Title

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

# Abstract

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

[**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 et al. 2017)](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.

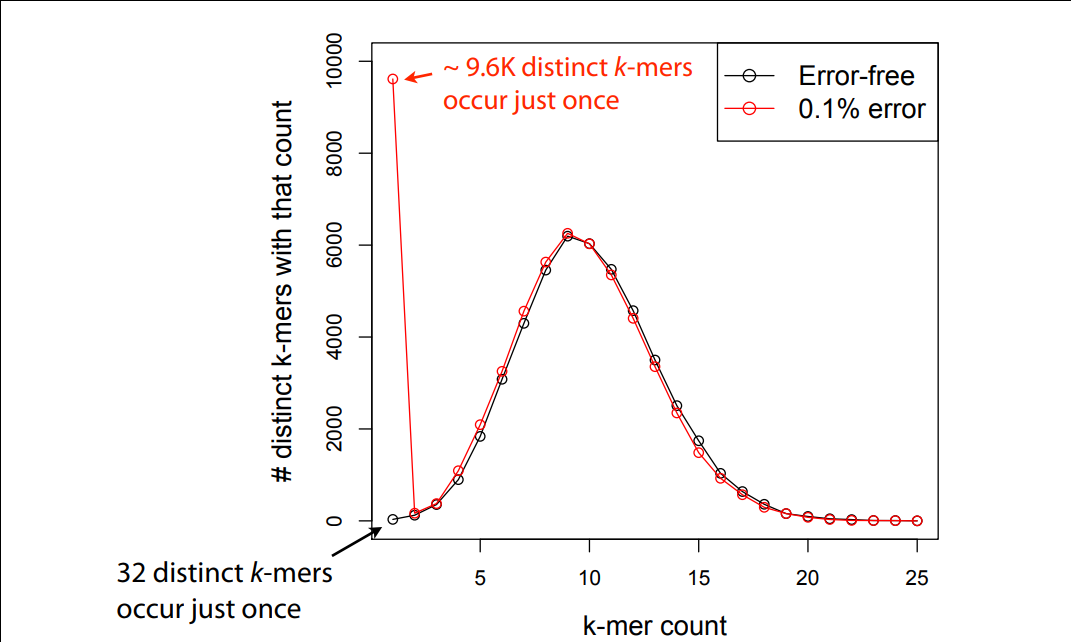
Briefly, this script is performed in 3 main steps:

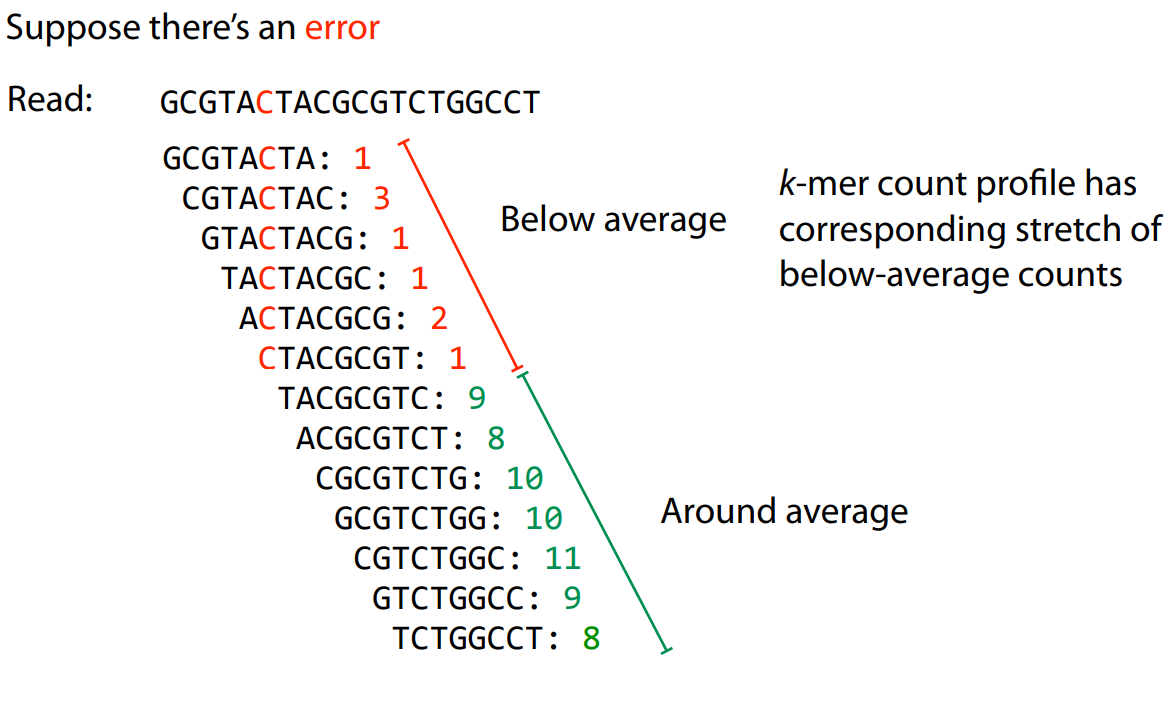
1. The removal of low quality and duplicate regions
2. The trimming of adaptors and synthetic artifacts
3. Error correction
   1. Normalisation if appropriate

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 [(Li et al. 2015)](https://paperpile.com/c/uaTuS8/5Gw7) and Tadwrapper [(Bushnell et al. 2017)](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 et al. 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/0ne0x7/1k4Wv), BBtools [(Bushnell, Rood, and Singer 2017)](https://paperpile.com/c/0ne0x7/tru6) and Minimap2 [(Li 2018)](https://paperpile.com/c/0ne0x7/hiEEM). Here, Bowtie2 was used to map reads back to the assembled contigs of samples taken from the same depth, resulting in 6 SAM files for each assembly. SAM files were then filtered for reads with a 95 % identity using BamM [(Rabosky et al. 2014)](https://paperpile.com/c/0ne0x7/vOIQ). Reads mapping to a lower percentage identity were discarded. A 95 % identity was chosen as this is the general boundary used to delineate species.

