## Table of Contents

## 1. Abstract

Three possibly novel bacterial genomes were found through an HGAP assembly of PacBio reads originating from the diatom *Skeletonema marinoi* strain *ST 54.* The genomes were confirmed to be bacterial through circularisation using both *Amos minimus2* and 16S rRNA matching using BLAST. Phylogenetic analysis of the genomes together with >3000 other bacterial genomes resulted in a taxonomic classification . Phylogenetic analysis of the genomes together with >3000 other bacterial genomes resulted in a taxonomic classification to in one case order and in two cases genus level. The taxonomy was further investigated by creating phylogenetic trees of the taxonomic order the genome of interest pertained to using bacterial reference genomes obtained from the NCBI ftp database. These analyses suggest that: 1) one genome is closely related to or a member of the genus *Kordia*
2) one to the genus *Parvibaculum* 3) a third genome without affinities to previously sequenced publicly available genomes but is suggested to belong to the bacterial order *Cellvibrionales*.

## 2. Background

### ***2.1 Bacterial genomes***

Since the first bacterial genome was sequenced in 1995, the pace at which bacterial sequences have been added to sequence databases like GenBank has unsurprisingly increased exponentially. This has come with the decreased cost of generating data using new sequencing technologies. A downside with next or second generation sequencing techniques like illumina is that since there is a very high coverage required for completion of a genome and the read lengths are short, a disproportionately large amount of reads are required to cover the whole genome to a sufficient level. This has lead to a large number of bacterial genomes in permanent draft status that are merely considered “good enough” to be released [1]. The new single molecule sequencing methods like PacBio and Nanopore [2] which generate longer reads have the potential to put an end to this trend, however.

Looking at the bacterial genomes currently in GenBank, the typical one is ~5 Mb of which ~88% encodes ~5000 proteins. This can vary a lot however; The largest being over 14.6 Mb encoding nearly 12 000 genes while the smallest is 122 kb and encodes only 137 proteins. [1]

## ***2.2 Diatoms and their environmental importance***

Diatoms (*Bacillariophyceae*), a large class of algae, are among the most common types of phytoplankton. They are unicellular but often form colonies of various shapes. A distinguishing feature of diatoms is that they are encapsulated in silica ($SiO_2$) cell-walls referred to as frustules. The shape of these frustules can vary wildly between diatome species but are commonly bilaterally symmetrical.

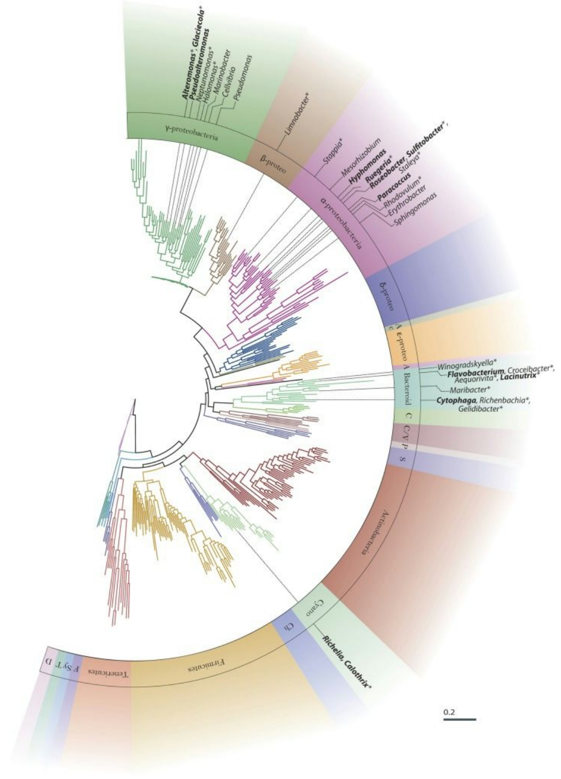
In order to be able to photosynthesise, phytoplankton like diatoms need to be close to the surface to receive sunlight. The silicified wall of diatoms make up about 50% of the dry weight of the cells and has a molecular density of 2.600 while seawater has a molecular density of between 1.021 and 1.028. This means that the cells need some way of gaining buoyancy to avoid sinking. That is done by the vacuole in the cells, which contents have a density of 1.0202, lower than seawater. This is achieved by the cell through keeping the ion content inside the vacuole high in low molecular weight molecules like Na and $NH_4$ and low in high molecular weight molecules like K and no $SO_4$. When the cells die, however, they are unable to uphold this differential ion gradient and will sink to the ocean floor, building up the sediment layer.[5]

Diatoms are primary producers and are estimated to be responsible for 45% of oceanic primary production. As such, they are a good indicator of marine environmental health. The main factors limiting diatom growth are availability of nitrate, phosphate and silicic acid while the main limiting nutrients for other phytoplankton are nitrate and phosphate. This results in the total ocean primary production being mainly controlled by nitrate and phosphate availability while the fraction of diatoms within the primary producer population are controlled by silicic acid abundance.[6]

## ***2.3 Importance of Bacteria to diatom life***

For Diatoms, like so many other organisms, interaction with microbes is a vital part of life. It is known that diatoms and bacteria have cohabitated and interacted for over 200 million years and that the diatoms have, through horizontal gene transfer, acquired hundreds of genes from bacteria over this time. Interestingly, the bacterial taxa that have successfully developed interactions with diatoms are fairly few and are mostly situated in the *Alpha- Beta- Gamma-proteobacteria* and *Bacteroidetes* groups (Figure 1).[3, 11, 12]

The area close around the diatom cell where its exudate diffuses is called the “phycosphere”. Within the phycosphere, bacteria are able to live and utilise the micro nutrients excreted by the diatom cells. This diffusive boundary layer can exist and be maintained because on very small scales, turbulence is an insignificant factor. Bacteria remain in this zone either through chemotactic movement (run and tumble or run and reverse) or attachment to the diatom cells. The action of the bacteria attaching to the cell is thought to depend partly on diatom production of transparent exopolymer particles (TEP) which the bacteria can colonise to feed on[13] and partly on exopolysaccharide (EPS) production by the bacteria. [3, 14]



One of the primary ways symbiotic bacteria can assist a diatom is to remineralise dissolved organic matter into carbon dioxide required by the diatom to photosynthesise. Some other benefits that the bacteria can provide the diatom are: Vitamin production (primarily B12)[15], making iron more available for uptake [16], nitrogen-fixation [18] and oxidative stress relief [19]. In turn, the diatom cell provides the bacteria with nutrients like dissolved organic carbon through its exudate [3, 17, 14]

There are also other bacteria which predate and parasitise on the diatom cells by causing stress or triggering cell lysis. For example, the flavobacterium *Kordia algicida* has been shown to, through quorum sensing, release a protease that acts against a subset of diatoms (including members of the genus *Skeletonema)* which kills the cells.[4]

[redacted]

## *2.4 Skeletonema marinoi*

Skeletonema marinoi is an ocean-living diatom. It forms chains/filaments as it grows where the frustules surrounding the cells are connected by spines. In motive water, these spines can quite easily break off resulting in the formation of a new chain which will continue to grow. These spines increase the total pre-cell surface area, increasing their hydrodynamic drag [7]. While the cells are alive, this helps the chains stay in the euphotic zone, closer to the surface, to get more sunlight for photosynthesis.[5, 6]

[redacted]

*Skeletonema marinoi*, diatom
Photo: Anna Godhe

Skeletonema marinoi is one of the phytoplankton species that can give rise to algal spring-blooms in temperate waters. Spring blooms are thought to be caused by a variety of different phenomena (mainly winter storms) that result in a destratification which mixes nutrients from the deeper waters and sediment into the euphotic zone. It has been observed that diatoms of the genus *Skeletonema* and other large diatoms grow more rapidly when exposed to higher levels of $CO_2$. This raises concerns about the ocean acidification being brought about by global climate change.[9]

## *2.5 Phylotaxonomy*

Traditionally, taxonomic assignment of species has been done through analysis of organismal morphologies and traits. This method relies on a number of predecided upon definite character traits that are used in order to determine which taxonomic group an organism belongs to. If these traits are not representative of how evolution has happened, inaccuracies will result. It has since been argued that a more robust method is to define taxa based on phylogenetic inference, since this avoids conflicts between taxon names and their actual evolutionary relationships [20]. Some of the issues that can arise in phylogenetic inference, when using morphological character states to classify organisms, are paraphyletic groups, which are groups that consist of the descendants of a common ancestor but don’t include all descendents of that ancestor, and polyphyletic groups, which are groups which are characterised by homoplasies (similar traits that have not arisen by inheritance from a common ancestor) [10].

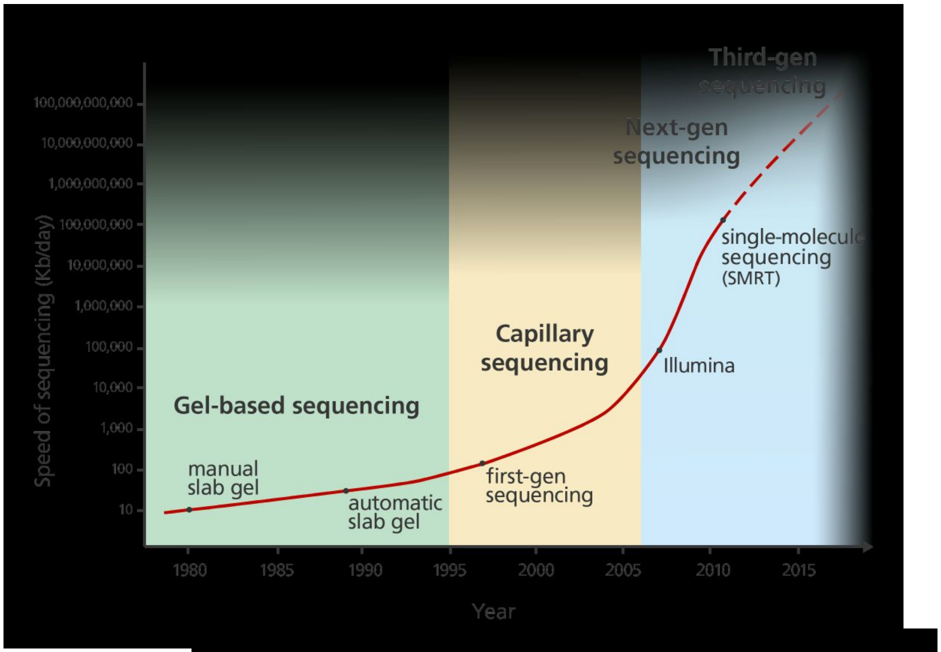
In the early days of molecular taxonomy, single homologous genes from species of interest were used to generate phylogenetic trees. These trees are called gene-trees and may not be indicative of the phylogenetic relationship of the species as a whole. One of the reasons for this is that genes can undergo duplication events without leading to a speciation event. Other reasons can be that genes can be acquired through horizontal gene transfer, which is especially common in bacteria, or incomplete lineage sorting where alleles that are closely related end up in species that are more distantly related and vice versa. Related sexually reproducing species can also hybridise, which can given the right conditions lead to a separate intermediate specie with genes originating from both parental species.

In terms of bacteria, it can be difficult to differentiate between species because of the rapid pace at which they undergo evolution through horizontal gene transfer [23, 24]. Bacterial species were originally determined through morphology and responses to biochemical tests like the gram stain. After that, DNA-DNA Hybridisation (DDH) instead became the standard; a labour intensive procedure where the DNA from one of the species is labeled and mixed with the DNA from the species it is to be compared to so that they can form double strands. Through measuring the temperatures required for melting the dsDNA, it is possible to determine to what level the DNA from the two species hybridise. 70% hybridisation has the classic cutoff for a separate species but it has recently been suggested that 79% - 80% would be more appropriate [21]. Then, 16S rRNA similarity started to be used, since it is present in all bacteria and contains a number of hypervariable regions where mutations tend to accumulate over time. The cutoff to be declared a separate species is/was a less than 97% match to its closest relative. Two 16S rRNA sequences that display homology is sometimes not indicative of species similarity, however [22]. Now, it is possible to do phylotaxonomy on complete genomes or groups of conserved proteins in order to improve phylogeny and by extension also improve taxonomical classifications. [1]

## ***2.6 Sequencing methods***

### 2.6.1 *First generation sequencing*

Genetic sequencing projects were first made possible through the invention of so called Sanger sequencing by Frederick Sanger et al (1977) [25]. Sanger sequencing is a manually intensive technique where special ddNTP nucleotides terminate the elongation of a fragment. The fragments are then separated via gel electrophoresis on four different lanes where the band positions reveal the genetic sequence [25]. This type of gel-based sequencing was the dominant type of DNA sequencing method for about 25 years and is still used for smaller sequencing projects, usually



due to its upside of generating longer reads.

### 2.6.2 *Second generation sequencing*

#### 2.6.2.1 *Pyrosequencing (454)*

In the 1990's, other sequencing methods like pyrosequencing, which is based on sequencing by synthesis, were developed, which were able to more rapidly generate sequence data with the downside of short reads lengths. These types of sequencing are starting to fall out of use with Hoffman LaRoche announcing the discontinuation of its 454 pyrosequencing platform in 2013, with newer sequencing methods like the ones by Illumina and Pacific Biosciences gaining market share. [26]

#### 2.6.2.2 *Illumina*

Illumina is a San Diego based company with a large line of genomics products. Their sequencing technology is based on systems developed by Solexa which was acquired by Illumina. By January 2014, Illumina held a 70% share of the sequencing machine market [27] and it was estimated that over 90% of the DNA sequence data generated was from Illumina machines [28].

