itle

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

# Abstract

**90 % of the ocean’s biomass comprise of marine microorganisms** [**(Curtis A. Suttle 2007)**](https://paperpile.com/c/uaTuS8/14Tcv) **where the SAR11 bacteria is the most ubiquitous, comprising of 30 % of all surface water cells and one of the most widespread organisms on Earth** [**(Morris et al. 2002)**](https://paperpile.com/c/uaTuS8/0cjK)**. Its ability to harvest labile dissolved organic matter** [**(Carlson et al. 2004)**](https://paperpile.com/c/uaTuS8/RQrv) **and produce volatile organic compounds** [**(Sun et al. 2011; Carini et al. 2014)**](https://paperpile.com/c/uaTuS8/SjN3+oykW) **plays an important role in global biogeochemistry. Its small genome size and complexity provides insight into the microbiological evolution and function of streamlined cells** [**(Giovannoni et al. 2014)**](https://paperpile.com/c/uaTuS8/Nlyu)**. However, complete SAR11 whole genomes sequences are relatively rare due to their difficulty in culturing** [**(Rappé et al. 2002; Carini et al. 2014; Carini et al. 2013)**](https://paperpile.com/c/uaTuS8/LJ0G+Qkm6+B6qS)**, leading to gaps in knowledge of SAR11 diversity. Here we show the advantage of single-cell amplified genomes in genomic analysis of the SAR11 clade. 451 SAR11 bacteria were isolated and sequenced using SAG technology. We found evidence suggesting the presence of two new phylogenetic clades of SAR11, expanding the numbers of niche habitats SAR11 inhabits** [**(Giovannoni 2017)**](https://paperpile.com/c/uaTuS8/XGgJU)**. Additionally, genomic analysis confirms the existence of multiple hypervariable regions** [**(Wilhelm et al. 2007; Grote et al. 2012)**](https://paperpile.com/c/uaTuS8/T9Vh+3ODS) **containing horizontally transferred genes . Analysis of these HVRs revealed genes related to nutrient uptake and phage defense, consistent with previous studies advocating for a King of the Mountain ecological strategy** [**(Giovannoni et al. 2013)**](https://paperpile.com/c/uaTuS8/vkFQh)**. Lastly, identification of viral signatures within SAR11 SAGs lead to confirmation of existing viral clades within SAR11. Our results demonstrate the value of SAGs to allow for clade wide studies of the SAR11 pangenome. We anticipate the usage of SAGs to become a powerful alternative for the study of difficult to isolate and culture organisms.** Isolation of additional SAR11 genomes would lead to further understanding of SAR11 phylogenetics and studies of its genomic variations leading to its successful ecological strategy. Lack of SAR11 SAGs within clade 5 may indicate that clade 5 within the SAR11 group, consistent with other studies and should instead be a sister clade [(Viklund et al. 2013; Thrash et al. 2011; Ferla et al. 2013)](https://paperpile.com/c/uaTuS8/37NT+kD0q+VWAa).

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/uaTuS8/14Tcv) 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/uaTuS8/enzNc). This biomass consists mainly of heterotrophic and eutrophic plankton [(Proctor and Fuhrman 1990)](https://paperpile.com/c/uaTuS8/NGzoE), 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/uaTuS8/EiECN). 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/uaTuS8/14Tcv). 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.

[**Abstract**](#_wb62gpwwgpj7) **1**

[**1.0 Introduction/Background**](#_t7jk6rojtx31) **3**

[1.1 What are marine microorganisms](#_trksp175l43v) 3

[1.2 Where do they exist in the oceans](#_g4ttn5d76tkn) 4

[1.3 What population dynamics exist](#_z4qmi1v2jwjr) 4

[1.3 Why are they important to study](#_dnd46qp91dy2) 5

[1.4 What remains to be studied](#_8b9ap94hogse) 5

[1.5 What existing techniques are there to study these microbes](#_mj9zotvqqvn) 5

[1.5.1 SAGs and what are they](#_d96j47qw6h1b) 5

[1.5.2 Marine Metagenomics](#_mr3e7jxmma3z) 5

[1.XX How do we obtain genomic sequences from metagenomes and SAGs](#_v8of71bgsocf) 7

[1.XX How do we assemble a genome](#_ugx1xagkgyv) 7

[1.XX importance of read quality control](#_gv0v056ae0n4) 8

[1.XX How do we assess genome completeness](#_wa2rrlfnpxkf) 10

[**2.XX Chapter 1 Background**](#_jx8mci9l3zx2) **10**

[2.XX Where do the marine metagenomes come from](#_lrt9lbkkwuf6) 10

[X.XX Why is it important to obtain metagenomes](#_zij4dd752wqe) 11

[X.XX How do we obtain MAGs](#_jorwyrw0z1je) 11

[X.XX Binning by coverage](#_ygij3d35fl7b) 11

[X.XX Binning by sequence composition](#_to6ghh7x3znc) 11

[**X.XX CHAPTER 1 METHOD**](#_9mie4hnz1reo) **12**

[X.XX How were the reads obtained](#_v905q0tr030i) 12

[X.XX Read preprocessing and assembly](#_31qvavv9wuvi) 12

[X.XX Binning by Coverage](#_i4fjqxumge92) 13

[X.XX Visualisation of Binning Algorithims](#_ix3y6shfqev6) 13

[X.XX Assessing the quality of MAGs](#_qwvpguc94oir) 13

[X.XX Binning by binning software](#_mtlpk97ked9j) 14

[X.XX Binning by taxonomy](#_8c71rmsx8gml) 17

[X.XX Binning by k-mer counting](#_4f565ps62bqk) 21

[**2.3 CHAPTER 1 DISCUSSION**](#_clnxev650v4x) **28**

[**3.0 CHAPTER 2 - SAGs Phylogenetics and Genome analysis**](#_waiz2wcm0f0r) **28**

[3.1 451 SAR11 SAGs isolated and sequenced](#_avcnsac9zo5c) 28

[3.2 CHAPTER 1 METHOD](#_vlr285cc6uoj) **29**

[3.2.1 Preprocessing](#_7xa6sh79vk3u) 29

[3.2.2 Quality control on reads](#_xmxhjn5awq9r) 29

[3.2.2.1 SPAdes Assembly](#_egqbfwmnkgla) 29

[3.2.2.2 Genomic analysis](#_6kgrwfyx0kl4) 29

[3.2.3 Phylogenetics](#_s68z4jk4lr38) 30

[X.XX kmer counting](#_4qh6w7wvk62r) 40

[**3.3 CHAPTER 2 DISCUSSION**](#_jb4u59c23n4o) **44**

[Problems with MDA](#_rtd1rpikfuay) 44

[Improvements to MDA:](#_pb3xfx78lt1u) 44

[pico/nanoliter reactions suppress amplification bias](#_tvays93bpzs1) 44

[PicoPLEX (https://www.nature.com/articles/s41467-017-00128-z)](#_wk29uh36i2z5) 44

[MALBAC (both these alternate methods reduce amplification bias (https://www.nature.com/articles/s41467-017-00128-z)](#_puvg9exe2po9) 44

[Usage of metagenomic composite material as internal reference](#_z1pseybn8wfg) 44

[Combining data = higher chance of contamination](#_6aizfxx9pje6) 44

[Loss of actual data like SNPs or species specific genes?](#_l43pblslaw8p) 44

[**X.XX Chapter 3 - SAG viral signatures**](#_1jvmwquo2rtu) **45**

[**X.XX What are viral signatures**](#_ypx94j9mvvq5) **45**

[**X.XX CHAPTER 4 - SAGS and Variable region**](#_jh0wsr7ujji) **47**

[**5.2 CHAPTER 4 METHOD**](#_j8rxfy56oq9v) **48**

[**5.3 CHAPTER 4 DISCUSSION**](#_53cprhqz3lxq) **49**

[**6.0 THESIS DISCUSSION**](#_y1q40nf0uh6c) **50**

[**THESIS BIBLIOGRAPHY**](#_9nkxfng6fpxt) **53**

[**Appendices**](#_37g53aiefuml) **55**

# 1.0 Introduction/Background

## 1.1 What are marine microorganisms

* Marine defined by living in marine environment
  + Marine environment - Sea, Ocean, brackwater, coastal estuary
* Microorganism
  + Microscopic living organisms, too small to see by the naked eye
  + Highly diverse
  + Single cell or multicellular
  + Types - Eukaryotes, prokaryotes, viruses
    - Eukaryotes
      * Fungi, algae, microscopic animals
    - Prokaryotes
      * Bacteria and Archaea
    - Viruses
      * Contention - not living
      * But huge impact on microorganisms

Broadly marine ecosystems are saltwater systems that support aquatic life. They are usually differentiated by coastal and open ocean habitats - an environmental area inhabited by living species. Coastal regions generally include brackwater zones and tidal regions up to the continental shelf, areas of shallow water before the drop into deeper open oceans. These contain a vast majority of marine life. Open oceans generally refer to the regions beyond the continental shelf. Within the open oceans, the pelagic and demersal zones exist for differentiating deep and shallow water zones. The pelagic zone and greatly affected by ocean currents where demersal zones tend to be more stable with the exception of seasonal upwelling events.

Microorganisms are defined as single-cell or collections of cells usually too small for the naked eye to see. Such organisms usually consist of small free-living units of life and are the biological foundation on which higher life forms rely on. Broadly, these microorganisms can be divided into 2 groups, eukaryotes and prokaryotes. Eukaryotic microorganisms consist of fungi, algae and some microscopic animals. These are generally considered more complex lifeforms. Prokaryotes are split into two clades - groups of organisms descended from a common ancestor. These clades consist of bacteria and archaea, generally a less complex lifeform. Lastly and most contentiously are viruses, particles existing freely in the environment as virions. They are contentious because they do not have all the characteristics of life, mainly reproduction as they require a host to replicate. However, they are the most numerous biological entity on earth and their influence on the world's biosphere cannot be disregarded.

## 1.2 Where do they exist in the oceans

* What are the different types of bacteria
  + Explicit mention of SAR11 and importance - talk more broad over specifics
  + What it does and where it is and why is can exist
* Roughly where do each clade exist
* Types of systems within the oceans
  + How does this impact microorganism abundance

A multitude of factors can influence where an organism lives. In marine ecosystems, abiotic factors like sunlight, nutrients and temperature define organism habitat boundaries. Pelagic habitats are greatly influenced by ocean currents. They transport nutrients for these organisms to survive. Most microorganisms do not have the means for free movement and therefore are at the mercy of ocean currents to provide transport into suitable zones for survival. This takes the form of nutrients and shallower waters for photosynthetic phytoplankton. Ocean currents form due to the changes in temperature and salinity of water. Cooler and saltier water is denser and therefore sinks with warmer and lower salinity water doing the opposite. Winds and the gravitational pull of the sun and moon also influence currents in the form of tides and waves. The Coriolis effect also influences currents, creating subtropical ocean gyres - circulating ocean currents. As the earth rotates eastwards, moving water columns possess this momentum and also moves eastward. Water columns propelled by northerly and southerly trade winds move bodies of water north and south. This would alter its trajectory eastward as well was forward, creating a clockwise movement in the northern hemisphere and an anti-clockwise movement in the southern hemisphere.