## 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/0ne0x7/dCN6D) can be used. CheckM uses conserved genes present in all bacteria or archaea and looks for them within a queried genome. The percentage presence of these genes can give an idea into the completeness of MAG without prior knowledge of the taxonomy of the bin. However, with the known taxonomy, CheckM allows for a more detailed analysis. Different taxonomic groups have different numbers of conserved genes, with lower taxonomic classifications having a higher amount of conserved genes within its grouping. With a higher number of conserved genes, a more detailed analysis of the sample can be done. This would give a more in depth analysis as more specific and higher numbers of genes can be used. For example, there exists 104 genes that are conserved amongst all bacteria [(Parks et al. 2015)](https://paperpile.com/c/0ne0x7/dCN6D) but potentially hundreds more genes more specific to that organism.

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

## 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/0ne0x7/YYNpD), BinSanity [(Graham, Heidelberg, and Tully 2017)](https://paperpile.com/c/0ne0x7/X0LTW), CONCOCT [(Alneberg et al. 2014)](https://paperpile.com/c/0ne0x7/LHcUn) and MaxBin [(Wu et al. 2014)](https://paperpile.com/c/0ne0x7/AMKmA) are used. All of them have their own proprietary algorithm but result in the same process of binning contigs with the usage of SAM files.

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

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.

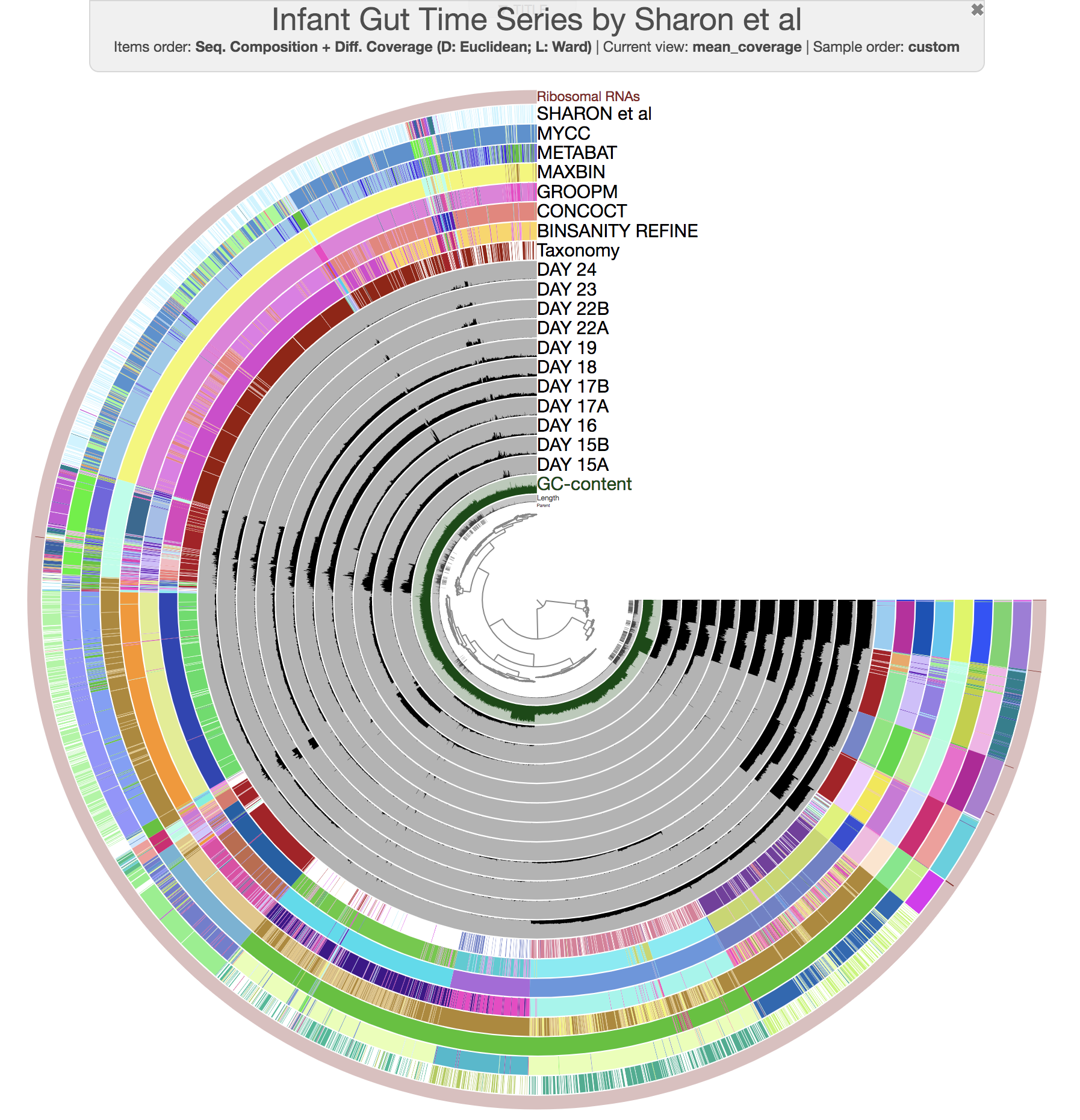
## X.XX Binning by taxonomy

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

### 1.1.4 Visualisation of metagenomes

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





## 

## 

## 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/0ne0x7/MxXC) cluster data points on a scatter plot based on neighbour distance. This allows for the automation and visualisation of clusters to aid extraction of MAGs.

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

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

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 then 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 not 0 were displayed. Although completion statics was high, redundancy statics was high too indicating a highly contamination bin with multiple species within a bin. This was not deemed as a successful binning.

