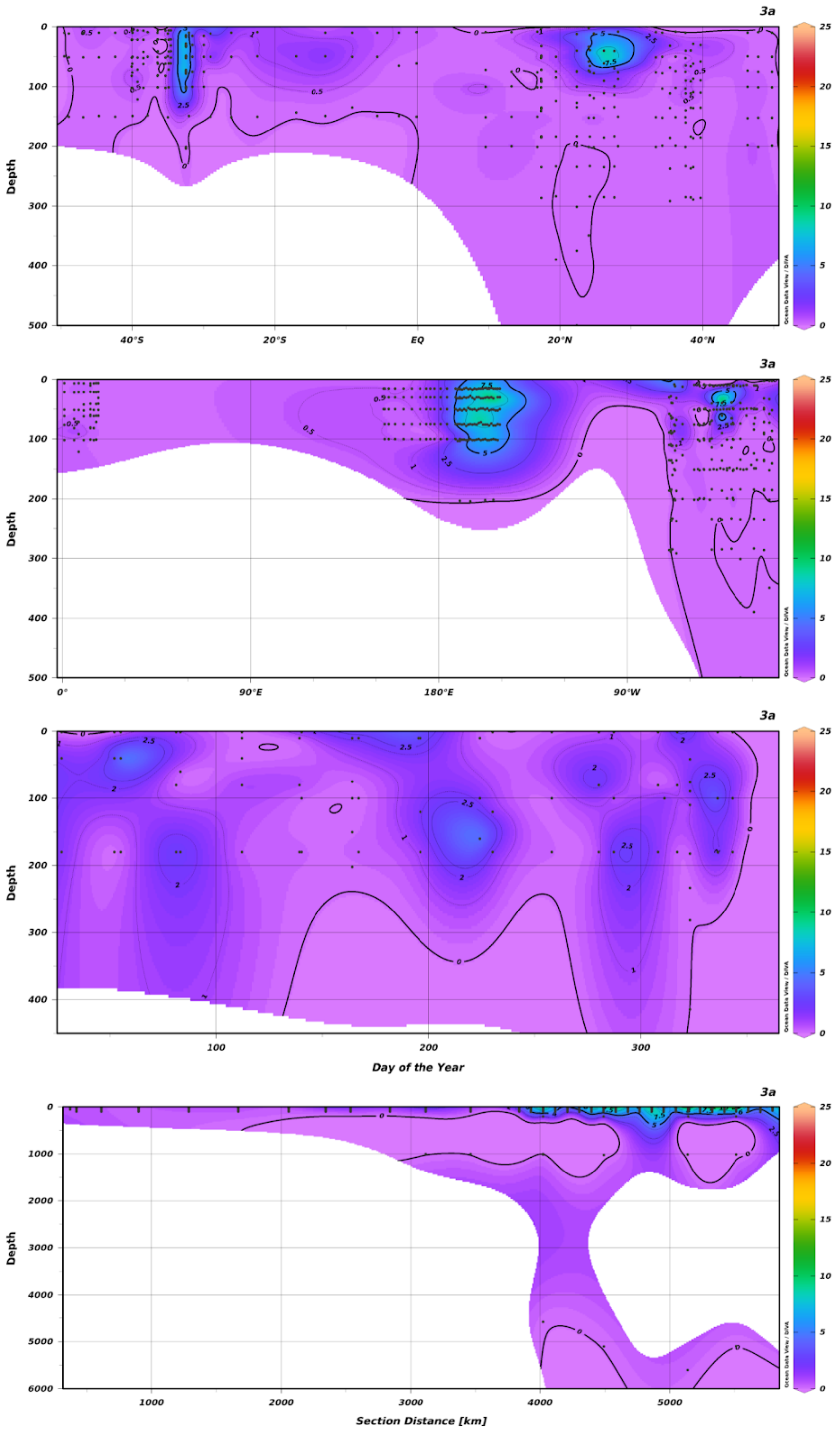


Figure 7.12 Mapping of SAR11 Clade IIIa reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade IIIa generally present in surface water above 100m and more abundance in the open ocean.