## 1.3 What population dynamics exist

* Viral prey and predation dynamics
* Piggyback the winner
* King of the hill

## 1.3 Why are they important to study

* Energy food web and energy cycle?
* What is their impact on global systems?
  + How do they regulate global oxygen and carbon cycles

## 1.4 What remains to be studied

* Gene content throughout ocean systems
* Exact population abundance and proportions
* Population dynamics and interactions

## 1.5 What existing techniques are there to study these microbes

### 1.5.1 SAGs and what are they

* + Genomes from isolated single cells are that whole genome amplified and sequenced
  + Organisms usually only have 1 to a few copies of their genomic DNA. Not enough to sequence so need to amplify with current sequencing tech
  + Longer the amplicon (sequence that is the source of/is amplified)
* How does SAG technology work
  + Isolated
  + Lysed
  + Multiple Displacement Amplification (MDA)
    - <https://en.wikipedia.org/wiki/Multiple_displacement_amplification>
    - Annealing random hexamer primers
    - 30 degrees isothermal reaction
    - DNA synthesis via phi29 DNA polymerase
    - Make multiple copies of DNA template via strand displacement reaction (between primers?)
    - 12 kb in length up to 100kb
* Why do we isolate them?
  + Application in microbiomes
  + Microbes unable to be cultivated, therefore any tool without cultivation can get the uncultivatable
  + Can target specific organism if can isolate and find
* What are the advantages of isolating them this way
  + Without reference, how to determine % complete/contamination

### 1.5.2 Marine Metagenomics

* Metagenomes are all genetic material recovered from environmental samples
* Used to discover genomic content and population dynamics
* Snapshot of a population's genetic material at a given space and time
* Great for presence absence studies
  + Possible abundance studies
  + More Macroscale approach to populations
  + More limited in ability to look at individual strains, more higher classifications of phylogeny (class level)
* What are marine metagenomes
  + Genetic material recovered from a sample
  + Environmental sample is take, sometimes amplified and sequenced
  + Allows for whole populations to be capture in one sample
  + Doesn’t require cultivate of organisms ~90% cant be cultivate anyway
  + Allows for presence absence or population abundance studies
  + Allows for gene wide studies in population … etc
* Disadvantages of it
  + Tend to by highly fragmented
  + Not all DNA may be recovered leading to incomplete genes
  + Relies on good methods al lysis and access genes
  + Organisms of similar genetic content are hard to separate
  + Captures env DNA
  + Requires specialist bioinformatics tools to assemble
  + Difficult to assess quality of assembly
  + Needs binning assembly to obtain WGS
* Link between SAGS and metagenomics
  + SAGS and metagenomes can complement each other when looking at genetic studies
  + SAGS isolated from the same sample metagenomics allow for metagenomics to “fill in the gaps”
* Where do my marine metagenomes come from
  + 12 samples of marine metagenomes awas obtain from the Bermuda Atlantic Time series at two differing depths (80m and 200m) and over a different time periods. This would allow for comparisons of marine populations of spatial and temporal gradients. This would indicate if there is a different in population abundance and gene content over a diel cycle.
* Why do we care about marine metagenomes
  + Can get population abundance - snapshot of the wild rather then taking a look at the zoo after capture and conservation
  + Population dynamics, structure and composition, down to a genetic level
  + Allows for analysis of population without having to culture
    - Unable to culture ~90% of organisms
* What is it
* Where is it
* Why are we using it
* Is this representational?
* Why 80m and 200m

## 

## 1.XX How do we obtain genomic sequences from metagenomes and SAGs

Genomic sequences is defined as any kind of DNA or RNA used to store information relating to an organism genes. Extraction of the Whole Genome Sequence (WGS) of an organism usually involves its capture and extraction of its DNA. Broadly, in micro-organisms samples are isolated and usually cultured to produce enough DNA for DNA sequencing technology, machines that can translate DNA sequences into an informatic form that can be used for analysis. However, many organisms have not been documented to be culturable, and instead DNA amplification methods exist to increase the amount of DNA material to allow for sequencing and enough coverage, repeats of a certain area or the WGS of an organism to ensure the correct sequence of the organisms DNA is obtained as DNA sequencing technology are not 100% accurate. With enough coverage, a consensus sequence over erroneous areas would theoretically compensate for a non-perfect accuracy. [Amplification of DNA however is not without its caveats](#_rtd1rpikfuay).

With even the smallest organism being 1.3 Mps in size, sequencing an organisms from start to end would take a long time, nevermind including the number of repeats needed to provide roughly a 50 fold coverage to compensation for erroneous areas. Therefore parallel sequencing is used where an organism's genome is split into multiple fragments and sequenced in parallel. This drastically increases the time take to sequence a whole genome to a reasonable amount of time. This is a popular technique in next generation sequencing technologies alike illumina, where 150-300 bp sized fragments are amplified and sequenced in parallel, allowing for WGS to be sequenced quickly. Other methods like the third generation method by Oxford Nanopore uses the change in electrical potential across a pore on a membrane when genetic material passes through allows for longer reads to be used but suffers from a lower accuracy rate.

## 1.XX How do we assemble a genome

Sequencing assembly is the process where reads from various DNA sequencing technology is ordered, aligned and merged together to form a longer consensus sequence resulting in a contiguous assembly of reads (contig). This bioinformatic process is required as DNA sequencing technology produces short sequences called reads that are representations of small fragments of the genetic code of an organism, similar to jigsaw pieces in a puzzle. These resulting “pieces” or reads need to be “put” or assembled together to form the complete “puzzle” or assembly.

There exists two methods or assembly of reads, one being de-novo assembly and the other a mapping assembly. A de-novo assembly refers to the assembly of reads without a reference, relying purely the composition of the reads alone. With a mapping assembly, reads are “mapped” to a reference, where a read is compared to an existing template or reference and its position within an assembly based on that. A mapping assembly can only be used where a reference is available, being impossible to perform on species that are new or novel. De-novo sequence is the most popular sequencing technology but suffers from multiple problems mainly assembling repeat regions and a higher computational costs having to compare each read against each other.

De-novo assemblies briefly use two types of assembly algorithms for assembly of reads. The greedy algorithm or overlap-layout-consensus (OLC) method where reads are compared for overlaps and joined together based on these overlaps. This produces a consensus sequence but is often computationally intensive. Graph based methods include the string and De Brujin method. In a De Brujin graph, reads are reduced down to a smaller specific sizes, of size *k*. These resulting k-mers become nodes within a graph. Overlaps of these nodes are then denoted by a connecting edge. The assembler will then assembly the sequence based on the “path” that the nodes and edges align to. This is less computational intensive and performs better at repeat regions. Assembliers like SPAdes use De Bruijn graphs to perform their assembly.

## 1.XX importance of read quality control

With the rise of next generation and third generation sequencing methods, large amounts of DNA sequences can be collected quickly and cheaply. However, these methods are not infallible, as errors can occur in the DNA sequencing process due to XXX. With the illumina platform, error rates usually exist in the ~99 percentile, but with long reads, these error rates can rise to within the 90 percentile. With a large amount of genetic material, even a very low error rate can mean a large amount of incorrect reads. One of the smallest marine organisms, SAR11 having a genome size of 1.3 Mps can still result in 13 Kbs in erroneous reads. Therefore read error correction becomes important in removing reads with a low quality to avoid propagation of these erroneous reads within the final assembly of an organism.

Within this report, an in house quality control script written by [(Bushnell, Rood, and Singer 2017a)](https://paperpile.com/c/uaTuS8/1wta) is used as a form of quality assurance of the resulting reads before assembly. The quality control is written in the form of a snakemake file [(Köster and Rahmann 2012)](https://paperpile.com/c/uaTuS8/bDxm), a computer workflow management system written in Python allowing for scalable data management. This allows for multiple samples to be efficiently managed within predefined computer resources.

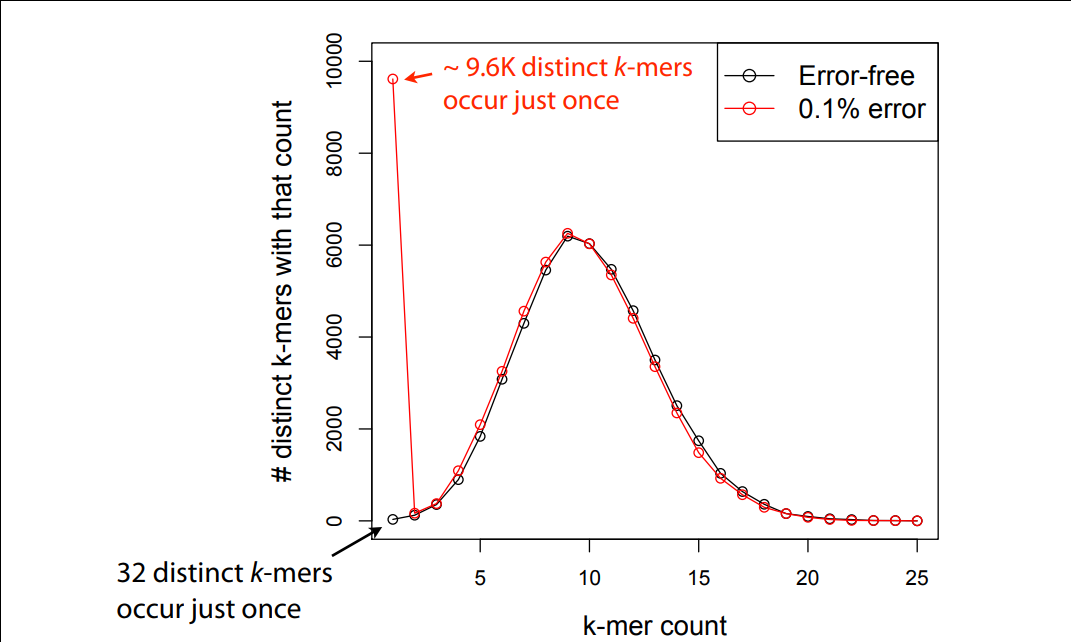
Briefly, this script is performed in 3 main steps:

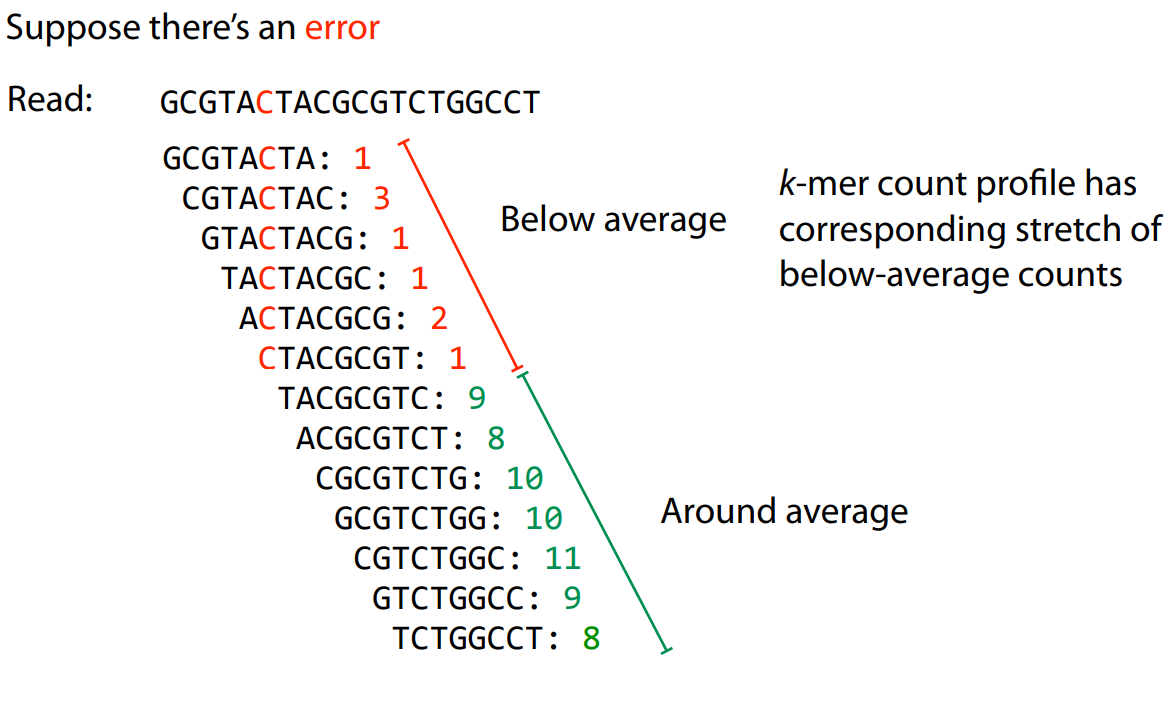
1. The removal of low quality and duplicate regions
2. The trimming of adaptors and synthetic artifacts
3. Error correction
   1. Normalisation if appropriate

Low quality reads occur where bases at the ends of DNA sequences are identified potentially incorrectly and indicated as such via confident scores. These untrusted regions are “trimmed” from the ends of DNA reads due to their propensity for assembly software to assemble samples based on the wrong order due to the incorrect reads, resulting is unassembled or incorrectly assembled assemblies.