The Illumina technology is also known as sequencing by synthesis. The genomic material is first fragmented into small pieces and ligated to different adapter sequences at each end. These sequences are then loaded into flow cells (a glass slide with lanes) where the 5'-ends hybridise to oligos which are complements of those adapters. Each bound fragment is then amplified into a clonal cluster by "bridge-amplification", where the complement of the 3'-adapter hybridises to a second type of oligo fixed to the bottom of the lane in the flow cell, after the original template has been used to polymerise its reverse complement and washed away. DNA polymerase then forms a double bridge which is denatured into two single strands attached to the lane floor; one forward and one reverse. This process being repeated results in the amplification of the DNA fragments. After amplification, the reverse strands are washed away and their oligos are blocked.

The sequencing is then done by adding a sequence primer that binds to the 5' end of the DNA fragment. Fluorescently tagged nucleotides are then added which compete for incorporation into the DNA replication. When a nucleotide is incorporated, a fluorescent tag is cleaved off which when excited by a lightsource fluoresces with a colour specific to that nucleotide. This can then be detected by an image sensor and recorded. Since the DNA strands are clustered and the bases will be incorporated at the same time, the fluorescence colour of one nucleotide incorporation in a cluster will be amplified to be visible to image sensors. The replication will continue until the selected number of nucleotide incorporation cycles are completed. The number of cycles will thus be the read length for this run. This is happening hundreds of millions of times in parallel, giving rise to the large amount of data per sequencing run. Then, the read product is washed away and a new sequence index “1” primer is added, which gives the read an indexing tag at the 5’-end of the sequence. This tag is used to identify and match the sequence to its pair and whether it is forward or reverse. The 3’-end of the sequence is then deprotected and the sequence folds over into a bridge. Index tag “2”, at the 3’ end of the sequence is then read. The strand is then completely polymerised to form a double bridge and then separated and linearised. The original template is then cleaved and washed off. The reverse sequence then remains. The second primer is then added and all the steps done for the forward strand is repeated for the reverse.

In silico, the samples are separated in groups dictated by their index tags. Overlapping reads are then clustered by overlapping regions and paired with their reverse partners in order to create contiguous sequences (contigs). [29, 30]

### 2.6.3 Third generation sequencing

#### 2.6.3.1 *Pacific Biosciences*

Pacific Biosciences (PacBio) is the Silicon Valley based company who developed the Single Molecule, Real-Time (SMRT) nucleotide sequencing technology. This technology enables sequencing of much longer reads (~5000 bp to ~20 000 bp) than other methods which require extensive fragment shearing of DNA before sequencing. The down-side of PacBio sequencing is that the reads are somewhat more error-prone (indel errors) than illumina sequenced reads and they also require more raw DNA input volume.

SMRT sequencing is done on SMRT cells (plates) on which there are thousands of ZMWs (Zero Mode Waveguides), extremely small wells through which light coming through from the bottom can only penetrate a very short distance (20 - 30 nm); due to the small diameter of the ZMW. A single DNA-polymerase + template complex is fixed to the bottom of each ZMW and phospho-linked nucleotides with different-coloured fluorophores are added. Whenever one of these nucleotides are held in the light-detection volume at the bottom of the ZMW during polymerisation, it will give off a pulse of light of the colour of the corresponding base as the fluorophore is cleaved off by the DNA polymerase. This is done in parallel over all the ZMWs on the SMRT-cell, generating a large amount of sequence data. [31]

Depending on which type of data that was sequenced, there are a variety of different assemblers that can be used for PacBio-generated data. When there is a high coverage of sequence data (>50x), PacBio-only *de novo* assemblies can be made. If the coverage is lower, hybrid assemblies where short reads are also used is recommended. The HGAP assembler is the best performing assembler for PacBio-only assemblies near the minimum coverage [32] (see Methods).

# 3. Materials & Methods

### *3.1 PacBio sequencing*

The input data was 22 SMRT-cells, 11 from *Skeletonema* cultures not treated with antibiotics and 11 from cultures that were treated with antibiotics. The 22 SMRT-cells contained 458 649 reads with a mean read length of 5984 bp and an N50 read length of 8137 bp. The total number of base pairs sequenced was 2 744 584 126.

### *3.1 Sequence assembly using HGAP.3 and the PacBio SmrtPortal*

A total of seven assemblies with varying *Minimum Seed Read Lengths* were run with the PacBio SmrtPortal HGAP (Hierarchical Genome Assembly Process) Assembler which is designed to be used for complete shotgun assembly of bacterial genomes. [34]

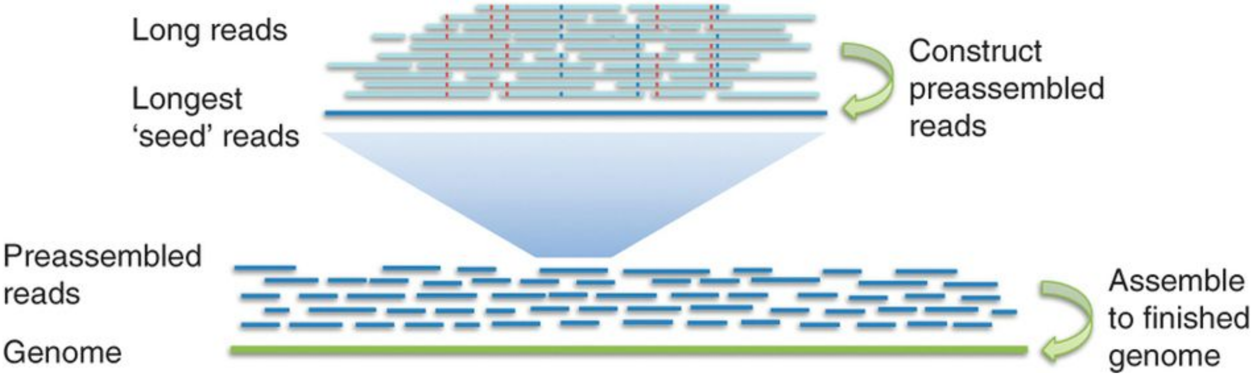


*Figure 4. HGAP assembly procedure. The longest reads are used to assemble seed reads which then assemble into large, bacterial genome-sized, fragments.* [34]

*Minimum Seed Read Length* is the minimum length where uncorrected reads are used as "seeds". These seeds then recruit the shorter reads for error correction and a consensus sequence is then generated, or assembled. These preassembled reads are then used in the next assembly step by the Celera assembler software, to generate unitigs, which can be several megabases in size (Figure 4).

The HGAP assembler also includes a correction/polishing step called Quiver which finds the maximum-likelihood template sequence when provided PacBio reads from the template. This is done through first prescribing a certain probability of a read through applying a conditional random field model. It then also looks at quality value covariates provided by the basecaller in order to achieve more a more accurate consensus sequence. Quiver doesn't use the alignment provided by the mapper, except on a macro scale for grouping the reads [33].

## ***3.2 Bacterial genome circularisation***

Since bacterial genomes are generally circular while the contiguous sequences created by the assemblers in a text file are linear, circularising the sequences is a way to ensure that the entire bacterial genomes has been correctly assembled without any missing fragments.

### *3.2.1 Circularising contigs from assembly with Amos minimus2*

In order to see if any of the unitigs in the assembly could be full genomes, tools from Amos v3.1.0 were used.

First, breaks were introduced into the unitigs of the assembly by, in a text-editor like Vim, going to the middle of a unitig sequence, making a new line and inserting "`>Break`" in order to indicate that these are now two separate sequences.

Then, the chopped up assembly was converted to the Amos .afg file format, a text format which normally holds read and consensus information together:

```
$ toAmos -s polished_assembly.fasta -o circularized.afg
```

Then the *Amos minimus2* module, which merges one or two sequence sets , was used to attempt to merge our manually broken up sequence set:

```
$ minimus2 circularized
```

Two different settings were used for genome circularisation; default and "relaxed". The relaxed setting was used when the default setting failed to produce output due to the sequence not meeting their requirements.

*Table 1. Amos minimus2 settings. MINID: Minimum percentage match required to create circularisation overlap. MAXTRIM: The maximum length (nt) allowed to be trimmed in order to make a good end-end match. OVERLAP: How long (nt) the overlap has to be for the circularisation to complete.*

| Setting | **MINID** | **MAXTRIM** | **OVERLAP** |
|---|---|---|---|

| | | | |
|---|---|---|---|
| *Default* | 94% | 20 | 40 |
| *“Relaxed”* | 70% | 10000 | 100 |

The resulting circularized.fasta file was quiver-corrected via the “RS_Resequencing.1” module available via SmrtPortal, a process which maps back the original reads to the circularised genome and quiver-corrects them in order to highlight possible errors or coverage abnormalities.

### *3.2.2 Manual recircularisation*

Unitigs that were not circularised in the *Amos minimus2* analysis were instead attempted to be manually circularised. This was done using blastn+ v.2.2.28 to align the unitigs back to the reference assembly in order to identify any overlap at the ends of the sequences:

```
$ blastn -query unitig_0.fasta -db 16536_polished_assembly.fasta -outfmt 5 -out
unitig_0_to_16536_BLASTn.xml
```

The resulting blast report in xml format was visualised with the Korilog BlastViewer application v2.5 (no longer available). This was done in order to see if any unitig had ends that potentially overlapped and hence could be circularised. The coordinates and motifs of these overlaps were then noted. In the text-editor Vim, a piece of the end of the sequence, upstream of the blast hit, was cut out. Then, using SeaView [36], these two segments were lined up to overlap and create a contiguous sequence with representation by both the “start” and “end” pieces over the regions they both occupy. After this, a consensus sequence of the two aligned pieces was created which results in a sequence where the original terminal ends of the unitig are now on the inside of the unitig (Figure 5).

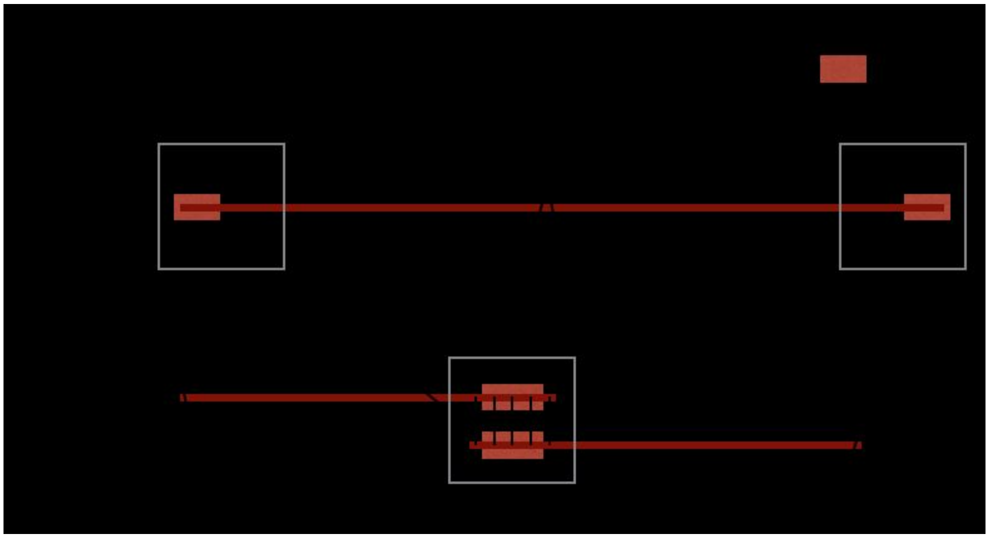


*Figure 5. Simplified visualisation of circularisation. A manually introduced break in the unitig is introduced after which the two previously terminal pieces are annealed together to reform the unitig.*

The alignment process consists of putting as many hyphen characters (“-”) in front of the “start” of the sequence as needed to get it to line up with the overlap at the “end” of the end piece (see Fig X). This was done in VIM and python:

Workflow using Vim:

1. Place marker at start of sequence.
2. Input number of hyphens with number keys.
3. Press “i” for insert. Press “-”, then press escape.

This only works if there are less than about 250 000 hyphens to be input. Otherwise it will be too slow and the Python method should be used instead.

Workflow using Python:
First the sequence header of the sequence that needs to be “pushed forward” was pasted into a new file. Then, on the command line:

```
$ python -c “print ‘-’ * 500000” >> newfile.fasta
```

This method  appends 500000 hyphens to the end of the file. Then fastaparser (**fp.py**; https://github.com/mtop/ngs/blob/master/fp.py) was used to extract the sequence from the file where a cut was made (the end of the genome) and appended it to the file containing hyphens:

```
$ fp.py --seq “sequenceheader” sourcefile.fa >> newfile.fasta
```

The extra header added by **fp.py** was the removed by searching in vim, using the search command "**/>**". The modified sequence (now starting with 50000 hyphen characters) was then appended to a file containing the "end" of the unitig. This file was then manually checked in seaview to see that the ends matched and also if there were additional indel errors that needed to be corrected by inputting more hyphens.