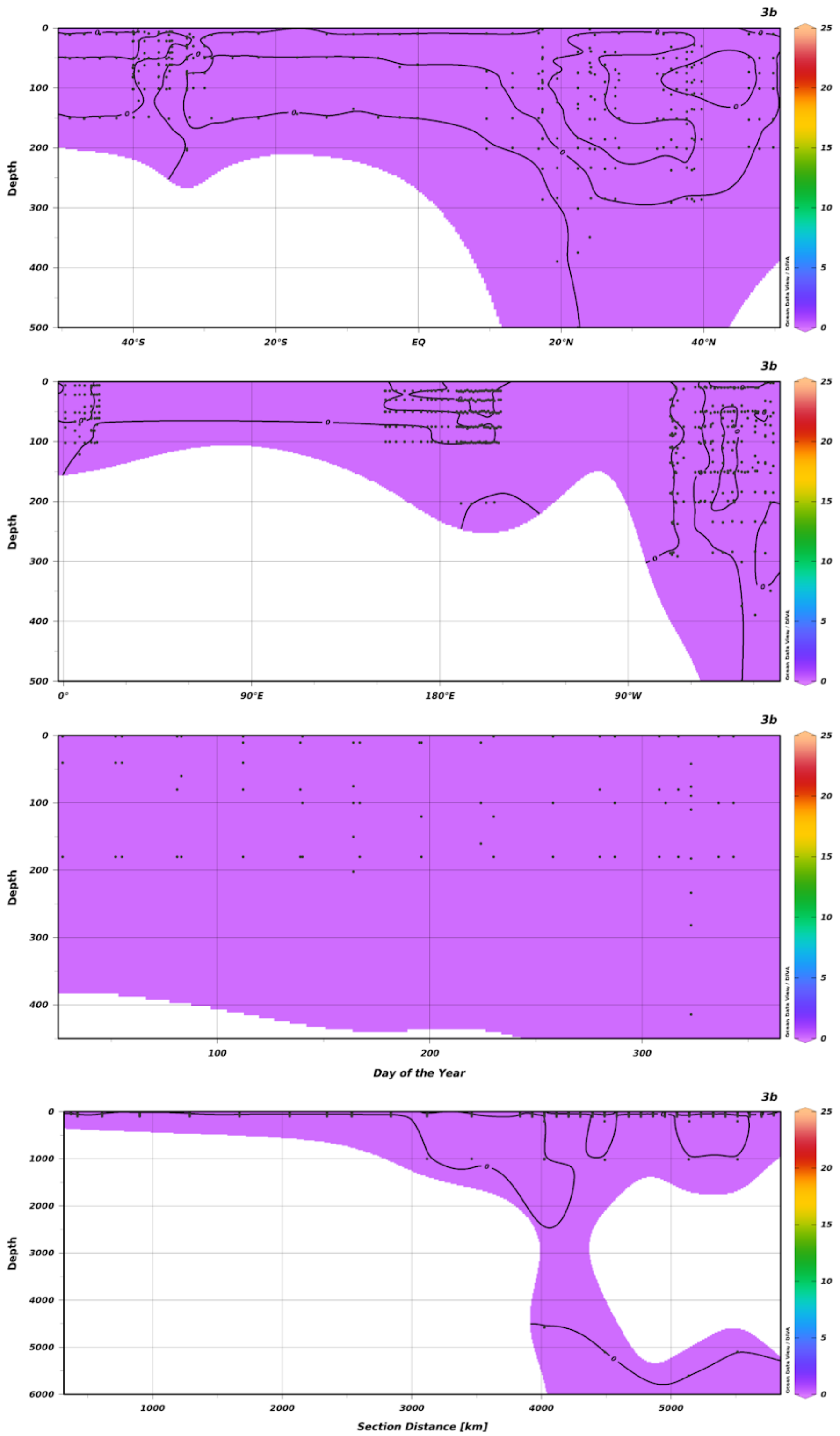


Figure 7.13 Mapping of SAR11 Clade IIIb reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade IIIb is characterised as freshwater and is not present in marine samples.