It therefore becomes apparent that a n\_neighbours below the default of 15 but higher then 2 would 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 was high, redundancy statics was high too indicating a highly contamination 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

* 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

## 3.1 451 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 2017)](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. Kmers of intervals of 10 up to 99 and 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. Statistics of assemblies are located here. (XXX link to table in appendices) Several methods were used to improved the assemblies and the above method was summarised as the most complete.

## 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 establish 5 clades of SAR11. Further work on each SAG would then be relatable to its known ecological biosphere.

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 clades is split into 5 existing grouping with multiple subgroups. 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.

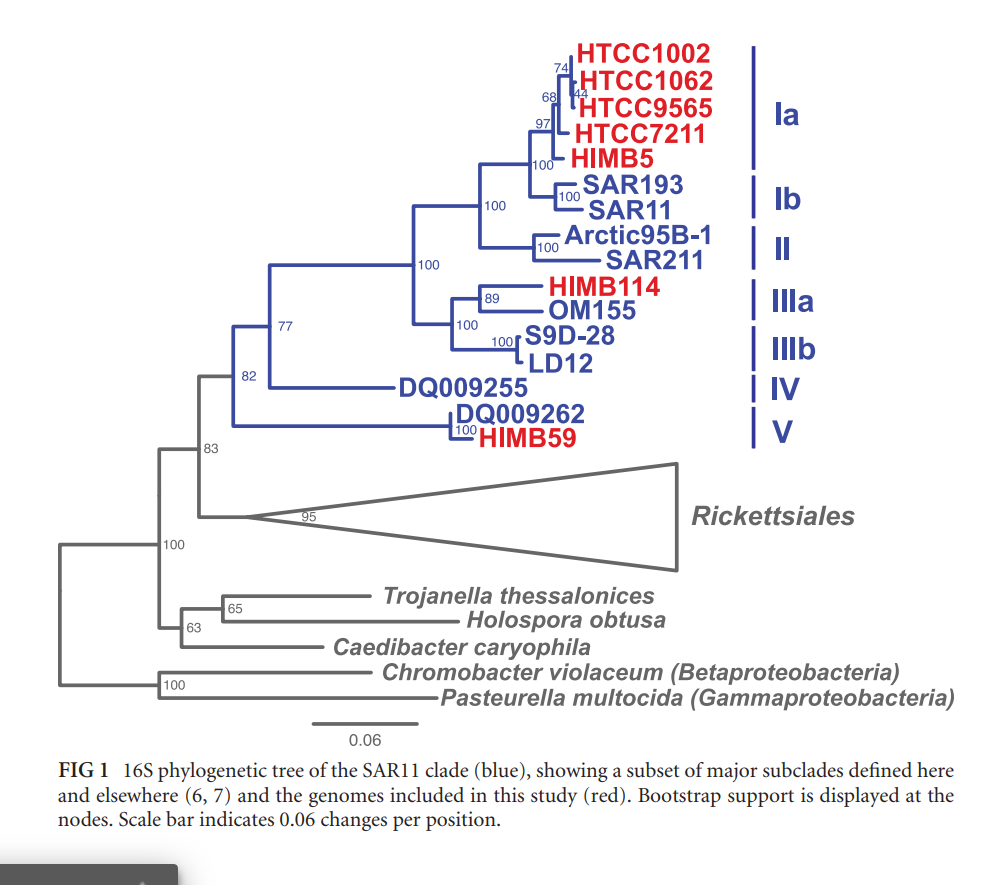




Fig X.XXX 16S rRNA Phylogenetic tree of 373 SAR11 SAGs. Clades are labelled in colour with bold colours indicating a reference 16S rRNA from the NCBI database. Lighter colours indicate SAR11 SAGs that are probable members of their respective coloured clade. Nodes are collapsed to aid visualisation with 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.

