6 Month Review Progress Report

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

Lead supervisor: Ben Temperton

Co-supervisor: Ivana Gudelj

# Abstract / Background

70 % of the world’s surface is covered by ocean; its impact on the global carbon cycle, climate change, and acid-base biochemistry remains crucial to our understanding of the natural world. Considering 90 % of the ocean’s biomass comprise of marine microorganisms [(Curtis A. Suttle 2007a)](https://paperpile.com/c/0ne0x7/HjVx) their impact on global systems remain largely understudied. Marine microorganisms are critical in the energy cycle and are the foundation for marine life. This is kept in check with viruses lysing at least 20 % of this biomass daily [(C. A. Suttle 1994)](https://paperpile.com/c/0ne0x7/V5fP). This biomass consists mainly of heterotrophic and eutrophic plankton [(Proctor and Fuhrman 1990)](https://paperpile.com/c/0ne0x7/5eZ5), where subsequent virioplankton levels seems to mirror increases and decreases in these plankton levels describing a predator-prey like relationship [(Parsons et al. 2012)](https://paperpile.com/c/0ne0x7/NSqu). Theses viruses also hold previously unexplored prokaryotic genetic diversity and its relationship with both physical and biological factors is key to understanding marine biome’s population dynamics [(Curtis A. Suttle 2007a)](https://paperpile.com/c/0ne0x7/HjVx). However, studying these microorganisms remains challenging with a small fraction being culturable for *in situ* experimentation. Alternative study methods include obtaining genomes via metagenomics studies and Single-cell Amplified Genomes (SAGs). Both of these methods have advantages and drawbacks, with marine metagenomes being highly complex to analyse and SAGs suffering from both low coverage and bias. This study aims to explore both approaches and provide improvements to aid analysis.

This report is split into 4 chapters and is analysed separately. They are as follows:

1. Metagenomic analysis of marine prokaryotic samples from the Bermuda Atlantic Time Series over different temporal and spatial gradients.
2. Genetic and phylogenetic characterisation of 451 SAR11 Single-cell Amplified Genomes (SAGs) based on current knowledge of taxonomic arrangements.
3. Genetic analysis of viral signature protein clusters within SAR11 SAGs
4. Identification and analysis of Hypervariable Regions within the SAR11 clade.

# Chapter 1 - Metagenomic analysis of Marine Prokaryotes

## Background

Metagenomes offer insight into an ecosystem's organic population and its abundance, acting as a “snapshot” in space and time. This snapshot allows for tracking of all organisms within a population, not just a single organism in single genome techniques. This allows for changes in population abundance to be tracked over time if conditions change. A particular example of this is within this report where populations of marine bacteria and integrated prophages can be tracked over a diel cycle.

## Aims and Objectives

1. Separation of high quality Metagenomically Assembled Genomes (MAGs) from metagenomes
2. Analysis of cellular abundance changes over temporal and spatial gradients
3. Analysis of temporal viral signature abundance and correlations with cellular abundance changes over temporal and spatial planes

## Developed Methods

### 1.1.1 Binning by Coverage

To allow for this mapping process, a combination of bioinformatics programs are used to create a Sequence Alignment Map (SAM) file. This indicates the location and quality of a mapped query against a reference genome. Examples of mapping algorithms include Bowtie2 [(Langmead and Salzberg 2012)](https://paperpile.com/c/0ne0x7/1k4Wv), BBtools [(Bushnell, Rood, and Singer 2017)](https://paperpile.com/c/0ne0x7/tru6) and Minimap2 [(Li 2018)](https://paperpile.com/c/0ne0x7/hiEEM). Here, Bowtie2 was used to map reads back to the assembled contigs of samples taken from the same depth, resulting in 6 SAM files for each assembly. SAM files were then filtered for reads with a 95 % identity using BamM [(Rabosky et al. 2014)](https://paperpile.com/c/0ne0x7/vOIQ). Reads mapping to a lower percentage identity were discarded. A 95 % identity was chosen as this is the general boundary used to delineate species.

### 1.1.2 Binning by binning software

From these SAM files, a variety of binning software exists for grouping contigs into categories of similar coverage. For this process, multiple binning software was used and compared to produce the best groupings. Here MetaBAT2 [(Kang et al. 2019)](https://paperpile.com/c/0ne0x7/YYNpD), BinSanity [(Graham, Heidelberg, and Tully 2017)](https://paperpile.com/c/0ne0x7/X0LTW), CONCOCT [(Alneberg et al. 2014)](https://paperpile.com/c/0ne0x7/LHcUn) and MaxBin [(Wu et al. 2014)](https://paperpile.com/c/0ne0x7/AMKmA) are used. All of them have their own proprietary algorithm but result in the same process of binning contigs with the usage of SAM files.

Resulting bins allow for the grouping of similarly covered contigs together, but provide no insight into the identity of the bin in question. To solve this, there exists a multitude of taxonomic classifiers which, with the use of a database, can assist in providing taxonomic identity of the contigs in question based on coding regions. Briefly, coding regions within contigs are extracted from each contig and compared against a database of known genes. Similarly mapping or identical genes can then be identified on the contig, providing an identity for each contig. E.g if a gene for HTCC1062, a common bacteria in the SAR11 clade is identical to a coding region on a contig, it provides strong evidence that such contig is part of the HTCC1062 genome. With larger contigs, multiple genes within said contig can be analysed and a consensus derived. Generally, if five or more genes are determined to be from the same species, the whole contig is deemed as part of that species.

### 1.1.3 Binning by taxonomy

There are many ways to determine the identity of a gene, with the most well known example being BLAST [(Altschul et al. 1990)](https://paperpile.com/c/0ne0x7/vyIZ3). Upon the framework of BLAST, there exists a multitude of programs which both speed up the process, for example Diamond [(Buchfink, Xie, and Huson 2015)](https://paperpile.com/c/0ne0x7/uNkJT) and allow for the automation of taxonomic classification of whole genome sequences. Centrifuge [(Kim et al. 2016)](https://paperpile.com/c/0ne0x7/mQyK7), Kaiju [(Menzel, Ng, and Krogh 2016)](https://paperpile.com/c/0ne0x7/Gztnq) and CAT/BAT [(von Meijenfeldt et al. 2019)](https://paperpile.com/c/0ne0x7/6VlWq) are specialist metagenomic taxonomic classifiers and can be used for additional binning of assemblies by taxonomy rather then coverage. These programs work by pulling out protein coding regions [(Hyatt et al. 2010)](https://paperpile.com/c/0ne0x7/9KsW) and assigning them the corresponding taxonomic identity based on a database. This can also be used in conjunction with coverage to describe identities for each contig and by extension bin, or used as another classifier to be used in conjunction with coverage.