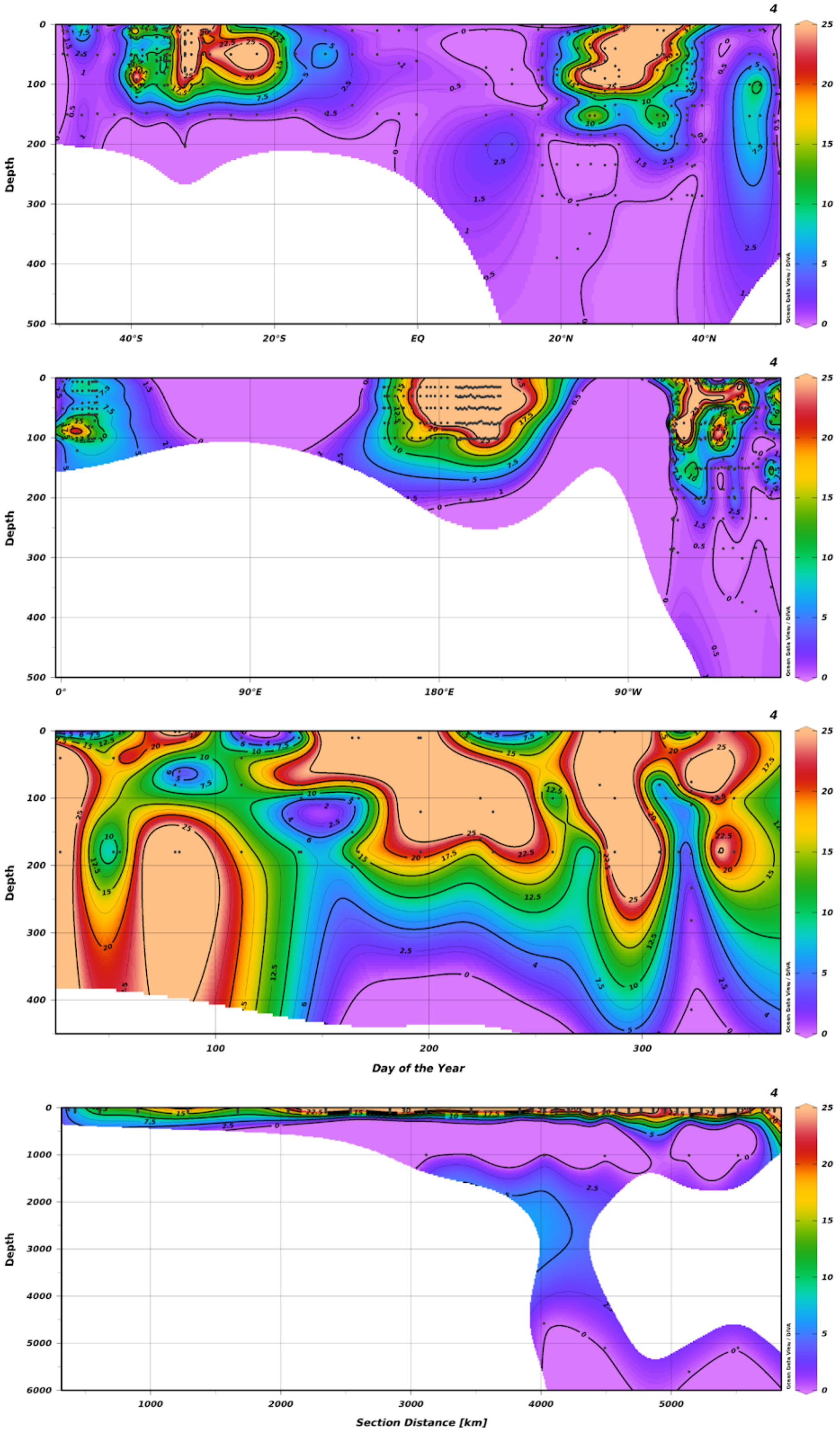


Figure 7.14 Mapping of SAR11 Clade IV reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade IV more present in open oceans generally above 100m in depth.

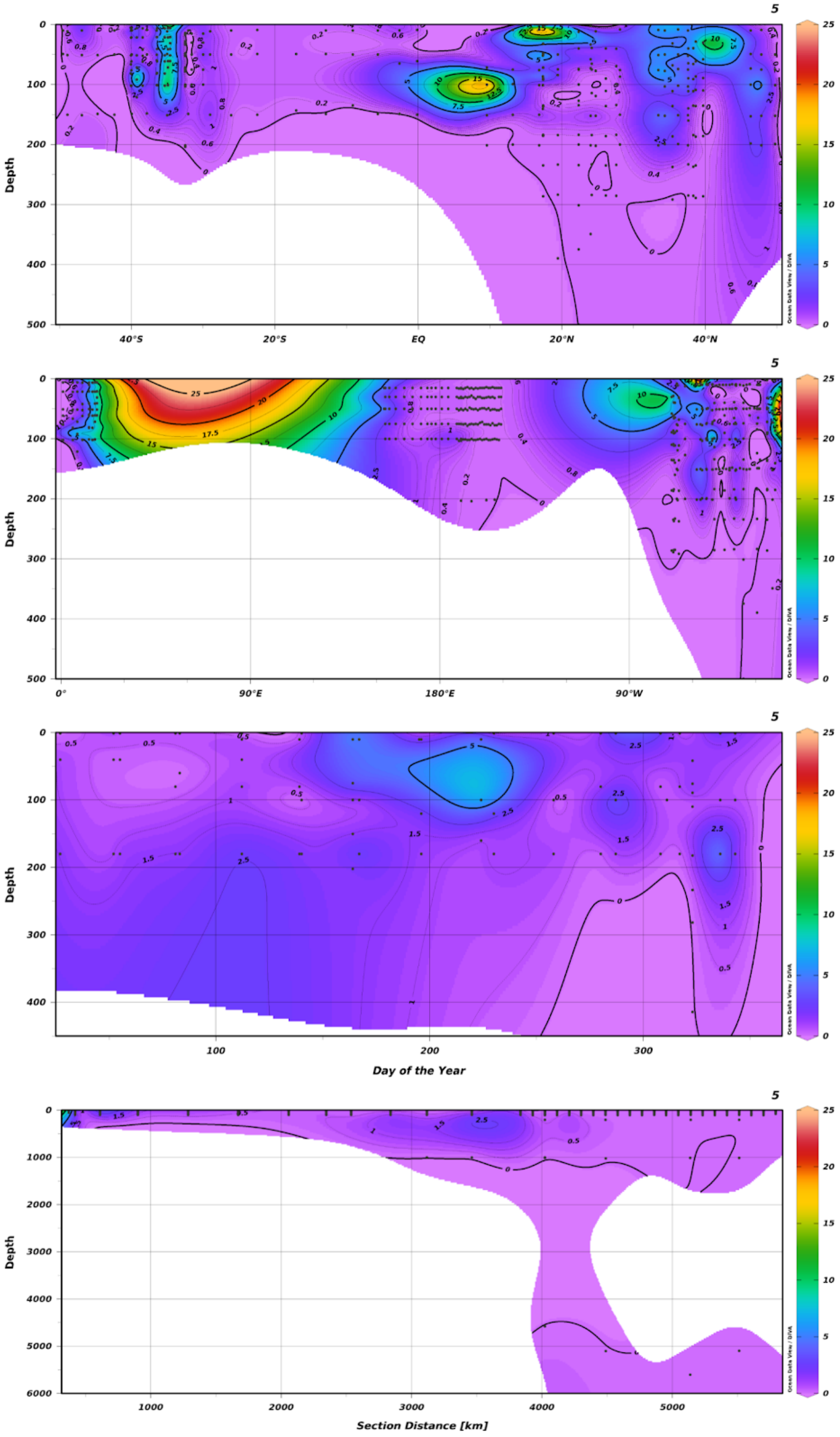


Figure 7.15 Mapping of SAR11 Clade V reference and SAGs established from WGS phylogenetic tree against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location. Clade V generally present above 200m.