Adaptors are short DNA sequences synthetically made and bound to the ends of DNA fragments before DNA sequencing. Adaptors perform a multitude of functions including primers for identification of forward and reverse reads for paired end sequencing and fixing of DNA fragments to flow cells to allow for DNA amplification in illumina sequencing technology. These sequencing are sequenced within the DNA sequencing process and are present within sequenced reads. These need to be removed due to being artificial in nature, and of the same sequence. This would cause misassembly as such sequences would assemble together due to being the same sequence. Since this sequence is known and on the ends of reads, these can be easily trimmed if the resulting sequence on the ends of reads match the adaptor sequence.

Error correction is the process in which reads with an incorrect nucleotide sequence is corrected via consensus or by k-mer counting. Generally, within an organism's genetic sequence, the distribution of k-mers remains the same. Therefore any sequence out of the average distribution has a change of being wrong. This can be detected by a lower average k-mer count, and either removed or corrected to the correct base that would result in a continuation of the average k-mer distribution. XXX





## 1.XX How do we assess genome completeness

# 2.XX Chapter 1 Background

## 2.XX Where do the marine metagenomes come from

The Bermuda Atlantic Time-series Study (BATS) site in the Sargasso Sea (Michaels and Knap, 1996) provides a unique opportunity to understand the virosphere and its relationships with physical and biological factors. The BATS is a long-term study site in a seasonal oligotrophic system. It records biological, chemical and physical data monthly and has been doing so for decades (Giovannoni and Vergin, 2012). Biological samples taken from this at differing depths can help us understand viral abundance in response to differing conditions. Previous studies have been successful in identifying viral abundance and classification (Parsons et al., 2012). This is important in understanding the ecological impact of such viruses. Its temporal and spatial variability in context with other physical and biological parameters.

Diel cycles also provide a critical parameter where viral abundance changes in the short term (Winter, Herndl and Weinbauer, 2004). A previous study (unpublished data) took samples at 8-hour period over a 72-hour timeframe. Subsequently, 24 metagenomes were established from the Sargasso Sea, separated into 12 cellular and 12 viral communities via 0.2um filter. The viral fraction is already undergoing analyses, however a comparison against the cellular fraction would be critical in understanding viral lifecycles. This is because virus can undergo lysogenic lifecycles (Suttle, 2007). The cellular proportion may contain viral sequences within its genome and it would be interesting to see if these proportion changed based on diel cycle, as well as the identity and source of these viral sequences. Overall this would look at how viral contigs within the cellular fraction compare to the viral free communities.

## X.XX Why is it important to obtain metagenomes

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.

## X.XX How do we obtain MAGs

MAGs are often obtain bioinformatically after sequencing of a metagenome. Individual genomes are separated or “binned” using specialised software. There exists multiple approaches for this process, mainly binning by coverage or by sequence compostion.

## X.XX Binning by coverage

To separate genomes from metagenomes, the most common and well documented method is by coverage. If multiple samples are taken from the same environmental location and time, expected species and their abundance would expected to be the same. (i.e. if you took two samples from the same place and time, you would expect them to be the same). Upon this assumption, the two resulting metagenomes would be expected to be somewhat identical. With this in mind, organisms exist in differing abundances within a population. Therefore, a disproportionate amount of reads are generated for each organism, with organisms at a higher abundance being more highly represented. Therefore, mapping reads back to assembled contigs would not only provide an idea to the abundance of such contig and possibly the organism it represents, contigs of a similar coverage would indicate this may be the same organism, if there is no DNA sequencing bias like amplification bias. With multiple samples, contigs with the same changes in coverage would further confirm such contigs as coming from the same organism. Contigs can then by grouped into “bins” representing a said organisms.

## X.XX Binning by sequence composition

Genetic material is made up of ATGCs in a specific order, often to code for proteins and other building blocks of life. Species often code of proteins specific to them and therefore can be differentiated upon this difference. For example an e.coli codes for protein XXX different to species XXX even though they are the same protein. Sequence compositions are more similar within species as they code things their own way. Therefore any process that counts the sequence composition should be able to differentiate species and therefore [kmer counting.](#_4qh6w7wvk62r)

# X.XX CHAPTER 1 METHOD

## X.XX How were the reads obtained

Physical extraction of marine metagenomic samples and sequencing was not performed in this report, but briefly, such method included water being obtained from the BATS station. XXXL of water was obtained and filtered to separate eukaryotes, prokaryotes and viruses. Filter size was dependent on the cell size of each fraction, Metagenomic samples were filtered for larger eukaryotic organisms. The viral and cellular fractions were separated in a similar fashion with the filters used as the medium of extraction of each fraction. Only the cellular fraction is analysed in this report.

Samples were obtained from the BATS over differing spatial and temporal gradients. Six samples were obtained from a depth of 80 meters below sea level and an additional six obtained at 200 meters. Samples were taken at 4 hour intervals completing a single diel cycle with 6 samples. Cellular samples were sequenced with Nextera DNA library preparation kits on the illumina next generation sequencing platform.

## X.XX Read preprocessing and assembly

Resulting reads were checked for quality using an **inhouse quality control script**, with the exception of read normalisation. (<https://github.com/ash-bell/masters_thesis/blob/master/QC_reads.smk>) Read normalisation is the process in which some reads are overexpressed within a sample due to amplification bias or a higher representation of a common species within a metagenome. Read normalisation allows for adjustment of read quantity around a standard bell curve, allowing for less represented reads to be more prevalent whilst maintaining existing proportions. This has been shown to improve single cell assembly but not metagenomic assembly. In addition, SPAdes includes a normalisation step within its assembly algorithm and presumably would expect non-normalised data.

After quality control, reads are assembled using a multitude of assembliers, chiefly SPAdes [(Bankevich et al. 2012)](https://paperpile.com/c/uaTuS8/OwCT), MEGAHIT [(D. Li et al. 2015)](https://paperpile.com/c/uaTuS8/5Gw7) and Tadwrapper [(Bushnell, Rood, and Singer 2017a)](https://paperpile.com/c/uaTuS8/1wta). The resulting assemblies were mapped back to their reads to determine the percentage of mapping reads. This determined the completion of each of the assemblies. Additional MetaQUAST [(Mikheenko, Saveliev, and Gurevich 2016)](https://paperpile.com/c/uaTuS8/XHmd) was used to assess the statistics of each assembly. The assembly with the largest N50 was determined to have the best assembly. It was concluded the SPAdes produced the best assembly using these parameters.

## X.XX 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/uaTuS8/PwQa), BBtools [(Bushnell, Rood, and Singer 2017b)](https://paperpile.com/c/uaTuS8/zRzs4) and Minimap2 [(H. Li 2018)](https://paperpile.com/c/uaTuS8/VYCg). 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/uaTuS8/XtTjT). Reads mapping to a lower percentage identity were discarded. A 95 % identity was chosen as this is the general boundary used to delineate species.

## X.XX Visualisation of Binning Algorithims

Starting from the innermost, within Anvi’o a dendrogram is used to display the metagenome in with hierarchical clustering. Hierarchical clustering is the process where the closest reads are combined and group together. This is repeated until all reads are grouped into 1 group. In the next layer, the parent layer is included. Anvi’o by default splits contigs longer then 2500bps. With assemblies containing contigs longer than 2500bps, the parent tab allows for keeping track of contigs that have been assembled together. These by default are kept together. The next layer out contains the Guanine-Cytosine (GC) content. Species generally have the same GC content within their entire genome and rarely deviate from it. This provides an additional binning conformational method. The tracks in grey represent the samples and their coverage from the SAM files. The colours in black indicate a higher coverage at a particular point along the contig. The coloured layers on the outside of the anvi’o plot can be any additional label provided for each contig. In this case, taxonomy was provided by kaiju, with different colours along the taxonomy track aligning to different clades. The additional layers are binning software, where contigs that have been assigned to a bin labeled in colour, with the same colours being contigs within the same bin. As can be seen within Fig X.XX (figure above). Clear division lines are apparent in all tracks and form a consensus for where bin boundaries are, providing further evidence of successful binning.

## X.XX 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/uaTuS8/52Xi) 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/uaTuS8/52Xi) 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.

## X.XX 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/uaTuS8/8ooV), BinSanity [(Graham, Heidelberg, and Tully 2017)](https://paperpile.com/c/uaTuS8/lLtR), CONCOCT [(Alneberg et al. 2014)](https://paperpile.com/c/uaTuS8/3Fj3) and MaxBin [(Wu et al. 2014)](https://paperpile.com/c/uaTuS8/nrBi) 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.

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

BinSanity Results



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 Results



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



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.

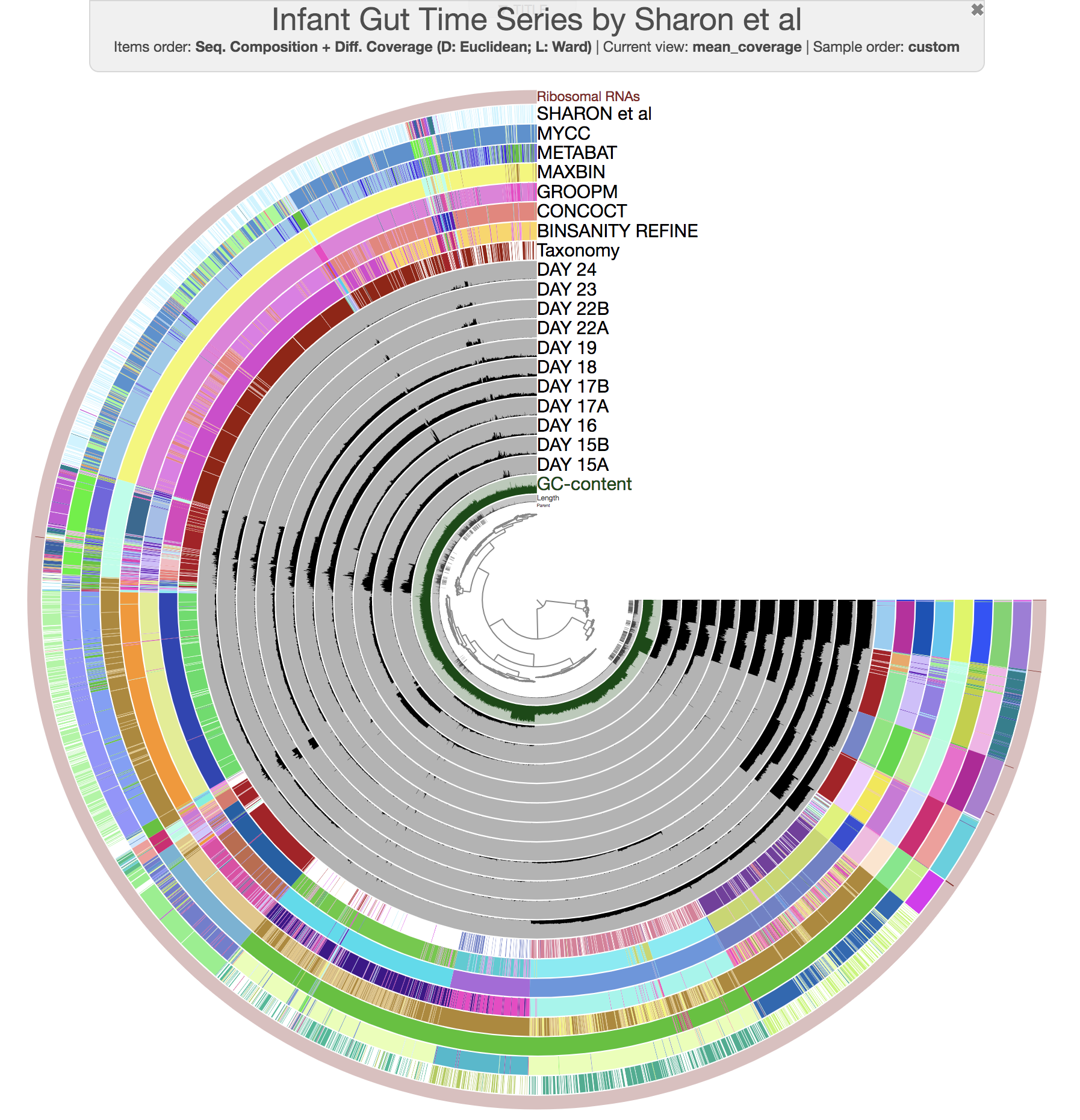
## X.XX Binning by taxonomy

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/uaTuS8/xp9N). 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/uaTuS8/xHpc) and allow for the automation of taxonomic classification of whole genome sequences. Centrifuge [(Kim et al. 2016)](https://paperpile.com/c/uaTuS8/TXx9), Kaiju [(Menzel, Ng, and Krogh 2016)](https://paperpile.com/c/uaTuS8/cjQS) and CAT/BAT [(von Meijenfeldt et al. 2019)](https://paperpile.com/c/uaTuS8/ByA6) 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/uaTuS8/uv25) 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/uaTuS8/gcnj) a visualisation platform for ‘omics data provides a great tool in its ability to visualise a whole metagenome and bins concisely.