## ***3.3 Mapping and Alignment***

### ***3.3.1 bowtie2***

(http://computing.bio.cam.ac.uk/local/doc/bowtie2.html#what-is-bowtie-2)
Bowtie2 is an aligner that is specialised for aligning reads of length 50bp up to 1000s bp to a longer reference sequence, like a genome. It was used for aligning illumina paired-end reads of insert sizes 150bp, 300bp and 650bp and two MP libraries with insert sizes 2500bp and 4300bp, to the bacterial unitigs assembled from PacBio reads.

First, a database was created out of an existing assembly:

```
$ bowtie2-build -f <assemblyfile.fasta> <dbname>
```

Then, the bowtie2 mapping analysis was run to map the Illumina reads to the contigs in the assembly, only saving aligned reads to output:

```
$ bowtie2 -x <dbname> -1 <forward read file> -2 <reverse read file> -S
<outputfile.sam> --no-unal
```

Which produced the mapping result in SAM (Sequence Alignment/Map) format.

### ***3.3.2 SAMtools***

(https://samtools.github.io)
SAMtools is a collection of applications used to handle SAM and BAM (Binary Alignment/Map) formatted files. It was first used to convert the SAM files to their binary equivalent (BAM) in order to make the data more easily readable for the computer and smaller in size:

```
$ samtools view -Sb inputfile.sam > outputfile.bam
```

In order for most of the sequence viewing software, like igv (Integrative Genomics Viewer) or Tablet Viewer, to be able to recognise and display BAM filed in a human viewable format, the BAM files needed to be sorted and indexed:

```
$ samtools sort input.bam output_sorted.bam

$ samtools index output_sorted.bam
```

Which created a BAI (Binary Alignment Index) file to go with the BAM file. For a more detailed tutorial of bowtie2 mapping, visit
https://github.com/alvaralmstedt/Tutorials/wiki/Bowtie2-mapping

### *3.3.3 Mauve*

(http://darlinglab.org/mauve/mauve.html)
Mauve is a multiple genome alignment and viewer software that is used to detect large-scale evolutionary events like inversions and rearrangements.

The progressiveMauve alignment function of Mauve v.2.4.0 was used to align the circularised genomes derived from HGAP assembly to a reference genome of a species with good BLAST hits to that genome in order to determine the probable completeness of it.

### *3.3.4 Blastp to non-redundant protein database*

Annotated protein prediction multi-fasta files were used as a query to search the non-redundant protein database [35]:

```
$ blastp -query pretein_predicitons.faa -db /dbs/nr -outfmt 5 -num_threads 16
-out results_nr.xml
```

which returns the hits in XML format (-outfmt 5). To make a list of only the name of the hits in the xml file and place them in a separate list, containing one name per line:

```
$ grep -A 2 "<Hit_num>1</Hit_num>" results_nr.xml | grep -e "\[" | sed -e
"s/.*\[//g" | sed -e "s/\].*//g" > hit_list.lst
```

To get a count of the number of occurrences of each of these names in the list the short python script count_names.py (X) was used:

```
$ count_names.py hit_list.lst > hit_list_count.lst
```

## *3.4 Scaffolding*

### *3.4.1 SOAPdenovo2*

SOAPdenovo2 is a program used to assemble Illumina reads into contigs or scaffolds, and is part of SOAP (Short Oligonucleotide Analysis Package). SOAPdenovo2's  "data prepare" module was used on the reference assembly in order to generate scaffolds from the unitigs assembled by the PacBio HGAP assembler. An Illumina mate-pair library with an insert-size of ~3kb was used to link the unitigs.

The settings provided to SOAP in the configuration file:

```
#maximal read length
max_rd_len=150

#SkeletonemaMates 1
[LIB]
#average insert size
avg_ins=3000
#Mate Pair end library, orientation "1"
reverse_seq=1

#Use this library for all steps
asm_flags=3

#in which order the reads are used while scaffolding, starting with shortest
inserts first
rank=1

#Using cutoff of 5
pair_num_cutoff=5

#Using a high mapping length value for reliable placement of reads
map_len=48

#Skeletonema MP 3kb
q1=/data02/skeletonema_MP/matepairlibrary_forward_1.fastq
q2=/data02/skeletonema_MP/matepairlibrary_reverse_2.fastq
```

### *3.4.2 PBJelly*

PBJelly, which is part of the PBSuite of applications, is a program made to align long reads, such as PacBio- or long 454 reads, to a draft assembly or scaffold using the read mapper BLASR [37]. PBJelly will attempt to close gaps, usually represented as a series of “N”, in the draft assembly or scaffold sequence.

The settings for gap-closing with PBJelly were provided in an xml file where input and output files, as well as BLASR settings were specified:

```
<jellyProtocol>
<reference>/path/to/reference.fasta</reference>
<outputDir>/path/to/outputdirectory/</outputDir>
<blasr>-minMatch 8 -minPctIdentity 70 -bestn 1 -nCandidates 20
-maxScore -500 -nproc 20 -noSplitSubreads</blasr>
```

```
<input baseDir="/path/to/readdirectory/">
<job>filtered_subreads_over500bp.fasta</job>
</input>
</jellyProtocol>
```

Then, the steps of the analysis were run sequentially (Setup, Mapping, Support, Extraction, Assembly and Output) in order to generate a scaffold with fewer and/or shorter gaps. A more detailed version of the process can be found at https://github.com/alvaralmstedt/Tutorials/wiki/Gap-closing-with-PBJelly

## ***3.5 Annotation***

### ***3.5.1 Prokka***

For annotating, a program designed for quick prokaryote annotation, Prokka [39], was used with the default settings:

```
$ prokka circularised_genome.fasta
```

This generates a number of result files (ex. .faa .gbk .gff files) containing gene-predictions made from the input sequence. Those can then be used for further phylogenetic analysis.

### ***3.5.2 BASys***

The BASys automated bacterial annotation web platform [38] was used to gain increased annotation depth and visualisation of the bacterial genomes. The circularised genomes were uploaded via the web-interface on the BASys page (https://www.basys.ca) and specified as circular and gram-stained according to their closest relative. After the analysis has been completed BASys sends an email containing a link where the annotation results are displayed.

## ***3.6 Phylogeny***

### ***3.6.1 PhyloPhlAn***

A program called PhyloPhlAn [41] was used for phylogenetic analysis. The program is made for taxonomic classification using phylogenetic analysis of whole microbial genomes. It downloads a large number of bacterial reference genomes (>3000) that can be used along with user-provided genomes from species of interest in order to create a large phylogenetic tree. The trees are inferred through “minimum-evolution principle” with heuristic neighbor- joining, minimum-evolution interchanges and subtree-pruning-regrafting, and approximated maximum likelihood joining applying FastTree on the concatenated alignments (default JTT+CAT model)" [41]. This means that a rough topology is first gotten through neighbor joining, a fast and “greedy” method which starts out with a star-shaped tree and creates nodes between sequences which are more similar than all others in the set (least distant), usually from a distance matrix generated through running a multiple sequence alignment [42]. The branch lengths are then shortened by applying the minimum-evolution [43] criterion, which assumes that the shortest total length between all pairs of terminal nodes is the best. The tree is then further improved through maximum likelihood rearrangements by looking at the JTT and CAT substitution models for how likely mutations are to occur and modifying the tree accordingly. [44] [45]

PhyloPhlAn can also be used to create phylogenetic trees containing only sequences that the user has provided.

Multi-fasta amino-acid gene prediction files (.faa) generated from Prokka were provided to PhyloPhlan in order to first generate a tree using the large number of default bacterial genomes:

```
$ phylophlan.py -i project_name --nproc 8
```

The results generated were trees of newick format [46] where the leaves of the tree are denoted by their IMG taxon ID (http://img.jgi.doe.gov/). In order to get the taxonomic names on the leaves of the trees, the following command was run on the newick tree:

```
$ IFS=$'\n'; for r in `cat /home/username/programs/data/ppafull.tax.txt`; do
id=`echo ${r} | cut -f1`; tax=`echo ${r} | cut -f2`; sed -i
"s/${id}/${id}_${tax}/g"
/home/username/programs/nsegata-phylophlan-f2d78771d71d/output/job_name/genome.
tree.int.nwk; done; unset IFS
```

A full tutorial on how to get PhyloPhlAn running on the cluster Albiorix at the University of Gothenburg is available at:

https://github.com/alvaralmstedt/Tutorials/wiki/Creating-bacterial-phylogenetic-trees-with-PhyloPhlAn

The resulting tree contains 3741 leaves and in order to get trees of more manageable sizes, the PhyloPhlAn software was run on gene predictions of a limited number of bacterial reference genomes downloaded from the NCBI ftp server (ftp://ftp.ncbi.nlm.nih.gov/).  The reference genomes that were downloaded were ones which belonged to the same taxonomic orders that each of the target sequences got placed into in the large phylogenetic tree. This was in order to achieve better and more localised phylogenetic resolution surrounding the bacterial sequences of interest. Since the NCBI ftp genome database is organised alphabetically, not taxonomically, a custom Python script was written. The script, genome_donwloader.py (see Results, Coding), parses the user-specified folders on the NCBI ftp-server and downloads the contents of all subfolders where genus-names matches with any entries in a user-provided list of species genera. The genera list is intended to be pre-created by the user by searching the NCBI taxonomy browser (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) for the taxonomic range of interest (e.g. family, order or phylum)  and copying the corresponding genera names into a plain text, one name per line.

After the reference genomes available for the respective orders of interest were collected, PhyloPhlan was then run on the amino-acid gene prediction files in order to generate newick trees:

```
$ phylophlan.py -u project folder --nproc 8
```

The resulting newick file was then viewed using either FigTree v.1.4.2 (X) or TreeGraph 2 v.2.7.1 (X).

### ***3.6.2 16S analysis***

In order to isolate species-specific sequences needed for the adapter design but also to get additional taxonomic information about the bacterial sequences, a large number (~4 984 915) of 16S-RNA SSU Parc (Small Sub-Units from Prokaryotes and Archaea) sequences from different species and strains were downloaded from the SILVA database (http://www.arb-silva.de). These sequences were then used as a BLASTn database to which the circularised bacterial genomes were BLASTed. The sequences that had the top scoring alignments were excised from the circularised genomes as their corresponding 16S genes.

To find a variable region within the 16S gene, smaller regions were excised using the two bacterial primers Bakt_341F (CCTACGGGNGGCWGCAG) and Bakt_805R (GACTACHVGGGTATCTAATCC) (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3176514/). The sections flanked by these two primers were then aligned with clustal omega v1.1.0 or muscle v3.8.31 in seaview in order to locate variable regions that could be used to uniquely identify each species. Primer regions of length 20bp were designed which were confirmed to be unique through regex searches to the full circularised genomes and BLASTn to the NR database.

## ***3.7 Code generated for this project***

Several times during this project there was a need for coding custom scripts in order to accomplish certain tasks for which, to our knowledge, no software had previously been created. The following two were the most comprehensive:

**mapping_filtering.sh:**
(https://github.com/alvaralmstedt/shell_scripts/blob/master/mapping_filtering.sh)
A tool written for separating read libraries into mapped/unmapped/half_mapped (one of a pair) reads, depending on whether they map to reference sequence(s). This is useful if there are contaminants in your libraries which need to be removed. You can then map your reads to the

known contaminants and get out only the reads that did not map to the contaminants. This can then be repeated for further filtering against other contaminating sequences, thereby cleaning the read libraries.

**genome_downloader.py** *(in need of refactoring)***:**
(https://github.com/alvaralmstedt/python_genome_ref_collab/blob/master/genome_downloader.py)
Is a currently working but still work in progress python script that parses the NCBI ftp server folder system and downloads data from a set of species whose genus have been specified in a text list supplied by the user. The user can easily create this list by searching the NCBI taxonomy database and copying genus names of interest into a text file. This is useful for example when a certain taxonomic group is of interest for for a phylogenetic analysis. Currently, this script is able to download sequences from the following ftp directories:

- ftp://ftp.wip.ncbi.nlm.nih.gov/genomes/Bacteria/
- ftp://ftp.wip.ncbi.nlm.nih.gov/genomes/Bacteria_DRAFT/
- ftp://ftp.wip.ncbi.nlm.nih.gov/genomes/ASSEMBLY_BACTERIA/
- ftp://ftp.wip.ncbi.nlm.nih.gov/genomes/Fungi/

Note: The reason it can not be used for downloading animal references is due to a different folder structure containing that data. This would be possible to add in the future but would require some refactoring of the script

## 4. Results

### ***4.1 Initial HGAP assemblies***

Seven different HGAP assemblies were run on data from 22 PacBio SMRT cells, with different seed read lengths, in order to determine the optimal value for producing the longest unitigs.