7.1.3 Phage and Host ecological mapping

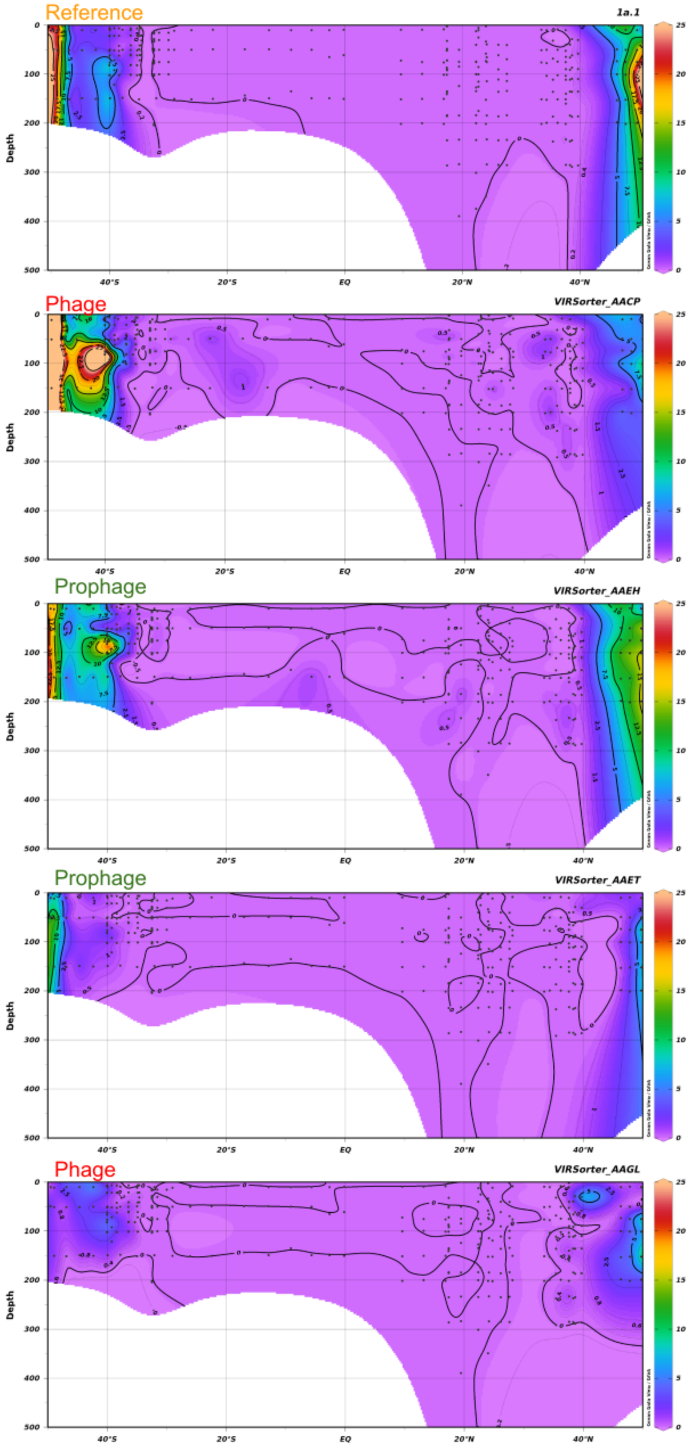


Figure 7.16 Mapping of SAR11 viral signatures produced from SAG data and host clade Ia.1 against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location.