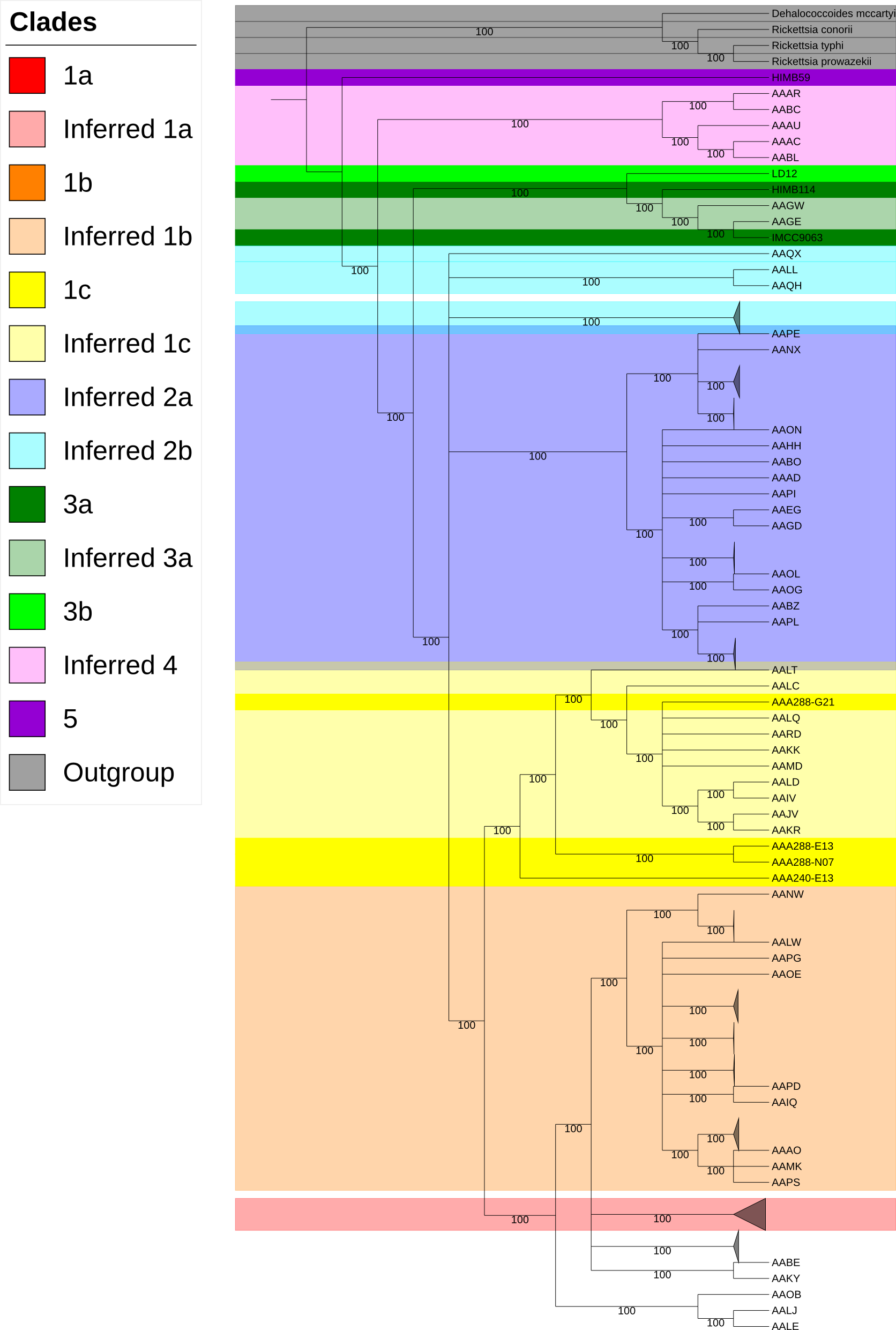


Fig X.XXX A WGS phylogenetic tree of SAR11 SAGs. Clades are displayed in colour with bold colours indicating reference genomes. Clades are collapsed to aid visualisation with grey triangles, indicative of the size of the collapsed node. Clades 2 and 4 do not have a reference genome and clades are inferred based on positioning with the phylogenetic tree. Non- coloured branches indicate branches of unknown identity based on other existing trees. Branches are collapsed with a bootstrapping value lower than 100.

From the WGS phylogenetic, we can be fairly confident of its structure due to the usage of a high bootstrapping number. However, without reference genomes for clades 2 and 4, it becomes difficult to inductively prove which node belongs to which 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, clades 2 and 4 are inferred in blues and pink respectively. However, clade 2 is show in have two splits with its phylogeny, with clade 2a being described as spring surface and DCM (XXX define DCM) where as clade 2b as spring upper mesopelagic. But more importantly, these split within a node whereas I have show that they are located on the same branch. Without a reference, this makes it difficult to correctly ascertain if this is a new branch, or that clade 2b should be its separate branch and not within clade 2 at all. Either way, this new branch is either indicative of a new classification of SAR11 or existing classifications need to be promoted from a subclade (2b) to its own clade. Lastly, clades not classified in colour indicate an additional clade within the 1a and 1b node. This clade is not small with 35 members of SAR11 SAGs. This would suggest there maybe a possible additional subclade within clade 1 here.

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 works and is in nature so we use it. 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, 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>).