## 

## 

## X.XX 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/uaTuS8/zgLC) 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/uaTuS8/M0Og) 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/uaTuS8/afFi) 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.

Settings for UMAP where explored mainly minimum distance (min\_dist) is how closely points are allowed to be plotted together. Since the goal was to cluster data, it was determined that the min\_dist would be set to 0 allowing for points of the same k-mer distribution to be superimposed on each other, with the hopes of tighter clustering of points. Metric (metric) was kept to euclidean as there was no indication data would not be on an euclidean metric.

Setting was the number of neighbours (n\_neighbours) was explored as this was what UMAP uses to understand the structure of the data. This is the number of neighbouring points UMAP looks at for each point, starting from the closest (eluidean) distance. A lower N\_neighbours would make UMAP look at more localised difference within the dataset, comparing how simular two points are. This would be benefical for small local changes but would be poor at looking at the global picture in how this point was related to a wider variety of points. In summary, a larger n\_neighbour would reveal global structure and a smaller n\_neighbour values how different two points are from each other. The n\_neighbours was explored and visualised below.





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.

* What is binning and why do we use it
  + Separation of genomes from metagenomes
    - Ability to look at species differences rather than population differences
    - Look at things like viral hosts changes rather than viruses that happen to be there
  + Allows us to look at indiv species rather than environment as a whole
* Show binning with ANVIO
  + Why doesn’t it work
* What are the binning methods and the pros and cons of each
  + Binning by coverage
  + Binning by similarity of taxonomy
    - Binning relies on multiple samples take and to bin by coverage, i.e. different samples from the same community are taken and things within sample are assumed to be of the same abundance therefore contigs that have same change in abundance ratio are the same species
    - Works well for simple communities, less so the more abundant the population
  + Other options include binning based on GC content, taxonomy of genes identified, kmer counting
    - GC content and Kmer counting, because organisms make genes in their own way, they have specific levels of GCs vs ATs across their genomes. Therefore species with same level of GC across genome should be the same. Kmer counting more complex one using all possible combination of 5 kmers and counts the number for each. Species should have similar levels of kmer count distribution.
* How do we check the quality of bins
  + Phylogenetics CheckM ortholog genes
    - Usage of gene taxonomy to establish identity
    - 16S vs whole genome MSA
  + Kmer counts and ANI
    - How does Kmer counting work
    - How does ANI work
  + CheckM universal orthologs
    - Selecting
  + N50 and L50
* Why are they so hard to pull apart

Short reads vs long reads vs hybrid assemblies

* Why are we using t-SNE plots and not other methods
  + What are the other methods and why do they work with other environments and not marine ones
  + What are t-SNE plots and how do they work and why do they work better with marine metagenomes

# 2.3 CHAPTER 1 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.

* Marine metagenomes are incredibly difficult to pull apart when using short-read lengths due to assemblers being unable to resolve a lot of the similar sequences
* Approaches to bin using coverage don’t work at all - don't know why?
  + Maybe too much content and mapping a contigs back to other samples is difficult as there are too many samples that are similar or not exactly the same as spades struggles a lot
* Long read and hybrid approaches remain good alternatives and help to resolve repeat and similar regions
* Lack of specialised long read metagenomics assemblers
* Assembly with long reads then co assembly with short reads or read polishing doesn't help
* Similar species may assemble together giving rise to chimeric species contigs that may look like a pan genome rather than individual species
  + Mapping reads back to contigs and re-assembling doesn’t show to reduce this problem
* Large amounts of contamination remain in bins indicate perhaps many genes of similar NT content are assembling together in clusters rather than species. May need a way of binning within bins
  + Binning within bins using ANI, GC content, Kmer content, taxonomic gene content or 16S doesn’t work. Why?

# 3.0 CHAPTER 2 - SAGs Phylogenetics and Genome analysis

## 3.1 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/uaTuS8/0pIME). Organisms usually have only 1 or a few copies of their genomic DNA [(Alberts et al. 2014)](https://paperpile.com/c/uaTuS8/h4GMK), 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/uaTuS8/gMcJN). 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/uaTuS8/XGgJU). Culturing this slow growing organism is difficult and SAGs provide an alternative to the genomic study of these ubiquitous organisms.

## 3.1 451 SAR11 SAGs isolated and sequenced

* Why SAR11 as “chosen/model organism”
  + 5 existing clades in SAR11
  + Very abundant in oceans
  + Look for phylum level genetic changes
* Why SAG is good choice of phylogenetics
  + Pure uncontaminated samples
  + Many of them
* What is phylogenetics
  + Why is it important
  + How is it usually done

# 3.2 CHAPTER 1 METHOD

## 3.2.1 Preprocessing

* Explicit on method pre- bioinformatics/analysis
* What type of MDA reaction was used
* Where and how was my SAR11 obtained

## 3.2.2 Quality control on reads

* Trim\_galore
* Bbmap suite for quality of reads
* Assembled using SPAdes
  + Has single cell mode
  + Best assembler
    - Try with other assemblies, best one based on CheckM
    - CheckM used what parameters
  + Settings for SPAdes
    - Resulted in X results in completeness and contamination
      * N50 number
      * % reads mapping back to genome

SAGs obtained from XXX where subjected to quality control via adaptor removal and read quality checks. Quality control on samples were performed via an inhouse script (<https://github.com/ash-bell/masters_thesis/blob/master/QC_reads.smk>) using the BBTools suite written by Brian Bushnell [(Bushnell, Rood, and Singer 2017a)](https://paperpile.com/c/uaTuS8/1wta). Briefly, low-quality regions are removed, adaptors are trimmed and error correction is performed on reads. After quality control, reads are assembled with specialist bioinformatics assembler SPAdes [(Bankevich et al. 2012)](https://paperpile.com/c/uaTuS8/OwCT).

## 3.2.2.1 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/uaTuS8/52Xi). 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.

## X.XX Assembly analysis

Assemblies were put into Bandage [(Wick et al. 2015)](https://paperpile.com/c/uaTuS8/2FiTJ), a program for visualisation of assembly graphs to review the assembly quality along with CheckM.

## 3.2.2.2 Genomic analysis

Genomic analysis of these SAGs was performed to look at gene coding regions. Due to the lack of completeness for most SAGs, only the top 90% was analysed as the rest of the SAGs were missing large proportions of its genetic material. To perform the genetic analysis Prokka [(Torsten Seemann 2014)](https://paperpile.com/c/uaTuS8/Ixvr), a prokaryotic genome annotation was use. Default search settings were used to produce annotated genbank files, however additional annotation options were added to complete the genbank file. Setting can be seen here (<https://github.com/ash-bell/masters_thesis/blob/master/prokka.sh>) An in house script (<https://github.com/ash-bell/masters_thesis/blob/master/genome_visualisation.py>) was used to visualise the genbank files in a visual format via BioPython [(Cock et al. 2009)](https://paperpile.com/c/uaTuS8/vFta).

## 3.2.3 Phylogenetics

* SAR11 is a class level clade distinction? Level clades.
  + Where do my SAGS sit in clade
  + 2 options, 16S and WGS MSA
    - Pros and cons of 16S
      * + Easier to obtain reference
        + Inaccurate
      * 16S NT seq aligned via MSA using mafft
        + Only 384 16S sequences found
        + Gaps in MSA removed via trimAL
        + NT Tree made in R via iqtree
        + Bootstrapping
        + Tree displayed in iTOL
      * WGS used gtdbtk
        + Uses protein coding region similarity to make protein MSA
        + AA Tree made via iqtree after gap removal in trimAL
        + Bootstrapping and display in iTOL
      * Align trees to each other
        + Still significant difference
        + Loss of a lot of genomic content due to SAG method
        + No consensus
      * Alternate method
        + ANI

All vs all ANI heatmap indicate groups similar to clades

Hard to visualise and pull out

* + - * + Kmer

Similar method

Kmer count of each fasta file condensed from multi-fasta to single fasta

Plot kmer count after normalisation to account for short/longer genomes

Transform normalisation?

Plot using t-SNE scatter plot

Dimension reduction

Indication of clustering

HDBSCAN pulls out clusters

16S and WGS same clusters?

ANI within bins to confirm similarity

Re-phylogenetic tree to see phylogeny within clade

No consensus

Binning of multiple references of different clades together

Testing of kmer method via splitting reference genomes into pieces revealed differing kmer content on different parts of genomes in some references

Possible references with high coverage = good kmer consensus

References like Clade 1a SAR11 have assemblies with less consensus and therefore more variation in genome?

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/uaTuS8/Spkl) 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/uaTuS8/7QM6) and visualised in iTOL [(Letunic and Bork 2007)](https://paperpile.com/c/uaTuS8/6N7y), a web based phylogenetic tree viewer.

Generally, phylogenetics measured by how related a species is to each other by a phylogenetic tree. Similarly grouped species sit within the same “branch” of a phylogenetic tree. More distantly related species are then located further away. To establish relatedness, conserved genes of all species like the 16S rRNA is used. This is because all Eukaryotes have a form of this 16S gene within their genome, and comparisons of similarity can then be performed between such individuals. A lack of conserved genes would prove difficult to measure species relatedness as there would be no method to ascertain species similarity on a genomic level. However, more closely related organisms are likely to have similar if not the same genes, as such genes are required to perform a specific function characterised by said organisms. This provides additional choices for genes to assess species relatedness and can be used in conjunction with multiple conserved genes to provide a broader assessment of relatedness. However, this needs to be used in consideration of outgroups. Outgroups are species defined as not simularily related to the samples in question and are used to visualise the tree in the existing tree of life. They can be used as “anchors” of sorts, linking the tree within the bigger picture of existing life forms. These can also be used in sorts as negative controls. Outgroups are picked based a some similarity but still major differences between samples. Should samples be located on the same branch or similar to outgroups, this would indicate samples may be contaminated, or less closely related within samples than previously thought. This performs as quality control for sequences and helps to anchor trees, whereas ingroups perform a similar function.

Ingroups are used positive controls and represent the species as a whole. These usually are existing samples or reference genomes of the expected samples and performs 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 be closely related. These are used to confirm sample identity.

Bootstrapping is an important metric in phylogeny and is used to ascertain how certain a branch within a tree is correctly placed. It is done by replicating proportions of the data and recording how often a branch is placed in the same location. Generally, a smaller proportion of the data is taken and the branch location is deduced. It is then recorded if the smaller proportion of the data is used, where is the branch placed. This can be done as many times as possible and is normally expressed as a percentage. It provides a metric to how sure the location of the branch is based on smaller variations with the data. Bootstrapping is normally expressed as a whole number overlayed onto the phylogenetic tree, with branches below a certain percentage collapsed to show unsurity of the low percentage branch. Accepted bootstrapping values expressing confidence of a branch are subjective and mostly depend the data provided. With data of low resolution or with multiple gaps, a 30 BS value would indicate that almost a third of replicates indicate this structure. However, with a multi-gene phylogeny with thousands of amino acid alignments and <90 BS would highly indicate a sign of contradiction. But generally accepted thresholds can include >70 BS as a “good” value and >50 BS as moderate. Values <50 are generally regarded as ambiguous and collapsed.

Parsimony is a principle similar to Occam’s razor, basically if all else is equal, the simplest explanation and more likely to be correct in comparison to more complex solutions. Maximum parsimony in phylogenetics also follows this principle where the tree with the simplest solutions is probably the most correct one. Since it is highly difficult to calculate the most parisony tree or the maximum parisony for said data, instead multiple trees using different models are made the parsimony calculated for each. The tree with the highest parsimony score is then selected and used as the final tree. However, because parisony rewards the smallest and simplest of trees, this may not accurately represent true evolution which maybe more complex. But without actually observing said evolution, it is impossible to deduce the exact evolution and makes finding the actual evolutionary change highly difficult.