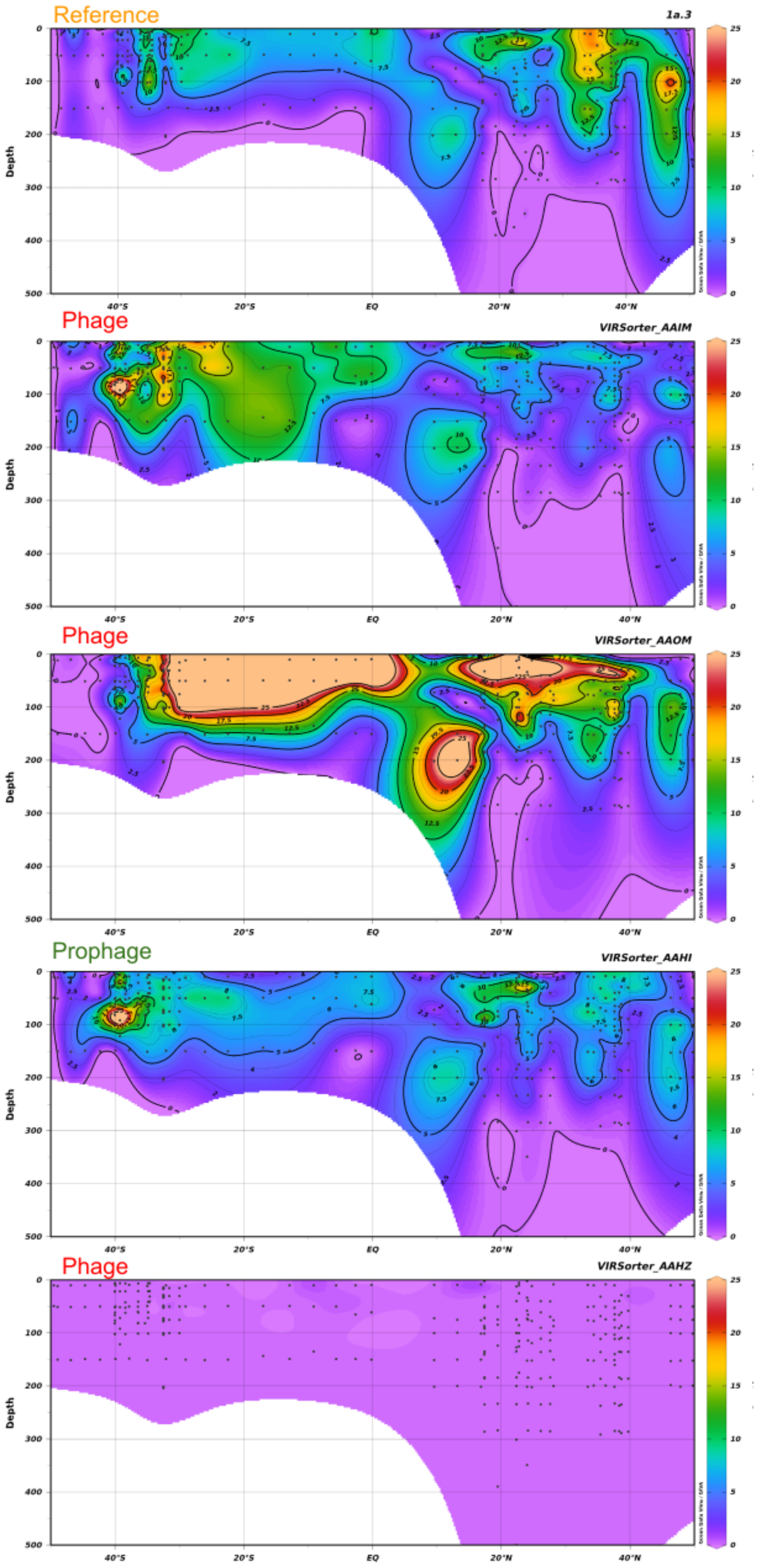
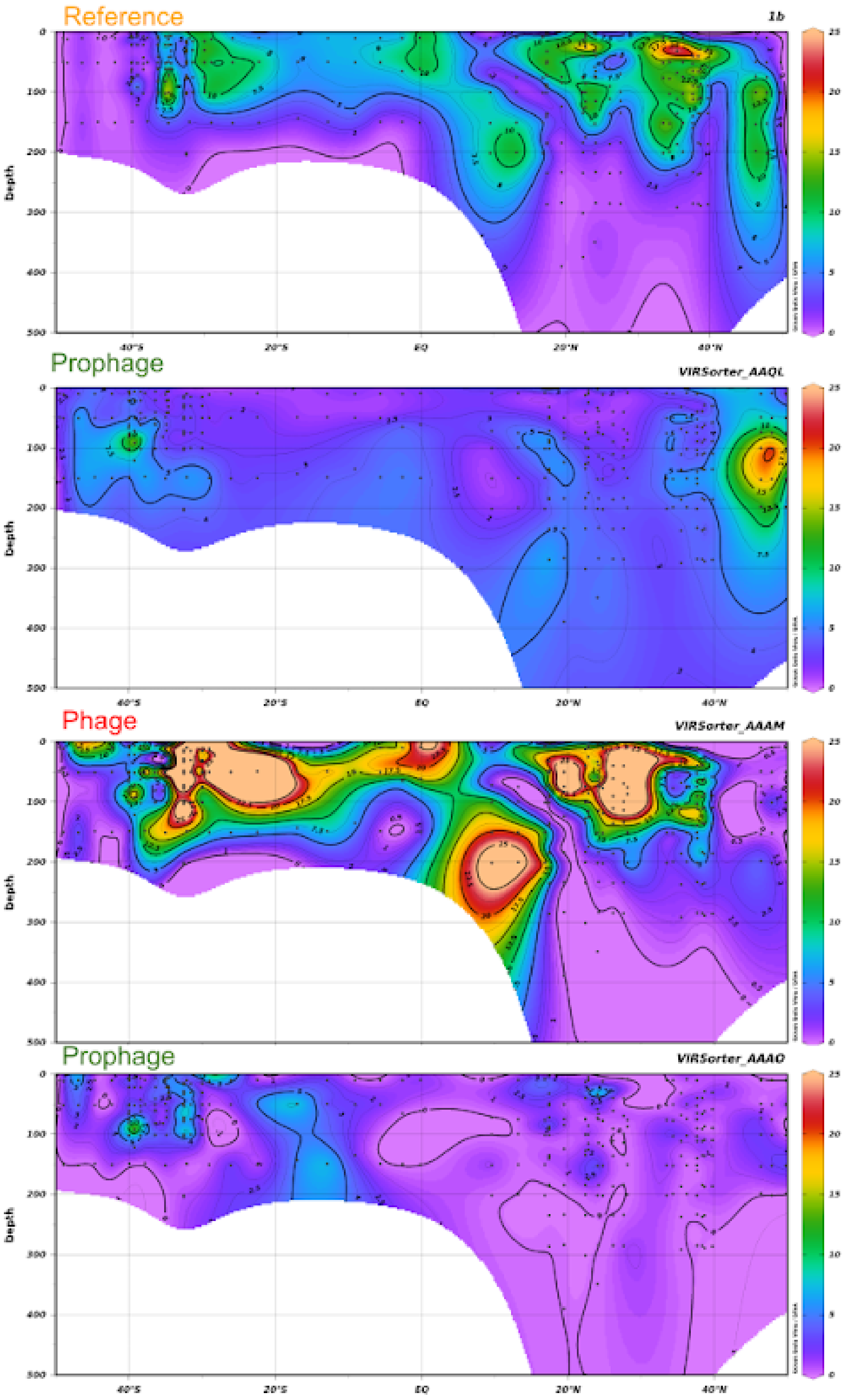


Figure 7.17 Mapping of SAR11 viral signatures produced from SAG data and host clade Ia.3 against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location.