### 1.1.4 Visualisation of metagenomes

A multitude of different algorithms are available to separation of a metagenome into its constituent Metagenomically Assembled Genomes (MAGs), but broadly follow two approaches. However, analysis of MAGs would benefit from an appropriate visualisation software. It would be beneficial to be able to look within bins made using coverage and to confirm with taxonomic classification of genes that bins are identifying along taxonomic lines. anvi’o [(Eren et al. 2015)](https://paperpile.com/c/0ne0x7/rmL7t) a visualisation platform for ‘omics data provides a great tool in its ability to visualise a whole metagenome and bins concisely.

### 1.1.5 Binning by k-mer counting

Additional ways of separating metagenomes into its constituent MAGs includes k-mer counting. Through a species genome, the sequence composition is likely to be the same throughout. Therefore tabulation of the k-mer count of each contig within a metagenome provides an additional mechanism for binning genomes. However, with many hundreds of k-mers, plotting each provides a difficult variable to visualise within a 2-D plot. Therefore dimension reduction techniques are used like the BH-tSNE to reduce the dimensions down into two, allowing for plotting on a 2-D scatter plot. It would be expected that organisms of a similar sequence composition would group together forming clusters within a scatter plot. Additional overlays with taxonomy previously described can also allow for confirmation of cluster formation and boundaries. Extraction of said clusters can therefore lead to MAGs. Metagenomes can contain many genomes and therefore clusters. Automation of cluster determination becomes a necessary method to aid in the extraction of MAGs. Algorithms like HDBSCAN [(McInnes, Healy, and Astels 2017)](https://paperpile.com/c/0ne0x7/MxXC) cluster data points on a scatter plot based on neighbour distance. This allows for the automation and visualisation of clusters to aid extraction of MAGs.

To trial the method of k-mer counting, samples where k-mer counted with a k value of 5. This was chosen as self-tested values above this have not been shown to beneficially increase binning ability and each increase in k value exponentially increases the number of k-mer values making this computationally intensive. Additionally, contigs of lengths smaller than 2500 bps were removed, as k-mer counting is more effective the longer the contig, and shorter contigs are more likely to lead to noise. 2500 is an arbitrary number, but reduced computational load drastically to more realistic timings and allowing for higher k numbers to be used.

Several python wrapper scripts written by [(Beaulaurier et al. 2019)](https://paperpile.com/c/0ne0x7/qFESH) first counts k-mers of reads, here repurposed for contigs, providing the k-mer distribution. K-mer counts were normally distributed due to differing contig sizes which would influence the number of k-mers, as there was a large distribution of contig sizes from 2500 bs to almost 200 kbs. Normalised k-mer counts were then visualised with UMAP [(McInnes, Healy, and Melville 2018)](https://paperpile.com/c/0ne0x7/iYXh) a dimension reduction algorithm that uses BH-tSNE plots to reduce the 512 k-mer distribution to 2 dimensions visualised on a 2D scatter plot. Lastly HDBSCAN is used to automate cluster determination. Exporting the clusters produced by HDBSCAN allows for binning of these clusters. Bins quality was assessed with anvi’o’s inbuilt bin assessment software based on established BUSCOs within each phylogenetic kingdom. CheckM could have been used to access bin quality but was deemed redundant as anvi’o’s bin assessment was adequate.

### 1.1.6 Assessing the quality of MAGs

Contigs that are grouped in bins and that have a consensus on taxonomy can then be extracted. These MAGs that are now extracted need to be assessed for quality. Although all of the related genetic material for an organism has been seperated into a bin, it is unclear how complete the genome is or how many organisms of the same or related strains are present. To check this, a tool called CheckM [(Parks et al. 2015)](https://paperpile.com/c/0ne0x7/dCN6D) can be used. CheckM uses conserved genes present in all bacteria or archaea and looks for them within a queried genome. The percentage presence of these genes can give an idea into the completeness of MAG without prior knowledge of the taxonomy of the bin. However, with the known taxonomy, CheckM allows for a more detailed analysis. Different taxonomic groups have different numbers of conserved genes, with lower taxonomic classifications having a higher amount of conserved genes within its grouping. With a higher number of conserved genes, a more detailed analysis of the sample can be done. This would give a more in depth analysis as more specific and higher numbers of genes can be used. For example, there exists 104 genes that are conserved amongst all bacteria [(Parks et al. 2015)](https://paperpile.com/c/0ne0x7/dCN6D) but potentially hundreds more genes more specific to that organism.

A genomes quality is represented by the percentage completeness and contamination/redundancy of these genes. This can be expressed as the number of conserved genes the MAG would be expected to have versus the amount it actually has. Percentage contamination is a key indicator if other strains or organisms are present within the MAG. If conserved genes are only expected to be present once within a genome, the presence of multiple copies would be a good indicator that genetic material from multiple strains or other species are present within the MAG. This would also be a good indicator of the quality of MAGs.

## Current Progress

### 1.2.1 Binning by specialised binning software

A variety of specialised binning software was used to attempt to bin the metagenome into its constituent MAGs.



BinSanity produced 17 bins with a completeness over 0 %. Although bins produced values of high completeness (>95 %), redundancy was also very high at values >100 % indicating a high contamination. This means bins would likely contain multiple species.



MetaBat2 produced 7 bins >0 % completeness with high amounts of contamination within its top completeness bins. Subsequent bins produced some success with bin\_7 producing 87.77 % completeness and 5.04 % redundancy. However, with only 7 bins, this does not produce many bins and therefore MAGs.



MaxBin produced 20 bins over >0 % completeness with a moderate quality. It produced bins with moderate contamination and completeness, with contamination being half the percentage of completeness. This does not lead to high quality MAGs but provides some success in some bins.

### 1.2.2 Visualisation of binning



Visualisation of binning is performed in anvi’o: a tool used to assist in manually binning metagenomes. There is a clear lack of consensus between binning software shown on the outside 3 tracks.

## 

Visualisation of bin\_017 from MaxBin showing a MAG. A lack of consensus between taxonomy and other binning software becomes apparent. The only consensus lies in contigs with a longer length, indicating longer contigs would assist dramatically in MAG construction.

### 1.2.3 k-mer counting





Both the default and high settings of n\_neighbours did reveal that there was little or no overall global structuring of the data. This would indicate that k-mer counts were unlikely to be drastically unique and clustering data based on global values would not be effective. Although there would be some clustering, the overall metagenome would not be binned sensibly. Therefore a lower n\_neighbours was chosen as localised difference between k-mer distribution would likely be a better factor for binning.



A lower n\_neighbours revealed clear cluster boundaries than with a higher n\_neighbours, although without any large clusters. Clusters identified by HDBSCAN were pulled out and resulting bins analysed for completeness and redundancy via anvi’o.



Bins with a completion value >0 % were displayed. Although completion statics were high, redundancy statics were also high indicating a highly contaminated bin with multiple species within. This was not deemed as a successful binning.

It therefore becomes apparent that an n\_neighbours below the default of 15 but higher than 2 would be a good setting to use to provide a good quality bins and therefore MAGs. n\_neighbours=5 was deems as the best setting as this provided clear clustering boundaries. This was analysed as above and results shown below.