With SAGs, Whole Genome Sequences (WGS) are provided to perform phylogenetic analysis. Therefore multi-gene phylogenetic seems like an attractive option considering the advantages of a broader analysis on a multitude of genes. However, SAGs suffer from GC bias during the amplification step, resulting a genomes with high coverage on proportions of their genomes and none in others. In fact on average, most SAGs are 40% complete (XXX fact check) and randomly amplified, providing difficulty in finding multi conserved genes. Additionally, in groups representing SAR11 clades may not have complete genomes and further limit choices of multi-gene targets. Alternative methods like 16S phylogeny where only the 16S rRNA gene is selected provides the same issue. Without the presence of this gene, comparisons within samples are difficult. However, reference ingroups provided online only provide the 16S gene, and therefore 16S phylogeny in conjunction with multi-gene phylogenetics is used.

A Multi Sequence Alignment (MSA) is a string of characters align according to each other. This is usually done to allow the comparison of the same gene from multiple samples. For example, 16S rRNA genes are isolated from samples and its nucleotide composition aligned against all other samples. Variations and similarities within the nucleotide sequence would reveal relatedness, where samples with a similar composition of nucleotides expected to be increasingly similar, as species with the same genetic sequences are regarded as the same. Alternatively, amino acid sequences can be used instead as protein coding regions are translated into their amino acid composition and aligned against each other. This is normally results in less variation within closely related species as variations within nucleotides compositions can still result in the same amino acid structure, as amino acids have more than one translated codon. However, smaller changes within nucleotides are not captured and may indicate slight differences between similar strains. Here, 16S rRNA phylogeny is performed at a nucleotide level to capture smaller nucleotide diversity and relatedness where multi-gene phylogeny is performed using amino acid composition to reduce computational time.

16S rRNA sequences was extracted using barrnap version 0.9 [(T. Seemann 2015)](https://paperpile.com/c/uaTuS8/PkcQ) with default settings and parameters. For a multi-gene phylogeny, the program GTDB-TK [(Parks et al. 2018)](https://paperpile.com/c/uaTuS8/Spkl) is used to produce a multi-gene MSA. Briefly, GTDB-TK works by identifying marker genes through prodigal [(Hyatt et al. 2010)](https://paperpile.com/c/uaTuS8/uv25) conserved by all bacteria/archaea, aligning marker genes to Hidden Markov Models (HMM) using HMMER [(Eddy 1998)](https://paperpile.com/c/uaTuS8/rSCv) and concatenation of these marker genes into an amino acid MSA. GTDB-TK also uses pplacer [(Matsen, Kodner, and Armbrust 2010)](https://paperpile.com/c/uaTuS8/yCuM) to place each genome within a pre-constructed tree based on maximum-likelihood. However this was not performed as construction of trees via IQ-TREE was preferred allowing for selection of the most parsimony tree.

A MSA file was created via MAFFT [(Katoh et al. 2002)](https://paperpile.com/c/uaTuS8/i4Xz) 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 were predicted to be similarly related species and therefore global over a local alignment was prefered. 1000 iterations was also prefered to ensure the best MSA outcome was used whilst still being computationally reasonable.

Other MSA algorithms MUSCLE [(Edgar 2004)](https://paperpile.com/c/uaTuS8/Z7de) and T-Coffee [(Wallace et al. 2006)](https://paperpile.com/c/uaTuS8/BTsp) were also used and compared. It was deduced MAFFT produced the best MSA through comparison with the tool trimAl [(Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009)](https://paperpile.com/c/uaTuS8/p8Wm). TrimAl removes poorly aligned regions within an MSA which can help with accuracy of future phylogenetic tree analysis. 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)](https://paperpile.com/c/uaTuS8/7QM6). There remains multiple algorithms to plot phylogenetic trees and with hundreds of options, it remains difficult to select the best tree, especially where there are hundreds of branches within each. IQ-TREE’s strength allows for comparisons against all these trees by calculating the parsimony score for each tree. Each score for each tree is recorded, along with its repeat iterations and most parsimonious tree is produced at the end. 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)](https://paperpile.com/c/uaTuS8/6N7y), an online phylogenetic tree viewer that allows detailed customisation of branch colours and trimming of lengths.

Currently the SAR11 clade is split into 5 existing groups with multiple sub-clades [(Thrash et al. 2014)](https://paperpile.com/c/uaTuS8/HdXVu). 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/uaTuS8/3ODS) 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.





Fig X.XXX 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 indicated a collapsed node containing additional branches. Nodes with bootstrapping values under 90 are collapsed to reduce ambiguity. Outgroup is coloured in grey.

Overall, it is obvious that multi-gene phylogeny is likely to be more accurate in its placement of organisms within a phylogenetic tree, also taking into advantage that the 16S gene is not required to be present, just a combination of some conserved genes within each organisms. This allows for not only longer MSA files where smaller local alignment differences can assist in further differentiation of each organisms. The main drawback to this method being the large computational time require to produce trees where MSA files of potential 5000 amino acids are used. Although trimming of such regions can be used to filter out poor quality regions with little or no variation, these still are computationally intensive. Although hypervariable regions are present within 16S rRNA, these hypervariable regions are unlikely to have large variations within species boundaries, and are usually used for differentiation of high taxa with larger variation within these hypervariable regions.

An advantage of 16S phylogeny is the presences of the 16S gene within all bacterial and archaea. This allows for comparisons of distantly related species but within closely related organisms, loses its advantage. However, with species poorly available in public databases, often only 16S genes are available due to the popularity of 16S genes as a genetic marker. The 16S region is highly conserved and therefore amplification and sequencing of this region is well documented. An example of this is clade 2 and 4 of the SAR11 phylogeny is only represented as 16S genes within the NCBI database, without their WGS. This is why within this report both the 16S rRNA phylogenetics and multi-gene phylogenetics are used for classification of SAR11 SAGs.

Through this tree it maybe to then extrapolate and combine branches were trees are similar. Where 16S branches align with the same ingroups in the WGS branches, with a high bootstrapping of >90 BS, it would be reasonable to suggest that these organisms are correctly identified.



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/uaTuS8/0ZFt) 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. FastANI used default settings, with script visible here ([https://github.com/ash-bell/masters\_thesis/blob/master/fastANI.sh](https://github.com/ash-bell/masters_thesis/blob/master/QC_reads.smkhttps://github.com/ash-bell/masters_thesis/blob/master/fastANI.sh)) A heatmap allows for a gradient of colour, showing clusters that are less or more closely related on a discontinuous scale (<https://github.com/ash-bell/masters_thesis/blob/master/ani_plot.py>).



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.

## X.XX kmer counting

Although plotting via ANI is very useful, delineated species boundaries on a discontinuous scale is subjective and difficult. Arbitrary delinearisation exist, with >95% dictating an intra-species relationship and <83% an interspecies relationship. However, with SAGs, where multiple genes are missing, this results are deduced based on small amounts of genes in common and can exacerbate bias in certain genes for and against these boundaries. Kmer counting provides an alternative method for categorisation and binning of SAGs. A k-mer refers to all the possible combination of letters or numbers with an existing string of characters. The k refers to a number and is the length of the chosen substring. Therefore if k was to equal 4, the k-mer count be referred to a 4-mer and all the possible combination of 4 characters within a string of letters or numbers would be the k-mer.

**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**

K-mer counting would then refer to the number of times each set of k-mer appears. This would be tabulated and given a value of each string of letters and numbers. Species normally have the same k-mer count throughout their genome and this can be used to classify similar genomes together. This process is used with the SAGs where a 5-mer count of each genome produces a set of values defining it k-mer composition and abundance. This is normally normalised, allowing for genomes of shorter or longer lengths to be compared as differing length can lead to differing numbers of k-mer counts. With these numerical counts, genomes can now be plotted in relation to each other to produce clusters where normalised k-mer counts of the same composition can be group together. This produces clustering on plot which can be then group and binned.

**Sequence:**

**AGATGCATGCA**

**Resulting 4-mers:**

**AGAT, GATG, ATGC, TGCA, GCAT, CATG**

**K-mer count:**

**AGAT: 1**

**GATG: 1**

**ATGC: 1**

**TGCA: 1**

**GCAT: 2**

**CATG: 2**

With k-mer counting, the larger the chosen k exponentially increases the number of k-mers. A 4-mer will produce 4^4 where there are 4 possible combinations of letters, ATGC and ^4 possible combinations if using a sting of 4 letters. This produces 256 possible combinations of k-mers, where a 5-mer would be 4^5 1024 (XXX check this) possible k-mers. With this many individual data points, it would be impossible to plot this on a 2 or 3D plot. Instead, tools for dimention redctionality are used to aid visualisation. BH-t SNE is used to reduce these counts down into 2 dimension. For this UMAP [(McInnes, Healy, and Melville 2018)](https://paperpile.com/c/uaTuS8/afFi) is used. This allows for visualisation of each SAG within a 2D plot. Clusters can be automatically determined via [(McInnes, Healy, and Astels 2017)](https://paperpile.com/c/uaTuS8/zgLC) which auto makes clusters. The resulting bins can be compared with ANI techniques and traditional phylogenetic trees to confirm a consensus identity.



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.



It becomes clear that kmer counting is not the perfect solution as there exists variations within k-mer counts of differing contigs throughout the genome, but this is not distributed randomly. Clear clusters are still apparent, but proportions of the genome remain different k-mer count wise. It also matter species to species as marine organisms 7211 and 1062 had a more varied k-mer count in comparison to *Clostridium difficile* who’s entire cluster contained all of its genomic content. Therefore k-mer counting would be seen as more effective for different bacteria, and apparently less so for marine organisms like SAR11.



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.

# 3.3 CHAPTER 2 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.

### Problems with MDA

* + Makes Chimeric artifacts - links non contiguous sequences
  + Results in biased coverage - highly variable, some none, some a lot
  + Over and under amplification of random regions
  + Even x1000 sequencing depth on average less than 50% of genome recovered (<https://www.nature.com/articles/s41467-017-00128-z>)
    - Solved by pooling same cell type
    - Using FISH
    - Post-sequence confirmation
  + Bias against high GC% regions
    - Use thermostable polymerases?
  + SNPs pose problems sequencing
    - Limited DNA extracted from cell ~80% genome coverage
    - Copy Number Variation and trying to assemble this with low coverage

### Improvements to MDA:

#### pico/nanoliter reactions suppress amplification bias

* + - Perform MDA in very very small quantities reducings amplification bias (<https://www.nature.com/articles/s41467-017-00128-z>
    - Perform on agarose gel - reduce amp bias (<https://www.nature.com/articles/s41467-017-00128-z>)
    - Protein priming - reduce amp bias (<https://www.nature.com/articles/s41467-017-00128-z>)
  + No method to reduce bias against high GC content

#### PicoPLEX (https://www.nature.com/articles/s41467-017-00128-z)

#### MALBAC (both these alternate methods reduce amplification bias (<https://www.nature.com/articles/s41467-017-00128-z>)

* + - But require thermocycling, multi-stage set up - basically MDA is one reaction on the lab bench and the rest need PCR and therefore increase chance of messing up/contamination
* Phylogenetic trees between 16S and WGS MSA prove to be quite different
  + Long sequences can show more variation and therefore more accurate trees?
  + Similar 16S may still indicate large differences in gene content
* Bioinformatics methods

#### Usage of metagenomic composite material as internal reference

#### Combining data = higher chance of contamination

#### Loss of actual data like SNPs or species specific genes?

* + Combine similar strains together to get WGS via ANI, phylogenetics, kmer count … etc
  + Diminishing returns. May need a lot of SAGs to get >95% completion but may need 2 to get over 50%
  + Random guess which SAG have complimentary reads other SAG is missing unless mapping to reference. If don’t know what SAG is, impossible to do.
  + Difficult to get over repeat regions with short reads, so use long reads or hybrid?