*Table 1. Read lengths for HGAP assemblies with different minimum seed read lengths. The selected reference assembly is highlighted in blue.*

| Assembly # | Minimum Seed Read Length (kb) | Longest contig(kb) | 2nd Longest(kb) | 3rd Longest(kb) | 4th Longest(kb) |
|---|---|---|---|---|---|
| 1 | 13 | 2225 | 1553 | 885 | 758 |
| 2 | 15 | 629 | 363 | 322 | 303 |
| 3 | 10 | 3639 | 3515 | 1987 | 1226 |
| 4 | 8 | 3612 | 3520 | 1987 | 1516 |
| 5 | 6 | 3623 | 3141 | 1987 | 1516 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 6 | 4 | 3646 | 3141 | 1717 | 504 |
| 7 | 2 | 3626 | 3141 | 1717 | 504 |

From these assemblies, its was determined that a min. seed read length of 8kb was optimal, yielding the longest unitigs, with the other minimum seed read lengths seeming to be either too long or too short for optimal assembly (Table 1 & 2). This assembly (SmrtPortal job nr. 16536), was determined to be the reference assembly to be used for subsequent analyses.

*Table 2. Continuation of Table 1.*

| **Assembly #** | **Job ID#** | **Unmapped subreads** | **Polished contigs** | **Mapped reads (total: 458649)** | **Total number of reads** |
|---|---|---|---|---|---|
| 1 | 16521 | 131071 | 440 | 338294 | 469365 |
| 2 | 16522 | 158751 | 355 | 314823 | 473574 |
| 3 | 16535 | 132672 | 168 | 336685 | 469357 |
| 4 | 16536 | 125121 | 215 | 343160 | 468281 |
| 5 | 16539 | 127229 | 225 | 341295 | 468524 |
| 6 | 16541 | 132367 | 200 | 336969 | 469336 |
| 7 | 16542 | 134653 | 194 | 334974 | 469627 |

## ***4.2 Assembly #16536 (reference)***

The assembly selected to be used as reference (job nr. 16536) comprised 215 unitigs. Of these, 8 were shown to originate from *Skeletonema marinoi* (one chloroplast unitig + 7 small presumable nuclear unitigs) with the rest having a bacterial origin.
The 7 largest of these unitigs were bacterial and of sizes expected for bacterial genomes. These were selected to try to circularise into complete genomes. The 8th largest contig was shown to be a full match to the *Skeletonema marinoi* chloroplast, which has previously been assembled from Illumina data (Table 3).

*Table 3. The eight longest unitigs produced by HGAP. Their taxonomic affinity was determined through BLASTx matching and the average coverage was calculated with fasta_analyzer.py [X]. aff. = species affinis, the most closely related to but not formally a member of, species found.*

| Length (bp) | Unitig name | BLASTx match | Status | Taxonomic affinity | Avg. Coverage |
|---|---|---|---|---|---|
| 3612930 | unitig_0\|quiver | Bact. | *Circularised* | *aff. Dasania marina* | 205.1 |
| 3522709 | unitig_1\|quiver | Bact. | *Circularised* | *aff. Kordia algicida OT-1* | 122.0 |
| 1987565 | unitig_2\|quiver | Bact. | *Circularised* | *aff. Kordia algicida OT-1* | 146.1 |
| 1518362 (total: 3857472) | unitig_5\|quiver | Bact. | Scaffolded (#2) + Circularised. | *aff. Parvibaculum lamentivorans DS-1T* | 71.7 |
| 900916 | unitig_6\|quiver | Bact. | Scaffolded (#4) | | 78.9 |
| 861214 | unitig_7\|quiver | Bact. | Scaffolded (#3) | | 73.5 |
| 576980 | unitig_8\|quiver | Bact. | Scaffolded (#1) | | 78.4 |
| 113723 | unitig_9\|quiver | Chloroplast | Circularised | *Skeletonema* | 58.2 |

**unitig_0**

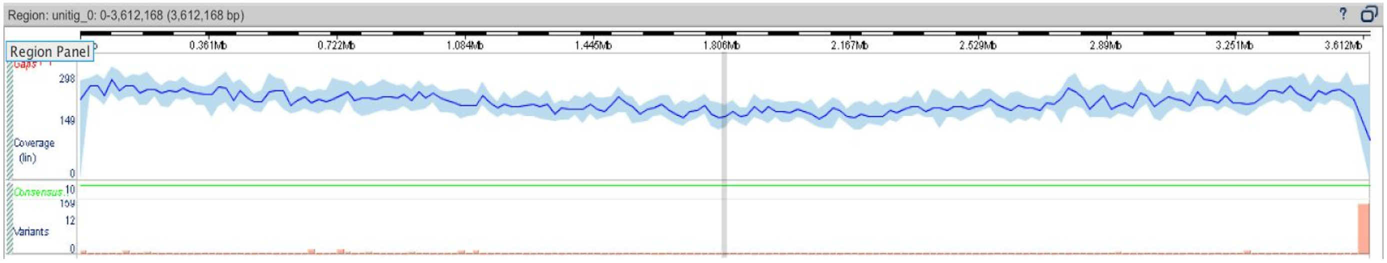


*Figure X. Coverage graph of unitig_0 from the reference assembly. Screencap from Pacific Biosciences SmrtViewer software. Dark blue line: median coverage, right blue area: span between min and max coverage, grey vertical line: zoom scale, green line: consensus quality value, green bars: sequence variants (insertions, deletions or substitutions).*

The largest unitig in the reference assembly was unitig_0 with a length of >3.6 Mb and an average coverage of 205.1 reads per base (Table 3). There was however a repetitive region near the end with many variants which presumably caused the coverage to drop and the HGAP assembler to stop the unitig building.

### unitig_1

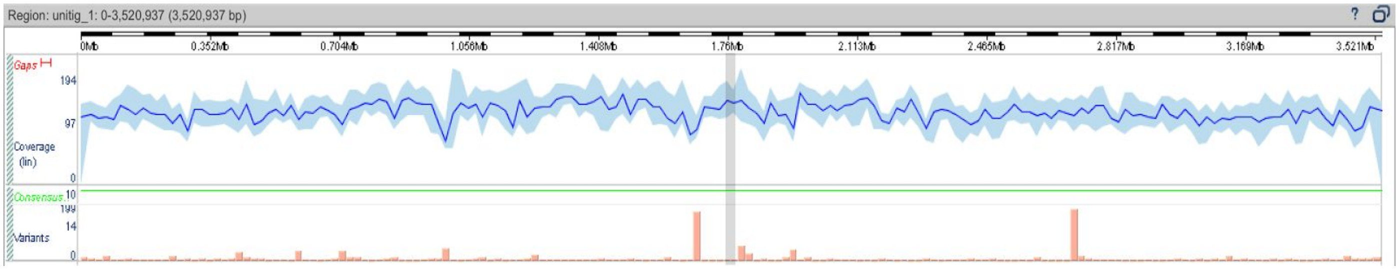


*Figure X. Coverage graph of unitig_1 from the reference assembly. Screencap from Pacific Biosciences SmrtViewer.*

The second longest (~3.52Mb) unitig in the reference was unitig_1 with an average coverage of 122.0 reads.

### unitig_2

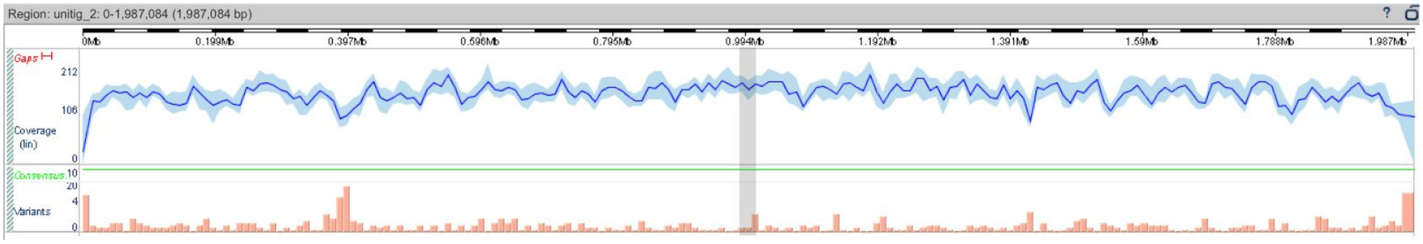


*Figure X. Coverage graph of unitig_2 from the reference assembly. Screencap from Pacific Biosciences SmrtViewer.*

The third largest (~1.99Mb) unitig in the reference was unitig_2 which had an average coverage of 146.1 reads with a sharp drop in coverage at the start of the unitig.

### unitig_5, 6, 7 & 8 (genome_4)

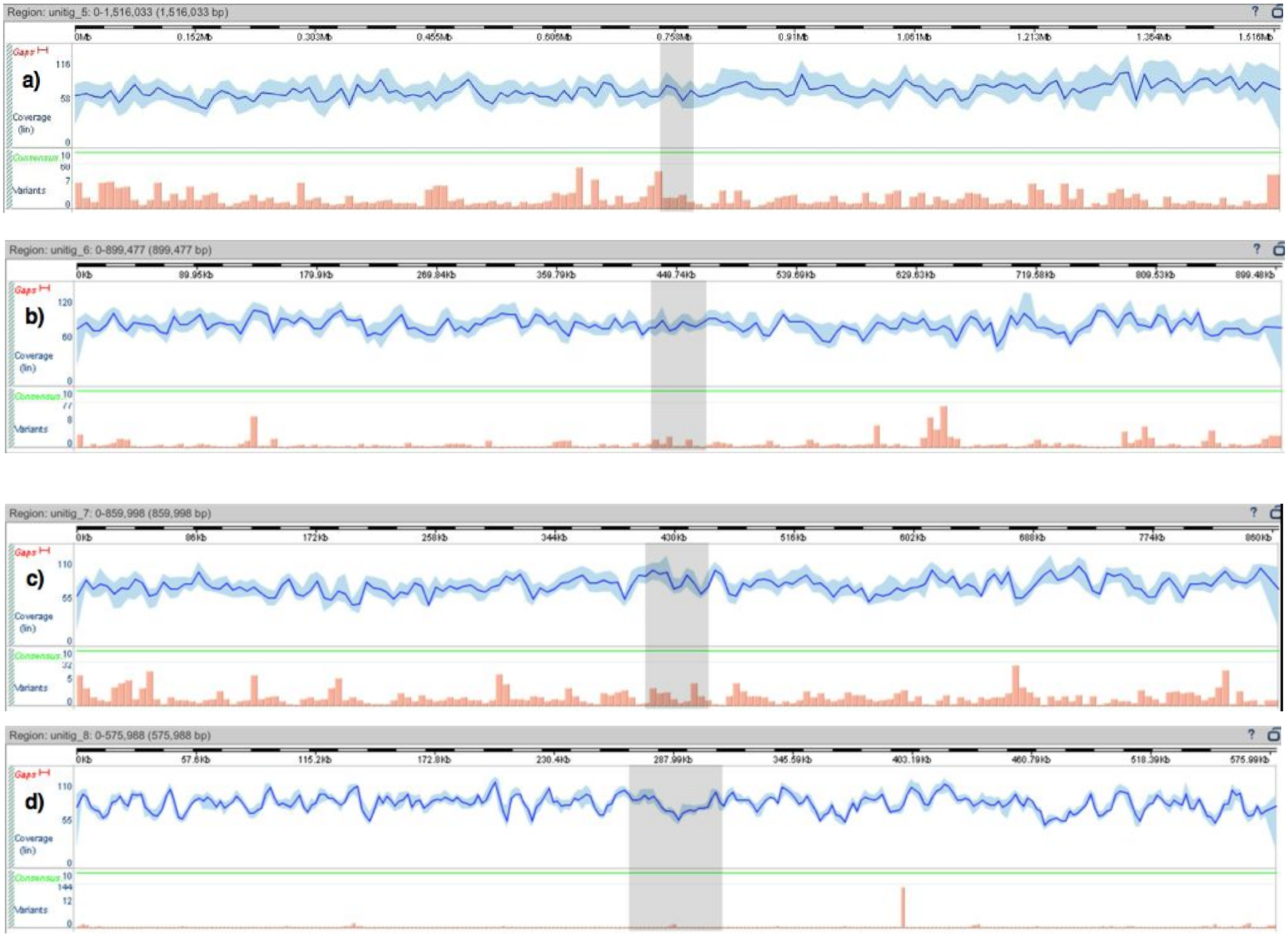


*Figure X. Coverage graphs of the four unitigs that comprise the genome here referred to as "genome_4". Screencap from Pacific Biosciences SmrtViewer. From top to bottom: (a) unitig_5, (b) unitig_6, (c) unitig_7 and (d) unitig_8.*

The fourth through seventh largest unitigs from the reference assembly were of lengths of ~1.52Mbp to ~577kb and all had a coverage between 70x and 80x.