Bins with a completion value not 0 were displayed. Although completion statics were high, redundancy statics were also high, indicating a highly contaminated bin with multiple species within a bin. This was not deemed as a successful binning.

## Discussion

Seperating a marine metagenome into its constituent MAGs is more challenging than previously thought. Existing methods of using coverage, taxonomy or k-mer frequencies all do not successfully separate a metagenome. No additional successful methods have been identified from existing literature. Existing methods may have failed due to poor assembly quality with low coverage. Longer contigs may help assembly of MAGs and derive a higher consensus between binning software. Long read technology should be explored as an alternative method for MAG assembly. Higher coverage of short reads may assist but, with indications of marine samples being highly similar, this may lead to reads being misassembled.

# Chapter 2 - SAR11 SAGs genomic analysis and phylogenetics

## Background

Single-cell Amplified Genomes (SAGs) are genomes isolated from single cell sorting techniques and are whole genome amplified and sequenced [(Eberwine et al. 2014)](https://paperpile.com/c/0ne0x7/aXf7). Organisms usually have only 1 or a few copies of their genomic DNA [(Alberts et al. 2014)](https://paperpile.com/c/0ne0x7/P8pI), which is not enough to sequence. Amplification of this genomic DNA allows for pure samples of an organism to be sequenced without the need for culturing [(Lasken 2007)](https://paperpile.com/c/0ne0x7/cUdt). This is particularly useful for organisms that are difficult to grow or are slow growing. 451 SAR11 SAGs are sequenced within this study. SAR11 is an aquatic bacteria and the most abundant microbe in the world’s oceans [(S. J. Giovannoni 2017)](https://paperpile.com/c/0ne0x7/MZFO). Culturing this slow growing organism is difficult and SAGs provide an alternative to the genomic study of these ubiquitous organisms.

## Aims and Objectives

1. Assemble SAR11 SAGs
2. Identify and confirm phylogenetic identity of all SAR11 SAGs
3. Identify new phylogenetic groups

## Method

### 2.1.1 SPAdes Assembly

Assembly with SPAdes was performed using version 3.13 in single cell mode. k-mers of intervals of 10 to 99 were used in conjunction with mismatch error correction mode. Resulting assemblies were assessed for quality via CheckM [(Parks et al. 2015)](https://paperpile.com/c/0ne0x7/dCN6D). Percentage completeness was used as the main metric to define quality of each assembly as well as percentage contamination, GC percentage deviation and N50. Several methods were used to improve the assemblies and the above method was summarised as the most complete.

### 2.1.2 Assembly analysis

Assemblies were put into Bandage [(Wick et al. 2015)](https://paperpile.com/c/0ne0x7/uiKR), a program for visualisation of assembly graphs to review the assembly quality along with CheckM.

### 2.1.3 Phylogenetic analysis

Phylogenetics was used to first describe the relationship of each SAG to each other. This would allow for categorisation of each SAG into the established 5 clades of SAR11. Further work on each SAG would then be relatable to its known ecological niche. Whole Genome Sequence (WGS) Multiple Sequence Alignment (MSA) files were generated via GTDB-Tk [(Parks et al. 2018)](https://paperpile.com/c/0ne0x7/b3gn) a phylogenetic algorithm using genes conserved across all bacteria to construct a MSA. An MSA is a genetic representation of the areas of similarity between different genomes to allow for comparisons. Trees were produced using IQ-Tree [(Nguyen et al. 2015)](https://paperpile.com/c/0ne0x7/RNYU) and visualised in iTOL [(Letunic and Bork 2007)](https://paperpile.com/c/0ne0x7/nhmI), a web based phylogenetic tree viewer.



Visualisation of a SAR11 SAG assembly shows sections have constructed well into contigs but not into complete contiguous assemblies. This would indicate some fraction of the genome would not be present, indicative of SAG assemblies with amplification bias [(Clingenpeel et al. 2015)](https://paperpile.com/c/0ne0x7/qF7y).



Currently the SAR11 clade is split into 5 existing groups with multiple sub-clades [(Thrash et al. 2014)](https://paperpile.com/c/0ne0x7/zx1N). Ingroup SAR11 references indicative of each of these groups can be used to infer phylogeny. References were taken from existing SAR11 clade groups [(Grote et al. 2012)](https://paperpile.com/c/0ne0x7/jGeV6) and included in this analysis to anchor branches within existing phylogenetic trees. Branches with low bootstrapping values can be collapsed to represent areas of ambiguous phylogeny.



From the WGS phylogenetic tree, we can be fairly confident of its structure due to the usage of a high bootstrapping numbers. However, without reference genomes for clade 4, it becomes difficult to infer if that node belong to that clade. However, based on other constructed SAR11 phylogenetic trees, it is possible to infer which splits are likely to lead to nodes of an unknown identity. In this case, clade 4 is highlighted in brown.

Areas of white indicate areas of the phylogenetic tree that cannot be inferred 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, a clade similar to clade 2 with 42 members (2c - Teal) and a clade similar to clade 1 with 32 members (1d - Olive).

Alternatives towards deducing clade level phylogenetic boundaries include Average Nucleotide Identity (ANI). ANI is an indicator of the similarity of two species by comparing nucleotides within coding regions. Species that are similar are expected to have similar compositions of nucleotides within their coding regions, along with similar gene lengths. An all vs all ANI can be calculated against each SAG to deduce these boundaries. FastANI [(Jain et al. 2018)](https://paperpile.com/c/0ne0x7/23EOO) is a bioinformatics tools that calculates the ANI of protein coding regions using k-mer counting. An all vs all ANI for each SAG is calculated and plotted into an ordered heatmap to reveal clusters of similarly group SAGs that are related to each. FastANI used default settings. A heatmap allows for a gradient of colour, showing clusters that are less or more closely related on a discontinuous scale.



Clustering of an ANIs within a heatmap are a good indicator if boundaries exist within a clade. Within an ANI heatmap, clusters should be identified with reference genome within each cluster providing identity of each grouping. This allows for identification of sample groupings and can reinforce existing categorisations from phylogenetic trees. However, with a discontinuous spectrum of ANI’s, this makes it difficult to see clade boundaries as these are blurred. Clade boundaries would be guesses at best. Additionally, sorting each ANI value based on reference did not lead to clear clustering. Instead clade 1a, the most common SAR11 within the 451 SAR11 SAGs was scattered throughout the plot leading to ambiguous identities to clusters. Cluster identity was therefore difficult to determine.



Clearer clustering can be seen with the usage of k-mer counting. 6 clusters appear quite 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.

To confirm this, bacteria from across the tree of life were taken and their genomes split into 10 kbp fragments. Samples of well characterised SAR11 species (HTCC7211, HTCC1062) were also included. Resulting contigs were then k-mer counted and dimension reduced via UMAP. The resulting BH-tSNE values were plotted into R using ggplot and visualised.