# X.XX Chapter 3 - SAG viral signatures

## X.XX What are viral signatures

It has long been established that viruses have a large impact on ecosystems [(Fuhrman 1999)](https://paperpile.com/c/uaTuS8/6Mri3). They are key factors in regulating bacterial and eukaryotic microorganism populations [(Curtis A. Suttle 2007b; Wommack and Colwell 2000)](https://paperpile.com/c/uaTuS8/RPZpI+21Amp). 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/uaTuS8/HbFXh+Zyzyh). Within single-cell organisms, they are responsible for the turnover of over 20 % of marine bacteria biomass daily [(Fuhrman 1999)](https://paperpile.com/c/uaTuS8/6Mri3) 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/uaTuS8/Syi6D+lQ1LD+8UqYu+eJzbY). 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/uaTuS8/k57He) and some bacterial genomes have up to 20 % of their genomes from viral origins [(Casjens et al. 2000)](https://paperpile.com/c/uaTuS8/KpbEH). 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/uaTuS8/SIG5k+J1KHm+hY8yT+BD84V). It has also been hypothesized this is how transfer of antimicrobial resistant or pathogenic genes [(Waldor and Friedman 2005)](https://paperpile.com/c/uaTuS8/Dv9J1) 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/uaTuS8/KFN2n).

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/uaTuS8/FGBHH). 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/uaTuS8/HpcVB+u5rXh). 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/uaTuS8/HpcVB+9jkm5+0s0nT+qe7u6+VcyBh+zWajT).

X.XX Why are they important

X.XX What is their impact

X.XX How do we find them

X.XX Chapter 4 - Method

What are the options and how they can go about it

### Extraction of viral signatures by VirSorter

VirSorter uses MetaGeneAnnotator [(Noguchi, Taniguchi, and Itoh 2008)](https://paperpile.com/c/uaTuS8/SwpbW) and hmmsearch [(Finn, Clements, and Eddy 2011)](https://paperpile.com/c/uaTuS8/LTz3) 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/uaTuS8/BDvmz+hFBUt) 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.

# X.XX CHAPTER 4 - SAGS and Variable region

* Grote et al paper indicate variable region between SAG 23s and 5s region in all SAR11 samples
  + This may indicate viral - cell relation
  + Lots of SAR11 genome from multiple clades would be able to confirm similar findings and allow for analysis of genes within region
    - What genes are hypervariable
    - How this would impact prey-predator relations
  + Q1 - Are there variable regions and are they only located between 23s - 5s
    - Could be multiple or none

### Ecological models

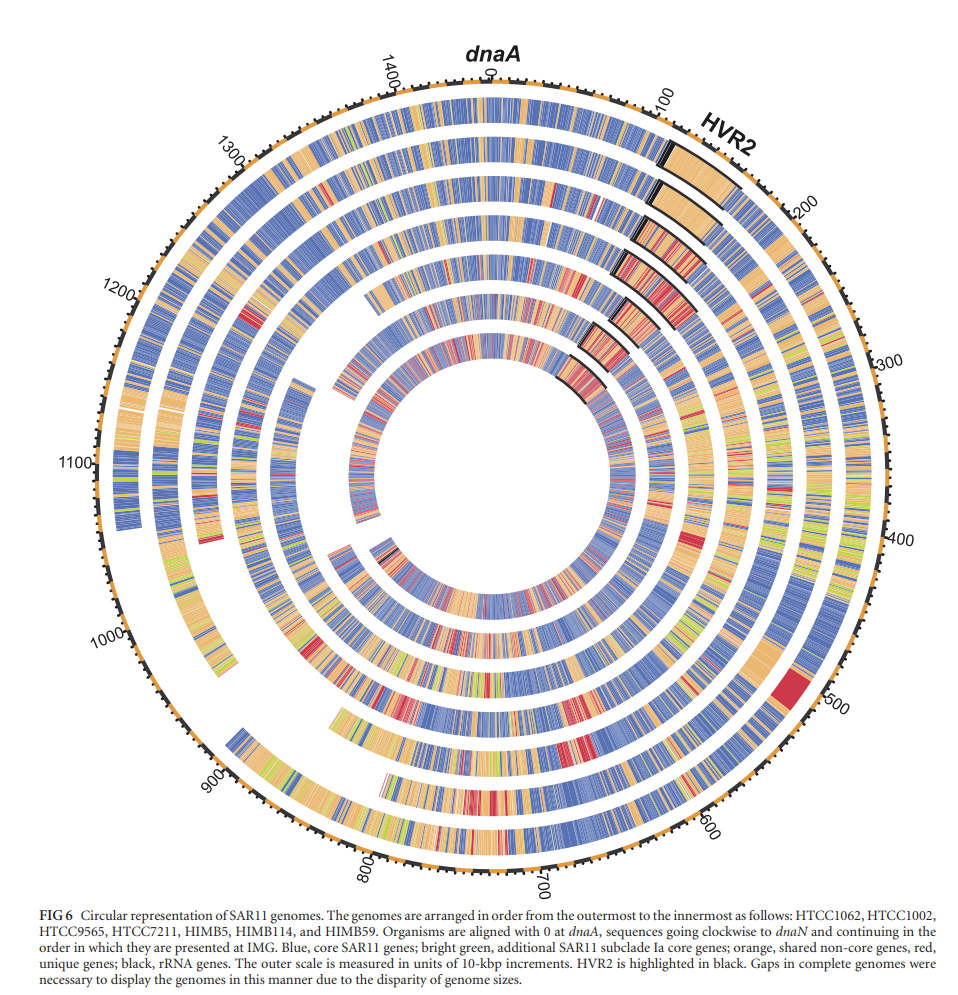
There are a variety of ecological models that explore the predator prey dynamic. Kill the Winner (KtW) hypothesis states that in the presence of a high density of prey, predator density will increase to match prey density and eventually bring predatory - prey density into equilibrium [(Avrani, Schwartz, and Lindell 2012; Winter et al. 2010)](https://paperpile.com/c/uaTuS8/CMSaC+4HCEW). An example of this is the regulation of gut microbiomes by phages [(Letarov and Kulikov 2009)](https://paperpile.com/c/uaTuS8/adSjP). 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/uaTuS8/h9bPp+SyDnf). An example of this is the diversity of bacteria populations maintained by viral predation [(Rodriguez-Valera et al. 2009)](https://paperpile.com/c/uaTuS8/dmOBk). 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/uaTuS8/Ut9qr). 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/uaTuS8/XGgJU). 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/uaTuS8/qwwUZ+rtko0). 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/uaTuS8/vkFQh).

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/uaTuS8/3ODS): 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/uaTuS8/3ODS). 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/uaTuS8/j4Ip), 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.

# 5.2 CHAPTER 4 METHOD

* + Map SAGs against biller metagenomes at 76% identity
    - Any SAR11 within biller would map but not to VR regions
    - Peaks and dips within coverage data would indicate VR region
    - Low coverage in VR region may indicate niche specific adaptation. I.e. only 500m SAR11 have this gene from biller data
    - Usage of thingy thing distribution on biller coverage table. If all coverage coming from below 5 sources, be able to tell this and see which 5 and if they have any relation to each other.
    - Didn’t run an all SAG vs SAG because can show class level different from each other but not niche specific. Better to see where biller maps then see which SAGs are from where
    - Because ends of contigs more likely to have low coverage due to less reads being able to map beyond the ends, either say within 250bp of contig end (size of biller read) = mean coverage for contig as unsure.
      * Alternative is to use Ns and allow for ambiguous mapping, but may lead to higher mapping then normal.
    - Usage of MAUVE to align contigs between contigs
      * Allow for genome alignment between multiple SAGs

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.

SAR11 SAGs were assembled quality controlled and assembled as per the previous chapter. First, assemblies needed to be aligned to each other. Although mapping could be performed without an all vs all alignment and coverage collated per contig, areas at the start and end of contigs would be artificially reduced in the number of mappings possible due to a non-contiguous sequence. These areas would then be artificially low in coverage and may provide a false positive for low coverage and therefore HVR. Alternatives include changing contig start and ends to median contig coverage to counteract this. However, this still posed the problem of having to compare each contig to each in an all vs all contig comparison. With some SAGs being poorly assembled with hundreds of contigs, this was deemed infeasible. Instead, assemblies were mapped to a reference genome, here chosen as HTCC7211. This provided ordering of contigs, allowing for alignment of contigs within an assembly to be the same in all SAGs, allowing coverage to be compared. However, with differing contig sizes, numbers and overlaps, this meant that contigs still needed to be placed aligned to each other, not one after each other. I.e. its contigs needed to be spaced apart so they aligned in base pair to base pair to simular contigs. To do this, Ns or non mapping nucleotides were introduced between contigs. As this contigs were mapped to HTCC7211, Ns between contigs were introduced to space apart this contigs from each other so they matched their mapping proportion of HTCC7211 and therefore by extension each other. This would allow for clades of SAR11 to be mapped against a suitable reference.

# 

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.





# 5.3 CHAPTER 4 DISCUSSION

* Is this an effective method to look for variable regions?
* Not really a discussion for this part yet as I have no findings
* SAGs are missing large parts of their genome, so not very good for micro scale studies
* Better to combine multiple together? But then lose VR region even if very closely related???
* What is the gene content in these regions?

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.

# 6.0 THESIS DISCUSSION

* SAGS and Metagenomics are both valid ways of studying marine populations. Each have their own drawbacks and strengths.
* SAGS are better at looking and WGS, metagenomics is better at looking at population dynamics
* Combination of the two allows for almost any kind of study into a marine environment. Both benefiting from the need not to culture samples.
  + Cannot study population dynamics in situ - changes in environment
* Only a snapshot in time, not a continuous trace on it changes
* If metagenomics can be separated on a per strain level, may remove need have SAGs unless only wanted to study one strain

**OLD PROPOSAL FOR RESEARCH PROJECT UNDER BT FOR REFERENCE**

**Title of Research Project:**

The impact of diel cycles and viral lysogenic lifestyles on marine cellular organisms.

**Background to research leading to research question**

REWRITE THIS + PAPERPILE IT

“””

With the earth’s oceans covering roughly 70% of the earth surface, it heavily influences the earth’s climate and provides huge amounts of energy globally that is consumed by a myriad of organisms. These microorganisms with the earth’s oceans produce up to half the earths oxygen and consist of 90% of the oceans biomass (Suttle, 2007). They are critical in providing nutrients and the foundation of marine food webs and the energy cycle. This seems to be kept in check with viruses killing at least 20% of this biomass daily (Suttle, 1994). This biomass consists mainly of heterotrophic and eutrophic plankton (Proctor and Fuhrman, 1990), where subsequent virioplankton levels seems to mirror increases and decreases in these plankton levels describing a predator-prey like relationship (Parsons et al., 2012). Theses viruses also hold previously unexplored prokaryotic genetic diversity and its relationship with both physical and biological factors key into understanding marine biome’s population dynamics (Suttle, 2007).

The Bermuda Atlantic Time-series Study (BATS) site in the Sargasso Sea (Michaels and Knap, 1996) provides a unique opportunity to understand the virosphere and its relationships with physical and biological factors. The BATS is a long-term study site in a seasonal oligotrophic system. It records biological, chemical and physical data monthly and has been doing so for decades (Giovannoni and Vergin, 2012). Biological samples taken from this at differing depths can help us understand viral abundance in response to differing conditions. Previous studies have been successful in identifying viral abundance and classification (Parsons et al., 2012). This is important in understanding the ecological impact of such viruses. Its temporal and spatial variability in context with other physical and biological parameters.

Diel cycles also provide a critical parameter where viral abundance changes in the short term (Winter, Herndl and Weinbauer, 2004). A previous study (unpublished data) took samples at 8-hour period over a 72-hour timeframe. Subsequently, 24 metagenomes were established from the Sargasso Sea, separated into 12 cellular and 12 viral communities via 0.2um filter. The viral fraction is already undergoing analyses, however a comparison against the cellular fraction would be critical in understanding viral lifecycles. This is because virus can undergo lysogenic lifecycles (Suttle, 2007). The cellular proportion may contain viral sequences within its genome and it would be interesting to see if these proportion changed based on diel cycle, as well as the identity and source of these viral sequences. Overall this would look at how viral contigs within the cellular fraction compare to the viral free communities.