## ***4.3 Circularisations***

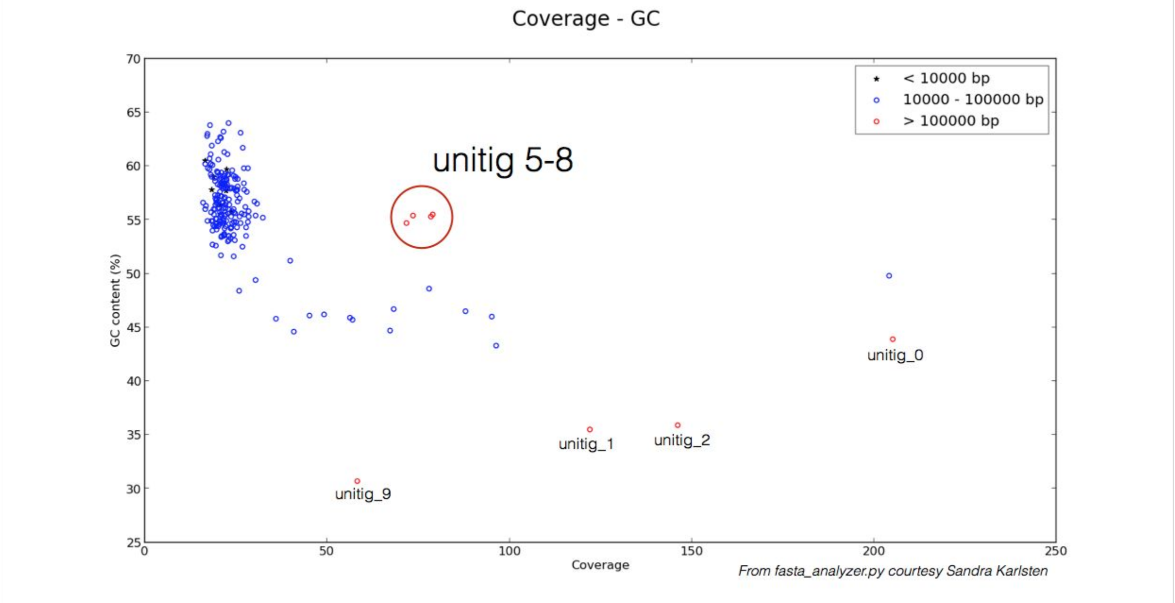
**unitig_1**

Circularisation of unitig_1 was done with *Amos minimus2* with the default settings (see Methods section 3.2.1).

**unitig_0 & 2**

Circularisation of unitig_0 and unitig_2 were initially done manually with vim/seaview after the *Amos minimus2* method failed due to unresolvable repetitive regions. The manual circularisations showed erratic drops and peaks in coverage when mapping back the PacBio reads, however, indicating improper circularisation (fig X). A slightly larger (27 807 bp longer) version of unitig_0 was found in assembly #16535 (see Table 2), which contained more of the terminal flanking repetitive regions. It was in all other regards identical to unitig_0 from assembly #16536. This unitig (unitig_3 from assembly #16535, still referred to as unitig_0 for consistency) and unitig_2 was then able to be circularised through *Amos minimus2* with more relaxed settings (see Table 1)

**unitig_5, 6, 7, 8**
The coverage to GC-ratio plot of the unitigs in the reference assembly was analysed in order to determine which unitigs may originate from the same source (see Figure X and Table 3). As can be seen, unitigs 5-8 form a small group with a GC content of about 55% and a coverage of ~ 75x.


Coverage - GC

*From fasta_analyzer.py courtesy Sandra Karlsten*

*Figure X. Coverage to GC-content plot of reference assembly (16536). From fasta_analyzer.py(X).*

These four unitigs were scaffolded (see Materials and Methods X; Results X) into a longer continuous fragment which was then able to be circularised using *Amos minimus*2

## ***4.4 Alignments***

### ***4.4.1 Blast***

Annotations of the circularised unitigs in protein multi-fasta format (.faa) generated from Prokka were used to perform a BLASTp search against the blast non-redundant protein database (nr)(See Methods section X). The hits against each unitig were quantified by occurrence in order to get more information about which species represented in the database most closely match the analysed bacterial genomes.

unitig_0

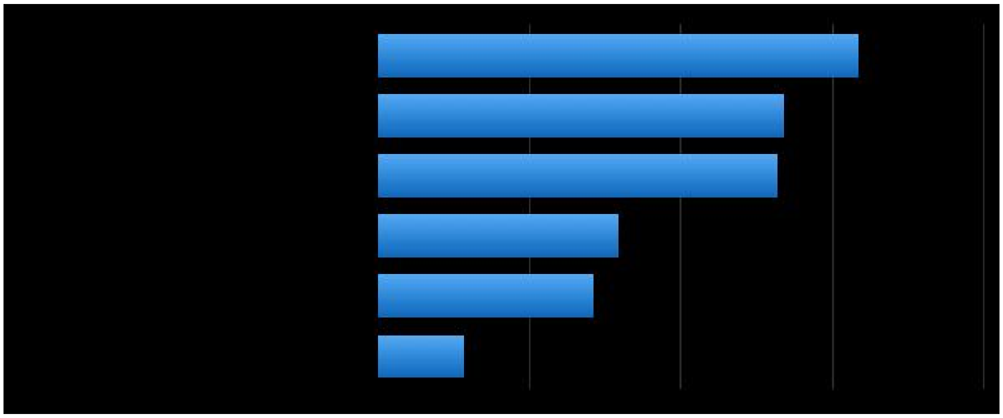


*Figure X. Top six identified species by number of BLASTp top-scoring HSPs against unitig_0.*

The marine gamma proteobacterium HTCC2143(http://www.ncbi.nlm.nih.gov/pubmed/20601481) & HTCC2148(http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2897341/) are strains of oligotrophic marine gammaproteobacteria found in waters of the coast of Oregon, USA.

unitig_1

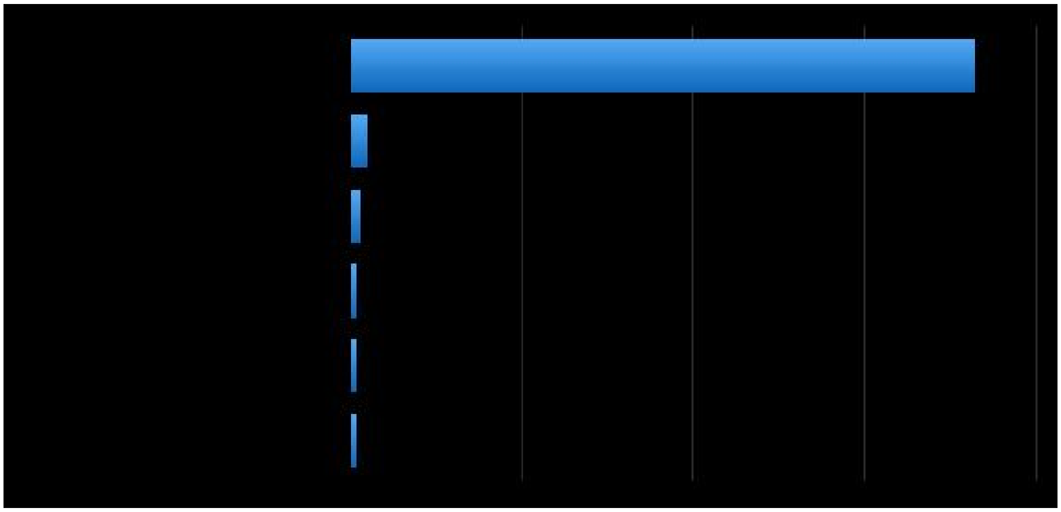


*Figure X. Top six identified species by number of BLASTp top-scoring HSPs against unitig_1*

unitig_2

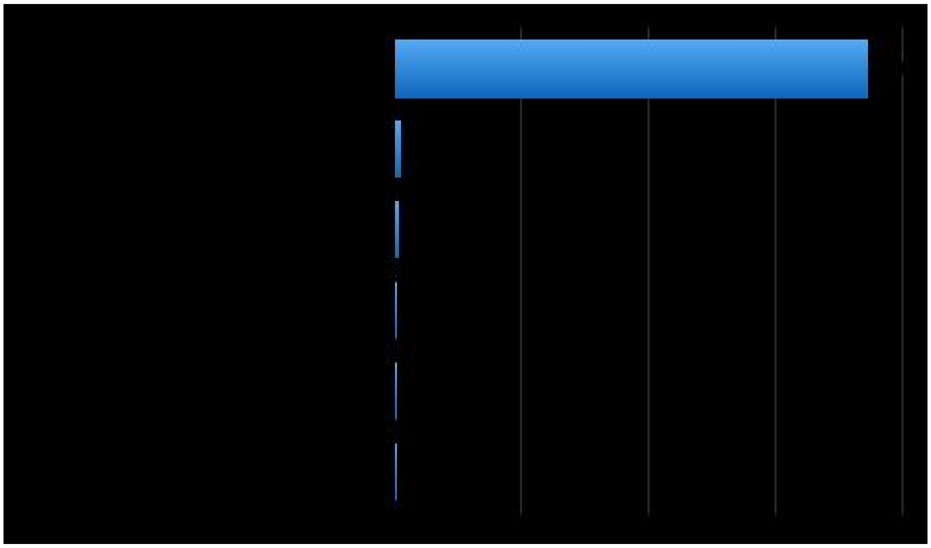


*Figure X. Top six identified species by number of BLASTp top-scoring HSPs against unitig_2*

unitig_1 and unitig_2 produced similar results; both matching by far the most to sequences from *Kordia algicida OT-1* with the remainder of the matches mostly being to other *flavobacteria* species.

genome_4

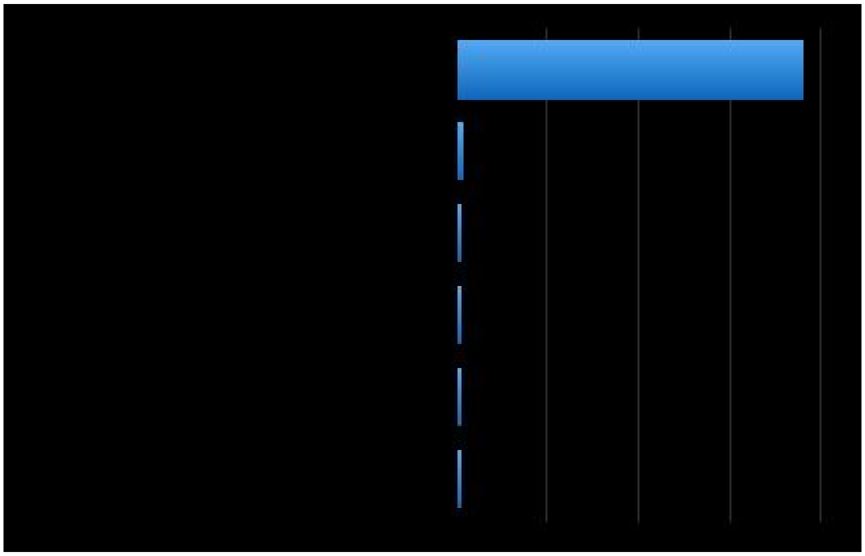


*Figure X. Top six identified species by number of BLASTp top-scoring HSPs against genome_4.*

## ***4.4.2 Mauve***

The initial blast result indicated that both unitig_1 and unitig_2 were closely related to *Kordia algicida* (see Fig. X). To further investigate this both fragments were aligned to a draft genome of *Kordia algicida OT-1* available from the NCBI ftp database(X). The result indicate that each of the two unitigs aligns to a different region of the *Kordia* draft genome.

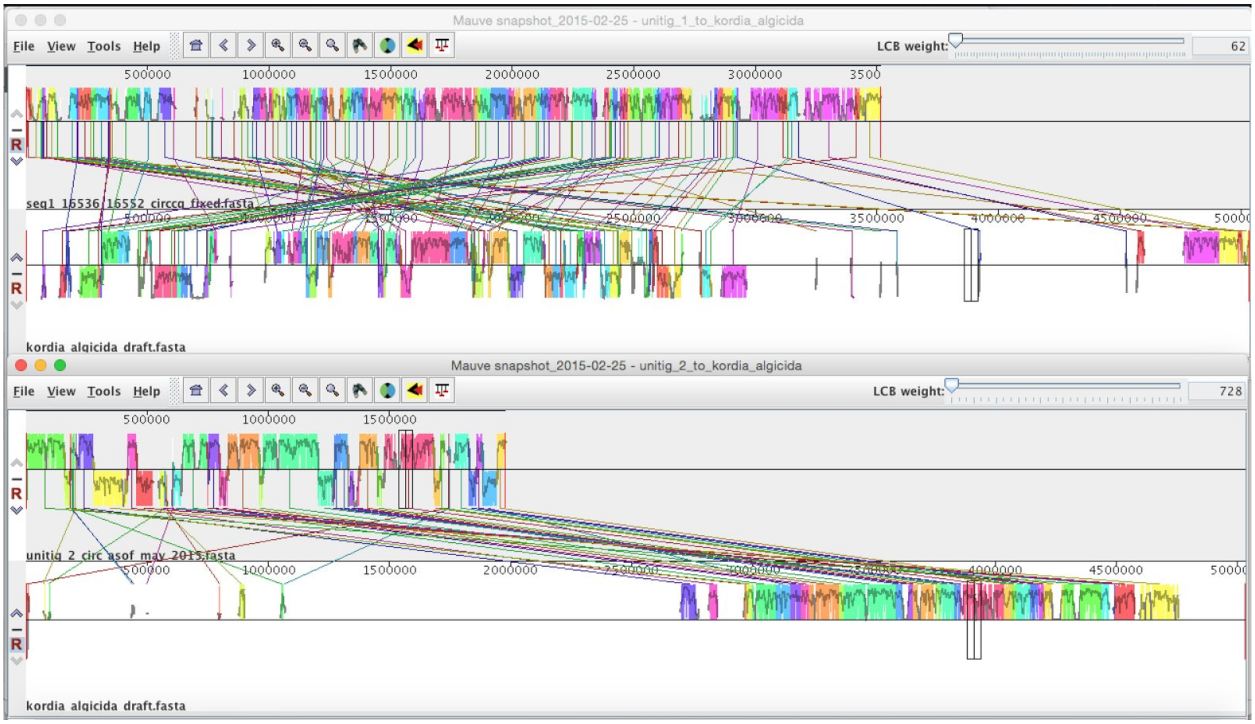


*Figure X. Alignments of unitig_1 (top) and unitig_2 (bottom) to Kordia algicida DRAFT reference genome indicates matches to two different regions of the reference genome.*

## ***4.5 Mapping & Scaffolding***

The four unitigs 5, 6, 7, 8 were suspected to be of the same origin because of their proximity to each other in the CG/Coverage plot (see X). To see if they were fragments of the same genome, only separated by repeated regions which broke the HGAP assembly, a scaffolding analysis was done using a Illumina mate-pair library with an insert size of ~3kb generated from a *Skeletonema marinoi* culture.