As can be seen in Fig. 2.6, some species cluster closely where others like SAR11 do not and have multiple clusters. This shows that different parts of the SAR11 genome have drastically different k-mer frequencies. Additionally this would also suggest that k-mer counting may not be a good metric for separation of SAR11 species from metagenomes, inline with the findings from the previous chapter.

## Discussion

SAR11 SAGs were constructed with the best possible assemblies deduced via assembly statistics like N50. Phylogenetic groups of SAR11 SAGs were produced using WGS and identity to existing SAR11 clades established along with possible new groups. However, without references from other SAR11 clades, it becomes unreliable to confirm the identity of sub-clades. A variety of different methods like ANI and k-mer counting were used to cluster clades to confirm phylogenetic identities but were not deemed to be more successful or confirm findings.

# Chapter 3 - SAR11 Viral signature

## Background

It has long been established that viruses have a large impact on ecosystems [(Fuhrman 1999)](https://paperpile.com/c/0ne0x7/hpfb). They are key factors in regulating bacterial and eukaryotic microorganism populations [(Curtis A. Suttle 2007b; Wommack and Colwell 2000)](https://paperpile.com/c/0ne0x7/ABUM+JhzB). They can infect both multicellular and single cell organisms often with detrimental effects to the host [(Rohwer and Thurber 2009; Weinbauer 2004)](https://paperpile.com/c/0ne0x7/CEvh+qrUL). Within single-cell organisms, they are responsible for the turnover of over 20 % of marine bacteria biomass daily [(Fuhrman 1999)](https://paperpile.com/c/0ne0x7/hpfb) and contribute enormously to the differentiation of the number of genes an organism may have through integration of temperate phages [(Breitbart et al. 2007; Sharon et al. 2011; Thompson et al. 2011; Anantharaman et al. 2014)](https://paperpile.com/c/0ne0x7/2ofn+Yaum+LGlQ+pTMH). Any study looking at single cell populations must consider the population of viruses within the same ecosystem.

Generally, viruses exist in two stages, integrated within genomes of their host or as free virions. These are referred to as temperate and non-temperate phages respectively. Many viruses can express both types of lifestyles, namely lysogenic and lytic lifestyles, or express only one type. Temperate phages are of interest due to their ability to integrate themselves within their hosts. The human genome itself consists of 8 % retrovirus like elements [(Lander et al. 2001)](https://paperpile.com/c/0ne0x7/eReu) and some bacterial genomes have up to 20 % of their genomes from viral origins [(Casjens et al. 2000)](https://paperpile.com/c/0ne0x7/Z0cV). This allows for additional genetic variation to be introduced into the host genome, allowing for additional genes that may confer additional advantages or be detrimental to the host [(Lindell et al. 2005; Roux et al. 2014; Sharon et al. 2009; Hurwitz, Brum, and Sullivan 2015)](https://paperpile.com/c/0ne0x7/JKRR+95Dk+V43p+ualz). It has also been hypothesized this is how transfer of antimicrobial resistant or pathogenic genes [(Waldor and Friedman 2005)](https://paperpile.com/c/0ne0x7/B9nW) are passed to other bacteria . Studying these genes may offer insight into the ability for these viruses to add advantages to the cell, increasing its fitness. Prophages can stop becoming viable 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)](https://paperpile.com/c/0ne0x7/anPP).

There are two mechanisms for obtaining prophage sequences: experimentally or computationally. Both rely on obtaining an organism with a prophage sequence. Within laboratory conditions, exposure to UV or other environmental stressors can trigger the prophage to start replicating and excise itself from the host [(Nanda, Thormann, and Frunzke 2015)](https://paperpile.com/c/0ne0x7/Hh4Z). This relies on the prophage being viable and receptive to excising itself in the presence of environmental triggers. Bioinformatically, organisms can be sequenced and viral sequences identified [(Roux et al. 2015; Zhou et al. 2011)](https://paperpile.com/c/0ne0x7/f6QP+o8X7). However, this is an estimation of viral signatures based on databases and viral hallmark characteristics.

Viral signatures can be broadly divided into two forms, a viral sequence and a prophage. Generally a viral sequence is an isolated viral sequence usually located on a single contig. This represents viruses captured during the DNA sequencing process and may be environmental DNA or free flowing DNA within an organism. Conversely, prophages are viral sequences that have been integrated into the host genome, becoming part of its transcriptome.

Obtaining viral signatures from existing genetic sequences usually relies on detection of genes already established to be viral. Some basic “viral-like” properties are identified such as the proportion of “AT” and “GC” nucleotides. A skew in the frequencies of these nucleotides can be an indicator of foreign genes. Additionally, attachment sites and disrupted or shorter genes can indicate the start and end of a prophage sequence from which the virus has integrated. Phages may have a duplicate gene for the location into which it integrates, allowing for DNA repair machinery to integrate [(Roux et al. 2015; Arndt et al. 2017; Bose and Barber 2006; Lima-Mendez et al. 2008; Akhter, Aziz, and Edwards 2012; Fouts 2006)](https://paperpile.com/c/0ne0x7/f6QP+acJT+gP34+Opoh+9jgQ+E9xN).

## Aim and Objectives

1. Identify viral signatures within SAR11 SAGs
2. Identify viral types using protein clustering
3. Identify genetic content of viral signatures
4. Comparison of viral signatures to existing SAR11 phages

## Method

### Extraction of viral signatures by VirSorter

VirSorter uses MetaGeneAnnotator [(Noguchi, Taniguchi, and Itoh 2008)](https://paperpile.com/c/0ne0x7/0xHy) and hmmsearch [(Finn, Clements, and Eddy 2011)](https://paperpile.com/c/0ne0x7/zLVa) to look for PFAM domains and viral domains using hidden markov models. Using a sliding window, it looks for viral hallmark genes and areas enriched for viral domains. Lower numbers of characterised PFAM domains and uncharacterised genes are also indicative of viral sequences. Shorter genes encoded on the same strand are also likely of an insertion event of a viral prophage. Circular sequences are also tend to be of phage origin. Based on these variables, a category is determined for the confidence of a viral sequence. If 80 % of a contig is deemed viral, then the whole contig is deemed viral, otherwise it is classified as a prophage. Categories are split into six: one to three are viral sequences and four to six are prophages. The lower the category number for each represent a high confidence the sequences is viral. For example, categories 1 and 4 are highly probable to be viral and 3 and 6 the least likely.

All 451 SAGs were analysed for the presence of viral sequences. Only categories 1, 2, 4 and 5 were retained as these represented a high confidence that sequences produced were viral. Additionally, sequence lengths below 10 kbps were removed as VirSorter has been documented to be less accurate at those ranges. Resulting phage sequences were annotated and displayed.

### Phylogeny of Viral Protein Clusters

Common genes exist between prokaryotic and eukaryotic organisms like 16S rRNA subunits. However, viral phylogeny is difficult to conclusively characterise as these do not share any specific genes in common. An alternative method looks at proteins within viral sequences and compares these to existing viruses. The presence and absence of these functional proteins may then indicate a typing of viruses. vcontact2 [(Bolduc et al. 2017; Jang et al. 2019)](https://paperpile.com/c/0ne0x7/JIV1+K2Mz) is an algorithm which allows for gene sharing networks to produce genome-based viral taxonomy.