“””

**Research Questions and context - expand upon my research question**

With the cellular fraction I would seek to:

* Recovery and assembly of cellular genomes from the metagenomic study.
  + This would consist of the assembly of the genomics of the cellular fraction using both short and long reads in combination to provide assemblies.
* Abundance and classification of cellular organisms and separation of viral and viral free cellular genomes.
* Identify and abundance of viral contigs within these genomes.
* Change in identify and abundance of viral and non-viral integrated genomes based on diel cycle, biological and physical data from BATS site.
* Comparison of free-viral fraction against viral contigs present in cellular fraction.
* Comparison of short read and hybrid metagenomic assembly for the cellular fraction.

**Significance of study**

Results from this study would first look at the applicability of using both short and long reads to assemble cellular genomes from a metagenomic study. Subsequent assembly would allow for an insight into the proportion of the viral community that undergoes a lysogenetic lifecycle under differing environmental conditions in the Sargasso Sea.

**Research method proposed along with timeline**

The recovery and assembly of viral and cellular genomes using short and long reads using method already established by Dr. Ben Temperton’s group. Assembled and co-assembled contigs will be binned into ‘Metagenomic Assembled Genomes’ using Metabat (Kang et al., 2015) and analysed for viral sequence in the cellular fraction using VirSorter (Roux et al., 2015) and VirFinder (Ren et al., 2017). Identification of tRNAs and CRISPR regions within the MAGs will enable identification of host-virus interactions between the cellular and viral fractions. Relative abundance of both cellular and viral sequences over a diel cycle will be assessed by read mapping.

**Process (results) including changes to method**

* Short read cellular contigs explored in anvio
* Reads mapped back to assembly for binning by coverage. With multiple tools including GroopM, Metabat … etc show largely fragmented bins with no overall consensus. Maxbin, binsanity
* Binning based on taxonomy, usage of centrifuge and kaiju database to blast genes against taxonomy database to identify host shows largely fragmented assemblies.
* Usage of HMMs pFams and COGs
* Usuage of hierarchical clustering
* CheckM for how good bins are
* CAT/BAT for classification
* Dustmaker
* vizbin
* Combining the 200m reads together + 80m co-assembly with spds - error correction seperate
* Long reads assembly with flye, canu, spades, wtdbg2
* Read poishing
* Vizbin - gtdbtk +checkm bins

**Bibliography**

Giovannoni, S. J. and Vergin, K. L. (2012) ‘Seasonality in ocean microbial communities’, Science, pp. 671–676. doi: 10.1126/science.1198078.

Kang, D. D. et al. (2015) ‘MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities’, PeerJ, 3, p. e1165. doi: 10.7717/peerj.1165.

Michaels, A. F. and Knap, A. H. (1996) ‘Overview of the U.S. JGOFS Bermuda Atlantic Time-series Study and the Hydrostation S program’, Deep Sea Research Part II: Topical Studies in Oceanography, 43(2–3), pp. 157–198. doi: 10.1016/0967-0645(96)00004-5.

Parsons, R. J. et al. (2012) ‘Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea’, ISME Journal, 6(2), pp. 273–284. doi: 10.1038/ismej.2011.101.

Proctor, L. M. and Fuhrman, J. A. (1990) ‘Viral mortality of marine bacteria and cyanobacteria’, Nature, 343(6253), pp. 60–62. doi: 10.1038/343060a0.

Ren, J. et al. (2017) ‘VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data’, Microbiome, 5(1), p. 69. doi: 10.1186/s40168-017-0283-5.

Roux, S. et al. (2015) ‘VirSorter: mining viral signal from microbial genomic data’, PeerJ, 3, p. e985. doi: 10.7717/peerj.985.

Suttle, C. A. (1994) ‘The significance of viruses to mortality in aquatic microbial communities’, Microbial Ecology, 28(2), pp. 237–243. doi: 10.1007/BF00166813.

Suttle, C. A. (2007) ‘Marine viruses - Major players in the global ecosystem’, Nature Reviews Microbiology, 5(10), pp. 801–812. doi: 10.1038/nrmicro1750.

Winter, C., Herndl, G. J. and Weinbauer, M. G. (2004) ‘Diel cycles in viral infection of bacterioplankton in the North Sea’, Aquatic Microbial Ecology, 35(3), pp. 207–216. doi: 10.3354/ame035207.

# THESIS BIBLIOGRAPHY

[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/uaTuS8/VcyBh)[10.1093/nar/gks406](http://dx.doi.org/10.1093/nar/gks406)[.](http://paperpile.com/b/uaTuS8/VcyBh)

[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/uaTuS8/h4GMK) <https://market.android.com/details?id=book-AGebDwAAQBAJ>[.](http://paperpile.com/b/uaTuS8/h4GMK)

[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/uaTuS8/3Fj3)[10.1038/nmeth.3103](http://dx.doi.org/10.1038/nmeth.3103)[.](http://paperpile.com/b/uaTuS8/3Fj3)

[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/uaTuS8/xp9N)[10.1016/S0022-2836(05)80360-2](http://dx.doi.org/10.1016/S0022-2836(05)80360-2)[.](http://paperpile.com/b/uaTuS8/xp9N)

[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/uaTuS8/eJzbY)[10.1126/science.1252229](http://dx.doi.org/10.1126/science.1252229)[.](http://paperpile.com/b/uaTuS8/eJzbY)

[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/uaTuS8/9jkm5)[10.1093/bib/bbx121](http://dx.doi.org/10.1093/bib/bbx121)[.](http://paperpile.com/b/uaTuS8/9jkm5)

[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/uaTuS8/CMSaC)[10.4161/mge.20031](http://dx.doi.org/10.4161/mge.20031)[.](http://paperpile.com/b/uaTuS8/CMSaC)

[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. https://doi.org/](http://paperpile.com/b/uaTuS8/OwCT)[10.1089/cmb.2012.0021](http://dx.doi.org/10.1089/cmb.2012.0021)[.](http://paperpile.com/b/uaTuS8/OwCT)

[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/uaTuS8/M0Og) <https://www.biorxiv.org/content/10.1101/619684v1.abstract>[.](http://paperpile.com/b/uaTuS8/M0Og)

[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/uaTuS8/j4Ip)[10.1038/sdata.2018.176](http://dx.doi.org/10.1038/sdata.2018.176)[.](http://paperpile.com/b/uaTuS8/j4Ip)

[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/uaTuS8/BDvmz)[10.7717/peerj.3243](http://dx.doi.org/10.7717/peerj.3243)[.](http://paperpile.com/b/uaTuS8/BDvmz)

[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/uaTuS8/0s0nT) <https://www.ncbi.nlm.nih.gov/pubmed/16922685>[.](http://paperpile.com/b/uaTuS8/0s0nT)

[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/uaTuS8/Syi6D) <http://www.jstor.org/stable/24860053>[.](http://paperpile.com/b/uaTuS8/Syi6D)

[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/uaTuS8/h9bPp)[10.1098/rspb.2014.1382](http://dx.doi.org/10.1098/rspb.2014.1382)[.](http://paperpile.com/b/uaTuS8/h9bPp)

[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/uaTuS8/xHpc)[10.1038/nmeth.3176](http://dx.doi.org/10.1038/nmeth.3176)[.](http://paperpile.com/b/uaTuS8/xHpc)

[Bushnell, Brian, Jonathan Rood, and Esther Singer. 2017a. “BBMerge – Accurate Paired Shotgun Read Merging via Overlap.” *PLOS ONE*. https://doi.org/](http://paperpile.com/b/uaTuS8/1wta)[10.1371/journal.pone.0185056](http://dx.doi.org/10.1371/journal.pone.0185056)[.](http://paperpile.com/b/uaTuS8/1wta)

[———. 2017b. “BBMerge--Accurate Paired Shotgun Read Merging via Overlap.” *PloS One* 12 (10): e0185056.](http://paperpile.com/b/uaTuS8/zRzs4) <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0185056>[.](http://paperpile.com/b/uaTuS8/zRzs4)

[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/uaTuS8/KFN2n)[10.1111/j.1365-2958.2004.04113.x](http://dx.doi.org/10.1111/j.1365-2958.2004.04113.x)[.](http://paperpile.com/b/uaTuS8/KFN2n)

[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. https://doi.org/](http://paperpile.com/b/uaTuS8/p8Wm)[10.1093/bioinformatics/btp348](http://dx.doi.org/10.1093/bioinformatics/btp348)[.](http://paperpile.com/b/uaTuS8/p8Wm)

[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/uaTuS8/KpbEH) <https://www.ncbi.nlm.nih.gov/pubmed/10672174>[.](http://paperpile.com/b/uaTuS8/KpbEH)

[Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. “Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.” *Bioinformatics*  25 (11): 1422–23. https://doi.org/](http://paperpile.com/b/uaTuS8/vFta)[10.1093/bioinformatics/btp163](http://dx.doi.org/10.1093/bioinformatics/btp163)[.](http://paperpile.com/b/uaTuS8/vFta)

[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/uaTuS8/0pIME) <https://www.ncbi.nlm.nih.gov/pubmed/24524134>[.](http://paperpile.com/b/uaTuS8/0pIME)

[Eddy, Sean. 1998. “HMMER: Profile HMMs for Protein Sequence Analysis.”](http://paperpile.com/b/uaTuS8/rSCv)

[Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/](http://paperpile.com/b/uaTuS8/Z7de)[10.1093/nar/gkh340](http://dx.doi.org/10.1093/nar/gkh340)[.](http://paperpile.com/b/uaTuS8/Z7de)

[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/uaTuS8/gcnj)[10.7717/peerj.1319](http://dx.doi.org/10.7717/peerj.1319)[.](http://paperpile.com/b/uaTuS8/gcnj)

[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/uaTuS8/LTz3)[10.1093/nar/gkr367](http://dx.doi.org/10.1093/nar/gkr367)[.](http://paperpile.com/b/uaTuS8/LTz3)

[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/uaTuS8/zWajT)[10.1093/nar/gkl732](http://dx.doi.org/10.1093/nar/gkl732)[.](http://paperpile.com/b/uaTuS8/zWajT)

[Fuhrman, J. A. 1999. “Marine Viruses and Their Biogeochemical and Ecological Effects.” *Nature* 399 (6736): 541–48. https://doi.org/](http://paperpile.com/b/uaTuS8/6Mri3)[10.1038/21119](http://dx.doi.org/10.1038/21119)[.](http://paperpile.com/b/uaTuS8/6Mri3)

[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/uaTuS8/XGgJU)[10.1146/annurev-marine-010814-015934](http://dx.doi.org/10.1146/annurev-marine-010814-015934)[.](http://paperpile.com/b/uaTuS8/XGgJU)

[Giovannoni, Stephen, Ben Temperton, and Yanlin Zhao. 2013. “Giovannoni et Al. Reply.” *Nature* 499 (7459): E4–5. https://doi.org/](http://paperpile.com/b/uaTuS8/vkFQh)[10.1038/nature12388](http://dx.doi.org/10.1038/nature12388)[.](http://paperpile.com/b/uaTuS8/vkFQh)

[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/uaTuS8/lLtR)[10.7717/peerj.3035](http://dx.doi.org/10.7717/peerj.3035)[.](http://paperpile.com/b/uaTuS8/lLtR)

[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/uaTuS8/3ODS)[10.1128/mBio.00252-12](http://dx.doi.org/10.1128/mBio.00252-12)[.](http://paperpile.com/b/uaTuS8/3ODS)

[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/uaTuS8/BD84V)[10.1038/ismej.2014.143](http://dx.doi.org/10.1038/ismej.2014.143)[.](http://paperpile.com/b/uaTuS8/BD84V)

[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/uaTuS8/uv25)[10.1186/1471-2105-11-119](http://dx.doi.org/10.1186/1471-2105-11-119)[.](http://paperpile.com/b/uaTuS8/uv25)

[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/uaTuS8/0ZFt)[10.1038/s41467-018-07641-9](http://dx.doi.org/10.1038/s41467-018-07641-9)[.](http://paperpile.com/b/uaTuS8/0ZFt)

[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/uaTuS8/hFBUt) <https://www.biorxiv.org/content/10.1101/533240v1.abstract>[.](http://paperpile.com/b/uaTuS8/hFBUt)

[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/uaTuS8/8ooV)[10.7287/peerj.preprints.27522v1](http://dx.doi.org/10.7287/peerj.preprints.27522v1)[.](http://paperpile.com/b/uaTuS8/8ooV)

[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. https://doi.org/](http://paperpile.com/b/uaTuS8/i4Xz)[10.1093/nar/gkf436](http://dx.doi.org/10.1093/nar/gkf436)[.](http://paperpile.com/b/uaTuS8/i4Xz)