Reference
1b
Depth
0
100
200
300
400
500
40°S
20°S
EQ
20°N
40°N
25
20
15
10
5
0
Prophage
VIRSorter_AAQL
Phage
VIRSorter_AAAM
Prophage
VIRSorter_AAAO

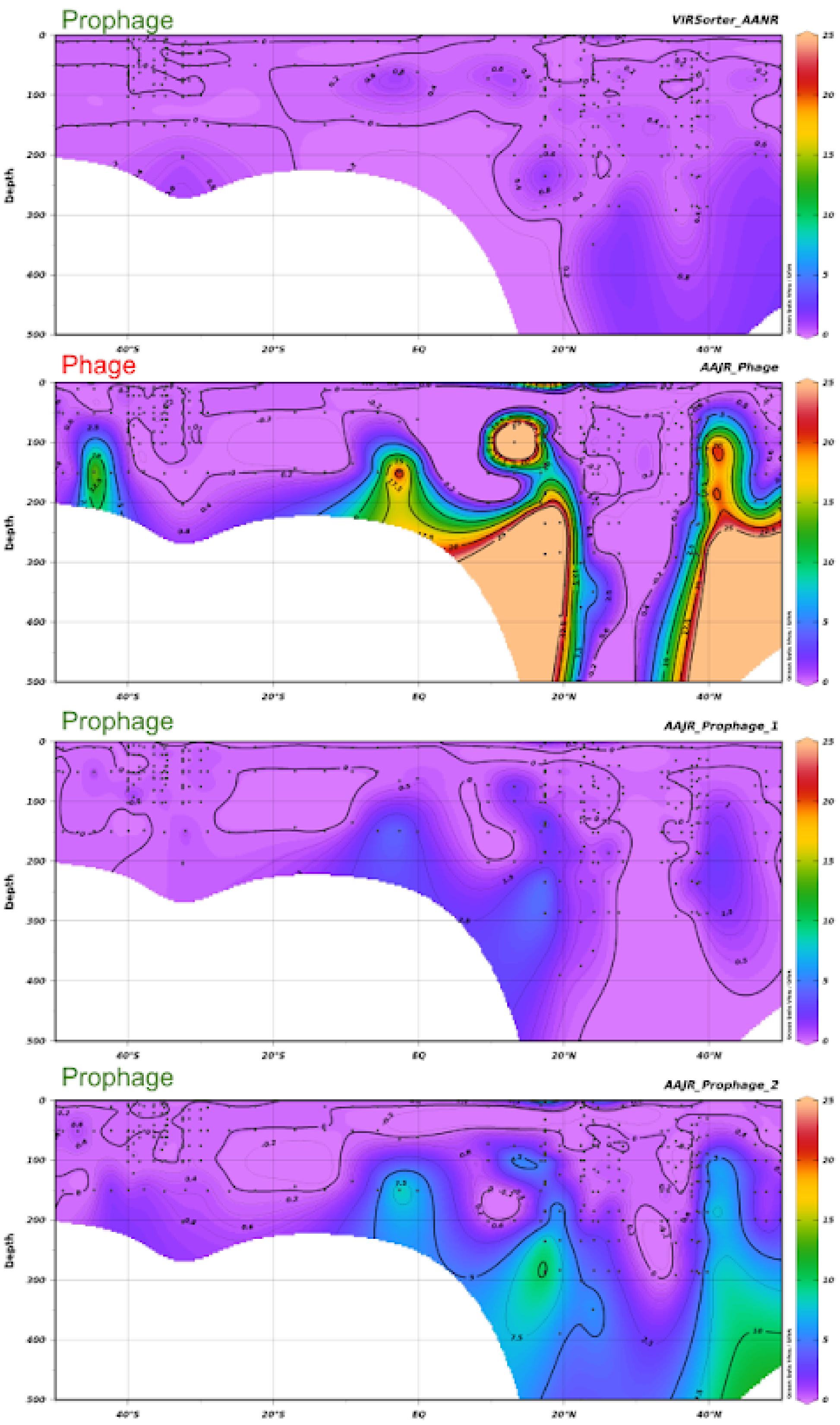


Figure 7.18 Mapping of SAR11 viral signatures produced from SAG data and host clade Ib against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location.