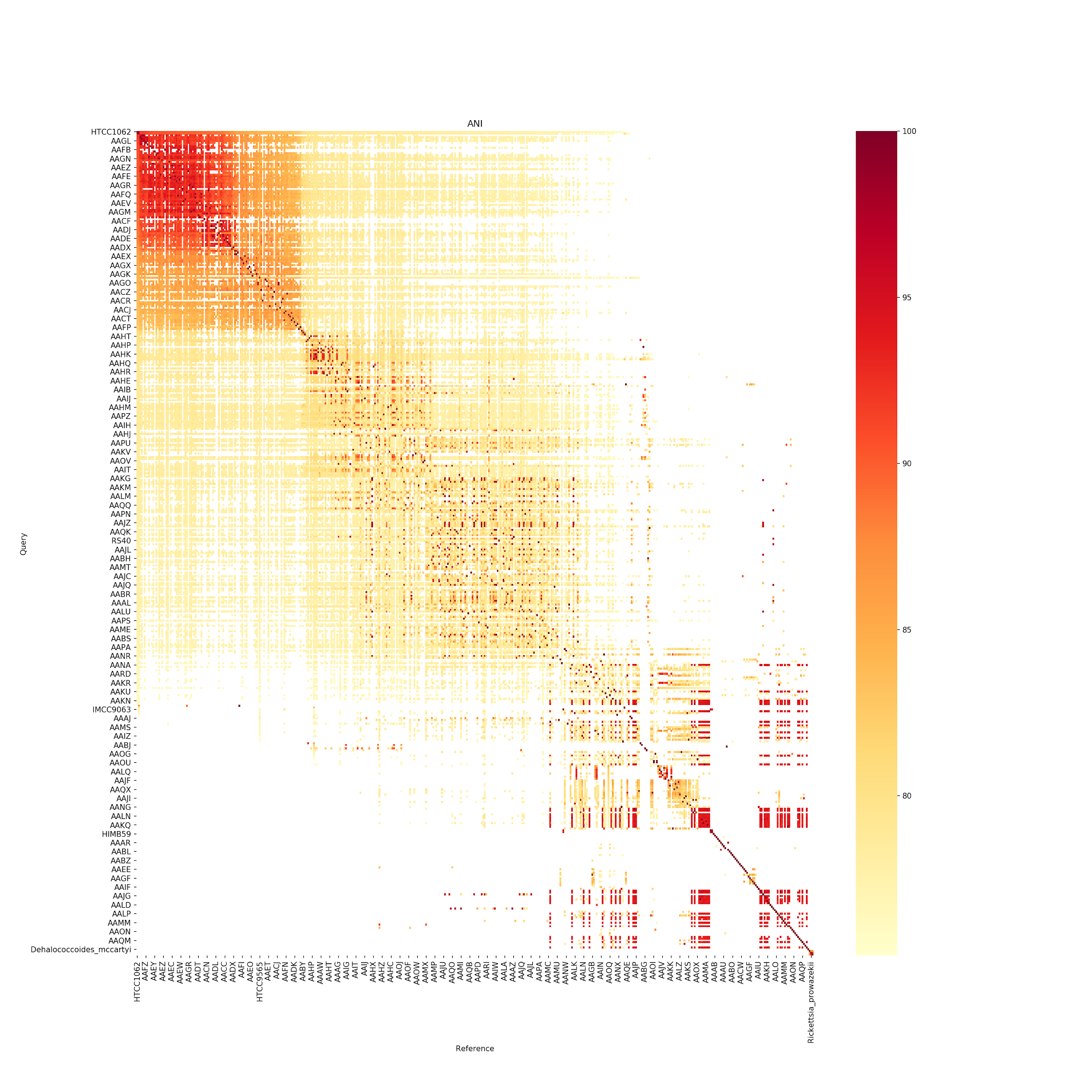


Fig X.XXX A trimmed selection of an ANI heatmap plot of and all vs all SAR11 SAGs and reference genomes showing overall structure of the data. A darker colour indicates a higher ANI and lighter colour a lower ANI on a discontinuous scale. ANI values below 75 are not shown and are displayed in white. Query genomes are labelled on the x-axis, where reference genomes are on the y-axis. Clustering of a darker colour indicate a grouping of similar SAR11 SAGs ANI values. Reference genomes are used to provide identity to clusters. Enlarged plot ([https://github.com/ash-bell/masters\_thesis/blob/master/fastANI\_heatmap.png](https://github.com/ash-bell/masters_thhttps://github.com/ash-bell/masters_thesis/blob/master/fastANI_heatmap.pngesis/blob/master/QC_reads.smkhttps://github.com/ash-bell/masters_thesis/blob/master/fastANI.sh))

Clustering of an ANIs within a heatmap are a good indicator if boundaries exist within a clade. Within this plot, clear clusters can be seen with reference genomes 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 and guesses a best. A good example of this is within clade 1 where it is difficult to tell if there exists additional subclades within the cluster and if so, where they begin and end.

## 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 et al. 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 et al. 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.