## Discussion

21 viral signatures were identified by VirSorter and clustered based on their protein content to reveal their viral type. Protein clusters were visualised in Cytoscape. Resulting protein clusters will be used to provide identities to 21 isolated viral signatures. Additionally, viruses will be analysed for their protein content and compared to existing SAR11 phages. Based on phage host ecological niches, trends in viral taxonomic groups or gene content may be drawn.

# Chapter 4 - Hypervariable Regions within the SAR11 Clade

## Background

### Ecological models

There are a variety of ecological models that explore the predator prey dynamic. Kill the Winner (KtW) hypothesis states that in the presence of a high density of prey, predator density will increase to match prey density and eventually bring predatory - prey density into equilibrium [(Avrani, Schwartz, and Lindell 2012; Winter et al. 2010)](https://paperpile.com/c/0ne0x7/82Uf+gHGQ). An example of this is the regulation of gut microbiomes by phages [(Letarov and Kulikov 2009)](https://paperpile.com/c/0ne0x7/1SFu). This indicates prey and predatory numbers are proportionally linked and will change based on prey density. This is a common strategy in fast growing and large communities of bacteria.

The Red Queen Hypothesis states that an organism must constantly evolve to survive against constantly evolving opposing organisms [(Brockhurst et al. 2014; Van Valen 1973)](https://paperpile.com/c/0ne0x7/TaCe+eWxT). An example of this is the diversity of bacteria populations maintained by viral predation [(Rodriguez-Valera et al. 2009)](https://paperpile.com/c/0ne0x7/CyMY). These are described as defensive specialists and a large proportion of resources is spent on adapting itself rather than replication [(Våge et al. 2014)](https://paperpile.com/c/0ne0x7/cTLf). Both models work in conjunction with each other and seldom exist in a vacuum without each other. SAR11 is a ubiquitous organism but with a slow replication rate [(S. J. Giovannoni 2017)](https://paperpile.com/c/0ne0x7/MZFO). SAR11 has a streamlined genome and undergoes extremely high rates of homologous recombination and is able to share its gene content [(Sun and Luo 2018; Vergin et al. 2007)](https://paperpile.com/c/0ne0x7/biZ2+KN2M). 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 [(S. Giovannoni, Temperton, and Zhao 2013)](https://paperpile.com/c/0ne0x7/8nyP).

The King of the Mountain (KotM) hypothesis states that SAR11 is successful because it shares its genes that contribute to its success at a rate higher than phages can evolve to predate on it. Its high abundance allows for a large diversity of genes to co-evolve and its high homologous recombination rates allow for successful genes to be passed quickly within a population. To instigate this, it has been hypothesised that SAR11 contains Hypervaraible regions (HVRs) [(Grote et al. 2012)](https://paperpile.com/c/0ne0x7/jGeV6): areas within a genome that have higher than normal evolution rates in comparison to the entire genome as a whole.

Although it may seem counterintuitive for a streamlined genome to include HVRs, it is thought to be an advantage in SAR11 as significant parts of its genome are dedicated to nutrient uptake. Regions of HVRs are under higher evolutionary pressure and any genes conferring additional fitness would allow for these more successful genes to confer an immediate advantage. If it has high recombination rates with other SAR11, these genes would allow it to outcompete other microbes and reinforce its dominance as the most widespread surface water bacteria.

In addition to HVRs containing nutrient uptake genes, genes conferring resistance to phage infection would explain its high abundance, thereby resisting the KtW model to a degree. Genes conferring an advantage to phage avoidance or resistance to attachment would help support evidence for its dominance.



It has been shown that SAR11 has a HVR amongst all its 1a and 5th clade [(Grote et al. 2012)](https://paperpile.com/c/0ne0x7/jGeV6). However, there is no evidence for HVRs being present in all SAR11 species. Whole genome sequences of SAR11s are rare due to its slow growth rate and difficulty to culture. In SAG techniques, the confirmation that HVRs are present in all SAR11 species would confirm the KotH hypothesis. Areas of the genome that do not have any similarity to other SAR11s would be indicative of HVRs. Although it would be logical to do an all vs all mapping of all the SAGs, poor quality of SAG assemblies and uneven coverage may give false positives for regions of high variability. Instead, metagenomic marine samples can be used as an alternative to map against the SAGs. Theoretically, they would also contain SAR11 genomes and mapping would still occur. Metagenomic samples come from the Biller dataset [(Biller et al. 2018)](https://paperpile.com/c/0ne0x7/dQIO), termed subsequently as “Biller”, comprising of five terabases of metagenomic data from 610 sampling sites across a range of depths and times. This provided the added advantage of looking at spatial and temporal distribution of mapping. Each Biller dataset is a representation of the metagenomic content at that location and time. If coverage only occurs within certain Biller datasets, this would represent genetic content only present within those locations or time. This would indicate specific habitats where certain sub-clades of SAGs are present and allow for pinpointing where certain phylogenetic groups of SAR11 SAGs are.

## Aim and Objectives

1. Identify HVRs within the SAR11 clade
2. Identity flanking genes of HVRs
3. Identity if HVRs are upregulated
4. Identify genetic content of HVR

## Method

WGS for SAR11 clades 1a, 1b, 1c, 2, 3a, 3b and 5 are publically available. These are referred to as reference organisms and used as representatives of the genomes of each of these clades. To identify HVRs, organisms of the same group are mapped against reads from the Biller metagenome set. The resulting coverage of each reference genome was plotted as a filled scatter plot to indicate areas of the genome were no mapping occurred. Large areas of continuous low coverage are then identified as HVRs.

# 

Areas of low coverage were extracted and its protein coding regions deduced. Regions indicating the start and end of a HVR are explored looking for the presence of conserved genes. This would indicate proportions of the genome that are common between all species of SAR11. The resulting regions would be areas that could undergo inter- and intra-species homologous recombination. This would mean regions within this conserved region could be passed between species of SAR11 contain these flanking genes.





## Discussion

HVR2 has been identified within clades 1a, 1b and 3a. It has also been confirmed that HVR2 within all tested clades belongs to the same region within each species of SAR11. This indicates that this HVR is common within all species of tested SAR11. Other clades like 5 and 3b have shown little or no mapping, complicating deduction of coverage and therefore the HVR. However, HVR2 has been shown between the 23S and 5S rRNA proteins in all tested species. As these RNA subunits are conserved amongst all prokaryotes, these regions can be pulled out and its contents analysed.

Next steps include the identification of flanking genes within HVR1 and 3 as well as the protein coding region for each. Presence of genes related to phage defence and nutrient uptake would give evidence that these are HVRs and support the KotM hypothesis.