The unitigs were successfully scaffolded using SOAPdenovo in the order: unitig_8 - unitig_5 - unitig_7 - unitig_6; separated by different lengths of unknown bases, “N” (See figure X).

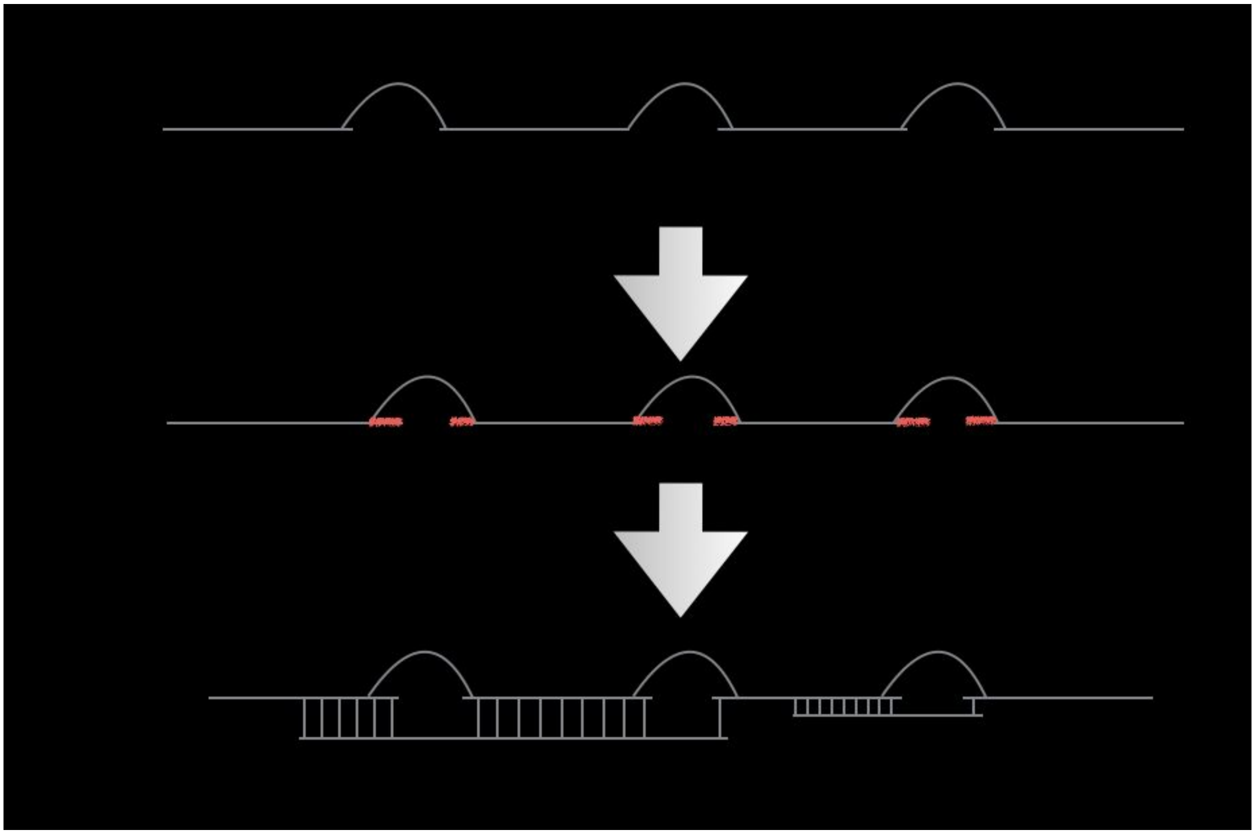


*Figure X. Workflow of scaffolding and gap-closing of genome_4.*

The length of the N-regions could be decreased by gap-closing with PBJelly, utilising the PacBio filtered subreads generated during the reference assembly process. The three gaps, of lengths 3134, 3498 and 1603 were respectively reduced to 1559, 1614 and 1370, a percentual decrease of 50.3%, 53.9% and 14.3%. This decrease in length of the N-regions was sufficient to completely close the gaps manually by bridging with contigs from an Illumina assembly from 300 bp and 600 bp libraries. The cut sites were confirmed by  observing overlapping predicted gene regions. This scaffold could then be circularised using the *Amos minimus2* method. The final length of the circularised scaffold “genome_4” was 3 861 475 bp, 4003 bp longer than the sum of the original contigs; 3 857 472 bp.

## ***4.6 Annotation***

*Table 4. Summary of results from Prokka annotation of bacterial genomes. CDS: Coding DNA Sequences, tRNA: transfer RNA genes, tmRNA: transfer messenger RNA genes, repeat regions: regions with repeating sequence elements.*

| genome | CDS | tRNA | tmRNA | repeat regions |
|---|---|---|---|---|
| unitig_0 | 3376 | 37 | 1 | n/a |
| unitig_1 | 2948 | 30 | 1 | 2 |
| unitig_2 | 1740 | 33 | n/a | n/a |
| genome_4 | 3795 | 42 | 2 | n/a |

*Table 5. Summary of results from BASys annotation of bacterial genomes.*

| genome | identified genes | annotated genes |
|---|---|---|
| unitig_0 | 3559 | 2128 |
| unitig_1 | n/a | n/a |
| unitig_2 | 1758 | 678 |
| genome_4 | 4165 | 2152 |

## ***4.7 Phylogenetic analysis***

Using the large library of bacterial genome data available via PhyloPhlAn (see Materials & Methods X), a large (3,171 microbial genomes) tree was constructed that included all the assembled and circularised bacterial genomes, in order to look at their phylogenetic position at a large scale (see https://github.com/alvaralmstedt/mthesis/blob/master/fulltree_bac_genomes_29_jun_2015.tree.int.nwk. This tree elucidated the general positions of the bacterial genomes in the bacterial tree of life (Table 6).

*Table 6. Summary of taxonomic affiliations derived from PhyloPhlAn large microbial tree of life. The affiliations are from the most closely related leaf on the tree. The * character indicates that the two equally close classified leaves were of orders pseudomonadales and alteromonadales.*

| fragment | phylum | class | order | family | genus |
|---|---|---|---|---|---|
| unitig_0 | Proteobacteria | Gammaproteobacteria | unclassified* | unclassified | Congregibacter |
| unitig_1 | Bacteroidetes | Flavobacteria | Flavobacteriales | Flavobacteriaceae | Kordia |
| unitig_2 | Bacteroidetes | Flavobacteria | Flavobacteriales | Flavobacteriaceae | Kordia |
| genome_4 | Proteobacteria | Alphaproteobacteria | Rhizobiales | Phyllobacteriaceae | Parvibaculum |

In order to get a more precise phylo-taxonomic placement for the bacterial genomes, and analysis of a smaller cohort of more closely related reference genomes was needed for increased resolution. These reference genomes were obtained from the NCBI ftp site (see Materials & Methods X).

All bacterial reference genomes and draft reference genomes from the taxonomic orders the assembled bacterial genomes belonged to were obtained and, using PhyloPhlan, a new phylogenetic tree was inferred. Leaves with names ending with ".concat" are neither completed genomes nor draft genomes, but instead concatenated contigs from assemblies of sequence libraries of the species name given in the leaf (ftp://ftp.wip.ncbi.nlm.nih.gov/genomes/ASSEMBLY_BACTERIA/).

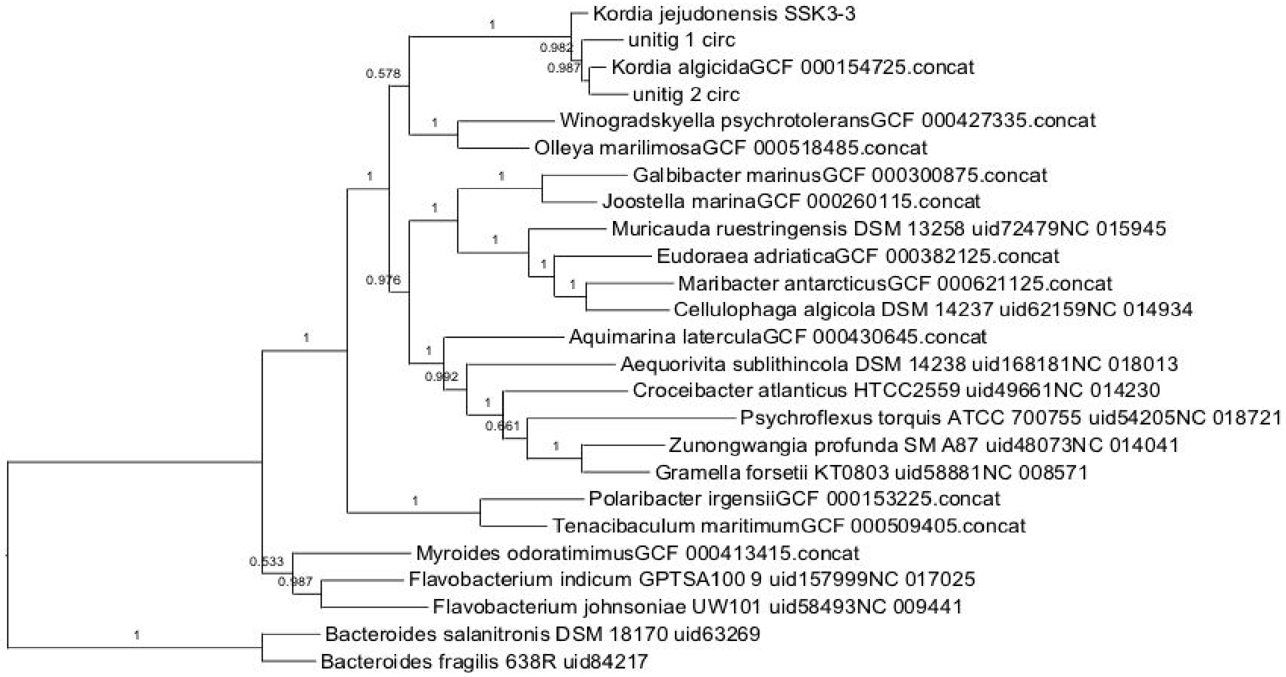

Kordia jejudonensis SSK3-3
unitig 1 circ
Kordia algicidaGCF 000154725.concat
unitig 2 circ
Winogradskyella psychrotoleransGCF 000427335.concat
Olleya marilimosaGCF 000518485.concat
Galbibacter marinusGCF 000300875.concat
Joostella marinaGCF 000260115.concat
Muricauda ruestringensis DSM 13258 uid72479NC 015945
Eudoraea adriaticaGCF 000382125.concat
Maribacter antarcticusGCF 000621125.concat
Cellulophaga algicola DSM 14237 uid62159NC 014934
Aquimarina laterculaGCF 000430645.concat
Aequorivita sublithincola DSM 14238 uid168181NC 018013
Croceibacter atlanticus HTCC2559 uid49661NC 014230
Psychroflexus torquis ATCC 700755 uid54205NC 018721
Zunongwangia profunda SM A87 uid48073NC 014041
Gramella forsetii KT0803 uid58881NC 008571
Polaribacter irgensiiGCF 000153225.concat
Tenacibaculum maritimumGCF 000509405.concat
Myroides odoratimimusGCF 000413415.concat
Flavobacterium indicum GPTSA100 9 uid157999NC 017025
Flavobacterium johnsoniae UW101 uid58493NC 009441
Bacteroides salanitronis DSM 18170 uid63269
Bacteroides fragilis 638R uid84217
1
0.982
0.987
0.578
1
1
1
1
1
1
1
0.976
1
1
0.992
1
0.661
1
1
0.533
0.987
1

*Figure X. Rooted subsampled tree showing the phylogenetic position of  unitig_1 and unitig_2 into the bacterial order Flavobacteriales. Two sequences from the order Bacteroides was used as outgroup.*

unitig_1 and unitig_2 are most closely related to the draft genome of *Kordia algicida* and an annotated assembly of *Kordia jejudonensis* (http://www.ncbi.nlm.nih.gov/Traces/wgs/?&val=LBMG01&size=100&size=all) within family *Flavobacteriaceae* (Figure X). This enhances the indications that these are two fragments comprising the genome of a single bacterial specie which were separated by a large repeated region which made it difficult for the assembler to bridge and circularise. The full tree in newick format containing all flavobacteriales species available on the NCBI genomes ftp (as of September 2015) can be found here(X).