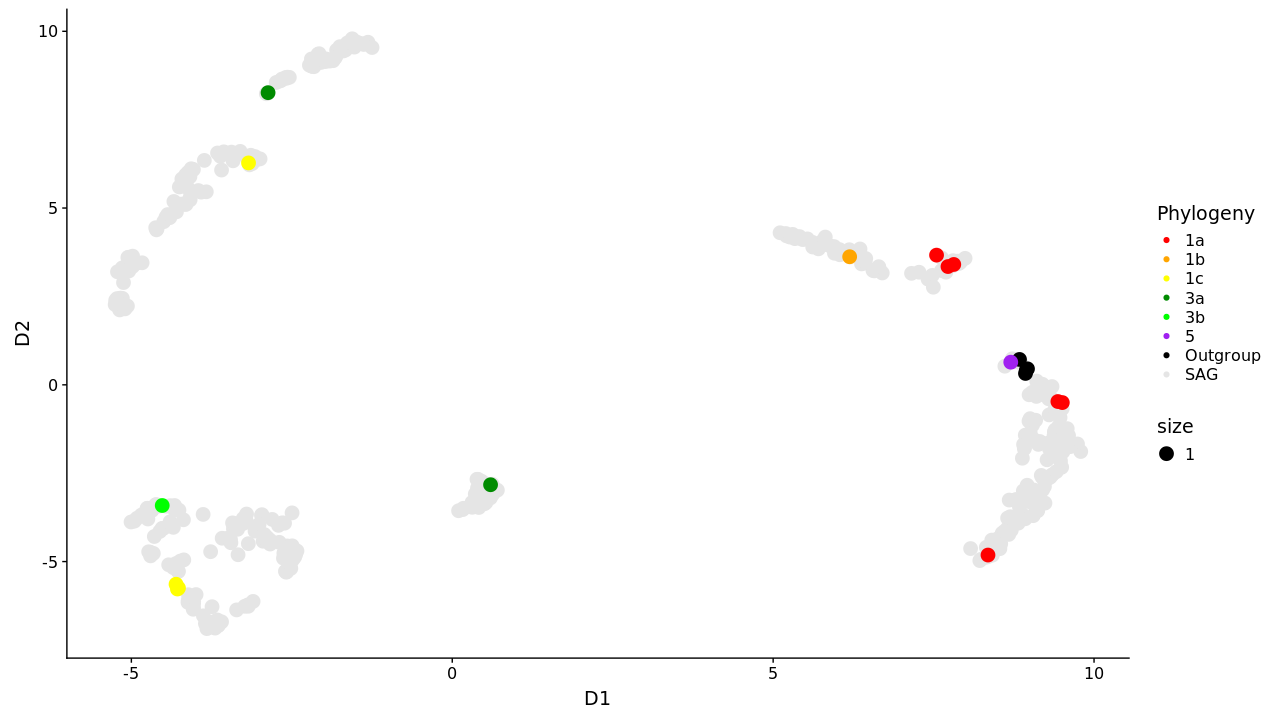
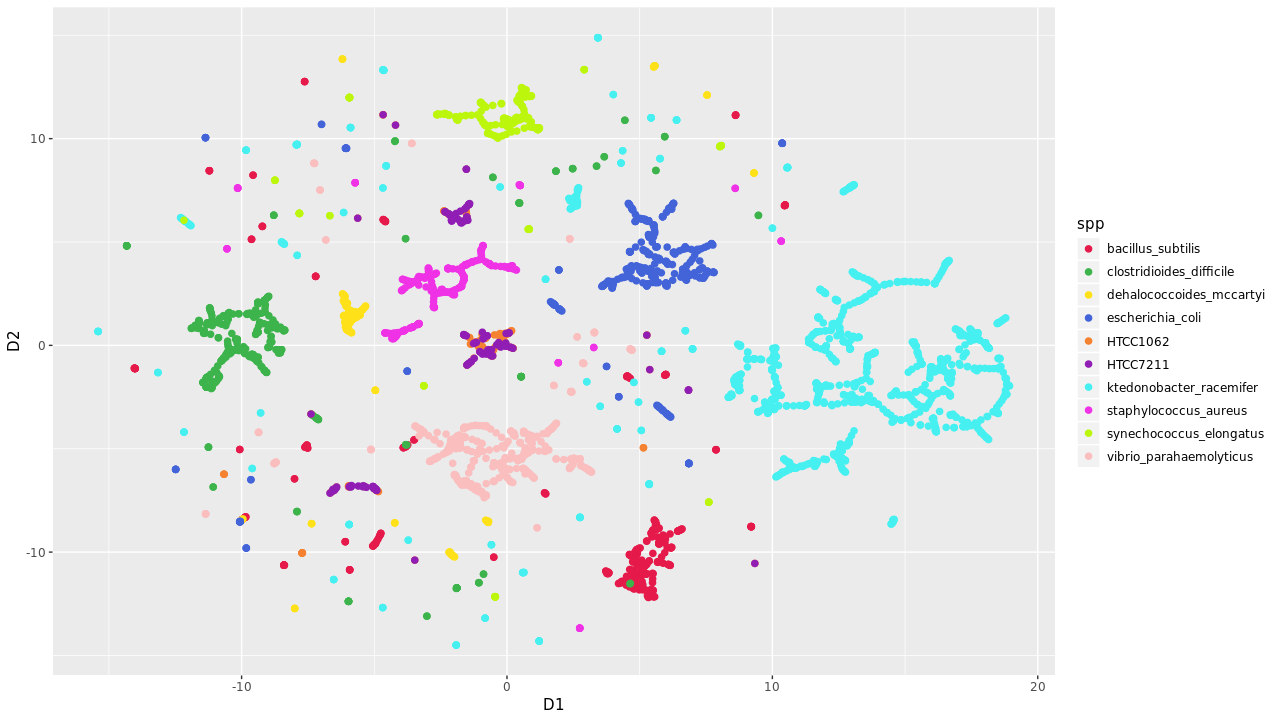
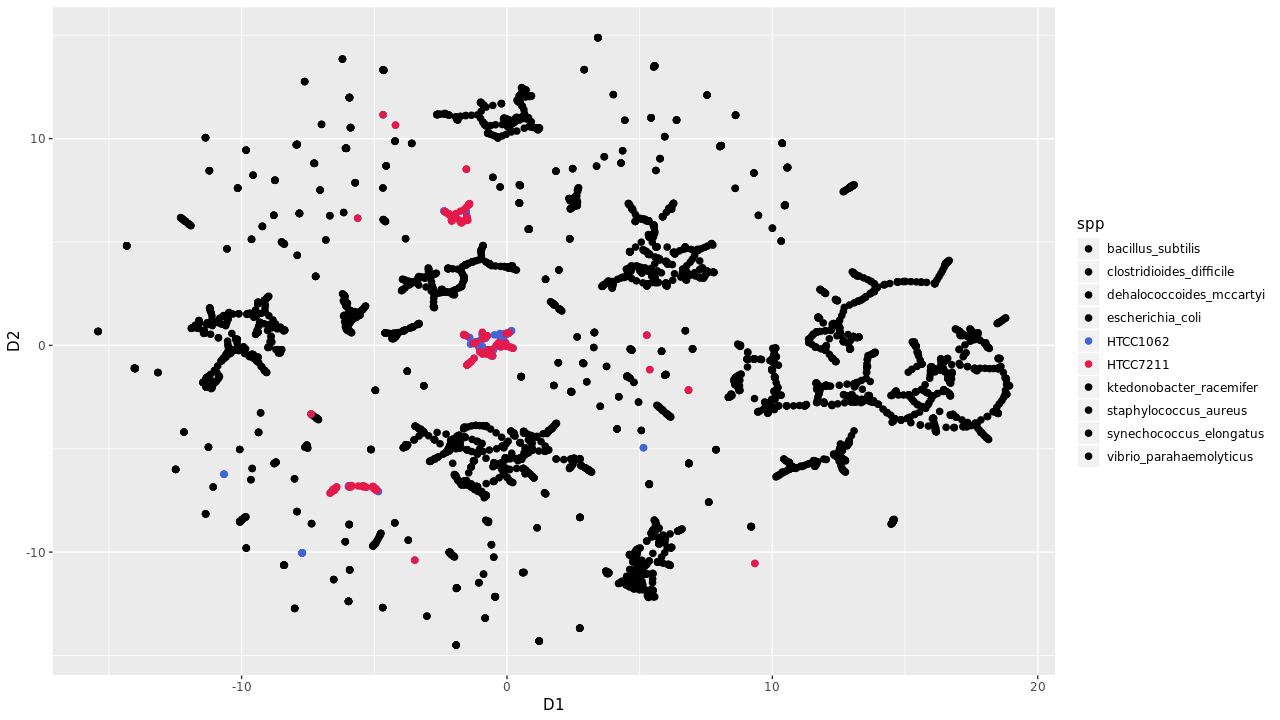


Fig X.XXX BH-tSNE scatter plot of WGS of SAR11 SAGs. SAG contigs were concatenated together forming a single contig fasta file. Kmer counting was performed and dimension reduction techniques in a BH-tSNE plot was performed via UMAP. SAGs where then plotted via their BH-tSNE values in R via ggplot [(Wickham, 2006)](https://paperpile.com/c/uaTuS8/aYCz) in the form of a scatter plot. Reference genomes are plotted in coloured dots.

Even 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. This proved interesting as produced a side experiment. Bacteria from across the tree of life were taken and their genomes split into 10kbp fragments. Samples of well characterised SAR11 species (7211, 1062) were also included. Resulting contigs were then k-mer counted and dimension reducted 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.



# 3.3 CHAPTER 2 DISCUSSION

### 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. They are key factors in regulating bacterial and eukaryotic microorganism populations. They can infect both multicellular and single cell organisms often with detrimental effects to the host. Within single-cell organisms, they are responsible for the turnover of over 3% of cyanobacteria biomass daily and contribute enormously to the differentiation of the number of genes an organism can access. With this in mind, any study looking at organism population must consider the population of viruses within the same ecosystem.