# Attended Trainings and Conferences

PRIMER-e Version 7 Training Course 29 October - 2 November 2018

Learning and Teaching Higher Education Stage 1 11 January 2019

BioCon Exeter Poster 10 May 2019

CLESCon Exeter Autumn 2019

Genome Science 2019 3 - 5 September 2019

# Timeline



# References

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/9jgQ)[10.1093/nar/gks406](http://dx.doi.org/10.1093/nar/gks406)[.](http://paperpile.com/b/0ne0x7/9jgQ)

[Alberts, Bruce, Alexander D. Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. 2014. *Molecular Biology of the Cell: Sixth International Student Edition*. W.W. Norton & Company.](http://paperpile.com/b/0ne0x7/P8pI) <https://market.android.com/details?id=book-AGebDwAAQBAJ>[.](http://paperpile.com/b/0ne0x7/P8pI)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/LHcUn)[10.1038/nmeth.3103](http://dx.doi.org/10.1038/nmeth.3103)[.](http://paperpile.com/b/0ne0x7/LHcUn)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/vyIZ3)[10.1016/S0022-2836(05)80360-2](http://dx.doi.org/10.1016/S0022-2836(05)80360-2)[.](http://paperpile.com/b/0ne0x7/vyIZ3)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/pTMH)[10.1126/science.1252229](http://dx.doi.org/10.1126/science.1252229)[.](http://paperpile.com/b/0ne0x7/pTMH)

[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/](http://paperpile.com/b/0ne0x7/acJT)[10.1093/bib/bbx121](http://dx.doi.org/10.1093/bib/bbx121)[.](http://paperpile.com/b/0ne0x7/acJT)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/82Uf)[10.4161/mge.20031](http://dx.doi.org/10.4161/mge.20031)[.](http://paperpile.com/b/0ne0x7/82Uf)

[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*.](http://paperpile.com/b/0ne0x7/qFESH) <https://www.biorxiv.org/content/10.1101/619684v1.abstract>[.](http://paperpile.com/b/0ne0x7/qFESH)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/dQIO)[10.1038/sdata.2018.176](http://dx.doi.org/10.1038/sdata.2018.176)[.](http://paperpile.com/b/0ne0x7/dQIO)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/JIV1)[10.7717/peerj.3243](http://dx.doi.org/10.7717/peerj.3243)[.](http://paperpile.com/b/0ne0x7/JIV1)

[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.](http://paperpile.com/b/0ne0x7/gP34) <https://www.ncbi.nlm.nih.gov/pubmed/16922685>[.](http://paperpile.com/b/0ne0x7/gP34)

[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.](http://paperpile.com/b/0ne0x7/2ofn) <http://www.jstor.org/stable/24860053>[.](http://paperpile.com/b/0ne0x7/2ofn)

[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/](http://paperpile.com/b/0ne0x7/TaCe)[10.1098/rspb.2014.1382](http://dx.doi.org/10.1098/rspb.2014.1382)[.](http://paperpile.com/b/0ne0x7/TaCe)

[Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60. https://doi.org/](http://paperpile.com/b/0ne0x7/uNkJT)[10.1038/nmeth.3176](http://dx.doi.org/10.1038/nmeth.3176)[.](http://paperpile.com/b/0ne0x7/uNkJT)

[Bushnell, Brian, Jonathan Rood, and Esther Singer. 2017. “BBMerge--Accurate Paired Shotgun Read Merging via Overlap.” *PloS One* 12 (10): e0185056.](http://paperpile.com/b/0ne0x7/tru6) <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0185056>[.](http://paperpile.com/b/0ne0x7/tru6)

[Canchaya, Carlos, Ghislain Fournous, and Harald Brüssow. 2004. “The Impact of Prophages on Bacterial Chromosomes.” *Molecular Microbiology* 53 (1): 9–18. https://doi.org/](http://paperpile.com/b/0ne0x7/anPP)[10.1111/j.1365-2958.2004.04113.x](http://dx.doi.org/10.1111/j.1365-2958.2004.04113.x)[.](http://paperpile.com/b/0ne0x7/anPP)

[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.](http://paperpile.com/b/0ne0x7/Z0cV) <https://www.ncbi.nlm.nih.gov/pubmed/10672174>[.](http://paperpile.com/b/0ne0x7/Z0cV)

[Clingenpeel, Scott, Alicia Clum, Patrick Schwientek, Christian Rinke, and Tanja Woyke. 2015. “Reconstructing Each Cell’s Genome within Complex Microbial Communities—dream or Reality?” *Frontiers in Microbiology* 5: 771.](http://paperpile.com/b/0ne0x7/qF7y) <https://www.frontiersin.org/articles/10.3389/fmicb.2014.00771>[.](http://paperpile.com/b/0ne0x7/qF7y)

[Eberwine, James, Jai-Yoon Sul, Tamas Bartfai, and Junhyong Kim. 2014. “The Promise of Single-Cell Sequencing.” *Nature Methods* 11 (1): 25–27.](http://paperpile.com/b/0ne0x7/aXf7) <https://www.ncbi.nlm.nih.gov/pubmed/24524134>[.](http://paperpile.com/b/0ne0x7/aXf7)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/rmL7t)[10.7717/peerj.1319](http://dx.doi.org/10.7717/peerj.1319)[.](http://paperpile.com/b/0ne0x7/rmL7t)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/zLVa)[10.1093/nar/gkr367](http://dx.doi.org/10.1093/nar/gkr367)[.](http://paperpile.com/b/0ne0x7/zLVa)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/E9xN)[10.1093/nar/gkl732](http://dx.doi.org/10.1093/nar/gkl732)[.](http://paperpile.com/b/0ne0x7/E9xN)

[Fuhrman, J. A. 1999. “Marine Viruses and Their Biogeochemical and Ecological Effects.” *Nature* 399 (6736): 541–48. https://doi.org/](http://paperpile.com/b/0ne0x7/hpfb)[10.1038/21119](http://dx.doi.org/10.1038/21119)[.](http://paperpile.com/b/0ne0x7/hpfb)

[Giovannoni, Stephen J. 2017. “SAR11 Bacteria: The Most Abundant Plankton in the Oceans.” *Annual Review of Marine Science* 9 (January): 231–55. https://doi.org/](http://paperpile.com/b/0ne0x7/MZFO)[10.1146/annurev-marine-010814-015934](http://dx.doi.org/10.1146/annurev-marine-010814-015934)[.](http://paperpile.com/b/0ne0x7/MZFO)

[Giovannoni, Stephen, Ben Temperton, and Yanlin Zhao. 2013. “Giovannoni et Al. Reply.” *Nature* 499 (7459): E4–5. https://doi.org/](http://paperpile.com/b/0ne0x7/8nyP)[10.1038/nature12388](http://dx.doi.org/10.1038/nature12388)[.](http://paperpile.com/b/0ne0x7/8nyP)

[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/](http://paperpile.com/b/0ne0x7/X0LTW)[10.7717/peerj.3035](http://dx.doi.org/10.7717/peerj.3035)[.](http://paperpile.com/b/0ne0x7/X0LTW)