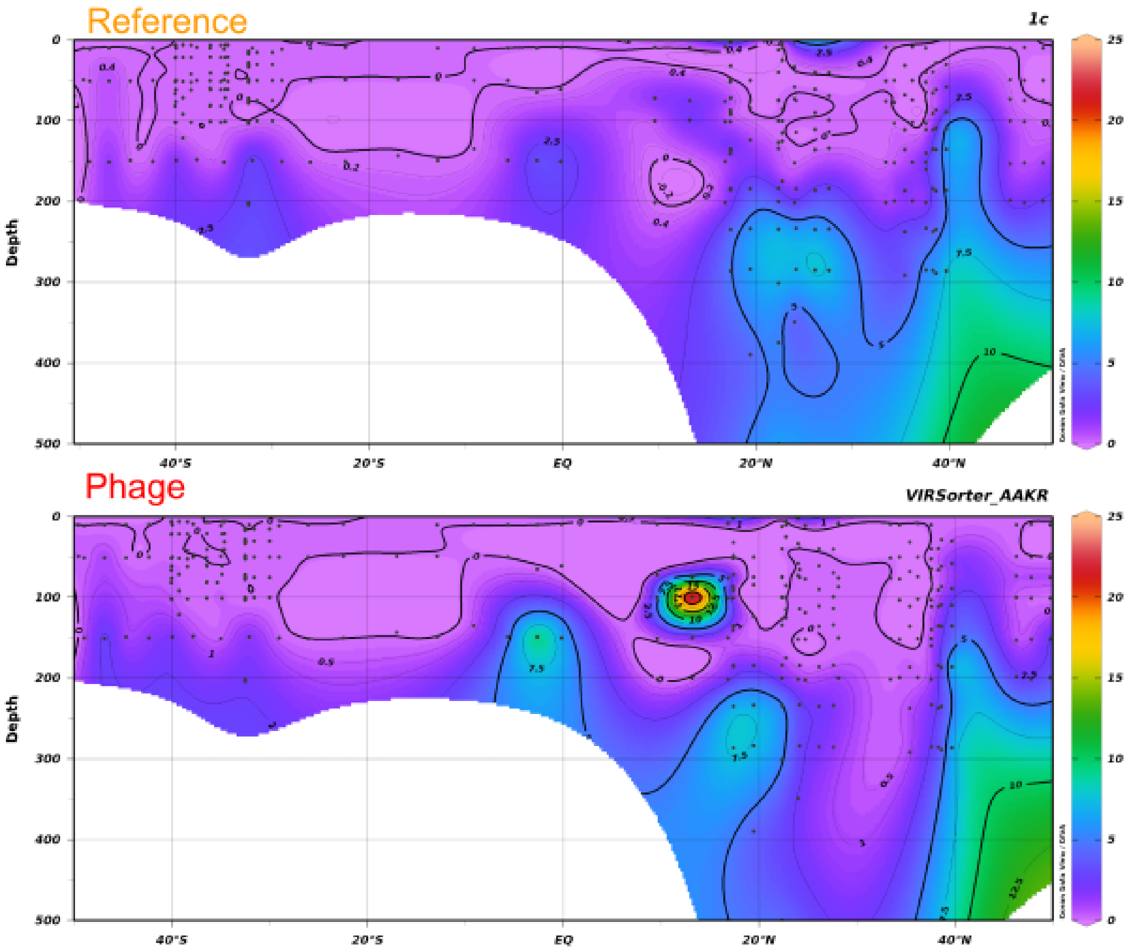


Figure 7.19 Mapping of SAR11 viral signatures produced from SAG data and host clade Ic against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location.