[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/uaTuS8/TXx9)[10.1101/gr.210641.116](http://dx.doi.org/10.1101/gr.210641.116)[.](http://paperpile.com/b/uaTuS8/TXx9)

[Köster, Johannes, and Sven Rahmann. 2012. “Snakemake—a Scalable Bioinformatics Workflow Engine.” *Bioinformatics*  28 (19): 2520–22. https://doi.org/](http://paperpile.com/b/uaTuS8/bDxm)[10.1093/bioinformatics/bts480](http://dx.doi.org/10.1093/bioinformatics/bts480)[.](http://paperpile.com/b/uaTuS8/bDxm)

[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/uaTuS8/k57He)[10.1038/35057062](http://dx.doi.org/10.1038/35057062)[.](http://paperpile.com/b/uaTuS8/k57He)

[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/uaTuS8/PwQa)[10.1038/nmeth.1923](http://dx.doi.org/10.1038/nmeth.1923)[.](http://paperpile.com/b/uaTuS8/PwQa)

[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/uaTuS8/gMcJN)[10.1016/j.mib.2007.08.005](http://dx.doi.org/10.1016/j.mib.2007.08.005)[.](http://paperpile.com/b/uaTuS8/gMcJN)

[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/uaTuS8/adSjP)[10.1111/j.1365-2672.2009.04143.x](http://dx.doi.org/10.1111/j.1365-2672.2009.04143.x)[.](http://paperpile.com/b/uaTuS8/adSjP)

[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/uaTuS8/6N7y)[10.1093/bioinformatics/btl529](http://dx.doi.org/10.1093/bioinformatics/btl529)[.](http://paperpile.com/b/uaTuS8/6N7y)

[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. https://doi.org/](http://paperpile.com/b/uaTuS8/5Gw7)[10.1093/bioinformatics/btv033](http://dx.doi.org/10.1093/bioinformatics/btv033)[.](http://paperpile.com/b/uaTuS8/5Gw7)

[Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics*  34 (18): 3094–3100. https://doi.org/](http://paperpile.com/b/uaTuS8/VYCg)[10.1093/bioinformatics/bty191](http://dx.doi.org/10.1093/bioinformatics/bty191)[.](http://paperpile.com/b/uaTuS8/VYCg)

[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/uaTuS8/qe7u6)[10.1093/bioinformatics/btn043](http://dx.doi.org/10.1093/bioinformatics/btn043)[.](http://paperpile.com/b/uaTuS8/qe7u6)

[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/uaTuS8/SIG5k)[10.1038/nature04111](http://dx.doi.org/10.1038/nature04111)[.](http://paperpile.com/b/uaTuS8/SIG5k)

[Matsen, Frederick A., Robin B. Kodner, and E. Virginia Armbrust. 2010. “Pplacer: Linear Time Maximum-Likelihood and Bayesian Phylogenetic Placement of Sequences onto a Fixed Reference Tree.” *BMC Bioinformatics* 11 (October): 538. https://doi.org/](http://paperpile.com/b/uaTuS8/yCuM)[10.1186/1471-2105-11-538](http://dx.doi.org/10.1186/1471-2105-11-538)[.](http://paperpile.com/b/uaTuS8/yCuM)

[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/uaTuS8/zgLC) <https://www.theoj.org/joss-papers/joss.00205/10.21105.joss.00205.pdf>[.](http://paperpile.com/b/uaTuS8/zgLC)

[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/uaTuS8/afFi) <http://arxiv.org/abs/1802.03426>[.](http://paperpile.com/b/uaTuS8/afFi)

[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/uaTuS8/ByA6)[10.1101/530188](http://dx.doi.org/10.1101/530188)[.](http://paperpile.com/b/uaTuS8/ByA6)

[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/uaTuS8/cjQS)[10.1038/ncomms11257](http://dx.doi.org/10.1038/ncomms11257)[.](http://paperpile.com/b/uaTuS8/cjQS)

[Mikheenko, Alla, Vladislav Saveliev, and Alexey Gurevich. 2016. “MetaQUAST: Evaluation of Metagenome Assemblies.” *Bioinformatics*  32 (7): 1088–90. https://doi.org/](http://paperpile.com/b/uaTuS8/XHmd)[10.1093/bioinformatics/btv697](http://dx.doi.org/10.1093/bioinformatics/btv697)[.](http://paperpile.com/b/uaTuS8/XHmd)

[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/uaTuS8/FGBHH)[10.1128/JB.02230-14](http://dx.doi.org/10.1128/JB.02230-14)[.](http://paperpile.com/b/uaTuS8/FGBHH)

[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/uaTuS8/7QM6)[10.1093/molbev/msu300](http://dx.doi.org/10.1093/molbev/msu300)[.](http://paperpile.com/b/uaTuS8/7QM6)

[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/uaTuS8/SwpbW)[10.1093/dnares/dsn027](http://dx.doi.org/10.1093/dnares/dsn027)[.](http://paperpile.com/b/uaTuS8/SwpbW)

[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/uaTuS8/Spkl)[10.1038/nbt.4229](http://dx.doi.org/10.1038/nbt.4229)[.](http://paperpile.com/b/uaTuS8/Spkl)

[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/uaTuS8/52Xi)[10.1101/gr.186072.114](http://dx.doi.org/10.1101/gr.186072.114)[.](http://paperpile.com/b/uaTuS8/52Xi)

[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/uaTuS8/EiECN)[10.1038/ismej.2011.101](http://dx.doi.org/10.1038/ismej.2011.101)[.](http://paperpile.com/b/uaTuS8/EiECN)

[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/uaTuS8/NGzoE)[10.1038/343060a0](http://dx.doi.org/10.1038/343060a0)[.](http://paperpile.com/b/uaTuS8/NGzoE)

[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/uaTuS8/XtTjT) <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12199>[.](http://paperpile.com/b/uaTuS8/XtTjT)

[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/uaTuS8/dmOBk)[10.1038/nrmicro2235](http://dx.doi.org/10.1038/nrmicro2235)[.](http://paperpile.com/b/uaTuS8/dmOBk)

[Rohwer, Forest, and Rebecca Vega Thurber. 2009. “Viruses Manipulate the Marine Environment.” *Nature* 459 (7244): 207–12. https://doi.org/](http://paperpile.com/b/uaTuS8/HbFXh)[10.1038/nature08060](http://dx.doi.org/10.1038/nature08060)[.](http://paperpile.com/b/uaTuS8/HbFXh)

[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/uaTuS8/HpcVB)[10.7717/peerj.985](http://dx.doi.org/10.7717/peerj.985)[.](http://paperpile.com/b/uaTuS8/HpcVB)

[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/uaTuS8/J1KHm)[10.7554/eLife.03125](http://dx.doi.org/10.7554/eLife.03125)[.](http://paperpile.com/b/uaTuS8/J1KHm)

[Seemann, T. 2015. “Barrnap.” Github.](http://paperpile.com/b/uaTuS8/PkcQ)

[Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics*  30 (14): 2068–69. https://doi.org/](http://paperpile.com/b/uaTuS8/Ixvr)[10.1093/bioinformatics/btu153](http://dx.doi.org/10.1093/bioinformatics/btu153)[.](http://paperpile.com/b/uaTuS8/Ixvr)

[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/uaTuS8/hY8yT)[10.1038/nature08284](http://dx.doi.org/10.1038/nature08284)[.](http://paperpile.com/b/uaTuS8/hY8yT)

[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/uaTuS8/lQ1LD)[10.1038/ismej.2011.2](http://dx.doi.org/10.1038/ismej.2011.2)[.](http://paperpile.com/b/uaTuS8/lQ1LD)

[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/uaTuS8/qwwUZ)[10.1128/AEM.02545-17](http://dx.doi.org/10.1128/AEM.02545-17)[.](http://paperpile.com/b/uaTuS8/qwwUZ)

[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/uaTuS8/enzNc)[10.1007/BF00166813](http://dx.doi.org/10.1007/BF00166813)[.](http://paperpile.com/b/uaTuS8/enzNc)

[Suttle, Curtis A. 2007a. “Marine Viruses — Major Players in the Global Ecosystem.” *Nature Reviews Microbiology*. https://doi.org/](http://paperpile.com/b/uaTuS8/14Tcv)[10.1038/nrmicro1750](http://dx.doi.org/10.1038/nrmicro1750)[.](http://paperpile.com/b/uaTuS8/14Tcv)

[———. 2007b. “Marine Viruses—major Players in the Global Ecosystem.” *Nature Reviews. Microbiology* 5 (10): 801.](http://paperpile.com/b/uaTuS8/RPZpI) <https://www.nature.com/articles/nrmicro1750>[.](http://paperpile.com/b/uaTuS8/RPZpI)

[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/uaTuS8/8UqYu)[10.1073/pnas.1102164108](http://dx.doi.org/10.1073/pnas.1102164108)[.](http://paperpile.com/b/uaTuS8/8UqYu)

[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/uaTuS8/HdXVu)[10.1038/ismej.2013.243](http://dx.doi.org/10.1038/ismej.2013.243)[.](http://paperpile.com/b/uaTuS8/HdXVu)

[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/uaTuS8/Ut9qr)[10.1371/journal.pone.0101415](http://dx.doi.org/10.1371/journal.pone.0101415)[.](http://paperpile.com/b/uaTuS8/Ut9qr)

[Van Valen, Leigh. 1973. “A NEW EVOLUTIONARY LAW.” *Evolutionary Theory* 1: 1–30.](http://paperpile.com/b/uaTuS8/SyDnf) <https://ebme.marine.rutgers.edu/HistoryEarthSystems/HistEarthSystems_Fall2010/VanValen%201973%20Evol%20%20Theor%20.pdf>[.](http://paperpile.com/b/uaTuS8/SyDnf)

[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/uaTuS8/rtko0)[10.1111/j.1462-2920.2007.01361.x](http://dx.doi.org/10.1111/j.1462-2920.2007.01361.x)[.](http://paperpile.com/b/uaTuS8/rtko0)

[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/uaTuS8/Dv9J1)[10.1016/j.mib.2005.06.001](http://dx.doi.org/10.1016/j.mib.2005.06.001)[.](http://paperpile.com/b/uaTuS8/Dv9J1)

[Wallace, Iain M., Orla O’Sullivan, Desmond G. Higgins, and Cedric Notredame. 2006. “M-Coffee: Combining Multiple Sequence Alignment Methods with T-Coffee.” *Nucleic Acids Research* 34 (6): 1692–99. https://doi.org/](http://paperpile.com/b/uaTuS8/BTsp)[10.1093/nar/gkl091](http://dx.doi.org/10.1093/nar/gkl091)[.](http://paperpile.com/b/uaTuS8/BTsp)

[Weinbauer, Markus G. 2004. “Ecology of Prokaryotic Viruses.” *FEMS Microbiology Reviews* 28 (2): 127–81. https://doi.org/](http://paperpile.com/b/uaTuS8/Zyzyh)[10.1016/j.femsre.2003.08.001](http://dx.doi.org/10.1016/j.femsre.2003.08.001)[.](http://paperpile.com/b/uaTuS8/Zyzyh)

[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/uaTuS8/2FiTJ)[10.1093/bioinformatics/btv383](http://dx.doi.org/10.1093/bioinformatics/btv383)[.](http://paperpile.com/b/uaTuS8/2FiTJ)

[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/uaTuS8/4HCEW)[10.1128/MMBR.00034-09](http://dx.doi.org/10.1128/MMBR.00034-09)[.](http://paperpile.com/b/uaTuS8/4HCEW)

[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/uaTuS8/21Amp)[10.1128/mmbr.64.1.69-114.2000](http://dx.doi.org/10.1128/mmbr.64.1.69-114.2000)[.](http://paperpile.com/b/uaTuS8/21Amp)

[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/uaTuS8/nrBi)[10.1186/2049-2618-2-26](http://dx.doi.org/10.1186/2049-2618-2-26)[.](http://paperpile.com/b/uaTuS8/nrBi)

[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/uaTuS8/u5rXh)[10.1093/nar/gkr485](http://dx.doi.org/10.1093/nar/gkr485)[.](http://paperpile.com/b/uaTuS8/u5rXh)

# Appendices