[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/](http://paperpile.com/b/0ne0x7/jGeV6)[10.1128/mBio.00252-12](http://dx.doi.org/10.1128/mBio.00252-12)[.](http://paperpile.com/b/0ne0x7/jGeV6)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/ualz)[10.1038/ismej.2014.143](http://dx.doi.org/10.1038/ismej.2014.143)[.](http://paperpile.com/b/0ne0x7/ualz)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/9KsW)[10.1186/1471-2105-11-119](http://dx.doi.org/10.1186/1471-2105-11-119)[.](http://paperpile.com/b/0ne0x7/9KsW)

[Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. “High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries.” *Nature Communications*. https://doi.org/](http://paperpile.com/b/0ne0x7/23EOO)[10.1038/s41467-018-07641-9](http://dx.doi.org/10.1038/s41467-018-07641-9)[.](http://paperpile.com/b/0ne0x7/23EOO)

[Jang, H. B., B. Bolduc, O. Zablocki, J. Kuhn, and S. Roux. 2019. “Gene Sharing Networks to Automate Genome-Based Prokaryotic Viral Taxonomy.” *BioRxiv*.](http://paperpile.com/b/0ne0x7/K2Mz) <https://www.biorxiv.org/content/10.1101/533240v1.abstract>[.](http://paperpile.com/b/0ne0x7/K2Mz)

[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/](http://paperpile.com/b/0ne0x7/YYNpD)[10.7287/peerj.preprints.27522v1](http://dx.doi.org/10.7287/peerj.preprints.27522v1)[.](http://paperpile.com/b/0ne0x7/YYNpD)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/mQyK7)[10.1101/gr.210641.116](http://dx.doi.org/10.1101/gr.210641.116)[.](http://paperpile.com/b/0ne0x7/mQyK7)

[Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921. https://doi.org/](http://paperpile.com/b/0ne0x7/eReu)[10.1038/35057062](http://dx.doi.org/10.1038/35057062)[.](http://paperpile.com/b/0ne0x7/eReu)

[Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/](http://paperpile.com/b/0ne0x7/1k4Wv)[10.1038/nmeth.1923](http://dx.doi.org/10.1038/nmeth.1923)[.](http://paperpile.com/b/0ne0x7/1k4Wv)

[Lasken, Roger S. 2007. “Single-Cell Genomic Sequencing Using Multiple Displacement Amplification.” *Current Opinion in Microbiology* 10 (5): 510–16. https://doi.org/](http://paperpile.com/b/0ne0x7/cUdt)[10.1016/j.mib.2007.08.005](http://dx.doi.org/10.1016/j.mib.2007.08.005)[.](http://paperpile.com/b/0ne0x7/cUdt)

[Letarov, A., and E. Kulikov. 2009. “The Bacteriophages in Human- and Animal Body-Associated Microbial Communities.” *Journal of Applied Microbiology*. https://doi.org/](http://paperpile.com/b/0ne0x7/1SFu)[10.1111/j.1365-2672.2009.04143.x](http://dx.doi.org/10.1111/j.1365-2672.2009.04143.x)[.](http://paperpile.com/b/0ne0x7/1SFu)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/nhmI)[10.1093/bioinformatics/btl529](http://dx.doi.org/10.1093/bioinformatics/btl529)[.](http://paperpile.com/b/0ne0x7/nhmI)

[Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics*  34 (18): 3094–3100. https://doi.org/](http://paperpile.com/b/0ne0x7/hiEEM)[10.1093/bioinformatics/bty191](http://dx.doi.org/10.1093/bioinformatics/bty191)[.](http://paperpile.com/b/0ne0x7/hiEEM)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/Opoh)[10.1093/bioinformatics/btn043](http://dx.doi.org/10.1093/bioinformatics/btn043)[.](http://paperpile.com/b/0ne0x7/Opoh)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/JKRR)[10.1038/nature04111](http://dx.doi.org/10.1038/nature04111)[.](http://paperpile.com/b/0ne0x7/JKRR)

[McInnes, Leland, John Healy, and Steve Astels. 2017. “Hdbscan: Hierarchical Density Based Clustering.” *The Journal of Open Source Software* 2 (11): 205.](http://paperpile.com/b/0ne0x7/MxXC) <https://www.theoj.org/joss-papers/joss.00205/10.21105.joss.00205.pdf>[.](http://paperpile.com/b/0ne0x7/MxXC)

[McInnes, Leland, John Healy, and James Melville. 2018. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *arXiv [stat.ML]*. arXiv.](http://paperpile.com/b/0ne0x7/iYXh) <http://arxiv.org/abs/1802.03426>[.](http://paperpile.com/b/0ne0x7/iYXh)

[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/](http://paperpile.com/b/0ne0x7/6VlWq)[10.1101/530188](http://dx.doi.org/10.1101/530188)[.](http://paperpile.com/b/0ne0x7/6VlWq)

[Menzel, Peter, Kim Lee Ng, and Anders Krogh. 2016. “Fast and Sensitive Taxonomic Classification for Metagenomics with Kaiju.” *Nature Communications*. https://doi.org/](http://paperpile.com/b/0ne0x7/Gztnq)[10.1038/ncomms11257](http://dx.doi.org/10.1038/ncomms11257)[.](http://paperpile.com/b/0ne0x7/Gztnq)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/Hh4Z)[10.1128/JB.02230-14](http://dx.doi.org/10.1128/JB.02230-14)[.](http://paperpile.com/b/0ne0x7/Hh4Z)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/RNYU)[10.1093/molbev/msu300](http://dx.doi.org/10.1093/molbev/msu300)[.](http://paperpile.com/b/0ne0x7/RNYU)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/0xHy)[10.1093/dnares/dsn027](http://dx.doi.org/10.1093/dnares/dsn027)[.](http://paperpile.com/b/0ne0x7/0xHy)

[Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. “A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life.” *Nature Biotechnology* 36 (10): 996–1004. https://doi.org/](http://paperpile.com/b/0ne0x7/b3gn)[10.1038/nbt.4229](http://dx.doi.org/10.1038/nbt.4229)[.](http://paperpile.com/b/0ne0x7/b3gn)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/dCN6D)[10.1101/gr.186072.114](http://dx.doi.org/10.1101/gr.186072.114)[.](http://paperpile.com/b/0ne0x7/dCN6D)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/NSqu)[10.1038/ismej.2011.101](http://dx.doi.org/10.1038/ismej.2011.101)[.](http://paperpile.com/b/0ne0x7/NSqu)

[Proctor, Lita M., and Jed A. Fuhrman. 1990. “Viral Mortality of Marine Bacteria and Cyanobacteria.” *Nature* 343 (6253): 60–62. https://doi.org/](http://paperpile.com/b/0ne0x7/5eZ5)[10.1038/343060a0](http://dx.doi.org/10.1038/343060a0)[.](http://paperpile.com/b/0ne0x7/5eZ5)