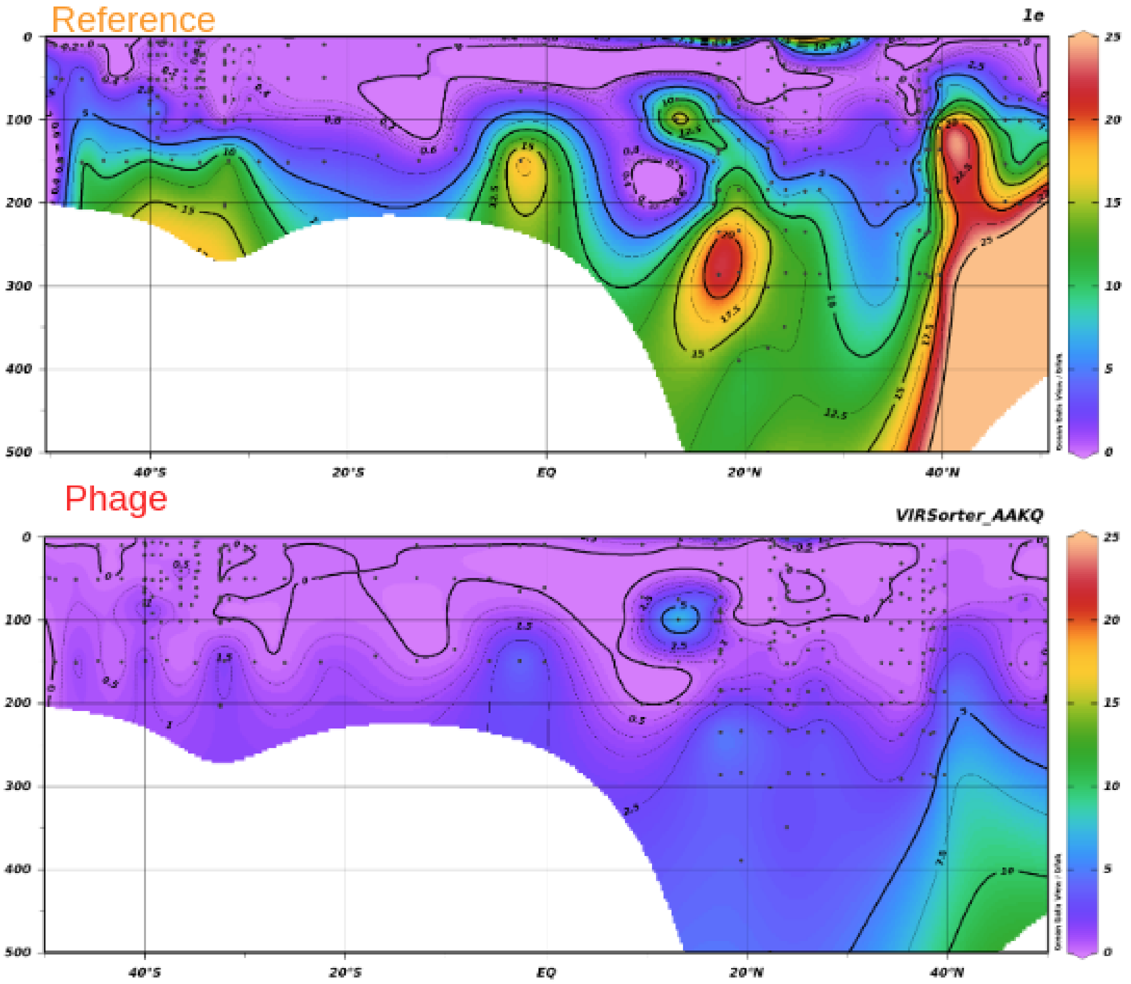


Figure 7.20 Mapping of SAR11 viral signatures produced from SAG data and host clade Ie against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location.

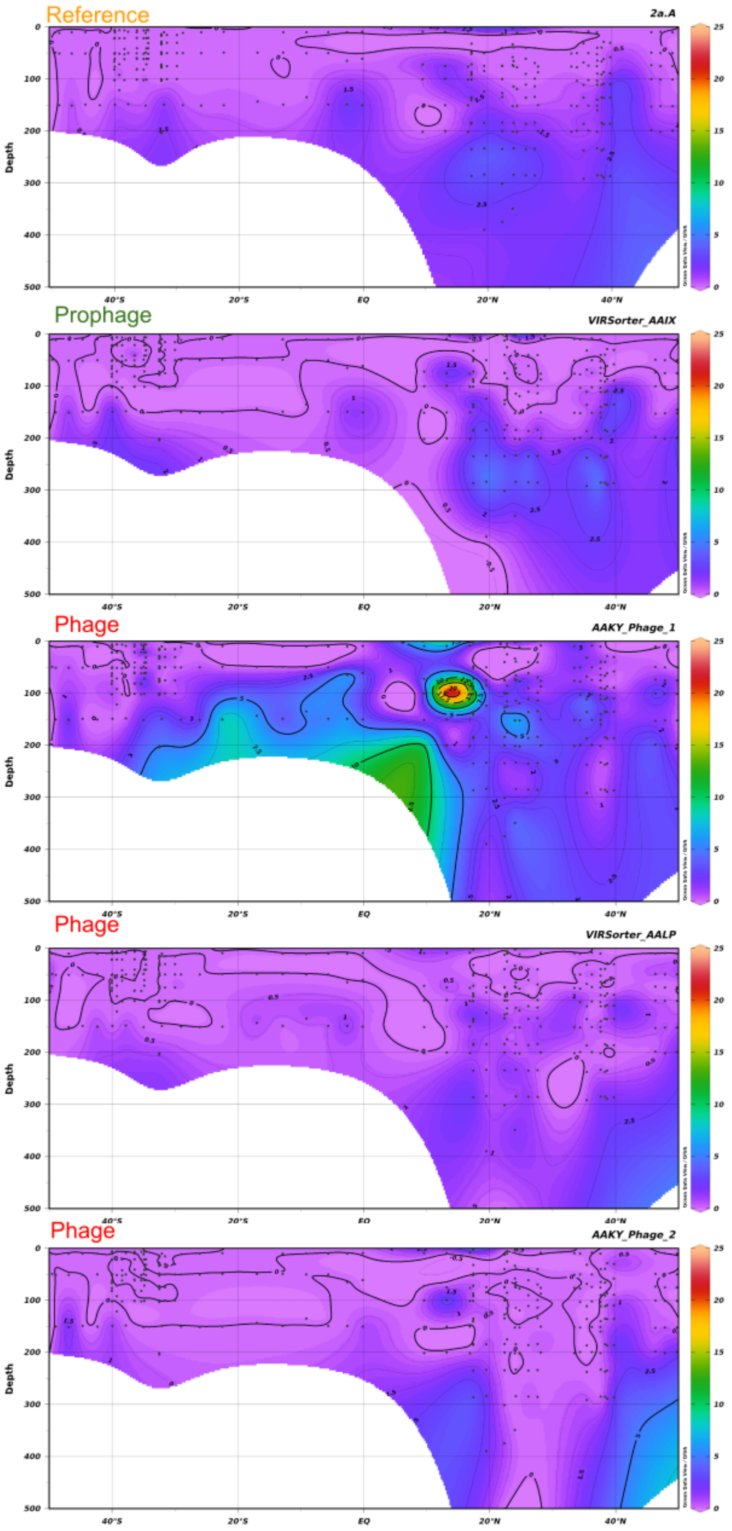


Figure 7.21 Mapping of SAR11 viral signatures produced from SAG data and host clade IIa.A against the (Biller et al. 2018) dataset. Mapping is measured as percentage genome coverage against each metagenome and its physical location.