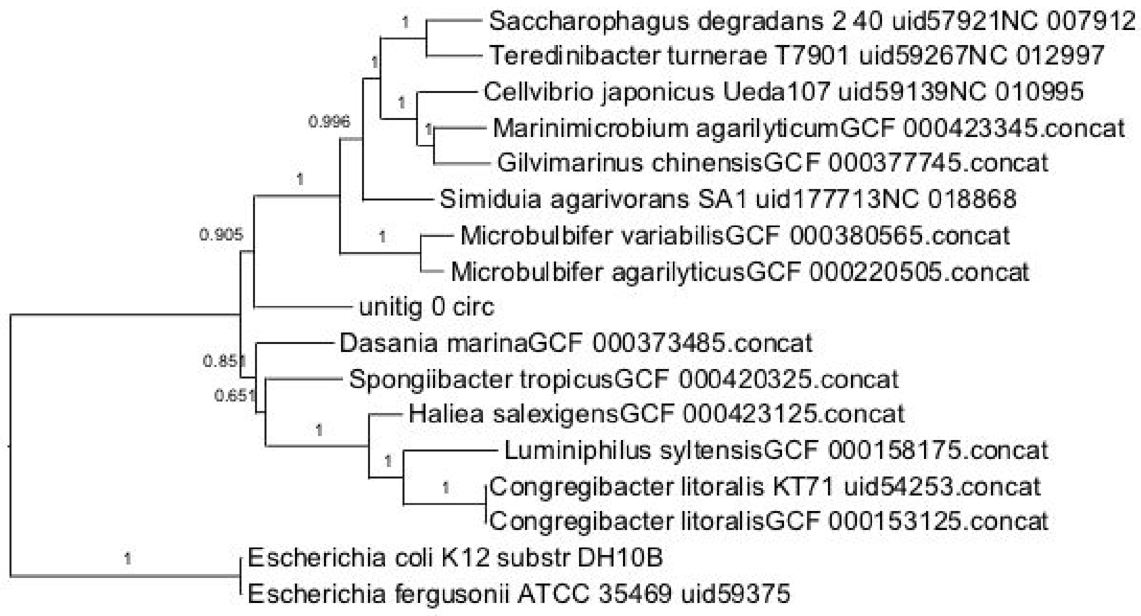

Saccharophagus degradans 2 40 uid57921NC 007912
Teredinibacter turnerae T7901 uid59267NC 012997
Cellvibrio japonicus Ueda107 uid59139NC 010995
Marinimicrobium agarilyticumGCF 000423345.concat
Gilvimarinus chinensisGCF 000377745.concat
Simiduia agarivorans SA1 uid177713NC 018868
Microbulbifer variabilisGCF 000380565.concat
Microbulbifer agarilyticusGCF 000220505.concat
unitig 0 circ
Dasania marinaGCF 000373485.concat
Spongiibacter tropicusGCF 000420325.concat
Haliea salexigensGCF 000423125.concat
Luminiphilus syltensisGCF 000158175.concat
Congregibacter litoralis KT71 uid54253.concat
Congregibacter litoralisGCF 000153125.concat
Escherichia coli K12 substr DH10B
Escherichia fergusonii ATCC 35469 uid59375
1
1
1
1
0.996
1
1
0.905
0.851
0.651
1
1
1
1

*Figure X. Subsampled tree of unitig_0 inserted into the bacterial order Cellvibrionales. Outgroup: Escherichia.*

Phylogenetic analysis of unitig_0 showed that the genome is phylogenetically located within the *Cellvibrionales* order. There was no close species match, making it hard to identify family or genus affiliations. The full tree, in newick format, containing all *Cellvibrionales* species available on the NCBI ftp (as of September 2015) can be found here.

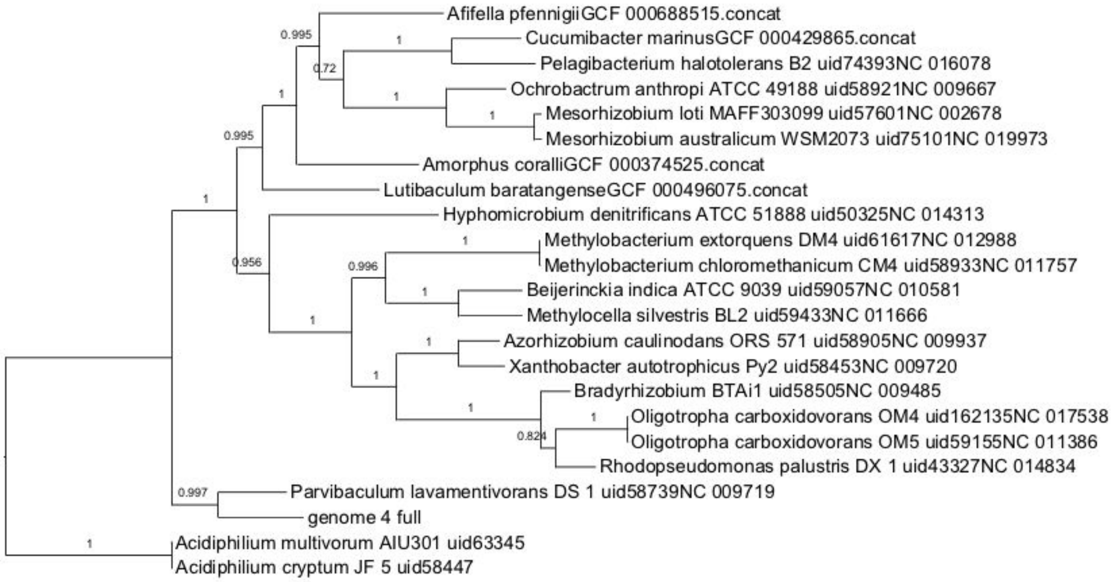

Afifella pfennigiiGCF 000688515.concat
Cucumibacter marinusGCF 000429865.concat
Pelagibacterium halotolerans B2 uid74393NC 016078
Ochrobactrum anthropi ATCC 49188 uid58921NC 009667
Mesorhizobium loti MAFF303099 uid57601NC 002678
Mesorhizobium australicum WSM2073 uid75101NC 019973
Amorphus coralliGCF 000374525.concat
Lutibaculum baratangenseGCF 000496075.concat
Hyphomicrobium denitrificans ATCC 51888 uid50325NC 014313
Methylobacterium extorquens DM4 uid61617NC 012988
Methylobacterium chloromethanicum CM4 uid58933NC 011757
Beijerinckia indica ATCC 9039 uid59057NC 010581
Methylocella silvestris BL2 uid59433NC 011666
Azorhizobium caulinodans ORS 571 uid58905NC 009937
Xanthobacter autotrophicus Py2 uid58453NC 009720
Bradyrhizobium BTAi1 uid58505NC 009485
Oligotropha carboxidovorans OM4 uid162135NC 017538
Oligotropha carboxidovorans OM5 uid59155NC 011386
Rhodopseudomonas palustris DX 1 uid43327NC 014834
Parvibaculum lavamentivorans DS 1 uid58739NC 009719
genome 4 full
Acidiphilium multivorum AIU301 uid63345
Acidiphilium cryptum JF 5 uid58447
0.995
1
0.72
1
1
0.995
1
1
0.956
0.996
1
1
1
1
1
0.824
1
0.997
1

*Figure X. Subsampled tree of genome_4 inserted into the bacterial order Rhizobiales. Outgroup: Acidiphilum.*

The closest relative of genome_4 was identified as *Parvibaculum lavamentivorans* within the *Rhodobiaceae* family. The full tree, in newick format, containing all Rhizobiales reference genomes available on the NCBI ftp database (as of September 2015) can be found here.

## ***4.8 16S rRNA analysis***

In order to find unique regions for primer design in the bacterial genomes, the variable regions of the 16S were selected as targets for capturing the bacterial species of interest. This is needed for amplification of the correct sequences when preparing new DNA samples for sequencing.

The blast analysis also revealed more about the identities of the bacterial genomes' identities (Table 7).

*Table 7. Summary of BLAST results from SILVA 123 16S SSU to bacterial genomes*

| Genome/unitig | 16S start position | 16S end position | 16S gene hit length | Closest match family | Best match to species | Match E - value | Match identities (percentage) |
|---|---|---|---|---|---|---|---|
| unitig_0 | 894232 | 895623 | 1392 | Halieaceae | *uncultured bacterium (JN018457)* | 0.0 | 1385/1392 (99%) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| unitig_1 | 1007861 | 1009381 | 1521 | Flavobacteriaceae | *Kordia algicida(ABIB 01000004)* | 0.0 | 1478/1522 (97%) |
| unitig_2 | 405093 | 403573 | 1521 | Flavobacteriaceae | *Kordia algicida(ABIB 01000004)* | 0.0 | 1478/1522 (97%) |
| genome_4 | 3524965 | 3523486 | 1480 | Rhodobiaceae | *uncultured alphaprote-obacterium(BAOK01000002)* | 0.0 | 1480/1480 (100%) |

## 5. Discussion

### ***5.1 Kordia sp.***

When unitig_1 and unitig_2 were able to be circularised independently, it was assumed that they were two related but separated bacterial genomes from looking at early blast match reports. When looking at at the Phylogeny, 16S analysis and Mauve analysis, it seems more likely that these are two large fragment of the same bacterial genome. Since mapping data supports the reference assembly, it can be presumed that the reason these two fragments were not assembled into a single fragment and that they were then able to be circularised independently is that the genome contains two repeat regions with the same or very similar repeat patterns.

This species (whose genome looks to be comprised of both unitig_1 and unitig_2) is a close relative of the parasitic *Kordia algicida*, which is a bacteria that kills algal cells for their own benefit when they reach a certain population size threshold; i.e. quorum sensing (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3117869/). Unfortunately, there are not many members of the *Kordia* genus that have complete reference genome sequences available, making pinpointing an exact phylogenetic position difficult.
The original *Kordia algicida OT-1* strain which was sequenced and added to the NCBI draft genome reference database was taken from the waters of Masan Bay, South Korea during a red bloom of *Skeletonema costatum*. This is an interesting organism because its algicidal activity can be one of the ways to limit harmful algal blooms, an occurrence which may become more common as the CO2 levels rise in the oceans due to climate change. The other Kordia species with assembled genetic data publicly available, *Kordia jejudonensis*, was found at a freshwater outlet on the Korean island Jeju. It was confirmed to be a separate species from *K. algicida* through 16S rRNA analysis. The other *Kordia* species described so far; *Kordia perrisulae* (http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.022764-0) and

*Kordia aquimaris* (http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.056051-0) have been found in east asian waters with *Kordia antarctica* (http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.052738-0) being found in coastal seawater off the antarctic peninsula. Hence, this is to my knowledge the first time a *Kordia* species has been described from European waters. The size of *Kordia algicida* draft genome (~5.03Mb) is smaller than unitig_1 and unitig_2 put together (~5.5Mb), meaning there could be significant differences between the two.

## ***5.2 Parvibaculum sp.***

The bacterial genome with the working name genome_4 was comprised of four unitigs (unitig_5, unitig_6, unitig_7 & unitig_8) separated by repeated regions which were, through GC/coverage analysis, suspected to be from the same origin. The essential 16S rRNA gene was also only found in unitig_8 but not in the other three unitigs, enforcing this hypothesis. After successfully scaffolding and gap closing these unitigs, the resulting genome was able to be circularised (see). Additionally, mapping and gene predictions which overlap joining regions indicate that there are no additional fragments missing from this genome.

Genome_4 was shown, through both phylogenetic analysis and 16S rRNA sequence comparison, to be a close relative of *Parvibaculum lavamentivorans*. This species is a bacteria heavily related to degradation of hydrocarbons such as linear alkylbenzenesulfonates (LAS)(http://www.ncbi.nlm.nih.gov/pubmed/16075201, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3368416/ ). *P. lavamentivorans* is commonly found in settings where some hydrocarbon pollution has occurred, such as in biofilms on rocks in areas where there have been diesel spills. The relation of this bacteria to *Skeletonema* or diatoms remains unclear, however.

## ***5.3 Unitig_0***

The genome that proved the most difficult to taxonomically classify was the one with the working name unitig_0. It has been shown that unitig_0 belongs to the bacterial class *Gammaproteobacteria*. Both the BLASTp alignments and Phylogenetic analysis showed *Dasania marina* as a result for a possible relative to unitig_0. However, unlike the other genomes, there have been no clear close relative to unitig_0. This is an interesting target for a possible new genus, although the completeness of the genome needs to be determined first.

## ***5.4 Conclusions***

In order to proceed with the identification of these bacterial genomes, a number of things need to be done, the main one being an assessment of their completeness. This can be done by comparison of generated gene predictions of the genomes with a set of genes known to be essential to all (bacterial) life, or what is known as core genes (as opposed to pan genes). If there are core genes missing from the genomes, we can assume there are sections missing. If this happens, more data will be needed to build out the contigs. There is nothing currently indicating that this is the case, however. We are currently attempting to design primers targeting variable regions in the 16S rRNA gene which are unique to these species in order to get targeted libraries.