Generally, viruses exist in two stages, as integrated within genomes of their host or free virions. These are referred to as temperate and non-temperate phages respectively. Many viruses can express both types of lifestyles, naimly lysogenic and lytic lifestyles, or express only one type of lifestyle. Temperate phages are of interest due to their ability to integrate themselves within their hosts. It has been hypothesized that >20% of bacterial genomes originate from phages. 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. It has also been hypothesized this is how transfer of antimicrobial resistant genes are passed to other bacteria. Studying these genes may offer insight into the ability for these viruses to add advantages to the cell of be mechanisms for them into intergreate or the areas in which they do be regions or interest. Prophages can stop becoming viable due to the loss of transcriptional/essential genes and therefore confer perminate DNA mutations and additional genes without being pathogenic.

There are two mechanisms for obtaining prophage sequences, mainly experimentally or computationally. Both reply 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. This replies on the prophage being viable and receptive to excising itself in the presence of environmental triggers. Bioinformatically, the organisms can be sequenced and viral sequences identified. However this is our best guess at what is viral but we have a lot of tools that determine what is a hallmark viral signature.

Viral signatures can be broadly differentiation 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 capture during the DNA sequencing process and may be environmental DNA or free flowing DNA within an organism. Prophages on the other hand are viral sequences that have been integrated into the host genome, become part of its transcriptome.

Obtaining viral signatures from existing genetic sequences usually rely on detection of genes already established to be viral. Some basic “viral-like” properties are searched for like the proportion of AT and GC, shorter protein lengths more common place in prophages. These look for attachment sites for where prophages attached and end. Look for disrupted genes and attahcment sites, places where the virus as integrated itself and maybe truncated genes. Differing nuclotide content, like AT and GC skew is different within these genes. Virusues have a duplicate gene for the location it wants to integrate into, allowing for DNA repair machinestry to intergte it there. Can use BLAST for know viral genes.

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

VirSorter

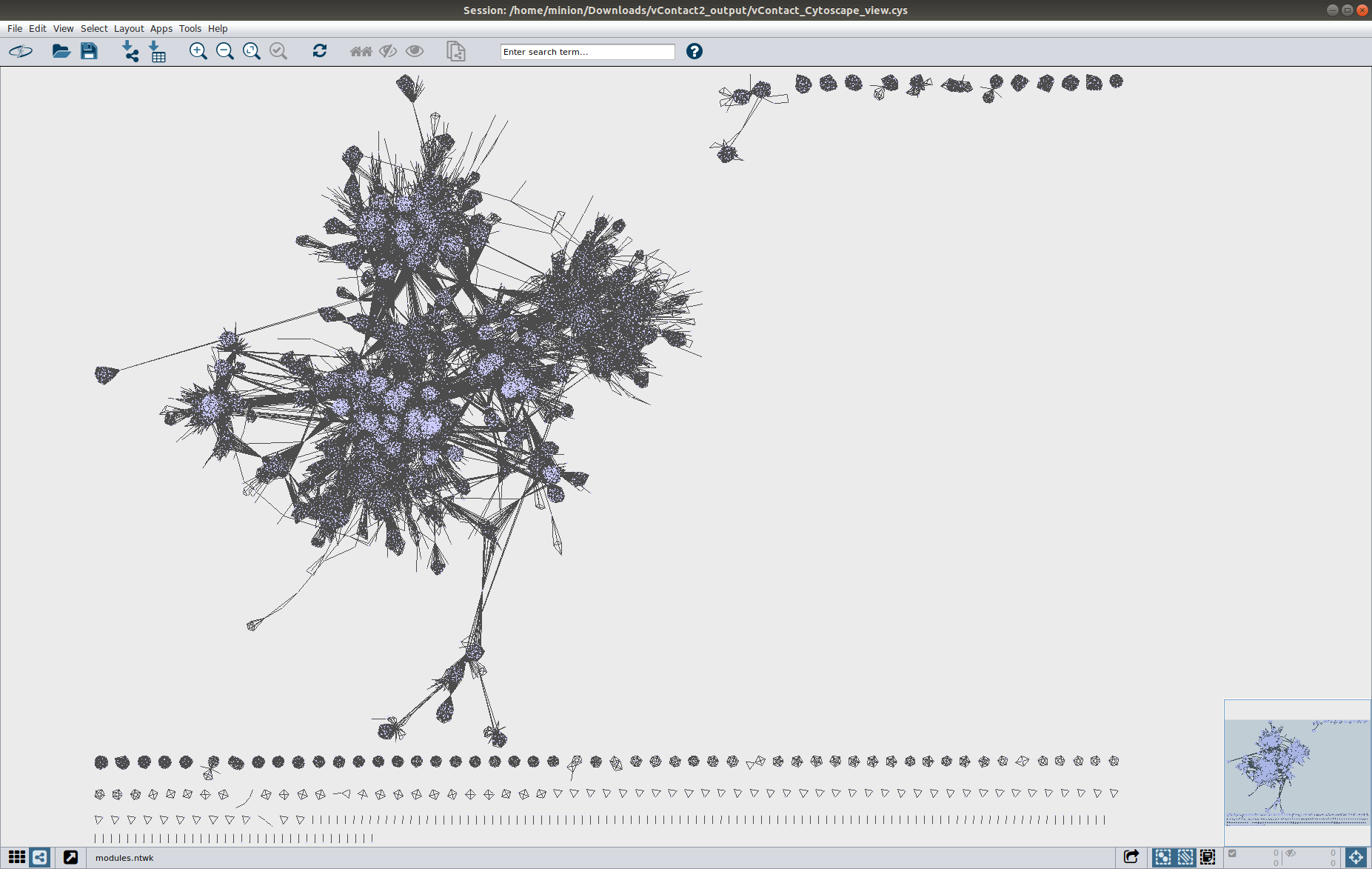
VirSorter uses MetaGeneAnnotator and hmmsearch to look for PFAM domains and viral domains using HMMs of established viral hallmark genes. Using a sliding window, it looks within this window for viral hallmark genes characterised by things like “capsid” or “portal” genes. It also looks for areas enriched for viral domains and lower PFAM domains then the average and have more then two genes. It also looks for genes that are shorter, hallmarks of viral genes as well as genes that are all encoded on the same strand, idicative of a insertion of a viral prophage along with many uncharacterised genes as visures are usually uncharactriesed. It also looks for circular sequences indicative of a phage. Based on these principles, a catagory is defned for areas that it is confident that are viral and by how much. If 80% of a contig is deemed viral, then the whole contig is deemed viral. If less then 80% is viral then it is classified as a prophage.