[Rabosky, Daniel L., Michael Grundler, Carlos Anderson, Pascal Title, Jeff J. Shi, Joseph W. Brown, Huateng Huang, and Joanna G. Larson. 2014. “BAMM Tools: An R Package for the Analysis of Evolutionary Dynamics on Phylogenetic Trees.” *Methods in Ecology and Evolution / British Ecological Society* 5 (7): 701–7.](http://paperpile.com/b/0ne0x7/vOIQ) <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12199>[.](http://paperpile.com/b/0ne0x7/vOIQ)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/CyMY)[10.1038/nrmicro2235](http://dx.doi.org/10.1038/nrmicro2235)[.](http://paperpile.com/b/0ne0x7/CyMY)

[Rohwer, Forest, and Rebecca Vega Thurber. 2009. “Viruses Manipulate the Marine Environment.” *Nature* 459 (7244): 207–12. https://doi.org/](http://paperpile.com/b/0ne0x7/CEvh)[10.1038/nature08060](http://dx.doi.org/10.1038/nature08060)[.](http://paperpile.com/b/0ne0x7/CEvh)

[Roux, Simon, Francois Enault, Bonnie L. Hurwitz, and Matthew B. Sullivan. 2015. “VirSorter: Mining Viral Signal from Microbial Genomic Data.” *PeerJ* 3 (May): e985. https://doi.org/](http://paperpile.com/b/0ne0x7/f6QP)[10.7717/peerj.985](http://dx.doi.org/10.7717/peerj.985)[.](http://paperpile.com/b/0ne0x7/f6QP)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/95Dk)[10.7554/eLife.03125](http://dx.doi.org/10.7554/eLife.03125)[.](http://paperpile.com/b/0ne0x7/95Dk)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/V43p)[10.1038/nature08284](http://dx.doi.org/10.1038/nature08284)[.](http://paperpile.com/b/0ne0x7/V43p)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/Yaum)[10.1038/ismej.2011.2](http://dx.doi.org/10.1038/ismej.2011.2)[.](http://paperpile.com/b/0ne0x7/Yaum)

[Sun, Ying, and Haiwei Luo. 2018. “Homologous Recombination in Core Genomes Facilitates Marine Bacterial Adaptation.” *Applied and Environmental Microbiology* 84 (11). https://doi.org/](http://paperpile.com/b/0ne0x7/biZ2)[10.1128/AEM.02545-17](http://dx.doi.org/10.1128/AEM.02545-17)[.](http://paperpile.com/b/0ne0x7/biZ2)

[Suttle, C. A. 1994. “The Significance of Viruses to Mortality in Aquatic Microbial Communities.” *Microbial Ecology* 28 (2): 237–43. https://doi.org/](http://paperpile.com/b/0ne0x7/V5fP)[10.1007/BF00166813](http://dx.doi.org/10.1007/BF00166813)[.](http://paperpile.com/b/0ne0x7/V5fP)

[Suttle, Curtis A. 2007a. “Marine Viruses — Major Players in the Global Ecosystem.” *Nature Reviews Microbiology*. https://doi.org/](http://paperpile.com/b/0ne0x7/HjVx)[10.1038/nrmicro1750](http://dx.doi.org/10.1038/nrmicro1750)[.](http://paperpile.com/b/0ne0x7/HjVx)

[———. 2007b. “Marine Viruses—major Players in the Global Ecosystem.” *Nature Reviews. Microbiology* 5 (10): 801.](http://paperpile.com/b/0ne0x7/ABUM) <https://www.nature.com/articles/nrmicro1750>[.](http://paperpile.com/b/0ne0x7/ABUM)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/LGlQ)[10.1073/pnas.1102164108](http://dx.doi.org/10.1073/pnas.1102164108)[.](http://paperpile.com/b/0ne0x7/LGlQ)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/zx1N)[10.1038/ismej.2013.243](http://dx.doi.org/10.1038/ismej.2013.243)[.](http://paperpile.com/b/0ne0x7/zx1N)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/cTLf)[10.1371/journal.pone.0101415](http://dx.doi.org/10.1371/journal.pone.0101415)[.](http://paperpile.com/b/0ne0x7/cTLf)

[Van Valen, Leigh. 1973. “A NEW EVOLUTIONARY LAW.” *Evolutionary Theory* 1: 1–30.](http://paperpile.com/b/0ne0x7/eWxT) <https://ebme.marine.rutgers.edu/HistoryEarthSystems/HistEarthSystems_Fall2010/VanValen%201973%20Evol%20%20Theor%20.pdf>[.](http://paperpile.com/b/0ne0x7/eWxT)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/KN2M)[10.1111/j.1462-2920.2007.01361.x](http://dx.doi.org/10.1111/j.1462-2920.2007.01361.x)[.](http://paperpile.com/b/0ne0x7/KN2M)

[Waldor, Matthew K., and David I. Friedman. 2005. “Phage Regulatory Circuits and Virulence Gene Expression.” *Current Opinion in Microbiology* 8 (4): 459–65. https://doi.org/](http://paperpile.com/b/0ne0x7/B9nW)[10.1016/j.mib.2005.06.001](http://dx.doi.org/10.1016/j.mib.2005.06.001)[.](http://paperpile.com/b/0ne0x7/B9nW)

[Weinbauer, Markus G. 2004. “Ecology of Prokaryotic Viruses.” *FEMS Microbiology Reviews* 28 (2): 127–81. https://doi.org/](http://paperpile.com/b/0ne0x7/qrUL)[10.1016/j.femsre.2003.08.001](http://dx.doi.org/10.1016/j.femsre.2003.08.001)[.](http://paperpile.com/b/0ne0x7/qrUL)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/uiKR)[10.1093/bioinformatics/btv383](http://dx.doi.org/10.1093/bioinformatics/btv383)[.](http://paperpile.com/b/0ne0x7/uiKR)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/gHGQ)[10.1128/MMBR.00034-09](http://dx.doi.org/10.1128/MMBR.00034-09)[.](http://paperpile.com/b/0ne0x7/gHGQ)

[Wommack, K. E., and R. R. Colwell. 2000. “Virioplankton: Viruses in Aquatic Ecosystems.” *Microbiology and Molecular Biology Reviews: MMBR* 64 (1): 69–114. https://doi.org/](http://paperpile.com/b/0ne0x7/JhzB)[10.1128/mmbr.64.1.69-114.2000](http://dx.doi.org/10.1128/mmbr.64.1.69-114.2000)[.](http://paperpile.com/b/0ne0x7/JhzB)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/AMKmA)[10.1186/2049-2618-2-26](http://dx.doi.org/10.1186/2049-2618-2-26)[.](http://paperpile.com/b/0ne0x7/AMKmA)

[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. https://doi.org/](http://paperpile.com/b/0ne0x7/o8X7)[10.1093/nar/gkr485](http://dx.doi.org/10.1093/nar/gkr485)[.](http://paperpile.com/b/0ne0x7/o8X7)