With concern to unitig_1 and unitig_2, the obvious thing that needs to be done is to join these two contigs into one large genomic fragment. Since they were able to be circularised individually, it seems as if the repeating regions flanking the unitigs, and thus leading to assembly stop, are similar. So far, attempts to do this manually have failed, although only a small amount of time has been dedicated to it.

# 6. References

## *6.1 Articles and Papers*

1. Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., … Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. Functional & Integrative Genomics, 15(2), 141–161. http://doi.org/10.1007/s10142-015-0433-4

2. Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Meth, 12(8), 733–735. Retrieved from http://dx.doi.org/10.1038/nmeth.3444

3. Amin, S. A., Parker, M. S., & Armbrust, E. V. (2012). Interactions between Diatoms and Bacteria. Microbiology and Molecular Biology Reviews : MMBR, 76(3), 667–684. http://doi.org/10.1128/MMBR.00007-12

4. Paul, C., & Pohnert, G. (2011). Interactions of the Algicidal Bacterium Kordia algicida with Diatoms: Regulated Protease Excretion for Specific Algal Lysis. PLoS ONE, 6(6), e21032. http://doi.org/10.1371/journal.pone.0021032

5. Phycology (Book) by Robert Edward Lee 4th ed. pp 391

6. Yool, A., and T. Tyrrell (2003), Role of diatoms in regulating the ocean's silicon cycle, Global Biogeochem. Cycles, 17, 1103, doi:10.1029/2002GB002018, 4.

7. Stellwagen Bank National Marine Sanctuary - Phytoplankton

8. Algae: An Introduction to Phycology (Book) By Christiaan Hoek, David Mann, H. M. Jahns pp 135
9. Oceanic Acidification: A Comprehensive Overview by Ronald Eisler pp 110

10. Queiroz, K. De. (1992). Phylogenetic definitions and taxonomic philosophy. Biology and Philosophy, (1990), 295–313. Retrieved from http://link.springer.com/article/10.1007/BF00129972

11. Grossart H-P, Levold F, Allgaier M, Simon M, Brinkhoff T. 2005. Marine diatom species harbour distinct bacterial communities. Environ. Microbiol. 7:860 – 873

12. Guannel ML, Horner-Devine MC, Rocap G. 2011. Bacterial commu- nity composition differs with species and toxigenicity of the diatom Pseu- do-nitzschia. Aquat. Microb. Ecol. 64:117–133.

13. Fukao T, Kimoto K, Kotani Y. 2010. Production of transparent exopolymer particles by four diatom species. Fish. Sci. 76:755–760.

14. Rinta-Kanto JM, Sun S, Sharma S, Kiene RP, Moran MA. 2012. Bacterial community transcription patterns during a marine phyto- plankton bloom. Environ. Microbiol. 14:228 –239

15. Croft MT, Lawrence AD, Raux-Deery E, Warren MJ, Smith AG. 2005. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. Nature 438:90 –93

16. Amin SA, et al. 2009. Photolysis of iron-siderophore chelates promotes bacterial-algal mutualism. Proc. Natl. Acad. Sci. U. S. A. 106:17071–17076

17. Cho BC, Azam F. 1988. Major role of bacteria in biogeochemical fluxes in the ocean's interior. Nature 332:441–443

18. Foster RA, et al. 2011. Nitrogen fixation and transfer in open ocean diatom-cyanobacterial symbioses. ISME J. 5:1484 –1493

19. Hünken M, Harder J, Kirst GO. 2008. Epiphytic bacteria on the Ant- arctic ice diatom Amphiprora kufferathii Manguin cleave hydrogen per- oxide produced during algal photosynthesis. Plant Biol. 10:519 –526.

20. de Queiroz, K.: 1988, 'Systematics and the Darwinian Revolution', Philosophy of Science 55, 238-259

21. Jan P Meier-Kolthoff et al. Complete genome sequence of DSM 30083T, the type strain (U5/41T) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. Standards in Genomic Sciences 2014, 9:2 doi:10.1186/1944-3277-9-2

22. Reeck G.R., Haën C., Teller D.C., Doolittle R.F., Fitch W.M., Dickerson R.E., Chambon P., McLachlan A.D., Margoliash E., Jukes T.H., et al. (1987) Cell 50:667

23. J. Peter Gogarten, W. Ford Doolittle and Jeffrey G. Lawrence. Prokaryotic Evolution in Light of Gene Transfer. Mol Biol Evol (2002) 19 (12): 2226-2238

24. Nesbo C. L., Y. Boucher, W. F. Doolittle, 2001 Defining the core of nontransferable prokaryotic genes: the euryarchaeal core J. Mol. Evol 53:340-350

25. F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec; 74(12): 5463–5467.

26. Julia Karow. Following Roche's Decision to Shut Down 454, Customers Make Plans to Move to Other Platforms. genomeweb. October 22, 2013

27. Eilene Zimmerman. 50 Smartest Companies: Illumina. MIT Technology Review. February 18, 2014

28. Antonio Regalado. EmTech: Illumina Says 228,000 Human Genomes Will Be Sequenced This Year. September 24, 2014

29. Illumina Inc. Intro to Sequencing by Synthesis: Industry-leading Data Quality (video). Youtube. April 2, 2014

30. Illumina Inc. An Introduction to Next-Generation Sequencing Technology. 21 April, 2015

31. Pacific Biosciences. Introduction to SMRT Sequencing. December 5, 2011

32. Ryan Wick. Large Genome Assembly with PacBio Long Reads. November 18, 2015

33. Pacific Biosciences. Quiver FAQ. November 26, 2015

34. Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., … Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Meth, 10(6), 563–569. Retrieved from http://dx.doi.org/10.1038/nmeth.2474

35. Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 33(Database Issue), D501–D504. http://doi.org/10.1093/nar/gki025

36. Manolo Gouy, Stéphane Guindon, and Olivier Gascuel SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building Mol Biol Evol (2010) 27 (2): 221-224 first published online October 23, 2009 doi:10.1093/molbev/msp259

37. Chaisson, M. J., & Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics, 13(1), 1–18. http://doi.org/10.1186/1471-2105-13-238

38. Van Domselaar, G. H., Stothard, P., Shrivastava, S., Cruz, J. A., Guo, A., Dong, X., … Wishart, D. S. (2005). BASys: a web server for automated bacterial genome annotation. Nucleic Acids Research, 33(Web Server issue), W455–W459. http://doi.org/10.1093/nar/gki593

39. Torsten Seemann Prokka: rapid prokaryotic genome annotation Bioinformatics (2014) 30 (14): 2068-2069 first published online March 18, 2014 doi:10.1093/bioinformatics/btu153

40. Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix Mol Biol Evol (2009) 26 (7): 1641-1650 first published online April 17, 2009 doi:10.1093/molbev/msp077

41. Segata, N., Börnigen, D., Morgan, X. C., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nature Communications, 4, 2304. http://doi.org/10.1038/ncomms3304

42. N Saitou and M Nei The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol (1987) 4 (4): 406-425

43. A Rzhetsky and M Nei A Simple Method for Estimating and Testing Minimum-Evolution Trees Mol Biol Evol (1992) 9 (5): 945

44. Morgan N Price. FastTree documentation at meta.microbesonline.org

45. Mikael Thollesson. Tree selection criteria at artedi.ebc.uu.se (Uppsala Universitet)

46. Gary Olsen. Gary Olsen's Interpretation of the "Newick's 8:45" Tree Format Standard. 1990.

## ***6.2 Bioinformatics software used***

Smrt Analysis Package: http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/ by Pacific Biosciences.

HGAP assembler (included in Smrt Analysis). Paper: Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., … Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Meth, 10(6), 563–569. Retrieved from http://dx.doi.org/10.1038/nmeth.2474

BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download Paper: Altschul SF1, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.

Korilog BlastViewer (discontinued): No longer available.

Integrative Genomics Viewer (igv): https://www.broadinstitute.org/software/igv/download Paper: Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration Brief Bioinform (2013) 14 (2): 178-192 first published online April 19, 2012 doi:10.1093/bib/bbs017

AMOS minimus2: http://sourceforge.net/projects/amos/ Paper: Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., & Pop, M. (2011). Next Generation Sequence Assembly with AMOS. Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.], CHAPTER, Unit11.8. http://doi.org/10.1002/0471250953.bi1108s33

seaview: http://doua.prabi.fr/software/seaview Paper: Manolo Gouy, Stéphane Guindon and Olivier Gascuel (2010) SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol 27 (2): 221-224. doi: 10.1093/molbev/msp259

SOAPdenovo: http://soap.genomics.org.cn/soapdenovo.html Paper: Luo et al.: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 2012 1:18.

bowtie2: http://sourceforge.net/projects/bowtie-bio/files/bowtie2/ Paper: Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

PBJelly (PBSuite): http://sourceforge.net/projects/pb-jelly/ Paper: English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. (2012) Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE 2012 7(11): e47768. doi:10.1371/journal.pone.0047768

SAMtools: http://sourceforge.net/projects/samtools/files/ Paper: Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009 25 (16): 2078-2079. doi: 10.1093/bioinformatics/btp352

Mauve: http://darlinglab.org/mauve/download.html Paper: Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Research, 14(7), 1394–1403. http://doi.org/10.1101/gr.2289704

BEDtools: https://github.com/arq5x/bedtools2/releases Paper: Aaron R. Quinlan and Ira M. Hall (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010 26 (6): 841-842. doi: 10.1093/bioinformatics/btq033

PhyloPhlAn: https://bitbucket.org/nsegata/phylophlan/downloads Paper: Segata, N., Börnigen, D., Morgan, X. C., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nature Communications 2013, 4, 2304. http://doi.org/10.1038/ncomms3304

Prodigal: https://github.com/hyattpd/prodigal/releases/ Paper: Doug Hyatt, 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 2010, 11:119 doi:10.1186/1471-2105-11-119

Prokka: http://www.vicbioinformatics.com/software.prokka.shtml Paper: Torsten Seemann (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014 30 (14): 2068-2069. doi: 10.1093/bioinformatics/btu153

Treegraph2: http://treegraph.bioinfweb.info Paper: Stöver B C, Müller K F: TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 2010, 11:7

Figtree: http://tree.bio.ed.ac.uk/software/figtree/ by Andrew Rambaut

seqtk: https://github.com/lh3/seqtk by Heng Li; Broad Institute, Cambridge, MA, USA

cap3: http://seq.cs.iastate.edu/cap3.html Paper: Xiaoqiu Huang and Anup Madan (1999) CAP3: A DNA Sequence Assembly Program. doi: 10.1101/gr.9.9.868 Genome Res. 1999. 9: 868-877

Phylosift: https://phylosift.wordpress.com/tutorials/downloading-phylosift/ Paper: Darling AE, Jospin G, Lowe E, Matsen FA IV, Bik HM et al. (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ 2:e243 http://dx.doi.org/10.7717/peerj.243

fp.py (fasta parser): https://github.com/mtop/ngs/blob/master/fp.py by Mats Töpel

fasta_analyzer.py: https://github.com/karlsten/summer-course/blob/master/fasta_analyzer.py by Sandra Karlsten

contig_average_coverage.py: https://github.com/karlsten/misc/blob/master/misc/contig_average_coverage.py by Sandra Karlsten

count_names.py: https://github.com/alvaralmstedt/shell_scripts/blob/master/count_names.py by Mats Töpel remove_colons.py: https://github.com/alvaralmstedt/python_genome_ref_collab/blob/master/remove_colons.py by Alvar Almstedt

genome_downloader.py:
https://github.com/alvaralmstedt/python_genome_ref_collab/blob/master/genome_downloader.py by Alvar Almstedt

covgc_plot.sh: https://github.com/alvaralmstedt/shell_scripts/blob/master/covgc_plot.sh by Alvar Almstedt

## ***6.3 Web-based tools***

BASys bacterial annotation system: https://www.basys.ca

BLAST on NCBI: http://blast.ncbi.nlm.nih.gov/Blast.cgi

Pfam: http://pfam.xfam.org

NCBI reference genome ftp: ftp://ftp.wip.ncbi.nlm.nih.gov/

NCBI Taxonomy browser: http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi

## ***6.4 Other software***

CLI: Vim, grep, sed, awk, cut, tr, ln, tar, gunzip, shopt, ps, screen, nano, less, ssh, sge, scp, git, X, python, java

Apple OS X Yosemite/El Capitan: Safari, Pages, Numbers, Keynote, Stickies, Dictionary, Preview, Mail

Other:  Firefox, Google Chrome, Google Drive, PyCharm, Skype, Slack, Mendeley, iTerm, Dropbox, Word Wrangler

## ***6.5 Credits***

Supervisor: Mats Töpel

Examinator: Anders Blomberg

Additional input was received from the following during this work: Sylvie Tesson, Tomas Larsson, Sandra Karlsten, Magnus Alm Rosenblad, Anna Godhe, Alexander Eiler, Olga Kourtchenko, Tobias Hofmann, Diana Amza, Lucas Sinclair, Esteban Fernandez Parada, Emil Karlsson, Simon Stenberg, Martin Zackrisson, Mathias Johansson.