How dies it work

What are the results

Visualising viral relationships

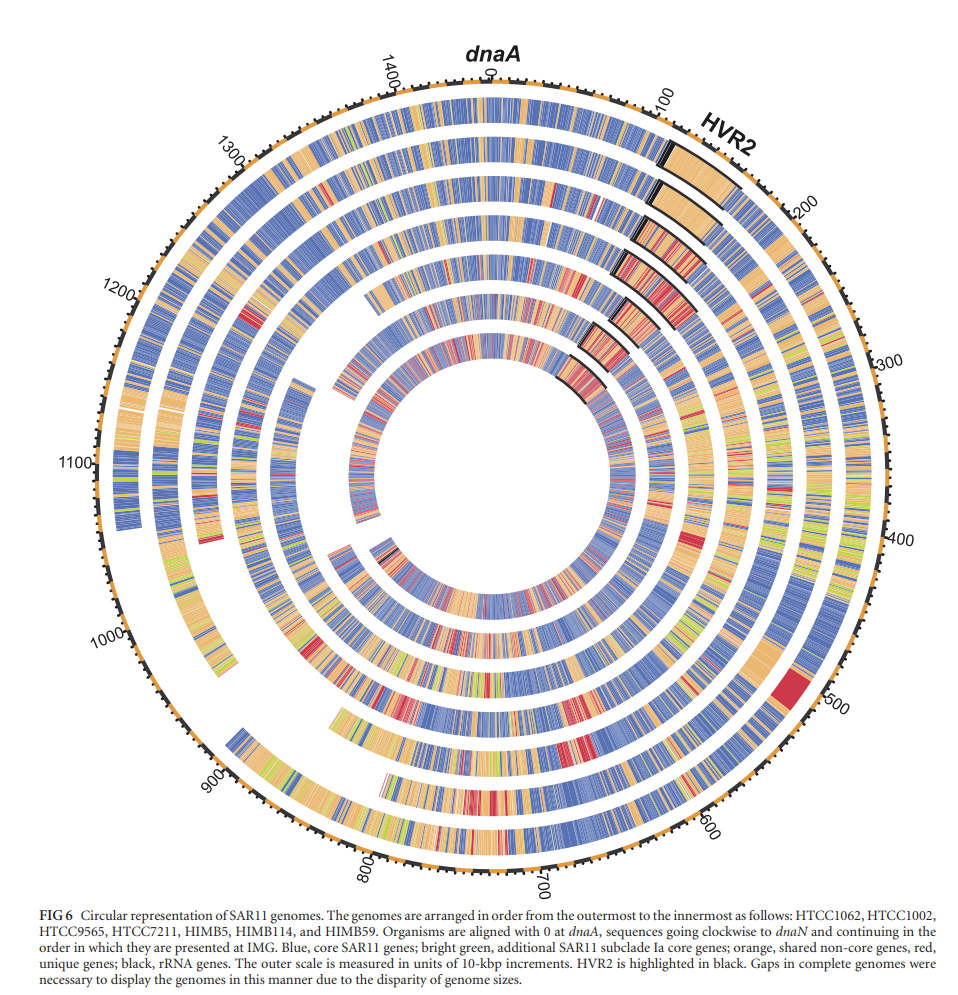
Vcontact2



# 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

Hypervariable regions are areas within a genome that have higher than normal mutation rates in comparison to the entire genome as a whole. This is cool because??? (XXX I don't actually know the benefits of a Hypervariable Region (HVR) unless it’s for genes that need to be changed constantly for some benefit))



It has been show that SAR11 has a HVR amongst all its 1a clade [(Grote et al. 2012)](https://paperpile.com/c/uaTuS8/3ODS). We wanted to confirm this finding with a larger dataset of 1a clade SAR11s and see if HRV2 was present in other SAGs outside of the 1a clade. To find the variable region, we needed to find areas of the genome that do not have any similarity to other SAR11s. Although it would be logical to do an all vs all mapping of all the SAGs, with the poor quality of SAG assemblies, the 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 within and mapping would still occur. Metagenomic samples come from the Biller dataset [(Biller et al. 2018)](https://paperpile.com/c/uaTuS8/j4Ip), 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. Regions coverages only from datasets in the same temporal or spatial plane would be indicative of niche habitats of SAR11s. This would show habitat where SAGs live in and allow for pinpointing where the SAGs are. So if from previous chapter if SAR11 SAGs all from clade 1c were confirmed to be in deep water it would confirm they are bathytype deep water stuff.

# 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

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.

# 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?

# 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

[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)

[Bar-On, Yinon M., Rob Phillips, and Ron Milo. 2018. “The Biomass Distribution on Earth.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (25): 6506–11. https://doi.org/](http://paperpile.com/b/uaTuS8/ODiF)[10.1073/pnas.1711842115](http://dx.doi.org/10.1073/pnas.1711842115)[.](http://paperpile.com/b/uaTuS8/ODiF)

[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)

[Bushnell, Brian, Jonathan Rood, and Esther Singer. 2017. “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)

[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)

[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)

[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)

[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)

[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)

[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)

[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)

[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)

[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)

[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)

[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)

[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)

# Appendices

